异步图书
www.epubit.com

Datawhale

机器学习公式详解

PUMPKIN BOOK

谢文睿 秦州 编著

中国工信出版集团 人民邮电出版社 POSTS & TELECOM PRESS

目录

版权信息

书名：机器学习公式详解

ISBN：978-7-115-55910-4

版　权

编　　著　谢文睿　秦 州

译　　郭　媛

责任编辑

人民邮电出版社出版发行　　北京市丰台区成寿寺路11号

邮编　100164　电子邮件　315@ptpress.com.cn

网址　http://www.ptpress.com.cn

读者服务热线：(010)81055410

反盗版热线：(010)81055315

内容提要

周志华老师的《机器学习》（俗称“西瓜书”）是机器学习领域的经典入门教材之一. 本书（俗称“南瓜书”）基于Datawhale成员自学“西瓜书”时记下的笔记编著而成，旨在对“西瓜书”中重、难点公式加以解析，以及对部分公式补充具体的推导细节.

全书共16章，与“西瓜书”章节、公式对应，每个公式的推导和解析都以本科数学基础的视角进行讲解，希望能够帮助读者达到“理工科数学基础扎实点的大二下学期学生”水平. 每章都附有相关阅读材料，以便有兴趣的读者进一步钻研探索.

本书思路清晰，视角独特，结构合理，可作为高等院校计算机及相关专业的本科生或研究生教材，也可供对机器学习感兴趣的研究人员和工程技术人员阅读参考.

序

虽然与本书的编著者素不相识、从未谋面，但是看过书稿之后，我便很乐意也感觉很荣幸有机会给这本书写序.

这是一本与众不同的书.

首先，确切地说，这是一本“伴侣书”. 类似于咖啡伴侣一样，这本书是周志华教授的“西瓜书”——《机器学习》的伴侣书， 它也有一个可爱的名字——“南瓜书”. “南瓜书”对“西瓜书”中的公式进行了解析，并补充了必要的推导过程；在推导公式的过程中有时候会需要一些先验知识，编著者也进行了必要的补充. 上述做法对学习机器学习时“知其然”并“知其所以然”非常重要. 现在能用一些机器学习工具来实现某个任务的人越来越多了，但是具有机器学习思维且了解其原理从而能够解决实际问题的能力在工作中更重要，具有这种能力的人也更具有竞争力.

其次，这是一本通过开源方式多人协作写成的书. 这种多人分工合作、互相校验、开放监督的方式，既保证了书的质量，也保证了写作的效率. 在我看来，这是一种站在读者角度且非常先进的生产方式，容易给读者带来很好的体验.

最后，我想说这是一本完全根据学习经历编著而成的书. 也就是说，这本书完全从读者学习的角度出发，分享编著者在学习中遇到的“坑”以及跳过这个“坑”的方法，这对初学者来说是非常宝贵的经验，

也特别能够引起他们的共鸣. 其实，每个人在学习一门新的课程时，都会有自己独特的经验和方法. 这种经验和方法的共享非常难能可贵. 在这里，理解公式便是编著者认为了解机器学习原理的最好方法，其实对于这一点我也深表赞同，因为在学习中我就是那种喜欢推导公式的典型代表，只有公式推导成功，才觉得对知识的原理理解得更深刻，否则总是觉得心里不踏实.

对于本书，我有几点阅读建议，供大家参考.

首先，这本 “南瓜书”要和“西瓜书”配套阅读，如果在阅读“西瓜书”时对公式疑惑或对概念理解不畅，可以通过“南瓜书”快速定位公式并进行推导，从而深入理解. 从这个意义来说，“南瓜书”可以看成是“西瓜书”的公式字典.

其次，阅读时一定要克服对公式的排斥或者畏惧心理. 公式是通过符号对原理本质的高度概括，是一种精简而美丽的数学语言. 推几个公式之后，相信读者会从中感觉到没有体验过的乐趣.

最后，这本书非常偏技术原理，看上去也有点儿枯燥，阅读时读者还是要事先做好克服困难的准备. 有时，即使编著者给出了推导过程，读者也不一定一眼就能理解，这就需要自己静下心来仔细研读. 只有这样，才有可能成为具有机器学习思维而不只是会用机器学习工具的人.

祝大家阅读愉快！

王　斌

小米AI实验室主任、NLP首席科学家

前　言

由于国内相关资料的匮乏，机器学习算法的公式推导历来都被认为是初学者的“噩梦”. 笔者两年前也受到了相同的困扰，但是在笔者师兄的鼓励下，笔者开始尝试做读书笔记， 经年累月遂有了编著本书的基本素材. 本书就是以笔者拜读周志华老师的《机器学习》（俗称“西瓜书”）时记下的笔 记为蓝本编著的. “西瓜书”作为机器学习领域的经典中文著作， 已经成为相关从业人员和学习者的必读 书目. 周老师为了兼顾更多读者，在“西瓜书”中尽可能少地使用数学知识. 然而这对笔者这 类对公式推导感兴趣的读者来说就颇费思量. 为此，本书便在“西瓜书”的基础上，对其中的 重难点公式进行一些补充. 具体地说，本书会对“西瓜书”中缺少推导细节的公式补充了详细的推导过程， 对不太易懂的公式补充解析.

全书的章节编排和“西瓜书”保持一致，共16章，各章中的内容都对应“西瓜书”中相应章节与公 式. 为了尽可能地降低阅读门槛，本书以本科数学视角编写，所以有本科数学基础的读者 基本都能畅读本书. 对于超过本科数学范围的数学知识，本书都会在相应章节附上详细讲 解的附注，以及具体的参考文献，读者可以按图索骥，拓展阅读. 由于本书主要是对“西瓜 书”进行的补充，所以在编写具体章节内容时，默认读者已经阅读过“西瓜书”相应章节.

本书需要搭配“西瓜书”一起阅读. 在阅读“西瓜书”的过程中，当遇

到推导不明白的公 式时再来查阅本书，效果最佳.

本书是由开源组织Datawhale的成员采用开源协作的方式完成，参与者包括2位主要编著 者（谢文睿和秦州）、6位编委会成员（贾彬彬、居凤霞、马晶敏、胡风范、周天烁和叶梁）、 12位特别贡献成员（awyd234、feijuan、Ggmatch、Heitao5200、huaqing89、LongJH、LilRachel、LeoLRH、 Nono17、spareribs、\linebreak sunchaothu和StevenLzq）.

本书可作为《机器学习》一书的配套读物，读者也可以 将其视为“一份现学现卖的读书笔记”. 由于编者水平有限，书中难免有所纰漏和表述不当的地方，还望各位读 者批评指正.

谢文睿

2020年12月27日

主要符号表

x 标量

$\boldsymbol{x}$ 向量

$\mathbf{x}$ 变量集

$\boldsymbol{A}$ 矩阵

$\mathbf{I}$ 单位阵

$\mathcal{X}$ 样本空间或状态空间

$\mathcal{D}$ 概率分布

D 数据样本（数据集）

$\mathcal{H}$ 假设空间

H 假设集

$\mathfrak{L}$ 学习算法

$(\cdot,\cdot,\cdot)$ 行向量

$(\cdot;\cdot;\cdot)$ 列向量

$(\cdot)^{\mathrm{T}}$ 向量或矩阵转置

$\{\cdots\}$ 集合

$|\{\cdots\}|$ 集合$\{\cdots\}$中元素个数

$\|\cdot\|_p$ L_p范数, p缺省时为L2范数

$P(\cdot), P(\cdot|\cdot)$ 概率质量函数, 条件概率质量函数

$p(\cdot), p(\cdot|\cdot)$ 概率密度函数, 条件概率密度函数

$\mathbb{E}_{\cdot\sim\mathcal{D}}[f(\cdot)$ 函数$f(\cdot)$对.在分布$\mathcal{D}$下的数学期望; 意义明确时将省略$\mathcal{D}$和(或).

$\sup(\cdot)$ 上确界

$\mathbb{I}(\cdot)$ 指示函数, 在.为真和假时分别取值为1, 0

$\mathrm{sgn}(\cdot)$ 符号函数, 在$\cdot<0,=0,>0$时分别取值为$-1,0,1$

第1章　绪论

式(1.1)

$$E_{ote}\left(\mathfrak{L}_a|X,f\right)=\sum_{h}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))P\left(h|X,\mathfrak{L}_a\right)$$

参见式 (1.2)

式(1.2）

$$\sum_{f}E_{\text{ote}}\left(\mathfrak{L}_a\mid X,f\right)=\sum_{f}\sum_{h}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))P\left(h\mid X,\mathfrak{L}_a\right) \quad ①$$

$$=\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\sum_{h}P\left(h\mid X,\mathfrak{L}_a\right)\sum_{f}\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x})) \quad ②$$

$$=\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\sum_{h}P\left(h\mid X,\mathfrak{L}_a\right)\frac{1}{2}2^{|\mathcal{X}|} \quad ③$$

$$=\frac{1}{2}2^{|\mathcal{X}|}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\sum_{h}P\left(h\mid X,\mathfrak{L}_a\right) \quad ④$$

$$=2^{|\mathcal{X}|-1}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\cdot 1 \quad ⑤$$

③→⑤显然成立

解析

①→②：

$$\begin{aligned}
&\sum_{f}\sum_{h}\sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x})\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))P(h|X,\mathfrak{L}_a)\\
=&\sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x})\sum_{f}\sum_{h}\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))P(h|X,\mathfrak{L}_a)\\
=&\sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x})\sum_{h}P(h|X,\mathfrak{L}_a)\sum_{f}\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))
\end{aligned}$$

②→③：首先要知道此时我们假设f是任何能将样本映射到{0,1}的函数.存在不止一个f时，f服从均匀分布，即每个f出现的概率相等.例如样本空间只有两个样本时，$\mathcal{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2\},|\mathcal{X}|=2$.那么所有可能的真实目标函数$f$如下：

$$\begin{aligned}
f_1:f_1(\boldsymbol{x}_1)=0,f_1(\boldsymbol{x}_2)=0;\\
f_2:f_2(\boldsymbol{x}_1)=0,f_2(\boldsymbol{x}_2)=1;\\
f_3:f_3(\boldsymbol{x}_1)=1,f_3(\boldsymbol{x}_2)=0;\\
f_4:f_4(\boldsymbol{x}_1)=1,f_4(\boldsymbol{x}_2)=1.
\end{aligned}$$

一共$2^{|\mathcal{X}|}=2^2=4$个可能的真实目标函数.所以此时通过算法$\mathfrak{L}_a$学习出来的模型$h(\boldsymbol{x})$对每个样本无论预测值为0还是1，都必然有一半的f与之预测值相等.例如，现在学出来的模型$h(\boldsymbol{x})$对$\boldsymbol{x}_1$的预测值为1，即$h(\boldsymbol{x}_1)=1$，那么有且只有f_3和f_4与$h(\boldsymbol{x})$的预测值相等，也就是有且只有一半的f与它预测值相等，所以$\sum_{f}\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))=\frac{1}{2}2^{|\mathcal{X}|}$.

值得一提的是，在这里我们假设真实的目标函数f服从均匀分布，但是实际情形并非如此，通常我们只认为能高度拟合已有样本数据的函数才是真实目标函数，例如，现在已有的样本数据为$\{(\boldsymbol{x}_1,0),(\boldsymbol{x}_2,1)\}$，那么此时$f_2$才是我们认为的真实目标函数，由于没有收集到或者压根不

存在$\{(\boldsymbol{x}_1,0),(\boldsymbol{x}_2,0)\},\{(\boldsymbol{x}_1,1),(\boldsymbol{x}_2,0)\},\{(\boldsymbol{x}_1,1),(\boldsymbol{x}_2,1)\}$这类样本，所以$f_1,f_3,f_4$都不算是真实目标函数.这也就是“西瓜书”式(1.3)下面的第3段中“骑自行车”的例子所想表达的内容.

第2章　模型评估与选择

式(2.20）

$$\mathrm{AUC}=\frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}-x_i)\cdot(y_i+y_{i+1})$$

解析

在解释AUC式之前，需要先弄清楚ROC曲线的具体绘制过程.下面我们就举个例子，按照“西瓜书”图2.4下方给出的绘制方法来讲解一下ROC曲线的具体绘制过程.

假设我们已经训练得到一个学习器$f(\boldsymbol{s})$，现在用该学习器来对8个测试样本（4个正例，4个反例，即$m^+=m^-=4$）进行预测，预测结果为：

$$\begin{gathered}(s_1,0.77,+),(s_2,0.62,-),(s_3,0.58,+),(s_4,0.47,+)\\(s_5,0.47,-),(s_6,0.33,-),(s_7,0.23,+),(s_8,0.15,-)\end{gathered}$$

此处用$\boldsymbol{s}$表示样本， 以和坐标(x,y)作出区分

其中，+和_分别表示样本为正例和为反例，数字表示学习器f预测该样本为正例的概率，例如对于反例$\boldsymbol{s}_2$来说， 当前学习器$f(\boldsymbol{s})$预测它是正例的概率为0.62.

上面给出的预测结果已经按照预测值从大到小排序

根据“西瓜书”上给出的绘制方法，首先需要对所有测试样本按照学习器给出的预测结果进行排序，接着将分类阈值设为一个不可能取到的最大值.显然，此时所有样本预测为正例的概率都一定小于分类阈值，那么预测为正例的样本个数为0，相应的真正例率和假正例率也都为0，所以我们可以在坐标$(0,0)$处标记一个点. 接下来需要把分类阈值从大到小依次设为每个样本的预测值，也就是依次设为0.77, 0.62, 0.58, 0.47, 0.33, 0.23,0.15，然后分别计算真正例率和假正例率，再在相应的坐标上标记点，最后再将各个点用直线连接, 即可得到ROC曲线.需要注意的是，在统计预测结果时，预测值等于分类阈值的样本也被算作预测为正例. 例如，当分类阈值为0.77时，测试样本 $\boldsymbol{s}_1$被预测为正例，由于它的真实标记也是正例，所以此时$\boldsymbol{s}_1$是一个真正例.为了便于绘图，我们将x轴（假正例率轴）的“步长”定为$\frac{1}{m^-}$，y轴（真正例率轴）的“步长”定为$\frac{1}{m^+}$. 根据真正例率和假正例率的定义可知，每次变动分类阈值时，若新增i个假正例，那么相应的x轴坐标也就增加$\frac{i}{m^-}$；若新增j个真正例，那么相应的y轴坐标也就增加$\frac{j}{m^+}$.按照以上讲述的绘制流程，最终我们可以绘制出如图2-1所示的ROC曲线.

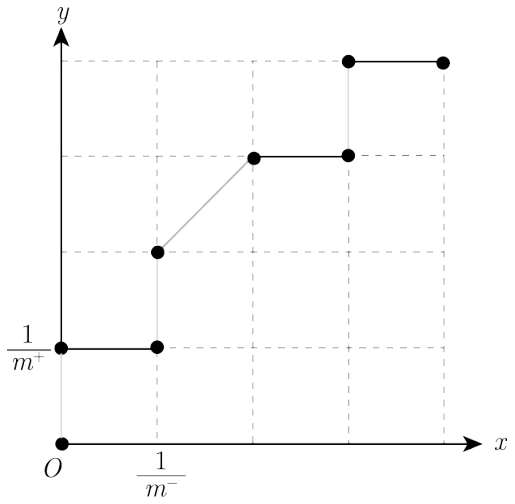

图2-1　ROC曲线示意

注：——表示红色线段；——表示蓝色线段；——表示绿色线段

在这里，为了能在解析式(2.21)时复用此图，我们没有写上具体的数值，转而用其数学符号代替.其中绿色线段表示在分类阈值变动的过程中只新增了真正例，红色线段表示只新增了假正例，蓝色线段表示既新增了真正例也新增了假正例.根据AUC值的定义可知，此时的AUC值其实就是所有红色线段和蓝色线段与x轴围成的面积之和.观察图2-1可知，红色线段与x轴围成的图形恒为矩形，蓝色线段与x轴围成的图形恒为梯形.由于梯形面积式既能算梯形面积，也能算矩形面积，所以无论是红色线段还是蓝色线段，其与x轴围成的面积都能用梯形公式来计算：

$$\frac{1}{2}\cdot(x_{i+1}-x_i)\cdot(y_i+y_{i+1}).$$

其中，$(x_{i+1}-x_i)$为“高”，y_i为“上底”，y_{i+1}为“下底”.那么对所有红色线段和蓝色线段与x轴围成的面积进行求和，则有

$$\sum_{i=1}^{m-1}\left[\frac{1}{2}\cdot(x_{i+1}-x_i)\cdot(y_i+y_{i+1})\right],$$

此即AUC.

式(2.21)

$$\ell_{rank}=\frac{1}{m^+m^-}\sum_{\boldsymbol{x}^+\in D^+}\sum_{\boldsymbol{x}^-\in D^-}\left(\mathbb{I}\left(f(\boldsymbol{x}^+)<f(\boldsymbol{x}^-)\right)+\frac{1}{2}\mathbb{I}\left(f(\boldsymbol{x}^+)=f(\boldsymbol{x}^-)\right)\right)$$

解析

按照我们上述对式(2.20)的解析思路，ℓ_{rank}可以看作是所有绿色线段和蓝色线段与y轴围成的面积之和，但从式(2.21)中很难一眼看出其面积的具体计算方式，因此我们进行恒等变形如下：

$$\begin{aligned}
\ell_{\text{rank}} &= \frac{1}{m^+m^-}\sum_{\boldsymbol{x}^+\in D^+}\sum_{\boldsymbol{x}^-\in D^-}\left(\mathbb{I}\left(f\left(\boldsymbol{x}^+\right)<f\left(\boldsymbol{x}^-\right)\right)+\frac{1}{2}\mathbb{I}\left(f\left(\boldsymbol{x}^+\right)=f\left(\boldsymbol{x}^-\right)\right)\right)\\
&= \frac{1}{m^+m^-}\sum_{\boldsymbol{x}^+\in D^+}\left[\sum_{\boldsymbol{x}^-\in D^-}\mathbb{I}\left(f\left(\boldsymbol{x}^+\right)<f\left(\boldsymbol{x}^-\right)\right)+\frac{1}{2}\cdot\sum_{\boldsymbol{x}^-\in D^-}\mathbb{I}\left(f\left(\boldsymbol{x}^+\right)=f\left(\boldsymbol{x}^-\right)\right)\right]\\
&= \sum_{\boldsymbol{x}^+\in D^+}\left[\frac{1}{m^+}\cdot\frac{1}{m^-}\sum_{\boldsymbol{x}^-\in D^-}\mathbb{I}\left(f\left(\boldsymbol{x}^+\right)<f\left(\boldsymbol{x}^-\right)\right)+\right.\\
&\qquad\left.\frac{1}{2}\cdot\frac{1}{m^+}\cdot\frac{1}{m^-}\sum_{\boldsymbol{x}^-\in D^-}\mathbb{I}\left(f\left(\boldsymbol{x}^+\right)=f\left(\boldsymbol{x}^-\right)\right)\right]\\
&= \sum_{\boldsymbol{x}^+\in D^+}\frac{1}{2}\cdot\frac{1}{m^+}\cdot\left[\frac{2}{m^-}\sum_{\boldsymbol{x}^-\in D^-}\mathbb{I}\left(f\left(\boldsymbol{x}^+\right)<f\left(\boldsymbol{x}^-\right)\right)+\right.\\
&\qquad\left.\frac{1}{m^-}\sum_{\boldsymbol{x}^-\in D^-}\mathbb{I}\left(f\left(\boldsymbol{x}^+\right)=f\left(\boldsymbol{x}^-\right)\right)\right].
\end{aligned}$$

在变动分类阈值的过程当中，如果有新增真正例，那么图2-1就会相应地增加一条绿色线段或蓝色线段，所以上式中的$\sum_{\boldsymbol{x}^+\in D^+}$可以看作是在累加所有绿色和蓝色线段，相应地，$\sum_{\boldsymbol{x}^+\in D^+}$后面的内容便是在求绿色线段或者蓝色线段与$y$轴围成的面积，即：

$$\frac{1}{2}\cdot\frac{1}{m^+}\cdot\left[\frac{2}{m^-}\sum_{\boldsymbol{x}^-\in D^-}\mathbb{I}\left(f(\boldsymbol{x}^+)<f(\boldsymbol{x}^-)\right)+\frac{1}{m^-}\sum_{\boldsymbol{x}^-\in D^-}\mathbb{I}\left(f(\boldsymbol{x}^+)=f(\boldsymbol{x}^-)\right)\right].$$

与式(2.20)中的求解思路相同，不论是绿色线段还是蓝色线段，其与y轴围成的图形面积都可以用梯形公式来进行计算，所以上式表示的依旧是一个梯形的面积公式.其中$\frac{1}{m^+}$即梯形的“高”，中括号内便是“上底+下底”，下面我们来分别推导一下“上底”（较短的底）和“下底”（较长的底）.

由于在绘制ROC曲线的过程中，每新增一个假正例时x坐标也就新增一个步长，所以对于“上底”，也就是绿色或者蓝色线段的下端点到y轴的距离，长度就等于$\frac{1}{m^-}$乘以预测值大于$f(\boldsymbol{x}^+)$的假正例的个数，即

$$\frac{1}{m^-}\sum_{\boldsymbol{x}^-\in D^-}\mathbb{I}\left(f(\boldsymbol{x}^+)<f(\boldsymbol{x}^-)\right);$$

而对于“下底”，长度就等于$\frac{1}{m^-}$乘以预测值大于等于$f(\boldsymbol{x}^+)$的假正例的个数，即

$$\frac{1}{m^-}\left(\sum_{\boldsymbol{x}^-\in D^-}\mathbb{I}\left(f(\boldsymbol{x}^+)<f(\boldsymbol{x}^-)\right)+\sum_{\boldsymbol{x}^-\in D^-}\mathbb{I}\left(f(\boldsymbol{x}^+)=f(\boldsymbol{x}^-)\right)\right).$$

式(2.27)

$$\overline{\epsilon} = \max \epsilon \quad \text{s.t.} \sum_{i=\epsilon_0 \times m+1}^{m} \binom{m}{i} \epsilon^i (1-\epsilon)^{m-i} < \alpha$$

解析

截至2018年12月“西瓜书”第1版第30次印刷，式(2.27)应当勘误为

$$\overline{\epsilon} = \min \epsilon \quad \text{s.t.} \sum_{i=\epsilon \times m+1}^{m} \binom{m}{i} \epsilon_0^i (1-\epsilon_0)^{m-i} < \alpha.$$

具体推导过程如下：由“西瓜书”中的上下文可知，对$\epsilon \leq \epsilon_0$进行假设检验，等价于本章附注中所述的对$p \leqslant p_0$进行假设检验，所以在“西瓜书”中求解最大错误率$\overline{\epsilon}$等价于在附注中求解事件最大发生频率$\frac{\overline{C}}{m}$. 由附注可知

$$\overline{C} = \min C \quad \text{s.t.} \sum_{i=C+1}^{m} \binom{m}{i} p_0^i (1-p_0)^{m-i} < \alpha.$$

所以

$$\frac{\overline{C}}{m} = \min \frac{C}{m} \quad \text{s.t.} \sum_{i=C+1}^{m} \binom{m}{i} p_0^i (1-p_0)^{m-i} < \alpha.$$

将上式中的$\frac{\overline{C}}{m}, \frac{C}{m}, p_0$等价替换为$\overline{\epsilon}, \epsilon, \epsilon_0$可得

$$\overline{\epsilon} = \min \epsilon \quad \text{s.t.} \sum_{i=\epsilon \times m+1}^{m} \binom{m}{i} \epsilon_0^i (1-\epsilon_0)^{m-i} < \alpha.$$

式**(2.41)**

$$E(f;D)=\mathbb{E}_D\left[(f(\boldsymbol{x};D)-y_D)^2\right] \quad ①$$

$$=\mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})+\bar{f}(\boldsymbol{x})-y_D\right)^2\right] \quad ②$$

$$=\mathbb{E}_D\left[(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x}))^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y_D\right)^2\right]+$$

$$\mathbb{E}_D\left[2(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x}))\left(\bar{f}(\boldsymbol{x})-y_D\right)\right] \quad ③$$

$$=\mathbb{E}_D\left[(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x}))^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y_D\right)^2\right] \quad ④$$

$$=\mathbb{E}_D\left[(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x}))^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y+y-y_D\right)^2\right] \quad ⑤$$

$$=\mathbb{E}_D\left[(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x}))^2\right]+\mathbb{E}_D\left[(\bar{f}(\boldsymbol{x})-y)^2\right]+\mathbb{E}_D\left[(y-y_D)^2\right]+$$

$$2\mathbb{E}_D\left[(\bar{f}(\boldsymbol{x})-y)(y-y_D)\right] \quad ⑥$$

$$=\mathbb{E}_D\left[(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x}))^2\right]+(\bar{f}(\boldsymbol{x})-y)^2+\mathbb{E}_D\left[(y_D-y)^2\right] \quad ⑦$$

①→②：减一个$\bar{f}(\boldsymbol{x})$再加一个$\bar{f}(\boldsymbol{x})$，属于简单的恒等变形

④→⑤：同①→②一样，减一个y再加一个y，属于简单的恒等变形

⑤→⑥：同②→③一样， 将最后一项利用期望的运算性质进行展开

解析

②→③：首先将中括号内的式子展开，有

$$\mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2+\left(\bar{f}(\boldsymbol{x})-y_D\right)^2+2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\left(\bar{f}(\boldsymbol{x})-y_D\right)\right],$$

然后根据期望的运算性质$\mathbb{E}[X+Y]=\mathbb{E}[X]+\mathbb{E}[Y]$可将上式化为

$$\begin{aligned}&\mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y_D\right)^2\right]+\\&\mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\left(\bar{f}(\boldsymbol{x})-y_D\right)\right].\end{aligned}$$

③→④：再次利用期望的运算性质将第3步得到的式子的最后一项展开，有

$$\begin{aligned}&\mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\left(\bar{f}(\boldsymbol{x})-y_D\right)\right]\\&=\mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\cdot\bar{f}(\boldsymbol{x})\right]-\mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\cdot y_D\right].\end{aligned}$$

首先计算展开后得到的第1项，有

$$\mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\cdot\bar{f}(\boldsymbol{x})\right]=\mathbb{E}_D\left[2f(\boldsymbol{x};D)\cdot\bar{f}(\boldsymbol{x})-2\bar{f}(\boldsymbol{x})\cdot\bar{f}(\boldsymbol{x})\right].$$

由于$\bar{f}(\boldsymbol{x})$是常量，所以由期望的运算性质：$\mathbb{E}[AX+B]=A\mathbb{E}[X]+B$（其中$A,B$均为常量）可得

$$\mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\cdot\bar{f}(\boldsymbol{x})\right]=2\bar{f}(\boldsymbol{x})\cdot\mathbb{E}_D\left[f(\boldsymbol{x};D)\right]-2\bar{f}(\boldsymbol{x})\cdot\bar{f}(\boldsymbol{x}).$$

由式(2.37)可知$\mathbb{E}_D\left[f(\boldsymbol{x};D)\right]=\bar{f}(\boldsymbol{x})$，所以

$$\mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\cdot\bar{f}(\boldsymbol{x})\right]=2\bar{f}(\boldsymbol{x})\cdot\bar{f}(\boldsymbol{x})-2\bar{f}(\boldsymbol{x})\cdot\bar{f}(\boldsymbol{x})=0.$$

接着计算展开后得到的第二项

$$\mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\cdot y_D\right]=2\mathbb{E}_D\left[f(\boldsymbol{x};D)\cdot y_D\right]-2\bar{f}(\boldsymbol{x})\cdot\mathbb{E}_D\left[y_D\right].$$

由于噪声和f无关，所以$f(\boldsymbol{x};D)$和y_D是两个相互独立的随机变量. 根据期望的运算性质$\mathbb{E}[XY]=\mathbb{E}[X]\mathbb{E}[Y]$（其中$X$和$Y$为相互独立的随机变

量）可得

$$\mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\cdot y_D\right]=2\mathbb{E}_D\left[f(\boldsymbol{x};D)\cdot y_D\right]-2\bar{f}(\boldsymbol{x})\cdot\mathbb{E}_D\left[y_D\right]$$

$$=2\mathbb{E}_D\left[f(\boldsymbol{x};D)\right]\cdot\mathbb{E}_D\left[y_D\right]-2\bar{f}(\boldsymbol{x})\cdot\mathbb{E}_D\left[y_D\right]$$

$$=2\bar{f}(\boldsymbol{x})\cdot\mathbb{E}_D\left[y_D\right]-2\bar{f}(\boldsymbol{x})\cdot\mathbb{E}_D\left[y_D\right]$$

$$=0,$$

所以

$$\mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\left(\bar{f}(\boldsymbol{x})-y_D\right)\right]$$

$$=\mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\cdot\bar{f}(\boldsymbol{x})\right]-\mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\cdot y_D\right]$$

$$=0+0$$

$$=0.$$

⑥→⑦：因为$\bar{f}(\boldsymbol{x})$和y均为常量，根据期望的运算性质，有⑥中的第2项

$$\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y\right)^2\right]=\left(\bar{f}(\boldsymbol{x})-y\right)^2.$$

同理有⑥中的最后一项

$$2\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y\right)\left(y-y_D\right)\right]=2\left(\bar{f}(\boldsymbol{x})-y\right)\mathbb{E}_D\left[y-y_D\right].$$

由于此时假定噪声的期望为零，即$\mathbb{E}_D\left[y-y_D\right]=0$，所以

$$2\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y\right)\left(y-y_D\right)\right]=2\left(\bar{f}(\boldsymbol{x})-y\right)\cdot 0=0.$$

附注

二项分布参数p的检验[1]

设某事件发生的概率为p，p未知.做m次独立试验，每次观察该事件是否发生，以X记该事件发生的次数，则X服从二项分布$B(m,p)$，现根据X检验如下假设：

$$H_0 : p \leqslant p_0;$$

$$H_1 : p > p_0.$$

由二项分布本身的特性可知：p越小，X取到较小值的概率越大.因此，对于上述假设，一个直观上合理的检验为

φ:当$X \leqslant C$时接受H_0,否则就拒绝H_0,

其中，$C \in N$表示事件最大发生次数.此检验对应的功效函数为

$$\begin{aligned}\beta_\varphi(p) &= P(X > C) \\ &= 1 - P(X \leqslant C) \\ &= 1 - \sum_{i=0}^{C} \binom{m}{i} p^i (1-p)^{m-i} \\ &= \sum_{i=C+1}^{m} \binom{m}{i} p^i (1-p)^{m-i}\end{aligned}$$

由于“p越小，X取到较小值的概率越大”可以等价表示为：$P(X \leqslant C)$是关于p的减函数，所以$\beta_\varphi(p) = P(X>C) = 1 - P(X \leqslant C)$是关于$p$的增函数，那么当$p \leqslant p_0$时，$\beta_\varphi(p_0)$即$\beta_\varphi(p)$的上确界.又根据参考文献

[1]中5.1.3的定义1.2可知，检验水平α默认取最小可能的水平，所以在给定检验水平α时，可以通过如下方程解得满足检验水平α的整数C：

更为严格的数学证明参见参考文献[1]中第二章习题7

$$\alpha = \sup\left\{\beta_{\varphi}(p)\right\}.$$

显然，当$p \leqslant p_0$时有

$$\begin{aligned}\alpha &= \sup\left\{\beta_{\varphi}(p)\right\} \\ &= \beta_{\varphi}\left(p_0\right) \\ &= \sum_{i=C+1}^{m}\left(\begin{array}{c} m \\ i \end{array}\right) p_0^i\left(1-p_0\right)^{m-i}.\end{aligned}$$

对于此方程，通常不一定正好解得一个使得方程成立的整数C，较常见的情况是存在这样一个$\bar{C}$使得

$$\sum_{i=\bar{C}+1}^{m}\left(\begin{array}{c} m \\ i \end{array}\right) p_0^i\left(1-p_0\right)^{m-i} < \alpha;$$

$$\sum_{i=\bar{C}}^{m}\left(\begin{array}{c} m \\ i \end{array}\right) p_0^i\left(1-p_0\right)^{m-i} > \alpha.$$

此时，C只能取$\bar{C}$或者$\bar{C}+1$.若C取$\bar{C}$，则相当于升高了检验水平α；若C取$\bar{C}+1$则相当于降低了检验水平α.具体如何取舍需要结合实际情况，但是通常为了减小犯第一类错误的概率，会倾向于令C取$\bar{C}+1$.

下面考虑如何求解$\bar{C}$.易证$\beta_{\varphi}(p_0)$是关于C的减函数，再结合上述关于$\bar{C}$的两个不等式易推得

$$\bar{C} = \min C \quad \text{s.t.} \quad \sum_{i=C+1}^{m} \binom{m}{i} p_0^i (1-p_0)^{m-i} < \alpha$$

参考文献

[1]陈希孺.概率论与数理统计[M].合肥：中国科学技术大学出版社,2009.

第3章　线性模型

式 (3.5)

$$\frac{\partial E_{(w,b)}}{\partial w} = 2\left(w\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m}(y_i - b)\,x_i\right)$$

解析

已知 $E_{(w,b)} = \sum_{i=1}^{m}(y_i - wx_i - b)^2$，所以

$$\frac{\partial E_{(w,b)}}{\partial w} = \frac{\partial}{\partial w}\left[\sum_{i=1}^{m}(y_i - wx_i - b)^2\right]$$

$$= \sum_{i=1}^{m}\frac{\partial}{\partial w}\left[(y_i - wx_i - b)^2\right]$$

$$= \sum_{i=1}^{m}\left[2\cdot(y_i - wx_i - b)\cdot(-x_i)\right]$$

$$= \sum_{i=1}^{m}\left[2\cdot\left(wx_i^2 - y_ix_i + bx_i\right)\right]$$

$$= 2\left(w\sum_{i=1}^{m}x_i^2 - \sum_{i=1}^{m}y_ix_i + b\sum_{i=1}^{m}x_i\right)$$

$$= 2\left(w\sum_{i=1}^{m}x_i^2 - \sum_{i=1}^{m}(y_i - b)\,x_i\right)$$

式**(3.6)**

$$\frac{\partial E_{(w,b)}}{\partial b} = 2\left(mb - \sum_{i=1}^{m}(y_i - wx_i)\right)$$

解析

已知 $E_{(w,b)} = \sum_{i=1}^{m}(y_i - wx_i - b)^2$，所以

$$\frac{\partial E_{(w,b)}}{\partial b} = \frac{\partial}{\partial b}\left[\sum_{i=1}^{m}(y_i - wx_i - b)^2\right]$$

$$= \sum_{i=1}^{m}\frac{\partial}{\partial b}\left[(y_i - wx_i - b)^2\right]$$

$$= \sum_{i=1}^{m}\left[2\cdot(y_i - wx_i - b)\cdot(-1)\right]$$

$$= \sum_{i=1}^{m}\left[2\cdot(b - y_i + wx_i)\right]$$

$$= 2\left(\sum_{i=1}^{m}b - \sum_{i=1}^{m}y_i + \sum_{i=1}^{m}wx_i\right)$$

$$= 2\left(mb - \sum_{i=1}^{m}(y_i - wx_i)\right).$$

式(3.7)

$$w=\frac{\sum_{i=1}^{m} y_i(x_i-\bar{x})}{\sum_{i=1}^{m} x_i^2-\frac{1}{m}\left(\sum_{i=1}^{m} x_i\right)^2}$$

解析

令式(3.5)等于0，有

$$0=w\sum_{i=1}^{m} x_i^2-\sum_{i=1}^{m}(y_i-b)x_i,$$

$$w\sum_{i=1}^{m} x_i^2=\sum_{i=1}^{m} y_i x_i-\sum_{i=1}^{m} b x_i.$$

由于令式(3.6)等于0可得 $b=\frac{1}{m}\sum_{i=1}^{m}(y_i-wx_i)$ ，又因为 $\frac{1}{m}\sum_{i=1}^{m} y_i=\bar{y}$ 且 $\frac{1}{m}\sum_{i=1}^{m} x_i=\bar{x}$ ，则$b=\bar{y}-w\bar{x}$，代入上式可得

$$w\sum_{i=1}^{m} x_i^2=\sum_{i=1}^{m} y_i x_i-\sum_{i=1}^{m}(\bar{y}-w\bar{x})x_i,$$

$$w\sum_{i=1}^{m} x_i^2=\sum_{i=1}^{m} y_i x_i-\bar{y}\sum_{i=1}^{m} x_i+w\bar{x}\sum_{i=1}^{m} x_i,$$

$$w\left(\sum_{i=1}^{m} x_i^2-\bar{x}\sum_{i=1}^{m} x_i\right)=\sum_{i=1}^{m} y_i x_i-\bar{y}\sum_{i=1}^{m} x_i,$$

$$w=\frac{\sum_{i=1}^{m} y_i x_i-\bar{y}\sum_{i=1}^{m} x_i}{\sum_{i=1}^{m} x_i^2-\bar{x}\sum_{i=1}^{m} x_i}.$$

将$\bar{y}\sum_{i=1}^{m}x_i=\frac{1}{m}\sum_{i=1}^{m}y_i\sum_{i=1}^{m}x_i=\bar{x}\sum_{i=1}^{m}y_i$和$\bar{x}\sum_{i=1}^{m}x_i=\frac{1}{m}\sum_{i=1}^{m}x_i\sum_{i=1}^{m}x_i=\frac{1}{m}\left(\sum_{i=1}^{m}x_i\right)^2$代入上式，即可得式(3.7)：

$$w=\frac{\sum_{i=1}^{m}y_i(x_i-\bar{x})}{\sum_{i=1}^{m}x_i^2-\frac{1}{m}\left(\sum_{i=1}^{m}x_i\right)^2}.$$

如果要想用Python来实现上式的话，上式中的求和运算只能用循环来实现.但是如果能将上式向量化，也就是转换成矩阵（即向量）运算的话，我们就可以利用诸如NumPy这种专门加速矩阵运算的类库来进行编写. 下面我们就尝试将上式进行向量化.

将$\frac{1}{m}\left(\sum_{i=1}^{m}x_i\right)^2=\bar{x}\sum_{i=1}^{m}x_i$代入分母可得

$$\begin{aligned}w&=\frac{\sum_{i=1}^{m}y_i\left(x_i-\bar{x}\right)}{\sum_{i=1}^{m}x_i^2-\bar{x}\sum_{i=1}^{m}x_i}\\&=\frac{\sum_{i=1}^{m}\left(y_ix_i-y_i\bar{x}\right)}{\sum_{i=1}^{m}\left(x_i^2-x_i\bar{x}\right)},\end{aligned}$$

又因为$\bar{y}\sum_{i=1}^{m}x_i=\bar{x}\sum_{i=1}^{m}y_i=\sum_{i=1}^{m}\bar{y}x_i=\sum_{i=1}^{m}\bar{x}y_i=m\bar{x}\bar{y}=\sum_{i=1}^{m}\bar{x}\bar{y}$且

$\sum_{i=1}^{m} x_i\bar{x} = \bar{x}\sum_{i=1}^{m} x_i = \bar{x}m\frac{1}{m}\sum_{i=1}^{m} x_i = m\bar{x}^2 = \sum_{i=1}^{m} \bar{x}^2$，则有

$$w = \frac{\sum_{i=1}^{m}\left(y_i x_i - y_i\bar{x} - x_i\bar{y} + \bar{x}\bar{y}\right)}{\sum_{i=1}^{m}\left(x_i^2 - x_i\bar{x} - x_i\bar{x} + \bar{x}^2\right)}$$

$$= \frac{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2}$$

若令$\boldsymbol{x} = (x_1, x_2, \cdots, x_m)^{\mathrm{T}}$，$\boldsymbol{x}_{\mathrm{d}} = (x_1 - \bar{x}, x_2 - \bar{x}, \cdots, x_m - \bar{x})^{\mathrm{T}}$为去均值后的$\boldsymbol{x}$；$\boldsymbol{y} = (y_1, y_2, \cdots, y_m)^{\mathrm{T}}$，$\boldsymbol{y}_{\mathrm{d}} = (y_1 - \bar{y}, y_2 - \bar{y}, \cdots, y_m - \bar{y})^{\mathrm{T}}$为去均值后的$\boldsymbol{y}$，代入上式可得

$\boldsymbol{x}$、$\boldsymbol{x}_{\mathrm{d}}$、$\boldsymbol{y}$、$\boldsymbol{y}_{\mathrm{d}}$均为$m$行1列的列向量

$$w = \frac{\boldsymbol{x}_{\mathrm{d}}^{\mathrm{T}}\boldsymbol{y}_{\mathrm{d}}}{\boldsymbol{x}_{\mathrm{d}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{d}}}.$$

式**(3.10)**

$$\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}} = 2\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y})$$

解析

将$E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{w}})$展开可得

$$E_{\hat{\boldsymbol{w}}} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{w}} - \hat{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} + \hat{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{w}},$$

对$\hat{\boldsymbol{w}}$求导可得

$$\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}} = \frac{\partial \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}}{\partial \hat{\boldsymbol{w}}} - \frac{\partial \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{w}}}{\partial \hat{\boldsymbol{w}}} - \frac{\partial \hat{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}}{\partial \hat{\boldsymbol{w}}} + \frac{\partial \hat{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{w}}}{\partial \hat{\boldsymbol{w}}}$$

$$= 0 - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} + \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\hat{\boldsymbol{w}}$$

$$= 2\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y}).$$

矩阵微分式 $\frac{\partial \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}$, $\frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}})\boldsymbol{x}$

式(3.27）

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m}\left(-y_i\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i + \ln\left(1 + \mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)\right)$$

解析

将式(3.26)代入式(3.25)可得

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m}\ln\left(y_i p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}) + (1-y_i)p_0(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})\right),$$

其中 $p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}) = \frac{\mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}{1+\mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}$，$p_0(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}) = \frac{1}{1+\mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}$，代入上式可得

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{m}\ln\left(\frac{y_i\mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i} + 1 - y_i}{1+\mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}\right)\\ &= \sum_{i=1}^{m}\left(\ln\left(y_i\mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i} + 1 - y_i\right) - \ln\left(1+\mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)\right)\end{aligned}$$

由于y_i=0或1，则

$$\ell(\boldsymbol{\beta}) = \begin{cases}\sum_{i=1}^{m}\left(-\ln\left(1+\mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)\right), & y_i = 0;\\ \sum_{i=1}^{m}\left(\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i - \ln\left(1+\mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)\right), & y_i = 1\end{cases}$$

两式综合可得

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left(y_i \boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i - \ln \left(1 + \mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i} \right) \right).$$

由于此式仍为极大似然估计的似然函数，所以最大化似然函数等价于最小化似然函数的相反数，即在似然函数前添加负号即可得式(3.27). 值得一提的是，若将式(3.26)改写为 $p(y_i|\boldsymbol{x}_i;\boldsymbol{w},b) = (p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}))^{y_i}(p_0(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}))^{1-y_i}$，再代入式(3.25)可得

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{m} \ln \left(\left(p_1 \left(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta} \right) \right)^{y_i} \left(p_0 \left(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta} \right) \right)^{1-y_i} \right) \\
&= \sum_{i=1}^{m} \left(y_i \ln \left(p_1 \left(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta} \right) \right) + (1 - y_i) \ln \left(p_0 \left(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta} \right) \right) \right) \\
&= \sum_{i=1}^{m} \left(y_i \left(\ln \left(p_1 \left(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta} \right) \right) - \ln \left(p_0 \left(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta} \right) \right) \right) + \ln \left(p_0 \left(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta} \right) \right) \right) \\
&= \sum_{i=1}^{m} \left(y_i \ln \left(\frac{p_1 \left(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta} \right)}{p_0 \left(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta} \right)} \right) + \ln \left(p_0 \left(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta} \right) \right) \right) \\
&= \sum_{i=1}^{m} \left(y_i \ln \left(\mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i} \right) + \ln \left(\frac{1}{1 + \mathrm{e}^{\boldsymbol{T}^{\mathrm{T}} \hat{\boldsymbol{x}}_i}} \right) \right) \\
&= \sum_{i=1}^{m} \left(y_i \boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i - \ln \left(1 + \mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i} \right) \right).
\end{aligned}$$

显然，此种方式更易推导出式(3.27).

式(3.30)

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^{m} \hat{\boldsymbol{x}}_i (y_i - p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}))$$

解析

此式可以进行向量化. 令$\hat{y}_i = p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta})$，代入上式得

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -\sum_{i=1}^{m} \hat{\boldsymbol{x}}_i \left(y_i - \hat{y}_i\right) \\ &= \sum_{i=1}^{m} \hat{\boldsymbol{x}}_i \left(\hat{y}_i - y_i\right) \\ &= \boldsymbol{X}^{\mathrm{T}}(\hat{\boldsymbol{y}} - \boldsymbol{y}) \\ &= \boldsymbol{X}^{\mathrm{T}} \left(p_1(\boldsymbol{X}; \boldsymbol{\beta}) - \boldsymbol{y}\right).\end{aligned}$$

式**(3.32)**

$$J = \frac{\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)\boldsymbol{w}}$$

解析

$$\begin{aligned}J &= \frac{\left\|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0 - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1\right\|_2^2}{\boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1\right)\boldsymbol{w}} \\ &= \frac{\left\|\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0 - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1\right)^{\mathrm{T}}\right\|_2^2}{\boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1\right)\boldsymbol{w}} \\ &= \frac{\left\|\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right)^{\mathrm{T}}\boldsymbol{w}\right\|_2^2}{\boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1\right)\boldsymbol{w}} \\ &= \frac{\left[\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right)^{\mathrm{T}}\boldsymbol{w}\right]^{\mathrm{T}}\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right)^{\mathrm{T}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1\right)\boldsymbol{w}}\end{aligned}$$

$$= \frac{\boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1\right)\left(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1\right)^{\mathrm{T}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1\right)\boldsymbol{w}}$$

式**(3.37)**

$$\boldsymbol{S}_b\boldsymbol{w}=\lambda\boldsymbol{S}_w\boldsymbol{w}$$

解析

由式(3.36)，可定义拉格朗日函数为

$$L(\boldsymbol{w},\lambda)=-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{w}+\lambda(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{w}-1),$$

对$\boldsymbol{w}$求偏导可得

$$\frac{\partial L(\boldsymbol{w},\lambda)}{\partial\boldsymbol{w}}=-\frac{\partial\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{w}\right)}{\partial\boldsymbol{w}}+\lambda\frac{\partial\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{w}-1\right)}{\partial\boldsymbol{w}}$$

$$=-\left(\boldsymbol{S}_b+\boldsymbol{S}_b^{\mathrm{T}}\right)\boldsymbol{w}+\lambda\left(\boldsymbol{S}_w+\boldsymbol{S}_w^{\mathrm{T}}\right)\boldsymbol{w}$$

$$=-2\boldsymbol{S}_b\boldsymbol{w}+2\lambda\boldsymbol{S}_w\boldsymbol{w}.$$

这是由于$\boldsymbol{S}_b=\boldsymbol{S}_b^{\mathrm{T}}$、$\boldsymbol{S}_w=\boldsymbol{S}_w^{\mathrm{T}}$

令上式等于**0**即可得

$$-2\boldsymbol{S}_b\boldsymbol{w}+2\lambda\boldsymbol{S}_w\boldsymbol{w}=\boldsymbol{0}$$
$$\boldsymbol{S}_b\boldsymbol{w}=\lambda\boldsymbol{S}_w\boldsymbol{w}$$

由于我们想要求解的只有$\boldsymbol{w}$，而拉格朗日乘子λ具体取值多少都无所谓，因此可以任意设定λ来配合我们求解$\boldsymbol{w}$. 注意到

$$\boldsymbol{S}_b\boldsymbol{w}=(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w},$$

如果我们令$\lambda = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}$，那么上式即可改写为

$$\boldsymbol{S}_b\boldsymbol{w} = \lambda(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1),$$

将其代入$\boldsymbol{S}_b\boldsymbol{w} = \lambda\boldsymbol{S}_b\boldsymbol{w}$即可解得

$$\boldsymbol{w} = \boldsymbol{S}_w^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1).$$

式**(3.38)**

$$\boldsymbol{S}_b\boldsymbol{w} = \lambda(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$$

参见式(3.37)

式**(3.39)**

$$\boldsymbol{w} = \boldsymbol{S}_w^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$$

参见式(3.37)

式**(3.43)**

$$\begin{aligned}\boldsymbol{S}_b &= \boldsymbol{S}_t - \boldsymbol{S}_w \\ &= \sum_{i=1}^{N} m_i \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}\right)^{\mathrm{T}}\end{aligned}$$

解析

由式(3.40)、式(3.41)、式(3.42)可得：

$$\begin{aligned}&\boldsymbol{S}_b = \boldsymbol{S}_t - \boldsymbol{S}_w \\ &= \sum_{i=1}^{m} \left(\boldsymbol{x}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}\right)^{\mathrm{T}} - \sum_{i=1}^{N}\sum_{\boldsymbol{x}\in X_i} \left(\boldsymbol{x} - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x} - \boldsymbol{\mu}_i\right)^{\mathrm{T}}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} \left(\sum_{\boldsymbol{x} \in X_i} \left((\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} - (\boldsymbol{x} - \boldsymbol{\mu}_i)\,(\boldsymbol{x} - \boldsymbol{\mu}_i)^{\mathrm{T}} \right) \right) \\
&= \sum_{i=1}^{N} \left(\sum_{\boldsymbol{x} \in X_i} \left((\boldsymbol{x} - \boldsymbol{\mu})\left(\boldsymbol{x}^{\mathrm{T}} - \boldsymbol{\mu}^{\mathrm{T}}\right) - (\boldsymbol{x} - \boldsymbol{\mu}_i)\left(\boldsymbol{x}^{\mathrm{T}} - \boldsymbol{\mu}_i^{\mathrm{T}}\right) \right) \right) \\
&= \sum_{i=1}^{N} \left(\sum_{\boldsymbol{x} \in X_i} \left(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} - \boldsymbol{x}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{x}^{\mathrm{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} + \boldsymbol{x}\boldsymbol{\mu}_i^{\mathrm{T}} + \boldsymbol{\mu}_i\boldsymbol{x}^{\mathrm{T}} - \boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}} \right) \right) \\
&= \sum_{i=1}^{N} \left(\sum_{\boldsymbol{x} \in X_i} \left(-\boldsymbol{x}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{x}^{\mathrm{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{x}\boldsymbol{\mu}_i^{\mathrm{T}} + \boldsymbol{\mu}_i\boldsymbol{x}^{\mathrm{T}} - \boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}} \right) \right) \\
&= \sum_{i=1}^{N} \left(- \sum_{\boldsymbol{x} \in X_i} \boldsymbol{x}\boldsymbol{\mu}^{\mathrm{T}} - \sum_{\boldsymbol{x} \in X_i} \boldsymbol{\mu}\boldsymbol{x}^{\mathrm{T}} + \sum_{\boldsymbol{x} \in X_i} \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \right. \\
&\left. \sum_{\boldsymbol{x} \in X_i} \boldsymbol{x}\boldsymbol{\mu}_i^{\mathrm{T}} + \sum_{\boldsymbol{x} \in X_i} \boldsymbol{\mu}_i\boldsymbol{x}^{\mathrm{T}} - \sum_{\boldsymbol{x} \in X_i} \boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}} \right) \\
&= \sum_{i=1}^{N} \left(-m_i\boldsymbol{\mu}_i\boldsymbol{\mu}^{\mathrm{T}} - m_i\boldsymbol{\mu}\boldsymbol{\mu}_i^{\mathrm{T}} + m_i\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + m_i\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}} + m_i\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}} - m_i\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}} \right) \\
&= \sum_{i=1}^{N} \left(-m_i\boldsymbol{\mu}_i\boldsymbol{\mu}^{\mathrm{T}} - m_i\boldsymbol{\mu}\boldsymbol{\mu}_i^{\mathrm{T}} + m_i\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + m_i\boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}} \right) \\
&= \sum_{i=1}^{N} m_i \left(-\boldsymbol{\mu}_i\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{\mu}_i^{\mathrm{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}} \right) \\
&= \sum_{i=1}^{N} m_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})\,(\boldsymbol{\mu}_i - \boldsymbol{\mu})^{\mathrm{T}}
\end{aligned}
$$

式(3.44)

$$
\max_{\boldsymbol{W}} \frac{\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{W})}{\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{W})}
$$

解析

此式是式(3.35)的推广形式，证明如下.

设$\boldsymbol{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_i,\cdots,\boldsymbol{w}_{N-1})\in\mathbb{R}^{d\times(N-1)}$，其中$\boldsymbol{w}_i\in\mathbb{R}^{d\times 1}$为$d$行1列的列向量，则

$$\begin{cases}\operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{W}\right)=\displaystyle\sum_{i=1}^{N-1}\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{w}_i,\\ \operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{W}\right)=\displaystyle\sum_{i=1}^{N-1}\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{w}_i.\end{cases}$$

所以式(3.44)可变形为

$$\max_{\boldsymbol{W}}\frac{\displaystyle\sum_{i=1}^{N-1}\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{w}_i}{\displaystyle\sum_{i=1}^{N-1}\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{w}_i}.$$

对比式(3.35)易知，上式即式(3.35)的推广形式.

式(3.45)

$$\boldsymbol{S}_b\boldsymbol{W}=\lambda\boldsymbol{S}_w\boldsymbol{W}$$

解析

同式(3.35)，此处也固定式(3.44)的分母为1，那么式(3.44)此时等价于如下优化问题

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & -\operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}_b \boldsymbol{W}\right) \\ \text{s.t.} \quad & \operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}_w \boldsymbol{W}\right) = 1 \end{aligned}$$

根据拉格朗日乘子法，可定义上述优化问题的拉格朗日函数

$$L(\boldsymbol{W}, \lambda) = -\operatorname{tr}(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}_b \boldsymbol{W}) + \lambda(\operatorname{tr}(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}_w \boldsymbol{W}) - 1).$$

根据矩阵微分式$\frac{\partial}{\partial \boldsymbol{X}} \operatorname{tr}\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{B} \boldsymbol{X}\right) = (\boldsymbol{B} + \boldsymbol{B}^{\mathrm{T}}) \boldsymbol{X}$对上式关于$\boldsymbol{W}$求偏导可得

$$\frac{\partial L(\boldsymbol{W}, \lambda)}{\partial \boldsymbol{W}} = -\frac{\partial\left(\operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}_b \boldsymbol{W}\right)\right)}{\partial \boldsymbol{W}} + \lambda \frac{\partial\left(\operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}_w \boldsymbol{W}\right) - 1\right)}{\partial \boldsymbol{W}}$$

$$= -\left(\boldsymbol{S}_b + \boldsymbol{S}_b^{\mathrm{T}}\right) \boldsymbol{W} + \lambda\left(\boldsymbol{S}_w + \boldsymbol{S}_w^{\mathrm{T}}\right) \boldsymbol{W}$$

$$= -2 \boldsymbol{S}_b \boldsymbol{W} + 2 \lambda \boldsymbol{S}_w \boldsymbol{W}.$$

这是由于$\boldsymbol{S}_b = \boldsymbol{S}_b^{\mathrm{T}}$且$\boldsymbol{S}_w = \boldsymbol{S}_w^{\mathrm{T}}$

令上式等于0即可得

$$\begin{aligned} -2 \boldsymbol{S}_b \boldsymbol{W} + 2 \lambda \boldsymbol{S}_w \boldsymbol{W} &= \boldsymbol{0} \\ \boldsymbol{S}_b \boldsymbol{W} &= \lambda \boldsymbol{S}_w \boldsymbol{W} \end{aligned}$$

第4章　决策树

式(4.1)

$$\mathrm{Ent}(D)=-\sum_{k=1}^{|\mathcal{Y}|}p_k\log_2 p_k$$

解析

下面证明$0\leqslant \mathrm{Ent}(D)\leqslant \log_2|\mathcal{Y}|$.

已知集合D的信息熵的定义为

$$\mathrm{Ent}(D)=-\sum_{k=1}^{|\mathcal{Y}|}p_k\log_2 p_k,$$

其中，$|\mathcal{Y}|$表示样本类别总数，p_k表示第k类样本所占的比例，有$0\leqslant p_k\leqslant 1,\sum_{k=1}^{n}p_k=1$.若令$|\mathcal{Y}|=n,p_k=x_k$，那么信息熵$\mathrm{Ent}(D)$就可以看作一个$n$元实值函数，即

$$\mathrm{Ent}(D)=f(x_1,\cdots,x_n)=-\sum_{k=1}^{n}x_k\log_2 x_k,$$

其中$0\leqslant x_k\leqslant 1,\sum_{k=1}^{n}x_k=1$.

下面考虑求该多元函数的最值. 首先我们先来求最大值，如果不考

虑约束$0 \leqslant x_k \leqslant 1$而仅考虑$\sum_{k=1}^{n} x_k = 1$，则对$f(x_1, \cdots, x_n)$求最大值等价于如下最小化问题：

$$\min \sum_{k=1}^{n} x_k \log_2 x_k$$

$$\text{s.t.} \sum_{k=1}^{n} x_k = 1.$$

显然，在$0 \leqslant x_k \leqslant 1$时，此问题为凸优化问题.对于凸优化问题来说，使其拉格朗日函数的一阶偏导数等于0的点即最优解.根据拉格朗日乘子法可知，该优化问题的拉格朗日函数为

$$L(x_1, \cdots, x_n, \lambda) = \sum_{k=1}^{n} x_k \log_2 x_k + \lambda \left(\sum_{k=1}^{n} x_k - 1 \right)$$

其中，λ为拉格朗日乘子.对$L(x_1, \cdots, x_n, \lambda)$分别关于$x_1, \cdots, x_n, \lambda$求一阶偏导数，并令偏导数等于0可得

$$\frac{\partial L\left(x_1, \cdots, x_n, \lambda\right)}{\partial x_1} = \frac{\partial}{\partial x_1} \left[\sum_{k=1}^{n} x_k \log_2 x_k + \lambda \left(\sum_{k=1}^{n} x_k - 1 \right) \right] = 0$$

$$= \log_2 x_1 + x_1 \cdot \frac{1}{x_1 \ln 2} + \lambda = 0$$

$$= \log_2 x_1 + \frac{1}{\ln 2} + \lambda = 0$$

$$\Rightarrow \lambda = -\log_2 x_1 - \frac{1}{\ln 2};$$

$$\frac{\partial L\left(x_1,\cdots,x_n,\lambda\right)}{\partial x_2}=\frac{\partial}{\partial x_2}\left[\sum_{k=1}^{n}x_k\log_2 x_k+\lambda\left(\sum_{k=1}^{n}x_k-1\right)\right]=0$$

$$\Rightarrow\lambda=-\log_2 x_2-\frac{1}{\ln 2};$$

...

$$\frac{\partial L\left(x_1,\cdots,x_n,\lambda\right)}{\partial x_n}=\frac{\partial}{\partial x_n}\left[\sum_{k=1}^{n}x_k\log_2 x_k+\lambda\left(\sum_{k=1}^{n}x_k-1\right)\right]=0$$

$$\Rightarrow\lambda=-\log_2 x_n-\frac{1}{\ln 2};$$

$$\frac{\partial L\left(x_1,\cdots,x_n,\lambda\right)}{\partial \lambda}=\frac{\partial}{\partial \lambda}\left[\sum_{k=1}^{n}x_k\log_2 x_k+\lambda\left(\sum_{k=1}^{n}x_k-1\right)\right]=0$$

$$\Rightarrow\sum_{k=1}^{n}x_k=1$$

.

整理一下可得

$$\begin{cases}\lambda=-\log_2 x_1-\frac{1}{\ln 2}=-\log_2 x_2-\frac{1}{\ln 2}=\cdots=-\log_2 x_n-\frac{1}{\ln 2},\\ \sum_{k=1}^{n}x_k=1.\end{cases}$$

由以上两个方程可以解得

$$x_1=x_2=\cdots=x_n=\frac{1}{n},$$

又因为x_k还需满足约束$0\leqslant x_k\leqslant 1$，显然$0\leqslant\frac{1}{n}\leqslant 1$，所以$x_1=x_2=\cdots=x_n=\frac{1}{n}$是满足所有约束的最优解，即当前最小化问题的最

小值点，同时也是$f(x_1,\cdots,x_n)$的最大值点.将$x_1=x_2=\cdots=x_n=\dfrac{1}{n}$代入$f(x_1,\cdots,x_n)$中可得

$$f\left(\frac{1}{n},\cdots,\frac{1}{n}\right)=-\sum_{k=1}^{n}\frac{1}{n}\log_2\frac{1}{n}=-n\cdot\frac{1}{n}\log_2\frac{1}{n}=\log_2 n,$$

所以$f(x_1,\cdots,x_n)$在满足约束$0\leqslant x_k\leqslant 1,\sum_{k=1}^{n}x_k=1$时的最大值为$\log_2 n$.

下面求最小值. 如果不考虑约束$\sum_{k=1}^{n}x_k=1$而仅考虑$0\leqslant x_k\leqslant 1$，则$f(x_1,\cdots,x_n)$可以看作$n$个互不相关的一元函数的和，即

$$f(x_1,\cdots,x_n)=\sum_{k=1}^{n}g(x_k),$$

其中，$g(x_k)=-x_k\log_2 x_k, 0\leqslant x_k\leqslant 1$.那么当$g(x_1),g(x_2),\cdots,g(x_n)$分别取到其最小值时，$f(x_1,\cdots,x_n)$也就取到了最小值.所以接下来考虑分别求$g(x_1),g(x_2),\cdots,g(x_n)$各自的最小值，由于$g(x_1),g(x_2),\cdots,g(x_n)$的定义域和函数表达式均相同，所以只需求出$g(x_1)$的最小值也就求出了$g(x_2),\cdots,g(x_n)$的最小值.下面考虑求$g(x_1)$的最小值，首先对$g(x_1)$关于$x_1$求一阶和二阶导数，有

$$g'(x_1)=\frac{\mathrm{d}(-x_1\log_2 x_1)}{\mathrm{d}x_1}=-\log_2 x_1-x_1\cdot\frac{1}{x_1\ln 2}=-\log_2 x_1-\frac{1}{\ln 2},$$

$$g''(x_1)=\frac{\mathrm{d}\left(g'(x_1)\right)}{\mathrm{d}x_1}=\frac{\mathrm{d}\left(-\log_2 x_1-\dfrac{1}{\ln 2}\right)}{\mathrm{d}x_1}=-\frac{1}{x_1\ln 2}.$$

显然，当$0 \leqslant x_k \leqslant 1$时$g''(x_1) = -\dfrac{1}{x_1 \ln 2}$恒小于0，所以$g(x_1)$是一个在其定义域范围内开口向下的凹函数，那么其最小值必然在边界取.分别取$x_1 = 0$和$x_1 = 1$，代入$g(x_1)$可得

计算信息熵时约定：若$x = 0$，则$x \log_2 x = 0$

$$g(0) = -0 \log_2 0 = 0$$

$$g(1) = -1 \log_2 1 = 0$$

所以，$g(x_1)$的最小值为0，同理可得$g(x_2), \cdots, g(x_n)$的最小值也都为0，即$f(x_1, \cdots, x_n)$的最小值为0. 但是， 此时仅考虑约束$0 \leqslant x_k \leqslant 1$，而未考虑$\sum_{k=1}^{n} x_k = 1$.若考虑约束$\sum_{k=1}^{n} x_k = 1$，那么$f(x_1, \cdots, x_n)$的最小值一定大于等于0.如果令某个$x_k = 1$，那么根据约束$\sum_{k=1}^{n} x_k = 1$可知$x_1 = x_2 = \cdots = x_{k-1} = x_{k+1} = \cdots = x_n = 0$，将其代入$f(x_1, \cdots, x_n)$可得

$$\begin{aligned} & f(0, 0, \cdots, 0, 1, 0, \cdots, 0) \\ = & -0 \log_2 0 - 0 \log_2 0 - \cdots - 0 \log_2 0 - 1 \log_2 1 - 0 \log_2 0 - \cdots - 0 \log_2 0 = 0. \end{aligned}$$

所以$x_k = 1, x_1 = x_2 = \cdots = x_{k-1} = x_{k+1} = \cdots = x_n = 0$一定是$f(x_1, \cdots, x_n)$在满足约束$\sum_{k=1}^{n} x_k = 1$和$0 \leqslant x_k \leqslant 1$的条件下的最小值点，此时$f$取到最小值0.

综上可知，当$f(x_1, \cdots, x_n)$取到最大值时：$x_1 = x_2 = \cdots = x_n = \dfrac{1}{n}$，此时样本集合纯度最低；当$f(x_1, \cdots, x_n)$取到最小值时：

$x_k = 1, x_1 = x_2 = \cdots = x_{k-1} = x_{k+1} = \cdots = x_n = 0$，此时样本集合纯度最高.

式**(4.2)**

$$\mathrm{Gain}(D, a) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v)$$

解析

此为信息增益的定义式.在信息论中信息增益也称为互信息（参见本章附注①），表示已知一个随机变量的信息后另一个随机变量的不确定性减少的程度.所以此式可以理解为，在已知属性a的取值后，样本类别这个随机变量的不确定性减小的程度.若根据某个属性计算得到的信息增益越大，则说明在知道其取值后样本集的不确定性减小的程度越大，即“西瓜书”上所说的“纯度提升”越大.

式**(4.6）**

$$\mathrm{Gini\_index}(D, a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Gini}(D^v)$$

解析

此为数据集D中属性a的基尼指数的定义，表示在属性a的取值已知的条件下，数据集D按照属性a的所有可能取值划分后的纯度.不过在构造CART决策树时并不会严格按照此式来选择最优划分属性，主要是因为CART决策树是一棵二叉树，如果用上面的式去选出最优划分属性，

无法进一步选出最优划分属性的最优划分点.常用的CART决策树的构造算法如下：

(1) 考虑每个属性a的每个可能取值v，将数据集D分为$a = v$和$a \neq v$两部分来计算基尼指数，即

$$\text{Gini\_index}(D, a) = \frac{|D^{a=v}|}{|D|}\text{Gini}(D^{a=v}) + \frac{|D^{a\neq v}|}{|D|}\text{Gini}(D^{a\neq v});$$

(2) 选择基尼指数最小的属性及其对应取值作为最优划分属性和最优划分点；

(3) 重复以上两步，直至满足停止条件.

下面以“西瓜书”中表4.2中西瓜数据集2.0为例来构造CART决策树，其中第一个最优划分属性和最优划分点的计算过程如下：以属性“色泽”为例，它有3个可能的取值：{青绿}，{乌黑}，{浅白}，若使用该属性的属性值是否等于“青绿”对数据集D进行划分，则可得到2个子集，分别记为D^1(色泽=青绿),D^2(色泽≠青绿).子集D^1包含编号$\{1, 4, 6, 10, 13, 17\}$共6个样例，其中正例占$p_1 = \frac{3}{6}$，反例占$p_2 = \frac{3}{6}$；子集D^2包含编号$\{2, 3, 5, 7, 8, 9, 11, 12, 14, 15, 16\}$共11个样例，其中正例占$p_1 = \frac{5}{11}$，反例占$p_2 = \frac{6}{11}$，根据式(4.5)可计算出用“色泽=青绿”划分之后得到基尼指数为

$$\text{Gini\_index}(D,\text{色泽=青绿})$$

$$= \frac{6}{17} \times \left(1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2\right) + \frac{11}{17} \times \left(1 - \left(\frac{5}{11}\right)^2 - \left(\frac{6}{11}\right)^2\right) = 0.497.$$

类似地，可以计算出不同属性取不同值的基尼指数如下：

Gini_index(D,色泽=乌黑)=0.456

Gini_index(D,色泽=浅白)=0.426

Gini_index(D,根蒂=蜷缩) =0.456

Gini_index(D,根蒂=稍蜷) =0.496

Gini_index(D,根蒂=硬挺) =0.439

Gini_index(D,敲声=浊响) =0.450

Gini_index(D,敲声=沉闷) =0.494

Gini_index(D,敲声=清脆) =0.439

Gini_index(D,纹理=清晰) =0.286

Gini_index(D,纹理=稍稀) =0.437

Gini_index(D,纹理=模糊) =0.403

Gini_index(D,脐部=凹陷) =0.415

Gini_index(D,脐部=稍凹) =0.497

Gini_index(D,脐部=平坦) =0.362

Gini_index(D,触感=硬挺) =0.494

Gini_index(D,触感=软粘) = 0.494.

特别地，对于属性“触感”，由于它的可取值个数为2，所以其实只需计算其中一个取值的基尼指数即可.根据上面的计算结果可知，Gini_index(D,纹理=清晰) =0.286最小，所以选择属性“纹理”为最优划分属性并生成根节点，接着以“纹理=清晰”为最优划分点生成D^1(纹理=清晰)、D^2(纹理≠清晰)两个子节点，对两个子节点分别重复上述步骤继续生成下一层子节点，直至满足停止条件.

以上便是CART决策树的构建过程，从构建过程可以看出，CART决策树最终构造出来的是一棵二叉树.CART除了决策树能处理分类问题以外，回归树还可以处理回归问题，附注②中给出了CART回归树的构造算法.

式**(4.7)**

$$T_a = \left\{ \frac{a^i + a^{i+1}}{2} | 1 \leq i \leq n-1 \right\}$$

解析

此式所表达的思想很简单，就是以每两个相邻取值的中点作为划分点.

下面以“西瓜书”中表4.3中西瓜数据集3.0为例来说明此式的用法.对于“密度”这个连续属性，已观测到的可能取值为

{0.243,0.245,0.343,0.360,0.403,0.437,
0.481,0.556,0.593,0.608,0.634,0.639,0.657,0.666,0.697,0.719,0.774}共17个值，根据式(4.7)可知，此时i依次取1到16，那么“密度”这个属性的候选划分点集合为

$$T_a = \{\frac{0.243+0.245}{2}, \frac{0.245+0.343}{2}, \frac{0.343+0.360}{2}, \frac{0.360+0.403}{2}, \frac{0.403+0.437}{2}, \frac{0.437+0.481}{2}, \frac{0.481+0.556}{2}, \frac{0.556+0.593}{2}, \frac{0.593+0.608}{2}, \frac{0.608+0.634}{2}, \frac{0.634+0.639}{2}, \frac{0.639+0.657}{2}, \frac{0.657+0.666}{2}, \frac{0.666+0.697}{2}, \frac{0.697+0.719}{2}, \frac{0.719+0.774}{2}\}.$$

式**(4.8)**

$$\begin{aligned}\mathrm{Gain}(D,a) &= \max_{t\in T_a} \mathrm{Gain}(D,a,t) \\ &= \max_{t\in T_a} \mathrm{Ent}(D) - \sum_{\lambda\in\{-,+\}} \frac{|D_t^\lambda|}{|D|}\mathrm{Ent}(D_t^\lambda)\end{aligned}$$

解析

此式是式(4.2)用于离散化后的连续属性的版本，其中T_a由式(4.7)计算得来，$\lambda \in \{-,+\}$表示属性a的取值分别小于等于和大于候选划分点t时的情形，即当$\lambda = -$时有$D_t^\lambda = D_t^{a\leqslant t}$，当$\lambda = +$时有$D_t^\lambda = D_t^{a>t}$.

附注

①互信息**[1]**

在解释互信息之前，需要先解释一下什么是条件熵.条件熵表示的是在已知一个随机变量的条件下，另一个随机变量的不确定性.具体地，假设有随机变量X和Y，且它们服从以下联合概率分布

$$P(X=x_i,Y=y_j)=p_{ij},\quad i=1,2,\cdots,n,\quad j=1,2,\cdots,m.$$

那么在已知X的条件下，随机变量Y的条件熵为

$$\mathrm{Ent}(Y|X)=\sum_{i=1}^{n}p_i\,\mathrm{Ent}(Y|X=x_i),$$

其中，$p_i=P(X=x_i)i=1,2,\cdots,n$.互信息定义为信息熵和条件熵的差，它表示的是已知一个随机变量的信息后使得另一个随机变量的不确定性减少的程度.具体地，假设有随机变量X和Y，那么在已知X的信息后，Y的不确定性减少的程度为

$$\mathrm{I}(Y;X)=\mathrm{Ent}(Y)-\mathrm{Ent}(Y|X).$$

此即互信息的数学定义.

②CART回归树[1]

假设给定数据集

$$D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_N,y_N)\},$$

其中$\boldsymbol{x}\in\mathbb{R}^d$为$d$维特征向量，$y\in\mathbb{R}$是连续型随机变量. 这是一个标准的回归问题的数据集,若把每个属性视为坐标空间中的一个坐标轴，则d个属性就构成了一个d维的特征空间，而每个d维特征向量$\boldsymbol{x}$就对应了d维的特征空间中的一个数据点.CART回归树的目标是将特征空间划分成若

干个子空间，每个子空间都有一个固定的输出值，也就是凡是落在同一个子空间内的数据点$\boldsymbol{x}_i$，它们所对应的输出值y_i恒相等，且都为该子空间的输出值.那么如何划分出若干个子空间呢？这里采用一种启发式的方法.

(1) 任意选择一个属性a，遍历其所有可能取值，根据下式找出属性a最优划分点v^*：

$$v^* = \arg\min_{v} \left[\min_{c_1} \sum_{\boldsymbol{x}_i \in R_1(a,v)} (y_i - c_1)^2 + \min_{c_2} \sum_{\boldsymbol{x}_i \in R_2(a,v)} (y_i - c_2)^2 \right],$$

其中，$R_1(a,v) = \{\boldsymbol{x} | \boldsymbol{x} \in D^{a \leqslant v}\}, R_2(a,v) = \{\boldsymbol{x} | \boldsymbol{x} \in D^{a > v}\}$，$c_1$和$c_2$分别为集合$R_1(a,v)$和$R_2(a,v)$中的样本$\boldsymbol{x}_i$对应的输出值$y_i$的均值，即

$$c_1 = \text{ave}(y_i | x \in R_1(a,v)) = \frac{1}{|R_1(a,v)|} \sum_{\boldsymbol{x}_i \in R_1(a,v)} y_i;$$

$$c_2 = \text{ave}(y_i | x \in R_2(a,v)) = \frac{1}{|R_2(a,v)|} \sum_{\boldsymbol{x}_i \in R_2(a,v)} y_i.$$

(2) 遍历所有属性，找到最优划分属性a^*，然后根据a^*的最优划分点v^*将特征空间划分为两个子空间，接着对每个子空间重复上述步骤，直至满足停止条件.这样就生成了一棵CART回归树，假设最终将特征空间划分为M个子空间$R_1, R_2, \cdots, R_M$，那么CART回归树的模型式可以表示为

$$f(\boldsymbol{x}) = \sum_{m=1}^{M} c_m \mathbb{I}(x \in R_m).$$

同理，其中的c_m表示的也是集合R_m中的样本$\boldsymbol{x}_i$对应的输出值y_i的均值.此式直观上的理解就是，对于一个给定的样本$\boldsymbol{x}_i$，首先判断其属于哪个子空间，然后将其所属的子空间对应的输出值作为该样本的预测值y_i.

参考文献

[1] 李航. 统计学习方法[M]. 北京：清华大学出版社, 2012.

第5章　神经网络

式(5.2)

$$\Delta w_i = \eta(y - \hat{y})x_i$$

解析

此式是感知机学习算法中的参数更新式，下面依次给出感知机模型、学习策略和学习算法的具体介绍[1].

感知机模型

已知感知机由两层神经元组成，故感知机模型的式可表示为

$$y = f\left(\sum_{i=1}^{n} w_i x_i - \theta\right) = f(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta),$$

其中，$\boldsymbol{x} \in \mathbb{R}^n$，为样本的特征向量，是感知机模型的输入；$\boldsymbol{w}, \theta$是感知机模型的参数，$\boldsymbol{w} \in \mathbb{R}^n$，为权重，$\theta$为阈值.假定$f$为阶跃函数，那么感知机模型的式可进一步表示为

用$\varepsilon(\cdot)$代表阶跃函数

$$y = \varepsilon\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta\right) = \begin{cases} 1, & \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta \geqslant 0; \\ 0, & \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta < 0. \end{cases}$$

由于n维空间中的超平面方程为

$$w_1x_1 + w_2x_2 + \cdots + w_nx_n + b = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0,$$

所以此时感知机模型式中的$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta$可以看作是$n$维空间中的一个超平面，将$n$维空间划分为$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta \geqslant 0$和$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta < 0$两个子空间，落在前一个子空间的样本对应的模型输出值为1，落在后一个子空间的样本对应的模型输出值为0，如此便实现了分类功能.

感知机学习策略

给定一个线性可分的数据集T（参见本章附注），感知机的学习目标是求得能对数据集T中的正负样本完全正确划分的分离超平面

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta = 0.$$

假设此时误分类样本集合为$M \subseteq T$，对任意一个误分类样本$(\boldsymbol{x}, y) \in M$来说，当$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta \geqslant 0$时，模型输出值为$\hat{y} = 1$，样本真实标记为$y = 0$；反之，当$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta < 0$时，模型输出值为$\hat{y} = 0$，样本真实标记为$y = 1$.综合两种情形可知，以下式恒成立：

$$(\hat{y} - y)\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta\right) \geq 0,$$

所以，给定数据集T，其损失函数可以定义为

$$L(\boldsymbol{w}, \theta) = \sum_{\boldsymbol{x} \in M} (\hat{y} - y)\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta\right).$$

显然，此损失函数是非负的. 如果没有误分类点，则损失函数值为0.而且，误分类点越少，误分类点离超平面越近，损失函数值就越小.因此，给定数据集T，损失函数$L(\boldsymbol{w}, \theta)$是关于$\boldsymbol{w}, \theta$的连续可导函数.

感知机学习算法

感知机模型的学习问题可以转化为求解损失函数的最优化问题，具体地，给定数据集

$$T = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\},$$

其中$\boldsymbol{x}_i \in \mathbb{R}^n, y_i \in \{0, 1\}$，求参数$\boldsymbol{w}, \theta$，使其为极小化损失函数的解：

$$\min_{\boldsymbol{w},\theta} L(\boldsymbol{w}, \theta) = \min_{\boldsymbol{w},\theta} \sum_{\boldsymbol{x_i} \in M} (\hat{y}_i - y_i)(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i - \theta),$$

其中$M \subseteq T$为误分类样本集合.若将阈值θ看作一个固定输入为-1的“哑节点”，即

$$-\theta = -1 \cdot w_{n+1} = x_{n+1} \cdot w_{n+1},$$

那么$\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i - \theta$可化简为

$$\begin{aligned} \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x_i} - \theta &= \sum_{j=1}^{n} w_j x_j + x_{n+1} \cdot w_{n+1} \\ &= \sum_{j=1}^{n+1} w_j x_j \\ &= \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x_i} \end{aligned}$$

其中$\boldsymbol{x_i} \in \mathbb{R}^{n+1}, \boldsymbol{w} \in \mathbb{R}^{n+1}$. 根据该式，可将要求解的极小化问题进一步简化为

$$\min_{\boldsymbol{w}} L(\boldsymbol{w}) = \min_{\boldsymbol{w}} \sum_{\boldsymbol{x_i} \in M} (\hat{y}_i - y_i) \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x_i},$$

假设误分类样本集合M固定，那么可以求得损失函数$L(\boldsymbol{w})$的梯度

$$\nabla_{\boldsymbol{w}} L(\boldsymbol{w}) = \sum_{\boldsymbol{x_i} \in M} (\hat{y}_i - y_i)\boldsymbol{x_i}.$$

感知机的学习算法具体采用的是随机梯度下降法，即在极小化过程中，不是一次使M中所有误分类点的梯度下降，而是一次随机选取一个误分类点并使其梯度下降.所以权重$\boldsymbol{w}$的更新式为

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \Delta \boldsymbol{w},$$

$$\Delta \boldsymbol{w} = -\eta(\hat{y}_i - y_i)\boldsymbol{x}_i = \eta(y_i - \hat{y}_i)\boldsymbol{x}_i.$$

相应地，$\boldsymbol{w}$中的某个分量w_i的更新式即式(5.2).

式**(5.10)**

$$g_j = -\frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j}$$

$$= -\left(\hat{y}_j^k - y_j^k\right) f'\left(\beta_j - \theta_j\right)$$

$$= \hat{y}_j^k\left(1 - \hat{y}_j^k\right)\left(y_j^k - \hat{y}_j^k\right)$$

参见式(5.12)

式**(5.12)**

$$\Delta\theta_j = -\eta g_j$$

解析

因为

$$\Delta\theta_j = -\eta \frac{\partial E_k}{\partial \theta_j},$$

又

$$\frac{\partial E_k}{\partial \theta_j} = \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \theta_j}$$

$$= \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \left[f\left(\beta_j - \theta_j\right) \right]}{\partial \theta_j}$$

$$= \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot f'\left(\beta_j - \theta_j\right) \times (-1)$$

$$= \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot f\left(\beta_j - \theta_j\right) \times \left[1 - f\left(\beta_j - \theta_j\right)\right] \times (-1)$$

$$= \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \hat{y}_j^k \left(1 - \hat{y}_j^k\right) \times (-1)$$

$$= \frac{\partial \left[\frac{1}{2} \sum_{j=1}^{l} \left(\hat{y}_j^k - y_j^k\right)^2 \right]}{\partial \hat{y}_j^k} \cdot \hat{y}_j^k \left(1 - \hat{y}_j^k\right) \times (-1)$$

$$= \frac{1}{2} \times 2 \left(\hat{y}_j^k - y_j^k\right) \times 1 \cdot \hat{y}_j^k \left(1 - \hat{y}_j^k\right) \times (-1)$$

$$= \left(y_j^k - \hat{y}_j^k\right) \hat{y}_j^k \left(1 - \hat{y}_j^k\right)$$

$$= g_j,$$

所以

$$\Delta\theta_j = -\eta \frac{\partial E_k}{\partial \theta_j} = -\eta g_j.$$

式**(5.13**）

$$\Delta v_{ih} = \eta e_h x_i$$

解析

因为

$$\Delta v_{ih} = -\eta \frac{\partial E_k}{\partial v_{ih}},$$

又

$$\frac{\partial E_k}{\partial v_{ih}} = \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \cdot \frac{\partial \beta_j}{\partial b_h} \cdot \frac{\partial b_h}{\partial \alpha_h} \cdot \frac{\partial \alpha_h}{\partial v_{ih}}$$

$$= \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \cdot \frac{\partial \beta_j}{\partial b_h} \cdot \frac{\partial b_h}{\partial \alpha_h} \cdot x_i$$

$$= \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \cdot \frac{\partial \beta_j}{\partial b_h} \cdot f'(\alpha_h - \gamma_h) \cdot x_i$$

$$= \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \cdot w_{hj} \cdot f'(\alpha_h - \gamma_h) \cdot x_i$$

$$= \sum_{j=1}^{l} (-g_j) \cdot w_{hj} \cdot f'(\alpha_h - \gamma_h) \cdot x_i$$

$$= -f'(\alpha_h - \gamma_h) \cdot \sum_{j=1}^{l} g_j \cdot w_{hj} \cdot x_i$$

$$= -b_h(1 - b_h) \cdot \sum_{j=1}^{l} g_j \cdot w_{hj} \cdot x_i$$

$$= -e_h \cdot x_i$$

所以

$$\Delta v_{ih} = -\eta \frac{\partial E_k}{\partial v_{ih}} = \eta e_h x_i.$$

式(5.14)

$$\Delta \gamma_h = -\eta e_h$$

解析

因为

$$\Delta \gamma_h = -\eta \frac{\partial E_k}{\partial \gamma_h},$$

又

$$\frac{\partial E_k}{\partial \gamma_h} = \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \cdot \frac{\partial \beta_j}{\partial b_h} \cdot \frac{\partial b_h}{\partial \gamma_h}$$

$$= \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \cdot \frac{\partial \beta_j}{\partial b_h} \cdot f'(\alpha_h - \gamma_h) \cdot (-1)$$

$$= -\sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \cdot w_{hj} \cdot f'(\alpha_h - \gamma_h)$$

$$= -\sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \cdot w_{hj} \cdot b_h(1 - b_h)$$

$$= \sum_{j=1}^{l} g_j \cdot w_{hj} \cdot b_h(1 - b_h)$$

$$= e_h,$$

所以

$$\Delta\gamma_h = -\eta\frac{\partial E_k}{\partial\gamma_h} = -\eta e_h.$$

式**(5.15)**

$$e_h = -\frac{\partial E_k}{\partial b_h}\cdot\frac{\partial b_h}{\partial\alpha_h}$$

$$= -\sum_{j=1}^{l}\frac{\partial E_k}{\partial\beta_j}\cdot\frac{\partial\beta_j}{\partial b_h}f'(\alpha_h-\gamma_h)$$

$$= \sum_{j=1}^{l} w_{hj}g_j f'(\alpha_h-\gamma_h)$$

$$= b_h(1-b_h)\sum_{j=1}^{l} w_{hj}g_j$$

参见式(5.13)

式**(5.20)**

$$E(\boldsymbol{s}) = -\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_{ij}s_i s_j - \sum_{i=1}^{n}\theta_i s_i$$

解析

Boltzmann机（Restricted Boltzmann Machin e，简称RBM）本质上是一个引入了隐变量的无向图模型，其能量可理解为

$$E_{\text{graph}} = E_{\text{edges}} + E_{\text{nodes}},$$

其中，E_{graph}表示图的能量，E_{edges}表示图中边的能量，E_{nodes}表示图中结点的能量.边能量由两连接结点的值及其权重的乘积确定， 即

$E_{\text{edge}_{ij}} = -w_{ij}s_is_j$；结点能量由结点的值及其阈值的乘积确定，即 $E_{\text{node}_i} = -\theta_i s_i$.图中边的能量为所有边能量之和为

$$E_{\text{edges}} = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E_{\text{edge}_{ij}} = -\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_{ij}s_is_j,$$

图中结点的能量为所有结点能量之和

$$E_{\text{nodes}} = \sum_{i=1}^{n} E_{\text{node}_i} = -\sum_{i=1}^{n}\theta_i s_i,$$

故状态向量$\boldsymbol{s}$所对应的Boltzmann机能量

$$E_{\text{graph}} = E_{\text{edges}} + E_{\text{nodes}} = -\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_{ij}s_is_j - \sum_{i=1}^{n}\theta_i s_i.$$

式**(5.22)**

$$P(\boldsymbol{v}|\boldsymbol{h}) = \prod_{i=1}^{d} P(v_i \,|\, \boldsymbol{h})$$

解析

受限Boltzmann机仅保留显层与隐层之间的连接.显层状态向量 $\boldsymbol{v} = (v_1; v_2; \cdots; v_d)$，隐层状态向量$\boldsymbol{h} = (h_1; h_2; \cdots; h_q)$.显层状态向量$\boldsymbol{v}$中的变量$v_i$仅与隐层状态向量$\boldsymbol{h}$有关，所以给定隐层状态向量$\boldsymbol{h}$，有 $v_1, v_2, \cdots, v_d$相互独立.

式**(5.23）**

$$P(\boldsymbol{h}|\boldsymbol{v}) = \prod_{j=1}^{q} P(h_i \,|\, \boldsymbol{v})$$

解析

由式(5.22)的解析同理可得，给定显层状态向量$\boldsymbol{v}$，有$h_1, h_2, \cdots, h_q$相互独立.

式(5.24）

$$\Delta w = \eta(\boldsymbol{v}\boldsymbol{h}^{\mathrm{T}} - \boldsymbol{v}'\boldsymbol{h}'^{\mathrm{T}})$$

解析

由式(5.20)可推导出受限Boltzmann机的能量函数

$$\begin{aligned} E(\boldsymbol{v}, \boldsymbol{h}) &= -\sum_{i=1}^{d}\sum_{j=1}^{q} w_{ij} v_i h_j - \sum_{i=1}^{d} \alpha_i v_i - \sum_{j=1}^{q} \beta_j h_j \\ &= -\boldsymbol{h}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{v} - \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{v} - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{h} \end{aligned}$$

其中

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_q \end{bmatrix} \in \mathbb{R}^{q \times d}$$

再由式(5.21)可知，RBM的联合概率分布

$$P(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z} \mathrm{e}^{-E(\boldsymbol{v}, \boldsymbol{h})},$$

其中Z为规范化因子

$$Z=\sum_{\boldsymbol{v},\boldsymbol{h}}\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}.$$

给定含m个独立同分布数据的数据集$V=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m\}$，记$\boldsymbol{\theta}=\{\boldsymbol{W},\boldsymbol{\alpha},\boldsymbol{\beta}\}$.学习RBM的策略是求出参数$\boldsymbol{\theta}$的值，使得如下对数似然函数最大化：

$$L(\boldsymbol{\theta})=\ln\left(\prod_{k=1}^{m}P\left(\boldsymbol{v}_k\right)\right)$$

$$=\sum_{k=1}^{m}\ln P\left(\boldsymbol{v}_k\right)$$

$$=\sum_{k=1}^{m}L_k(\boldsymbol{\theta}).$$

具体地，采用梯度上升法来求解参数$\boldsymbol{\theta}$，因此考虑求对数似然函数$L(\boldsymbol{\theta})$的梯度.对于V中的任意一个样本$\boldsymbol{v}_k$，其对应似然函数

$$L_k(\boldsymbol{\theta})=\ln P\left(\boldsymbol{v}_k\right)$$

$$=\ln\left(\sum_{\boldsymbol{h}}P\left(\boldsymbol{v}_k,\boldsymbol{h}\right)\right)$$

$$=\ln\left(\sum_{\boldsymbol{h}}\frac{1}{Z}\mathrm{e}^{-E(\boldsymbol{v}_k,\boldsymbol{h})}\right)$$

$$=\ln\left(\sum_{\boldsymbol{h}}\mathrm{e}^{-E(\boldsymbol{v}_k,\boldsymbol{h})}\right)-\ln Z$$

$$= \ln\left(\sum_{\boldsymbol{h}} \mathrm{e}^{-E(\boldsymbol{v}_k,\boldsymbol{h})}\right) - \ln\left(\sum_{\boldsymbol{v},\boldsymbol{h}} \mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}\right),$$

对$L_k(\boldsymbol{\theta})$求导有

$$\frac{\partial L_k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}}\left[\ln \sum_{\boldsymbol{h}} \mathrm{e}^{-E(\boldsymbol{v}_k,\boldsymbol{h})}\right] - \frac{\partial}{\partial \boldsymbol{\theta}}\left[\ln \sum_{\boldsymbol{v},\boldsymbol{h}} \mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}\right]$$

$$= -\frac{\sum_{\boldsymbol{h}} \mathrm{e}^{-E(\boldsymbol{v}_k,\boldsymbol{h})} \frac{\partial E(\boldsymbol{v}_k,\boldsymbol{h})}{\partial \boldsymbol{\theta}}}{\sum_{\boldsymbol{h}} \mathrm{e}^{-E(\boldsymbol{v}_k,\boldsymbol{h})}} + \frac{\sum_{\boldsymbol{v},\boldsymbol{h}} \mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})} \frac{\partial E(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}}{\sum_{\boldsymbol{v},\boldsymbol{h}} \mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}}$$

$$= -\sum_{\boldsymbol{h}} \frac{\mathrm{e}^{-E(\boldsymbol{v}_k,\boldsymbol{h})} \frac{\partial E(\boldsymbol{v}_k,\boldsymbol{h})}{\partial \boldsymbol{\theta}}}{\sum_{\boldsymbol{h}} \mathrm{e}^{-E(\boldsymbol{v}_k,\boldsymbol{h})}} + \sum_{\boldsymbol{v},\boldsymbol{h}} \frac{\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})} \frac{\partial E(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}}{\sum_{\boldsymbol{v},\boldsymbol{h}} \mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}}.$$

由于

$$\frac{\mathrm{e}^{-E(\boldsymbol{v}_k,\boldsymbol{h})}}{\sum_{\boldsymbol{h}} \mathrm{e}^{-E(\boldsymbol{v}_k,\boldsymbol{h})}} = \frac{\frac{\mathrm{e}^{-E(\boldsymbol{v}_k,\boldsymbol{h})}}{Z}}{\frac{\sum_{\boldsymbol{h}} \mathrm{e}^{-E(\boldsymbol{v}_k,\boldsymbol{h})}}{Z}} = \frac{\frac{\mathrm{e}^{-E(\boldsymbol{v}_k,\boldsymbol{h})}}{Z}}{\sum_{\boldsymbol{h}} \frac{\mathrm{e}^{-E(\boldsymbol{v}_k,\boldsymbol{h})}}{Z}} = \frac{P(\boldsymbol{v}_k,\boldsymbol{h})}{\sum_{\boldsymbol{h}} P(\boldsymbol{v}_k,\boldsymbol{h})} = P(\boldsymbol{h} \mid \boldsymbol{v}_k),$$

$$\frac{\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}}{\sum_{\boldsymbol{v},\boldsymbol{h}} \mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}} = \frac{\frac{\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z}}{\frac{\sum_{\boldsymbol{v},\boldsymbol{h}} \mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z}} = \frac{\frac{\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z}}{\sum_{\boldsymbol{v},\boldsymbol{h}} \frac{\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z}} = \frac{P(\boldsymbol{v},\boldsymbol{h})}{\sum_{\boldsymbol{v},\boldsymbol{h}} P(\boldsymbol{v},\boldsymbol{h})} = P(\boldsymbol{v},\boldsymbol{h}),$$

故

$$\frac{\partial L_k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\sum_{\boldsymbol{h}} P(\boldsymbol{h} \mid \boldsymbol{v}_k) \frac{\partial E(\boldsymbol{v}_k,\boldsymbol{h})}{\partial \boldsymbol{\theta}} + \sum_{\boldsymbol{v},\boldsymbol{h}} P(\boldsymbol{v},\boldsymbol{h}) \frac{\partial E(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}$$

$$= -\sum_{\boldsymbol{h}} P(\boldsymbol{h} \mid \boldsymbol{v}_k) \frac{\partial E(\boldsymbol{v}_k,\boldsymbol{h})}{\partial \boldsymbol{\theta}} + \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} P(\boldsymbol{v}) P(\boldsymbol{h} \mid \boldsymbol{v}) \frac{\partial E(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}$$

$$= -\sum_{\boldsymbol{h}} P\left(\boldsymbol{h} \mid \boldsymbol{v}_k\right) \frac{\partial E\left(\boldsymbol{v}_k, \boldsymbol{h}\right)}{\partial \boldsymbol{\theta}} + \sum_{\boldsymbol{v}} P(\boldsymbol{v}) \sum_{\boldsymbol{h}} P(\boldsymbol{h} \mid \boldsymbol{v}) \frac{\partial E(\boldsymbol{v}, \boldsymbol{h})}{\partial \boldsymbol{\theta}}.$$

$\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$包含三个参数，在这里我们仅以$\boldsymbol{W}$中的任意一个分量$w_{ij}$为例进行详细推导.首先将上式中的$\boldsymbol{\theta}$替换为$w_{ij}$可得

$$\frac{\partial L_k(\boldsymbol{\theta})}{\partial w_{ij}} = -\sum_{\boldsymbol{h}} P(\boldsymbol{h}|\boldsymbol{v}_k) \frac{\partial E(\boldsymbol{v}_k, \boldsymbol{h})}{\partial w_{ij}} + \sum_{\boldsymbol{v}} P(\boldsymbol{v}) \sum_{\boldsymbol{h}} P(\boldsymbol{h}|\boldsymbol{v}) \frac{\partial E(\boldsymbol{v}, \boldsymbol{h})}{\partial w_{ij}}.$$

根据式(5.23)可知

$$\sum_{\boldsymbol{h}} P(\boldsymbol{h} \mid \boldsymbol{v}) \frac{\partial E(\boldsymbol{v}, \boldsymbol{h})}{\partial w_{ij}}$$

$$= -\sum_{\boldsymbol{h}} P(\boldsymbol{h} \mid \boldsymbol{v}) h_i v_j$$

$$= -\sum_{\boldsymbol{h}} \prod_{l=1}^{q} P\left(h_l \mid \boldsymbol{v}\right) h_i v_j$$

$$= -\sum_{\boldsymbol{h}} P\left(h_i \mid \boldsymbol{v}\right) \prod_{l=1, l \neq i}^{q} P\left(h_l \mid \boldsymbol{v}\right) h_i v_j$$

$$= -\sum_{\boldsymbol{h}} P\left(h_i \mid \boldsymbol{v}\right) P\left(h_1, \cdots, h_{i-1}, h_{i+1}, \cdots, h_q \mid \boldsymbol{v}\right) h_i v_j$$

$$= -\sum_{h_i} P\left(h_i \mid \boldsymbol{v}\right) h_i v_j \sum_{h_1, \cdots, h_{i-1}, h_{i+1}, \cdots, h_q} P\left(h_1, \cdots, h_{i-1}, h_{i+1}, \cdots, h_q \mid \boldsymbol{v}\right)$$

$$= -\sum_{h_i} P\left(h_i \mid \boldsymbol{v}\right) h_i v_j \cdot 1$$

$$= -\left[P\left(h_i = 0 \mid \boldsymbol{v}\right) \cdot 0 \cdot v_j + P\left(h_i = 1 \mid \boldsymbol{v}\right) \cdot 1 \cdot v_j\right]$$

$$= -P\left(h_i = 1 \mid \boldsymbol{v}\right) v_j$$

同理可推得

$$\sum_{\boldsymbol{h}} P(\boldsymbol{h}|\boldsymbol{v}_k)\frac{\partial E(\boldsymbol{v}_k,\boldsymbol{h})}{\partial w_{ij}} = -P(h_i=1|\boldsymbol{v}_k)v_j^k,$$

将以上两式代回 $\frac{\partial L_k(\boldsymbol{\theta})}{\partial w_{ij}}$ 中可得

$$\frac{\partial L_k(\boldsymbol{\theta})}{\partial w_{ij}} = P(h_i=1|\boldsymbol{v}_k)v_j^k - \sum_{\boldsymbol{v}} P(\boldsymbol{v})P(h_i=1|\boldsymbol{v})v_j.$$

观察此式可知，通过枚举所有可能的$\boldsymbol{v}$来计算$\sum_{\boldsymbol{v}} P(\boldsymbol{v})P(h_i=1|\boldsymbol{v})v_j$的复杂度太高，因此可以考虑求其近似值来简化计算.具体地，RBM通常采用的是“西瓜书”上所说的“对比散度”（Contrastive Divergence，简称CD）算法.CD算法的核心思想是：用步长为s（通常设为1）的CD函数

读者可参阅“皮果提”发布于CSDN的文章《受限玻尔兹曼机（RBM）学习笔记（六）对比散度算法》

$$\mathrm{CD}_s(\boldsymbol{\theta},\boldsymbol{v}) = -\sum_{\boldsymbol{h}} P(\boldsymbol{h}|\boldsymbol{v}^{(0)})\frac{\partial E(\boldsymbol{v}^{(0)},\boldsymbol{h})}{\partial \boldsymbol{\theta}} + \sum_{\boldsymbol{h}} P(\boldsymbol{h}|\boldsymbol{v}^{(s)})\frac{\partial E(\boldsymbol{v}^{(s)},\boldsymbol{h})}{\partial \boldsymbol{\theta}}$$

近似代替

$$\frac{\partial L_k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\sum_{\boldsymbol{h}} P(\boldsymbol{h}|\boldsymbol{v}_k)\frac{\partial E(\boldsymbol{v}_k,\boldsymbol{h})}{\partial \boldsymbol{\theta}} + \sum_{\boldsymbol{v}} P(\boldsymbol{v})\sum_{\boldsymbol{h}} P(\boldsymbol{h}|\boldsymbol{v})\frac{\partial E(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}.$$

对于w_{ij}来说，即用

$$\mathrm{CD}_s(w_{ij},\boldsymbol{v}) = P(h_i=1|\boldsymbol{v}^{(0)})v_j^{(0)} - P(h_i=1|\boldsymbol{v}^{(s)})v_j^{(s)}$$

近似代替

$$\frac{\partial L_k(\boldsymbol{\theta})}{\partial w_{ij}} = P(h_i = 1|\boldsymbol{v}_k)v_j^k - \sum_{\boldsymbol{v}} P(\boldsymbol{v})P(h_i = 1|\boldsymbol{v})v_j.$$

令 $\Delta w_{ij} = \frac{\partial L_k(\boldsymbol{\theta})}{\partial w_{ij}}$，$\mathrm{RBM}(\boldsymbol{\theta})$表示参数为$\boldsymbol{\theta}$的RBM网络，则$\mathrm{CD}_s(w_{ij}, \boldsymbol{v})$的具体算法如图5-1所示.

输入：步长s;

数据集$V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_m\}$;

参数为$\boldsymbol{\theta}$的RBM网络RBM $(\boldsymbol{\theta})$.

过程：

1：初始化$\Delta w_{ij} = 0$

2：**for** $\boldsymbol{v} \in V$ **do**

3：　$\boldsymbol{v}^{(0)} = \boldsymbol{v}$

4：　**for**　$t = 1, 2, \cdots, s-1$　**do**

5：　　$\boldsymbol{h}^{(t)} = h\_given\_v(\boldsymbol{v}^{(t)}, \mathrm{RBM}(\boldsymbol{\theta}))$

6：　　$\boldsymbol{v}^{(t+1)} = v\_given\_h(\boldsymbol{h}^{(t)}, \mathrm{RBM}(\boldsymbol{\theta}))$

7：　**end**　**for**

8： **for** $i=1,2,\cdots,q;j=1,2,\cdots,d$ **do**

9： $\Delta w_{ij}=\Delta w_{ij}+\left[P(h_i=1|\boldsymbol{v}^{(0)})v_j^{(0)}-P(h_i=1|\boldsymbol{v}^{(s)})v_j^{(s)}\right]$

10： **end for**

11： **end for**

输出： Δw_{ij}

图5-1 CD算法

图5-1中函数$h_given_v(\boldsymbol{v},\mathrm{RBM}(\boldsymbol{\theta}))$表示在给定$\boldsymbol{v}$的条件下，从$\mathrm{RBM}(\boldsymbol{\theta})$中采样生成$\boldsymbol{h}$，同理，函数$v_given_h(\boldsymbol{h},\mathrm{RBM}(\boldsymbol{\theta}))$表示在给定$\boldsymbol{h}$的条件下，从$\mathrm{RBM}(\boldsymbol{\theta})$中采样生成$\boldsymbol{v}$.由于两个函数的算法可以互相类比推得，因此仅给出函数$h_given_v(\boldsymbol{v},\mathrm{RBM}(\boldsymbol{\theta}))$的具体算法，如图5-2所示.

综上可知，式(5.24)其实就是带有学习率η的Δw_{ij}的一种形式化表示.

输入：显层状态向量$\boldsymbol{v}$;

参数为$\boldsymbol{\theta}$的RBM网络$\mathrm{RBM}(\boldsymbol{\theta})$.

过程：

1: **for** $i=1,2,\cdots,q$ **do**

2: 随机生成$0\leqslant\alpha_i\leqslant 1$

3: $h_j = \begin{cases} 1, & \text{if } \alpha_i < P\left(h_i = 1 \mid \boldsymbol{v}\right); \\ 0, & \text{otherwise} \end{cases}$

4: **end for**

输出：$\boldsymbol{h} = (h_1; h_2; \cdots; h_q)$

图5-2 $h_given_v(\boldsymbol{v}, \mathrm{RBM}(\boldsymbol{\theta}))$算法

附注

数据集的线性可分**[1]**

给定一个数据集

$$T = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\},$$

其中$\boldsymbol{x}_i \in \mathbb{R}^n, y_i \in \{0, 1\}, i = 1, 2, \cdots, N$. 如果存在某个超平面

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0$$

能将数据集T中的正样本和负样本完全正确地划分到超平面两侧，即对所有$y_i = 1$的样本$\boldsymbol{x}_i$有$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b \geq 0$，对所有$y_i = 0$的样本$\boldsymbol{x}_i$有$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b < 0$，则称数据集$T$线性可分，否则称数据集$T$线性不可分.

参考文献

[1] 李航. 统计学习方法[M]. 北京：清华大学出版社, 2012.

第6章　支持向量机

式(6.9)

$$\boldsymbol{w}=\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i$$

解析

式(6.8)可作如下展开：

$$L(\boldsymbol{w},b,\boldsymbol{\alpha})=\frac{1}{2}\|\boldsymbol{w}\|^2+\sum_{i=1}^{m}\alpha_i\left(1-y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)\right)$$

$$=\frac{1}{2}\|\boldsymbol{w}\|^2+\sum_{i=1}^{m}\left(\alpha_i-\alpha_i y_i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i-\alpha_i y_i b\right)$$

$$=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\sum_{i=1}^{m}\alpha_i-\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i-\sum_{i=1}^{m}\alpha_i y_i b.$$

对$\boldsymbol{w}$和b分别求偏导数并分别令其等于**0**和0，有

$$\frac{\partial L}{\partial \boldsymbol{w}}=\frac{1}{2}\times 2\times\boldsymbol{w}+\mathbf{0}-\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{x}_i-\mathbf{0}=\mathbf{0}\Rightarrow\boldsymbol{w}=\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{x}_i$$

$$\frac{\partial L}{\partial b}=0+0-0-\sum_{i=1}^{m}\alpha_i y_i=0\Rightarrow\sum_{i=1}^{m}\alpha_i y_i=0$$

值得一提的是，上述求解过程遵循的是“西瓜书”附录B中式(B.7)左侧的那段话：“在推导对偶问题时，常通过将拉格朗日函数$L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu})$对$\boldsymbol{x}$

求导并令导数为0，来获得对偶函数的表达形式.”那么这段话背后的缘由是什么呢？在这里我们认为有两种说法可以进行解释：

(1) 对于强对偶性成立的优化问题，其主问题的最优解$\boldsymbol{x}^*$一定满足本章附注给出的KKT条件，而KKT条件中的条件(1)就要求最优解$\boldsymbol{x}^*$能使得拉格朗日函数$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$关于$\boldsymbol{x}$的一阶导数等于0；

证明参见参考文献[1]的5.5节

(2) 对于任意优化问题，若拉格朗日函数$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$是关于$\boldsymbol{x}$的凸函数，那么此时对$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$关于$\boldsymbol{x}$求导并令导数等于0解出来的点一定是最小值点. 根据对偶函数的定义，将最小值点代回$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$即可得到对偶函数.

显然，对于SVM来说，从以上任意一种说法都能解释得通.

式**(6.10)**

$$0 = \sum_{i=1}^{m} \alpha_i y_i$$

参见式(6.9)

式**(6.11**）

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j$$

$$\text{s.t.} \sum_{i=1}^{m} \alpha_i y_i = 0,$$

$$\alpha_i \geqslant 0, \quad i = 1, 2, \cdots, m$$

解析

将式(6.9)和式(6.10)代入式(6.8)即可将$L(\boldsymbol{w}, b, \boldsymbol{\alpha})$中的$\boldsymbol{w}$和$b$消去，再考虑式(6.10)的约束，就得到了式(6.6)的对偶问题：

$$\inf_{\boldsymbol{w},b} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i y_i \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i - \sum_{i=1}^{m}\alpha_i y_i b$$

$$= \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i - \boldsymbol{w}^{\mathrm{T}}\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i + \sum_{i=1}^{m}\alpha_i - b\sum_{i=1}^{m}\alpha_i y_i$$

$$= -\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i + \sum_{i=1}^{m}\alpha_i - b\sum_{i=1}^{m}\alpha_i y_i$$

由于$\sum_{i=1}^{m}\alpha_i y_i = 0$，所以上式最后一项可化为0，于是得

$$\inf_{\boldsymbol{w},b} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = -\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i + \sum_{i=1}^{m}\alpha_i$$

$$= -\frac{1}{2}\left(\sum_{i=1}^{m}\alpha_i y_i x_i\right)^{\mathrm{T}}\left(\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i\right) + \sum_{i=1}^{m}\alpha_i$$

$$= -\frac{1}{2}\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i^{\mathrm{T}}\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i + \sum_{i=1}^{m}\alpha_i$$

$$= \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j,$$

所以

$$\max_{\boldsymbol{\alpha}} \inf_{\boldsymbol{w},b} L(\boldsymbol{w},b,\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j.$$

式**(6.13)**

$$\begin{cases} \alpha_i \geqslant 0 \\ y_i f\left(\boldsymbol{x}_i\right) - 1 \geqslant 0 \\ \alpha_i\left(y_i f\left(\boldsymbol{x}_i\right) - 1\right) = 0 \end{cases}$$

参见式(6.9)中给出的第1点理由

式**(6.35**）

$$\min_{\boldsymbol{w},b,\xi_i} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t. } y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) \geqslant 1 - \xi_i,$$

$$\xi_i \geqslant 0, i = 1, 2, \cdots, m$$

解析

令

$$\max\left(0, 1 - y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right)\right) = \xi_i,$$

显然$\xi_i \geqslant 0$. 且当$1 - y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) > 0$时有

$$1 - y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) = \xi_i,$$

当$1 - y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) \leqslant 0$时有

$$\xi_i = 0.$$

综上可得

$$1 - y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) \leqslant \xi_i \Rightarrow y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) \geqslant 1 - \xi_i.$$

式**(6.37**）

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i$$

参见式(6.9)

式**(6.38**）

$$0 = \sum_{i=1}^{m} \alpha_i y_i$$

参见式(6.10)

式**(6.39)**

$$C = \alpha_i + \mu_i$$

解析

式(6.36)关于ξ_i求偏导并令其等于0可得

$$\frac{\partial L}{\partial \xi_i} = 0 + C \times 1 - \alpha_i \times 1 - \mu_i \times 1 = 0 \Rightarrow C = \alpha_i + \mu_i.$$

式**(6.40**）

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j$$

$$\text{s.t.} \sum_{i=1}^{m} \alpha_i y_i = 0,$$

$$0 \leqslant \alpha_i \leqslant C, \quad i = 1, 2, \cdots, m$$

解析

将式(6.37)~(6.39)代入式(6.36)可以得到式(6.35)的对偶问题，有

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{m} \xi_i + \sum_{i=1}^{m} \alpha_i \left(1 - \xi_i - y_i \left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b\right)\right) - \sum_{i=1}^{m} \mu_i \xi_i$$

$$= \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{m} \alpha_i \left(1 - y_i \left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b\right)\right) + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i \xi_i - \sum_{i=1}^{m} \mu_i \xi_i$$

$$= -\frac{1}{2} \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i^{\mathrm{T}} \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i + \sum_{i=1}^{m} \alpha_i + \sum_{i=1}^{m} C \xi_i - \sum_{i=1}^{m} \alpha_i \xi_i - \sum_{i=1}^{m} \mu_i \xi_i$$

$$= -\frac{1}{2} \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i^{\mathrm{T}} \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i + \sum_{i=1}^{m} \alpha_i + \sum_{i=1}^{m} \left(C - \alpha_i - \mu_i\right) \xi_i$$

$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j$$

$$= \min_{\boldsymbol{w}, b, \boldsymbol{\xi}} L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}),$$

所以

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\mu}} \min_{\boldsymbol{w}, b, \boldsymbol{\xi}} L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}) = \max_{\boldsymbol{\alpha}, \boldsymbol{\mu}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j$$

$$= \max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j .$$

又因为 $\alpha_i \geqslant 0$，$\mu_i \geqslant 0$，$C = \alpha_i + \mu_i$，消去μ_i可得等价约束条件

$$0 \leqslant \alpha_i \leqslant C, \quad i = 1, 2, \cdots, m.$$

式**(6.41)**

$$\begin{cases} \alpha_i \geqslant 0, \mu_i \geqslant 0 \\ y_i f(\boldsymbol{x}_i) - 1 + \xi_i \geqslant 0 \\ \alpha_i (y_i f(\boldsymbol{x}_i) - 1 + \xi_i) = 0 \\ \xi_i \geqslant 0, \mu_i \xi_i = 0 \end{cases}$$

参见式(6.13)

式**(6.52**）

$$\begin{cases} \alpha_i (f(\boldsymbol{x}_i) - y_i - \epsilon - \xi_i) = 0 \\ \hat{\alpha}_i \left(y_i - f(\boldsymbol{x}_i) - \epsilon - \hat{\xi}_i \right) = 0 \\ \alpha_i \hat{\alpha}_i = 0, \xi_i \hat{\xi}_i = 0 \\ (C - \alpha_i) \xi_i = 0, (C - \hat{\alpha}_i) \hat{\xi}_i = 0 \end{cases}$$

解析

将式(6.45)的约束条件全部恒等变形为小于等于0的形式可得

$$\begin{cases} f(\boldsymbol{x}_i) - y_i - \epsilon - \xi_i \leqslant 0, \\ y_i - f(\boldsymbol{x}_i) - \epsilon - \hat{\xi}_i \leqslant 0, \\ -\xi_i \leqslant 0, \\ -\hat{\xi}_i \leqslant 0. \end{cases}$$

由于以上四个约束条件的拉格朗日乘子分别为$\alpha_i, \hat{\alpha}_i, \mu_i, \hat{\mu}_i$，由本章

附注可知，以上四个约束条件可相应转化为KKT条件

$$\begin{cases} \alpha_i \left(f\left(\boldsymbol{x}_i\right)-y_i-\epsilon-\xi_i\right)=0, \\ \hat{\alpha}_i\left(y_i-f\left(\boldsymbol{x}_i\right)-\epsilon-\hat{\xi}_i\right)=0, \\ -\mu_i \xi_i=0 \Rightarrow \mu_i \xi_i=0, \\ -\hat{\mu}_i \hat{\xi}_i=0 \Rightarrow \hat{\mu}_i \hat{\xi}_i=0. \end{cases}$$

又由式(6.49)和式(6.50)有

$$\begin{cases} \mu_i=C-\alpha_i, \\ \hat{\mu}_i=C-\hat{\alpha}_i. \end{cases}$$

所以上述KKT条件可以进一步变形为

$$\begin{cases} \alpha_i \left(f\left(\boldsymbol{x}_i\right)-y_i-\epsilon-\xi_i\right)=0, \\ \hat{\alpha}_i\left(y_i-f\left(\boldsymbol{x}_i\right)-\epsilon-\hat{\xi}_i\right)=0, \\ (C-\alpha_i)\xi_i=0, \\ (C-\hat{\alpha}_i)\hat{\xi}_i=0. \end{cases}$$

又因为样本$(\boldsymbol{x}_i, y_i)$只可能处在间隔带的某一侧，即约束条件$f\left(\boldsymbol{x}_i\right)-y_i-\epsilon-\xi_i=0$和$y_i-f\left(\boldsymbol{x}_i\right)-\epsilon-\hat{\xi}_i=0$不可能同时成立，所以$\alpha_i$和$\hat{\alpha}_i$中至少有一个为0，即$\alpha_i\hat{\alpha}_i=0$.

在此基础上再进一步分析可知，如果$\alpha_i=0$，则根据约束$(C-\alpha_i)\xi_i=0$可知此时$\xi_i=0$.同理，如果$\hat{\alpha}_i=0$，则根据约束$(C-\hat{\alpha}_i)\hat{\xi}_i=0$可知此时$\hat{\xi}_i=0$.所以$\xi_i$和$\hat{\xi}_i$中也是至少有一个为0，即$\xi_i\hat{\xi}_i=0$.将$\alpha_i\hat{\alpha}_i=0, \xi_i\hat{\xi}_i=0$整合进上述KKT条件中即可得到式(6.52).

式**(6.60)**

$$\max_{\boldsymbol{w}} J(\boldsymbol{w}) = \frac{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_b^{\phi} \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_w^{\phi} \boldsymbol{w}}$$

类似于第3章的式(3.35)

式**(6.62**）

$$\boldsymbol{S}_b^{\phi} = \left(\boldsymbol{\mu}_1^{\phi} - \boldsymbol{\mu}_0^{\phi}\right)\left(\boldsymbol{\mu}_1^{\phi} - \boldsymbol{\mu}_0^{\phi}\right)^{\mathrm{T}}$$

类似于第3章的式(3.34)

式**(6.63)**

$$\boldsymbol{S}_w^{\phi} = \sum_{i=0}^{1} \sum_{\boldsymbol{x} \in X_i} \left(\phi(\boldsymbol{x}) - \boldsymbol{\mu}_i^{\phi}\right)\left(\phi(\boldsymbol{x}) - \boldsymbol{\mu}_i^{\phi}\right)^{\mathrm{T}}$$

类似于第3章的式(3.33)

式**(6.65**）

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i \phi\left(\boldsymbol{x}_i\right)$$

解析

由表示定理可知，此时二分类KLDA最终求得的投影直线方程总可以写成如下形式：

$$h(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i \kappa\left(\boldsymbol{x}, \boldsymbol{x}_i\right).$$

又因为直线方程的固定形式为

$$h(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}),$$

所以

$$\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i \kappa\left(\boldsymbol{x}, \boldsymbol{x}_i\right).$$

将$\kappa\left(\boldsymbol{x}, \boldsymbol{x}_i\right) = \phi(\boldsymbol{x})^{\mathrm{T}}\phi(\boldsymbol{x}_i)$代入可得

$$\begin{aligned}\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}) &= \sum_{i=1}^{m} \alpha_i \phi(\boldsymbol{x})^{\mathrm{T}}\phi\left(\boldsymbol{x}_i\right) \\ &= \phi(\boldsymbol{x})^{\mathrm{T}} \cdot \sum_{i=1}^{m} \alpha_i \phi\left(\boldsymbol{x}_i\right)\end{aligned}.$$

由于$\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x})$的计算结果为标量，而标量的转置等于其本身，所以

$$\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}) = \left(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x})\right)^{\mathrm{T}} = \phi(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{w} = \phi(\boldsymbol{x})^{\mathrm{T}} \sum_{i=1}^{m} \alpha_i \phi(\boldsymbol{x}_i),$$

即

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i \phi(\boldsymbol{x}_i).$$

式(6.66)

$$\hat{\boldsymbol{\mu}}_0 = \frac{1}{m_0}\boldsymbol{K}\mathbf{1}_0$$

解析

为了详细地说明此式的计算原理，下面首先举例说明，然后再在例

子的基础上延展出其一般形式.

假设此时仅有4个样本，其中第1和第3个样本的标记为0，第2和第4个样本的标记为1，那么此时有

$$m = 4, m_0 = 2, m_1 = 2;$$

$$X_0 = \{\boldsymbol{x}_1, \boldsymbol{x}_3\}, X_1 = \{\boldsymbol{x}_2, \boldsymbol{x}_4\};$$

$$\boldsymbol{K} = \begin{bmatrix} \kappa(\boldsymbol{x}_1, \boldsymbol{x}_1) & \kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) & \kappa(\boldsymbol{x}_1, \boldsymbol{x}_3) & \kappa(\boldsymbol{x}_1, \boldsymbol{x}_4) \\ \kappa(\boldsymbol{x}_2, \boldsymbol{x}_1) & \kappa(\boldsymbol{x}_2, \boldsymbol{x}_2) & \kappa(\boldsymbol{x}_2, \boldsymbol{x}_3) & \kappa(\boldsymbol{x}_2, \boldsymbol{x}_4) \\ \kappa(\boldsymbol{x}_3, \boldsymbol{x}_1) & \kappa(\boldsymbol{x}_3, \boldsymbol{x}_2) & \kappa(\boldsymbol{x}_3, \boldsymbol{x}_3) & \kappa(\boldsymbol{x}_3, \boldsymbol{x}_4) \\ \kappa(\boldsymbol{x}_4, \boldsymbol{x}_1) & \kappa(\boldsymbol{x}_4, \boldsymbol{x}_2) & \kappa(\boldsymbol{x}_4, \boldsymbol{x}_3) & \kappa(\boldsymbol{x}_4, \boldsymbol{x}_4) \end{bmatrix} \in \mathbb{R}^{4\times 4};$$

$$\mathbf{1}_0 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \in \mathbb{R}^{4\times 1}$$

$$\mathbf{1}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \in \mathbb{R}^{4\times 1}$$

所以

$$\hat{\boldsymbol{\mu}}_0 = \frac{1}{m_0}\boldsymbol{K}\mathbf{1}_0 = \frac{1}{2}\begin{bmatrix} \kappa(\boldsymbol{x}_1, \boldsymbol{x}_1) + \kappa(\boldsymbol{x}_1, \boldsymbol{x}_3) \\ \kappa(\boldsymbol{x}_2, \boldsymbol{x}_1) + \kappa(\boldsymbol{x}_2, \boldsymbol{x}_3) \\ \kappa(\boldsymbol{x}_3, \boldsymbol{x}_1) + \kappa(\boldsymbol{x}_3, \boldsymbol{x}_3) \\ \kappa(\boldsymbol{x}_4, \boldsymbol{x}_1) + \kappa(\boldsymbol{x}_4, \boldsymbol{x}_3) \end{bmatrix} \in \mathbb{R}^{4\times 1},$$

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{m_1}\boldsymbol{K}1_1 = \frac{1}{2}\begin{bmatrix} \kappa(\boldsymbol{x}_1,\boldsymbol{x}_2)+\kappa(\boldsymbol{x}_1,\boldsymbol{x}_4) \\ \kappa(\boldsymbol{x}_2,\boldsymbol{x}_2)+\kappa(\boldsymbol{x}_2,\boldsymbol{x}_4) \\ \kappa(\boldsymbol{x}_3,\boldsymbol{x}_2)+\kappa(\boldsymbol{x}_3,\boldsymbol{x}_4) \\ \kappa(\boldsymbol{x}_4,\boldsymbol{x}_2)+\kappa(\boldsymbol{x}_4,\boldsymbol{x}_4) \end{bmatrix} \in \mathbb{R}^{4\times 1}.$$

根据此结果易得$\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1$的一般形式为

$$\hat{\boldsymbol{\mu}}_0 = \frac{1}{m_0}\boldsymbol{K}\mathbf{1}_0 = \frac{1}{m_0}\begin{bmatrix} \sum_{\boldsymbol{x}\in X_0}\kappa(\boldsymbol{x}_1,\boldsymbol{x}) \\ \sum_{\boldsymbol{x}\in X_0}\kappa(\boldsymbol{x}_2,\boldsymbol{x}) \\ \vdots \\ \sum_{\boldsymbol{x}\in X_0}\kappa(\boldsymbol{x}_m,\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^{m\times 1},$$

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{m_1}\boldsymbol{K}1_1 = \frac{1}{m_1}\begin{bmatrix} \sum_{\boldsymbol{x}\in X_1}\kappa(\boldsymbol{x}_1,\boldsymbol{x}) \\ \sum_{\boldsymbol{x}\in X_1}\kappa(\boldsymbol{x}_2,\boldsymbol{x}) \\ \vdots \\ \sum_{\boldsymbol{x}\in X_1}\kappa(\boldsymbol{x}_m,\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^{m\times 1}.$$

式**(6.67)**

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{m_1}\boldsymbol{K}\mathbf{1}_1$$

参见式(6.66)的解析

式**(6.70**）

$$\max_{\boldsymbol{\alpha}} J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{N}\boldsymbol{\alpha}}$$

解析

此式是将式(6.65)代入式(6.60)后推得而来的，下面给出详细地推导过程.

首先将式(6.65)代入式(6.60)的分子可得

$$\begin{aligned}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_b^{\phi}\boldsymbol{w} &= \left(\sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)\right)^{\mathrm{T}}\boldsymbol{S}_b^{\phi}\sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)\\ &= \sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\boldsymbol{S}_b^{\phi}\sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right),\end{aligned}$$

其中

$$\begin{aligned}\boldsymbol{S}_b^{\phi} &= \left(\boldsymbol{\mu}_1^{\phi}-\boldsymbol{\mu}_0^{\phi}\right)\left(\boldsymbol{\mu}_1^{\phi}-\boldsymbol{\mu}_0^{\phi}\right)^{\mathrm{T}}\\ &= \left(\frac{1}{m_1}\sum_{\boldsymbol{x}\in X_1}\phi(\boldsymbol{x})-\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\phi(\boldsymbol{x})\right)\left(\frac{1}{m_1}\sum_{\boldsymbol{x}\in X_1}\phi(\boldsymbol{x})-\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\phi(\boldsymbol{x})\right)^{\mathrm{T}}\\ &= \left(\frac{1}{m_1}\sum_{\boldsymbol{x}\in X_1}\phi(\boldsymbol{x})-\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\phi(\boldsymbol{x})\right)\left(\frac{1}{m_1}\sum_{\boldsymbol{x}\in X_1}\phi(\boldsymbol{x})^{\mathrm{T}}-\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\phi(\boldsymbol{x})^{\mathrm{T}}\right).\end{aligned}$$

将其代入上式可得

$$\begin{aligned}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_b^{\phi}\boldsymbol{w} &= \sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\left(\frac{1}{m_1}\sum_{\boldsymbol{x}\in X_1}\phi(\boldsymbol{x})-\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\phi(\boldsymbol{x})\right)\\ &\quad\left(\frac{1}{m_1}\sum_{\boldsymbol{x}\in X_1}\phi(\boldsymbol{x})^{\mathrm{T}}-\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\phi(\boldsymbol{x})^{\mathrm{T}}\right)\sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)\\ &= \left(\frac{1}{m_1}\sum_{\boldsymbol{x}\in X_1}\sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\phi(\boldsymbol{x})-\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\phi(\boldsymbol{x})\right)\end{aligned}$$

$$\left(\frac{1}{m_1}\sum_{\boldsymbol{x}\in X_1}\sum_{i=1}^{m}\alpha_i\phi(\boldsymbol{x})^{\mathrm{T}}\phi\left(\boldsymbol{x}_i\right)-\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\sum_{i=1}^{m}\alpha_i\phi(\boldsymbol{x})^{\mathrm{T}}\phi\left(\boldsymbol{x}_i\right)\right).$$

由于$\kappa\left(\boldsymbol{x}_i,\boldsymbol{x}\right)=\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x})$为标量，所以其转置等于本身，即$\kappa\left(\boldsymbol{x}_i,\boldsymbol{x}\right)=\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\phi(\boldsymbol{x})=\left(\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x})\right)^{\mathrm{T}}=\phi(\boldsymbol{x})^{\mathrm{T}}\phi(\boldsymbol{x}_i)=\kappa\left(\boldsymbol{x}_i,\boldsymbol{x}\right)^{\mathrm{T}}$. 将其代入上式可得

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_b^{\phi}\boldsymbol{w}=\left(\frac{1}{m_1}\sum_{i=1}^{m}\sum_{\boldsymbol{x}\in X_1}\alpha_i\kappa\left(\boldsymbol{x}_i,\boldsymbol{x}\right)-\frac{1}{m_0}\sum_{i=1}^{m}\sum_{\boldsymbol{x}\in X_0}\alpha_i\kappa\left(\boldsymbol{x}_i,\boldsymbol{x}\right)\right)$$
$$\left(\frac{1}{m_1}\sum_{i=1}^{m}\sum_{\boldsymbol{x}\in X_1}\alpha_i\kappa\left(\boldsymbol{x}_i,\boldsymbol{x}\right)-\frac{1}{m_0}\sum_{i=1}^{m}\sum_{\boldsymbol{x}\in X_0}\alpha_i\kappa\left(\boldsymbol{x}_i,\boldsymbol{x}\right)\right).$$

设$\boldsymbol{\alpha}=(\alpha_1;\alpha_2;\cdots;\alpha_m)\in\mathbb{R}^{m\times 1}$，同时结合式(6.66)的解析可得到$\hat{\boldsymbol{\mu}}_0,\hat{\boldsymbol{\mu}}_1$的一般形式，上式可以化简为

$$\begin{aligned}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_b^{\phi}\boldsymbol{w}&=\left(\boldsymbol{\alpha}^{\mathrm{T}}\hat{\boldsymbol{\mu}}_1-\boldsymbol{\alpha}^{\mathrm{T}}\hat{\boldsymbol{\mu}}_0\right)\left(\hat{\boldsymbol{\mu}}_1^{\mathrm{T}}\boldsymbol{\alpha}-\hat{\boldsymbol{\mu}}_0^{\mathrm{T}}\boldsymbol{\alpha}\right)\\&=\boldsymbol{\alpha}^{\mathrm{T}}\left(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_0\right)\left(\hat{\boldsymbol{\mu}}_1^{\mathrm{T}}-\hat{\boldsymbol{\mu}}_0^{\mathrm{T}}\right)\boldsymbol{\alpha}\\&=\boldsymbol{\alpha}^{\mathrm{T}}\left(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_0\right)\left(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_0\right)^{\mathrm{T}}\boldsymbol{\alpha}\\&=\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\alpha}.\end{aligned}$$

以上便是式(6.70)分子部分的推导，下面继续推导式(6.70)的分母部分.

将式(6.65)代入式(6.60)的分母可得：

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_w^{\phi}\boldsymbol{w}=\left(\sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)\right)^{\mathrm{T}}\boldsymbol{S}_w^{\phi}\sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)$$

$$= \sum_{i=1}^{m} \alpha_i \phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}} \boldsymbol{S}_w^{\phi} \sum_{i=1}^{m} \alpha_i \phi\left(\boldsymbol{x}_i\right)$$

其中

$$\boldsymbol{S}_w^{\phi} = \sum_{i=0}^{1} \sum_{\boldsymbol{x} \in X_i} \left(\phi(\boldsymbol{x}) - \boldsymbol{\mu}_i^{\phi}\right) \left(\phi(\boldsymbol{x}) - \boldsymbol{\mu}_i^{\phi}\right)^{\mathrm{T}}$$

$$= \sum_{i=0}^{1} \sum_{\boldsymbol{x} \in X_i} \left(\phi(\boldsymbol{x}) - \boldsymbol{\mu}_i^{\phi}\right) \left(\phi(\boldsymbol{x})^{\mathrm{T}} - \left(\boldsymbol{\mu}_i^{\phi}\right)^{\mathrm{T}}\right)$$

$$= \sum_{i=0}^{1} \sum_{\boldsymbol{x} \in X_i} \left(\phi(\boldsymbol{x})\phi(\boldsymbol{x})^{\mathrm{T}} - \phi(\boldsymbol{x}) \left(\boldsymbol{\mu}_i^{\phi}\right)^{\mathrm{T}} - \boldsymbol{\mu}_i^{\phi} \phi(\boldsymbol{x})^{\mathrm{T}} + \boldsymbol{\mu}_i^{\phi} \left(\boldsymbol{\mu}_i^{\phi}\right)^{\mathrm{T}}\right)$$

$$= \sum_{i=0}^{1} \sum_{\boldsymbol{x} \in X_i} \phi(\boldsymbol{x})\phi(\boldsymbol{x})^{\mathrm{T}} - \sum_{i=0}^{1} \sum_{\boldsymbol{x} \in X_i} \phi(\boldsymbol{x}) \left(\boldsymbol{\mu}_i^{\phi}\right)^{\mathrm{T}} -$$

$$\sum_{i=0}^{1} \sum_{\boldsymbol{x} \in X_i} \boldsymbol{\mu}_i^{\phi} \phi(\boldsymbol{x})^{\mathrm{T}} + \sum_{i=0}^{1} \sum_{\boldsymbol{x} \in X_i} \boldsymbol{\mu}_i^{\phi} \left(\boldsymbol{\mu}_i^{\phi}\right)^{\mathrm{T}}$$

由于

$$\begin{aligned} \sum_{i=0}^{1} \sum_{\boldsymbol{x} \in X_i} \phi(\boldsymbol{x}) \left(\boldsymbol{\mu}_i^{\phi}\right)^{\mathrm{T}} &= \sum_{\boldsymbol{x} \in X_0} \phi(\boldsymbol{x}) \left(\boldsymbol{\mu}_0^{\phi}\right)^{\mathrm{T}} + \sum_{\boldsymbol{x} \in X_1} \phi(\boldsymbol{x}) \left(\boldsymbol{\mu}_1^{\phi}\right)^{\mathrm{T}} \\ &= m_0 \boldsymbol{\mu}_0^{\phi} \left(\boldsymbol{\mu}_0^{\phi}\right)^{\mathrm{T}} + m_1 \boldsymbol{\mu}_1^{\phi} \left(\boldsymbol{\mu}_1^{\phi}\right)^{\mathrm{T}} \end{aligned}$$

且

$$\begin{aligned}\sum_{i=0}^{1}\sum_{\boldsymbol{x}\in X_i}\boldsymbol{\mu}_i^{\phi}\phi(\boldsymbol{x})^{\mathrm{T}} &= \sum_{i=0}^{1}\boldsymbol{\mu}_i^{\phi}\sum_{\boldsymbol{x}\in X_i}\phi(\boldsymbol{x})^{\mathrm{T}} \\ &= \boldsymbol{\mu}_0^{\phi}\sum_{\boldsymbol{x}\in X_0}\phi(\boldsymbol{x})^{\mathrm{T}} + \boldsymbol{\mu}_1^{\phi}\sum_{\boldsymbol{x}\in X_1}\phi(\boldsymbol{x})^{\mathrm{T}}, \\ &= m_0\boldsymbol{\mu}_0^{\phi}\left(\boldsymbol{\mu}_0^{\phi}\right)^{\mathrm{T}} + m_1\boldsymbol{\mu}_1^{\phi}\left(\boldsymbol{\mu}_1^{\phi}\right)^{\mathrm{T}}\end{aligned}$$

所以

$$\begin{aligned}\boldsymbol{S}_w^{\phi} = &\sum_{\boldsymbol{x}\in D}\phi(\boldsymbol{x})\phi(\boldsymbol{x})^{\mathrm{T}} - 2\left[m_0\boldsymbol{\mu}_0^{\phi}\left(\boldsymbol{\mu}_0^{\phi}\right)^{\mathrm{T}} + m_1\boldsymbol{\mu}_1^{\phi}\left(\boldsymbol{\mu}_1^{\phi}\right)^{\mathrm{T}}\right] + \\ &m_0\boldsymbol{\mu}_0^{\phi}\left(\boldsymbol{\mu}_0^{\phi}\right)^{\mathrm{T}} + m_1\boldsymbol{\mu}_1^{\phi}\left(\boldsymbol{\mu}_1^{\phi}\right)^{\mathrm{T}} \\ = &\sum_{\boldsymbol{x}\in D}\phi(\boldsymbol{x})\phi(\boldsymbol{x})^{\mathrm{T}} - m_0\boldsymbol{\mu}_0^{\phi}\left(\boldsymbol{\mu}_0^{\phi}\right)^{\mathrm{T}} - m_1\boldsymbol{\mu}_1^{\phi}\left(\boldsymbol{\mu}_1^{\phi}\right)^{\mathrm{T}}.\end{aligned}$$

再将此式代回$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_b^{\phi}\boldsymbol{w}$可得

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_w^{\phi}\boldsymbol{w} = \sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\boldsymbol{S}_w^{\phi}\sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)$$

$$= \sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\left(\sum_{\boldsymbol{x}\in D}\phi(\boldsymbol{x})\phi(\boldsymbol{x})^{\mathrm{T}} - m_0\boldsymbol{\mu}_0^{\phi}\left(\boldsymbol{\mu}_0^{\phi}\right)^{\mathrm{T}} - m_1\boldsymbol{\mu}_1^{\phi}\left(\boldsymbol{\mu}_1^{\phi}\right)^{\mathrm{T}}\right)$$

$$\sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{\boldsymbol{x}\in D}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\phi(\boldsymbol{x})\phi(\boldsymbol{x})^{\mathrm{T}}\alpha_j\phi\left(\boldsymbol{x}_j\right) -$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}m_0\boldsymbol{\mu}_0^{\phi}\left(\boldsymbol{\mu}_0^{\phi}\right)^{\mathrm{T}}\alpha_j\phi\left(\boldsymbol{x}_j\right) -$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}m_1\boldsymbol{\mu}_1^{\phi}\left(\boldsymbol{\mu}_1^{\phi}\right)^{\mathrm{T}}\alpha_j\phi\left(\boldsymbol{x}_j\right),$$

其中，第1项

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{\boldsymbol{x}\in D}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\phi(\boldsymbol{x})\phi(\boldsymbol{x})^{\mathrm{T}}\alpha_j\phi\left(\boldsymbol{x}_j\right)\\
=&\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{\boldsymbol{x}\in D}\alpha_i\alpha_j\kappa\left(\boldsymbol{x}_i,\boldsymbol{x}\right)\kappa\left(\boldsymbol{x}_j,\boldsymbol{x}\right)\\
=&\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{\alpha},
\end{aligned}$$

第2项

$$\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}m_0\boldsymbol{\mu}_0^{\phi}\left(\boldsymbol{\mu}_0^{\phi}\right)^{\mathrm{T}}\alpha_j\phi\left(\boldsymbol{x}_j\right)$$

$$=m_0\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\boldsymbol{\mu}_0^{\phi}\left(\boldsymbol{\mu}_0^{\phi}\right)^{\mathrm{T}}\phi\left(\boldsymbol{x}_j\right)$$

$$=m_0\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\left[\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\phi(\boldsymbol{x})\right]\left[\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\phi(\boldsymbol{x})\right]^{\mathrm{T}}\phi\left(\boldsymbol{x}_j\right)$$

$$=m_0\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\left[\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\phi(\boldsymbol{x})\right]\left[\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\phi(\boldsymbol{x})^{\mathrm{T}}\phi\left(\boldsymbol{x}_j\right)\right]$$

$$=m_0\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\left[\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\kappa\left(\boldsymbol{x}_i,\boldsymbol{x}\right)\right]\left[\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\kappa\left(\boldsymbol{x}_j,\boldsymbol{x}\right)\right]$$

$$=m_0\boldsymbol{\alpha}^{\mathrm{T}}\hat{\boldsymbol{\mu}}_0\hat{\boldsymbol{\mu}}_0^{\mathrm{T}}\boldsymbol{\alpha},$$

同理，有第3项

$$\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}m_1\boldsymbol{\mu}_1^{\phi}\left(\boldsymbol{\mu}_1^{\phi}\right)^{\mathrm{T}}\alpha_j\phi\left(\boldsymbol{x}_j\right)=m_1\boldsymbol{\alpha}^{\mathrm{T}}\hat{\boldsymbol{\mu}}_1\hat{\boldsymbol{\mu}}_1^{\mathrm{T}}\boldsymbol{\alpha}.$$

将上述三项的化简结果代回再将此式代回$\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_b^{\phi} \boldsymbol{w}$可得

$$\begin{aligned}
\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_b^{\phi} \boldsymbol{w} &= \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{K}^{\mathrm{T}} \boldsymbol{\alpha} - m_0 \boldsymbol{\alpha}^{\mathrm{T}} \hat{\boldsymbol{\mu}}_0 \hat{\boldsymbol{\mu}}_0^{\mathrm{T}} \boldsymbol{\alpha} - m_1 \boldsymbol{\alpha}^{\mathrm{T}} \hat{\boldsymbol{\mu}}_1 \hat{\boldsymbol{\mu}}_1^{\mathrm{T}} \boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}^{\mathrm{T}} \left(\boldsymbol{K} \boldsymbol{K}^{\mathrm{T}} - m_0 \hat{\boldsymbol{\mu}}_0 \hat{\boldsymbol{\mu}}_0^{\mathrm{T}} - m_1 \hat{\boldsymbol{\mu}}_1 \hat{\boldsymbol{\mu}}_1^{\mathrm{T}} \right) \boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}^{\mathrm{T}} \left(\boldsymbol{K} \boldsymbol{K}^{\mathrm{T}} - \sum_{i=0}^{1} m_i \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \right) \boldsymbol{\alpha} \\
&= \alpha^{\mathrm{T}} \boldsymbol{N} \boldsymbol{\alpha}.
\end{aligned}$$

附注

KKT条件[2]

考虑一般的约束优化问题

$$\begin{aligned}
\min \quad & f(\boldsymbol{x}) \\
\text{s.t.} \quad & g_i(\boldsymbol{x}) \leqslant 0 \quad (i = 1, \cdots, m), \\
& h_j(\boldsymbol{x}) = 0 \quad (j = 1, \cdots, n).
\end{aligned}$$

其中，自变量$\boldsymbol{x} \in \mathbb{R}^n$. 设$f(\boldsymbol{x}), g_i(\boldsymbol{x}), h_j(\boldsymbol{x})$具有连续的一阶偏导数，$\boldsymbol{x}^*$是优化问题的局部可行解.若该优化问题满足任意一个约束限制条件（constrain t qualifications or regularity conditions），则一定存在$\boldsymbol{\mu}^* = (\mu_1^*; \mu_2^*; \cdots; \mu_m^*), \boldsymbol{\lambda}^* = (\lambda_1^*; \lambda_2^*; \cdots; \lambda_n^*)$，使得：

参见维基百科页面“Karush-kuhn-tucker conditions”

(1) $$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) = \nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(\boldsymbol{x}^*) + \sum_{j=1}^{n} \lambda_j^* \nabla h_j(\boldsymbol{x}^*) = 0;$$

(2) $h_j(\boldsymbol{x}^*) = 0;$

(3) $g_i(\boldsymbol{x}^*) \leqslant 0$;

(4) $\mu_i^* \geqslant 0$;

(5) $\mu_i^* g_i(\boldsymbol{x}^*) = 0$;

其中$L(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda})$为拉格朗日函数

$$L(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \mu_i g_i(\boldsymbol{x}) + \sum_{j=1}^{n} \lambda_j h_j(\boldsymbol{x}).$$

以上5条即KKT条件，严格数学证明参见参考文献[2]的4.2.1小节.

参考文献

[1] 王书宁. 凸优化[M].北京：清华大学出版社, 2013.

[2] 王燕军. 最优化基础理论与方法[M]. 上海：复旦大学出版社, 2011.

第7章　贝叶斯分类器

式(7.5)

$$R(c|\boldsymbol{x}) = 1 - P(c|\boldsymbol{x})$$

解析

由式(7.1)和式(7.4)可得

$$\begin{aligned} R\left(c_i \mid \boldsymbol{x}\right) =& 1 \cdot P\left(c_1 \mid \boldsymbol{x}\right) + 1 \cdot P\left(c_2 \mid \boldsymbol{x}\right) + \cdots + 1 \cdot P\left(c_{i-1} \mid \boldsymbol{x}\right) + 0 \cdot P\left(c_i \mid \boldsymbol{x}\right) + \\ & 1 \cdot P\left(c_{i+1} \mid \boldsymbol{x}\right) + \cdots + 1 \cdot P\left(c_N \mid \boldsymbol{x}\right), \end{aligned}$$

又$\sum_{j=1}^{N} P(c_j|\boldsymbol{x}) = 1$，则

$$R(c_i|\boldsymbol{x}) = 1 - P(c_i|\boldsymbol{x}),$$

此即式(7.5).

式(7.6）

$$h^*(\boldsymbol{x}) = \arg\max_{c \in \mathcal{Y}} P(c|\boldsymbol{x})$$

将式(7.5)代入式(7.3)即可推得此式

式(7.12）

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{|D_c|} \sum_{\boldsymbol{x} \in D_c} \boldsymbol{x}$$

参见式(7.13)

式**(7.13**）

$$\hat{\boldsymbol{\sigma}}_c^2 = \frac{1}{|D_c|} \sum_{\boldsymbol{x} \in D_c} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_c)(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_c)^{\mathrm{T}}$$

解析

根据式(7.11)和式(7.10)可知参数求解式为

$$\begin{aligned}\hat{\boldsymbol{\theta}}_c &= \arg\max_{\boldsymbol{\theta}_c} LL(\boldsymbol{\theta}_c) \\ &= \arg\min_{\boldsymbol{\theta}_c} - LL(\boldsymbol{\theta}_c) \\ &= \arg\min_{\boldsymbol{\theta}_c} - \sum_{\boldsymbol{x} \in D_c} \log P(\boldsymbol{x} \mid \boldsymbol{\theta}_c).\end{aligned}$$

由“西瓜书”上下文可知，此时假设概率密度函数$p(\boldsymbol{x}|c) \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)$，其等价于假设

$$P(\boldsymbol{x}|\boldsymbol{\theta}_c) = P(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_c|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_c)^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c)\right),$$

其中，d表示$\boldsymbol{x}$的维数，$\boldsymbol{\Sigma}_c = \boldsymbol{\sigma}_c^2$为对称正定协方差矩阵，$|\boldsymbol{\Sigma}_c|$表示$\boldsymbol{\Sigma}_c$的行列式.将其代入参数求解式可得

$$\left(\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\Sigma}}_c\right) = \arg\min_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} - \sum_{\boldsymbol{x} \in D_c} \log \left[\frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_c|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_c)^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c)\right)\right]$$

$$= \arg\min_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} - \sum_{\boldsymbol{x} \in D_c} \left[-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_c| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_c)^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c)\right]$$

$$= \arg\min_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} \sum_{\boldsymbol{x} \in D_c} \left[\frac{d}{2} \log(2\pi) + \frac{1}{2} \log |\boldsymbol{\Sigma}_c| + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_c)^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c) \right]$$

$$= \arg\min_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} \sum_{\boldsymbol{x} \in D_c} \left[\frac{1}{2} \log |\boldsymbol{\Sigma}_c| + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_c)^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c) \right].$$

假设此时数据集D_c中的样本个数为n，即$|D_c| = n$，则上式可以改写为

$$\left(\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\Sigma}}_c \right) = \arg\min_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} \sum_{i=1}^{n} \left[\frac{1}{2} \log |\boldsymbol{\Sigma}_c| + \frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_c)^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_c) \right]$$

$$= \arg\min_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} \frac{n}{2} \log |\boldsymbol{\Sigma}_c| + \sum_{i=1}^{n} \frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_c)^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_c).$$

为了便于分别求解$\hat{\boldsymbol{\mu}}_c$和$\hat{\boldsymbol{\Sigma}}_c$，在这里我们根据式$\boldsymbol{x}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{x} = \mathrm{tr}(\boldsymbol{A} \boldsymbol{x} \boldsymbol{x}^{\mathrm{T}})$和$\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$将上式中的最后一项作如下恒等变形：

$$\sum_{i=1}^{n} \frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_c)^{\mathrm{T}} \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_c)$$

$$= \frac{1}{2} \mathrm{tr} \left[\boldsymbol{\Sigma}_c^{-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu}_c) (\boldsymbol{x}_i - \boldsymbol{\mu}_c)^{\mathrm{T}} \right]$$

$$= \frac{1}{2} \mathrm{tr} \left[\boldsymbol{\Sigma}_c^{-1} \sum_{i=1}^{n} \left(\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} - \boldsymbol{x}_i \boldsymbol{\mu}_c^{\mathrm{T}} - \boldsymbol{\mu}_c \boldsymbol{x}_i^{\mathrm{T}} + \boldsymbol{\mu}_c \boldsymbol{\mu}_c^{\mathrm{T}} \right) \right]$$

$$= \frac{1}{2} \mathrm{tr} \left[\boldsymbol{\Sigma}_c^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} - n \bar{\boldsymbol{x}} \boldsymbol{\mu}_c^{\mathrm{T}} - n \boldsymbol{\mu}_c \bar{\boldsymbol{x}}^{\mathrm{T}} + n \boldsymbol{\mu}_c \boldsymbol{\mu}_c^{\mathrm{T}} \right) \right]$$

$$
\begin{aligned}
&= \frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}_c^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} - 2n\overline{\boldsymbol{x}}\boldsymbol{\mu}_c^{\mathrm{T}} + n\boldsymbol{\mu}_c\boldsymbol{\mu}_c^{\mathrm{T}} + 2n\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^{\mathrm{T}} - 2n\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^{\mathrm{T}}\right)\right] \\
&= \frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}_c^{-1}\left(\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} - 2n\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^{\mathrm{T}} + n\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^{\mathrm{T}}\right) + \left(n\boldsymbol{\mu}_c\boldsymbol{\mu}_c^{\mathrm{T}} - 2n\overline{\boldsymbol{x}}\boldsymbol{\mu}_c^{\mathrm{T}} + n\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^{\mathrm{T}}\right)\right)\right] \\
&= \frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}_c^{-1}\left(\sum_{i=1}^{n}\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)^{\mathrm{T}} + \sum_{i=1}^{n}\left(\boldsymbol{\mu}_c - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{\mu}_c - \overline{\boldsymbol{x}}\right)^{\mathrm{T}}\right)\right] \\
&= \frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}_c^{-1}\sum_{i=1}^{n}\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)^{\mathrm{T}}\right] + \frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}_c^{-1}\sum_{i=1}^{n}\left(\boldsymbol{\mu}_c - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{\mu}_c - \overline{\boldsymbol{x}}\right)^{\mathrm{T}}\right] \\
&= \frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}_c^{-1}\sum_{i=1}^{n}\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)^{\mathrm{T}}\right] + \frac{1}{2}\operatorname{tr}\left[n\cdot\boldsymbol{\Sigma}_c^{-1}\left(\boldsymbol{\mu}_c - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{\mu}_c - \overline{\boldsymbol{x}}\right)^{\mathrm{T}}\right] \\
&= \frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}_c^{-1}\sum_{i=1}^{n}\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)^{\mathrm{T}}\right] + \frac{n}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}_c^{-1}\left(\boldsymbol{\mu}_c - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{\mu}_c - \overline{\boldsymbol{x}}\right)^{\mathrm{T}}\right] \\
&= \frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}_c^{-1}\sum_{i=1}^{n}\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)^{\mathrm{T}}\right] + \frac{n}{2}\left(\boldsymbol{\mu}_c - \overline{\boldsymbol{x}}\right)^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}\left(\boldsymbol{\mu}_c - \overline{\boldsymbol{x}}\right).
\end{aligned}
$$

所以

$$
\left(\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\Sigma}}_c\right) = \arg\min_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c}\frac{n}{2}\log\left|\boldsymbol{\Sigma}_c\right| + \frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}_c^{-1}\sum_{i=1}^{n}\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)^{\mathrm{T}}\right] + \frac{n}{2}\left(\boldsymbol{\mu}_c - \overline{\boldsymbol{x}}\right)^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}\left(\boldsymbol{\mu}_c - \overline{\boldsymbol{x}}\right).
$$

观察上式可知，由于此时$\boldsymbol{\Sigma}_c^{-1}$和$\boldsymbol{\Sigma}_c$一样均为正定矩阵，所以当$\boldsymbol{\mu}_c - \bar{\boldsymbol{x}} \neq \mathbf{0}$时，上式最后一项为正定二次型.根据正定二次型的性质可知，此时上式最后一项的取值仅与$\boldsymbol{\mu}_c - \bar{\boldsymbol{x}}$相关，并有当且仅当$\boldsymbol{\mu}_c - \bar{\boldsymbol{x}} = \mathbf{0}$

时，上式最后一项取最小值0，此时可以解得

$$\hat{\boldsymbol{\mu}}_c = \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i.$$

将求解出来的$\hat{\boldsymbol{\mu}}_c$代回参数求解式可得新的参数求解式，有

$$\hat{\boldsymbol{\Sigma}}_c = \arg\min_{\boldsymbol{\Sigma}_c} \frac{n}{2}\log|\boldsymbol{\Sigma}_c| + \frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}_c^{-1}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathrm{T}}\right],$$

此时的参数求解式是仅与$\boldsymbol{\Sigma}_c$相关的函数.

为了求解$\hat{\boldsymbol{\Sigma}}_c$，在这里我们不加证明地给出一个引理：设$\boldsymbol{B}$为$p$阶正定矩阵，$n>0$为实数，则对所有$p$阶正定矩阵$\boldsymbol{\Sigma}$有

具体证明可搜索张伟平“多元正态分布参数的估计和数据的清洁与变换”课件

$$\frac{n}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\right] \geq \frac{n}{2}\log|\boldsymbol{B}| + \frac{pn}{2}(1-logn),$$

当且仅当$\boldsymbol{\Sigma} = \frac{1}{n}\boldsymbol{B}$时等号成立.根据此引理可知，当且仅当$\boldsymbol{\Sigma}_c = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathrm{T}}$时，上述参数求解式中$\arg\min$后面的式子取到最小值，那么此时的$\boldsymbol{\Sigma}_c$即我们想要求解的$\hat{\boldsymbol{\Sigma}}_c$.

式**(7.19)**

$$\hat{P}(c) = \frac{|D_c|+1}{|D|+N}$$

解析

从贝叶斯估计（参见本章附注①）的角度来说，拉普拉斯修正就等价于先验概率为Dirichlet分布（参见本章附注③）的后验期望值估计.为了接下来的叙述方便，我们重新定义一下相关数学符号.

设有包含m个独立同分布样本的训练集D，D中可能的类别数为k，其类别的具体取值范围为$\{c_1, c_2, \cdots, c_k\}$.若令随机变量$C$表示样本所属的类别，且$C$取到每个值的概率分别为$P(C=c_1)=\theta_1, P(C=c_2)=\theta_2, \cdots, P(C=c_k)=\theta_k$，那么显然$C$服从参数为$\boldsymbol{\theta}=(\theta_1, \theta_2, \cdots, \theta_k) \in \mathbb{R}^k$的Categorical分布（参见附注②），其概率质量函数为

$$P(C=c_i)=P(c_i)=\theta_i,$$

其中$P(c_i)=\theta_i$就是式(7.9)所要求解的$\hat{P}(c)$，下面我们用贝叶斯估计中的后验期望值估计来估计θ_i.根据贝叶斯估计的原理可知，在进行参数估计之前，需要先主观预设一个先验概率$P(\boldsymbol{\theta})$，通常为了方便计算后验概率$P(\boldsymbol{\theta}|D)$，我们会用似然函数$P(D|\boldsymbol{\theta})$的共轭先验作为我们的先验概率.显然，此时的似然函数$P(D|\boldsymbol{\theta})$是一个基于Categorical分布的似然函数，而Categorical分布的共轭先验为Dirichlet分布，所以只需要预设先验概率$P(\boldsymbol{\theta})$为Dirichlet分布，然后使用后验期望值估计就能估计出θ_i.

读者可搜索“共轭先验”（conjugate prior）和“共轭先验分布”（conjugate prior distribution）以了解更多

具体地，记D中样本类别取值为c_i的样本个数为y_i，则似然函数

$P(D|\boldsymbol{\theta})$可展开为

$$P(D|\boldsymbol{\theta}) = \theta_1^{y_1} \cdots \theta_k^{y_k} = \prod_{i=1}^{k} \theta_i^{y_i},$$

则有后验概率

$$\begin{aligned} P(\boldsymbol{\theta} \mid D) &= \frac{P(D \mid \boldsymbol{\theta}) P(\boldsymbol{\theta})}{P(D)} \\ &= \frac{P(D \mid \boldsymbol{\theta}) P(\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta}} P(D \mid \boldsymbol{\theta}) P(\boldsymbol{\theta})} \\ &= \frac{\prod_{i=1}^{k} \theta_i^{y_i} \cdot P(\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta}} \left[\prod_{i=1}^{k} \theta_i^{y_i} \cdot P(\boldsymbol{\theta}) \right]}. \end{aligned}$$

假设此时先验概率$P(\boldsymbol{\theta})$是参数为$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_k) \in \mathbb{R}^k$的Dirichlet分布，则$P(\boldsymbol{\theta})$可写为

$$P(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}.$$

将其代入$P(D|\boldsymbol{\theta})$可得

$$P(\boldsymbol{\theta} \mid D) = \frac{\prod_{i=1}^{k} \theta_i^{y_i} \cdot P(\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta}} \left[\prod_{i=1}^{k} \theta_i^{y_i} \cdot P(\boldsymbol{\theta}) \right]}$$

$$= \frac{\prod_{i=1}^{k} \theta_i^{y_i} \cdot \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}}{\sum_{\boldsymbol{\theta}} \left[\prod_{i=1}^{k} \theta_i^{y_i} \cdot \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right]}$$

$$= \frac{\prod_{i=1}^{k} \theta_i^{y_i} \cdot \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}}{\sum_{\boldsymbol{\theta}} \left[\prod_{i=1}^{k} \theta_i^{y_i} \cdot \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right] \cdot \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)}}$$

$$= \frac{\prod_{i=1}^{k} \theta_i^{y_i} \cdot \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}}{\sum_{\boldsymbol{\theta}} \left[\prod_{i=1}^{k} \theta_i^{y_i} \cdot \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right]}$$

$$= \frac{\prod_{i=1}^{k} \theta_i^{\alpha_i + y_i - 1}}{\sum_{\boldsymbol{\theta}} \left[\prod_{i=1}^{k} \theta_i^{\alpha_i + y_i - 1} \right]}.$$

此时若设$\boldsymbol{\alpha} + \boldsymbol{y} = (\alpha_1 + y_1, \alpha_2 + y_2, \cdots, \alpha_k + y_k) \in \mathbb{R}^k$，则根据Dirichlet分布的定义可知

$$P(\boldsymbol{\theta}; \boldsymbol{\alpha} + \boldsymbol{y}) = \frac{\Gamma\left(\sum_{i=1}^{k} (\alpha_i + y_i)\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i + y_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i + y_i - 1},$$

$$\sum_{\boldsymbol{\theta}} P(\boldsymbol{\theta}; \boldsymbol{\alpha} + \boldsymbol{y}) = \sum_{\boldsymbol{\theta}} \frac{\Gamma\left(\sum_{i=1}^{k} (\alpha_i + y_i)\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i + y_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i + y_i - 1},$$

$$1 = \sum_{\boldsymbol{\theta}} \frac{\Gamma\left(\sum_{i=1}^{k} (\alpha_i + y_i)\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i + y_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i + y_i - 1},$$

$$1 = \frac{\Gamma\left(\sum_{i=1}^{k} (\alpha_i + y_i)\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i + y_i)} \sum_{\boldsymbol{\theta}} \left[\prod_{i=1}^{k} \theta_i^{\alpha_i + y_i - 1} \right],$$

$$\frac{1}{\sum_{\boldsymbol{\theta}} \left[\prod_{i=1}^{k} \theta_i^{\alpha_i + y_i - 1} \right]} = \frac{\Gamma\left(\sum_{i=1}^{k} (\alpha_i + y_i)\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i + y_i)}.$$

将此结论代入$P(D|\boldsymbol{\theta})$可得

$$P(\boldsymbol{\theta} \mid D) = \frac{\prod_{i=1}^{k} \theta_i^{\alpha_i + y_i - 1}}{\sum_{\boldsymbol{\theta}} \left[\prod_{i=1}^{k} \theta_i^{\alpha_i + y_i - 1} \right]}$$

$$= \frac{\Gamma\left(\sum_{i=1}^{k}\left(\alpha_i + y_i\right)\right)}{\prod_{i=1}^{k}\Gamma\left(\alpha_i + y_i\right)} \prod_{i=1}^{k} \theta_i^{\alpha_i + y_i - 1}$$

$$= P(\boldsymbol{\theta}; \boldsymbol{\alpha} + \boldsymbol{y}).$$

综上可知，对于服从Categorical分布的$\boldsymbol{\theta}$来说，假设其先验概率$P(\boldsymbol{\theta})$是参数为$\boldsymbol{\alpha}$的Dirichlet分布时，得到的后验概率$P(\boldsymbol{\theta}|D)$是参数为$\boldsymbol{\alpha} + \boldsymbol{y}$的Dirichlet分布，通常我们称这种先验概率分布和后验概率分布形式相同的这对分布为共轭分布.在推得后验概率$P(\boldsymbol{\theta}|D)$的具体形式以后，根据后验期望值估计可得θ_i的估计值为

$$\theta_i = \mathbb{E}_{P(\boldsymbol{\theta}|D)}\left[\theta_i\right]$$

$$= \mathbb{E}_{P(\boldsymbol{\theta};\boldsymbol{\alpha}+\boldsymbol{y})}\left[\theta_i\right]$$

$$= \frac{\alpha_i + y_i}{\sum_{j=1}^{k}\left(\alpha_j + y_j\right)}$$

$$= \frac{\alpha_i + y_i}{\sum_{j=1}^{k}\alpha_j + \sum_{j=1}^{k} y_j}$$

$$= \frac{\alpha_i + y_i}{\sum_{j=1}^{k}\alpha_j + m}.$$

显然，式(7.9)是当$\boldsymbol{\alpha} = (1, 1, \cdots, 1)$时推得的具体结果，此时等价于我们主观预设的先验概率$P(\boldsymbol{\theta})$服从均匀分布，此即拉普拉斯修正.同理，当我们调整$\boldsymbol{\alpha}$的取值后，即可推得其他数据平滑的公式.

式**(7.20)**

$$\hat{P}\left(x_i|c\right) = \frac{|D_{c,x_i}| + 1}{|D_c| + N_i}$$

参见式(7.19)

式**(7.24**）

$$\hat{P}\left(c, x_i\right)=\frac{\left|D_{c, x_i}\right|+1}{|D|+N \times N_i}$$

参见式(7.19)

式**(7.25**）

$$\hat{P}\left(x_j | c, x_i\right)=\frac{\left|D_{c, x_i, x_j}\right|+1}{\left|D_{c, x_i}\right|+N_j}$$

参见式(7.20)

式**(7.27)**

$$\begin{aligned} P\left(x_1, x_2\right) &=\sum_{x_4} P\left(x_1, x_2, x_4\right) \\ &=\sum_{x_4} P\left(x_4 \mid x_1, x_2\right) P\left(x_1\right) P\left(x_2\right) \\ &=P\left(x_1\right) P\left(x_2\right) \end{aligned}$$

解析

在这里补充一下同父结构和顺序结构的推导.

同父结构：在给定父节点x_1的条件下x_3, x_4独立，有

$$
\begin{aligned}
P\left(x_3, x_4 \mid x_1\right) &= \frac{P\left(x_1, x_3, x_4\right)}{P\left(x_1\right)} \\
&= \frac{P\left(x_1\right) P\left(x_3 \mid x_1\right) P\left(x_4 \mid x_1\right)}{P\left(x_1\right)} \\
&= P\left(x_3 \mid x_1\right) P\left(x_4 \mid x_1\right)
\end{aligned}
$$

顺序结构：在给定节点x的条件下y, z独立，有

$$
\begin{aligned}
P(y, z \mid x) &= \frac{P(x, y, z)}{P(x)} \\
&= \frac{P(z) P(x \mid z) P(y \mid x)}{P(x)} \\
&= \frac{P(z, x) P(y \mid x)}{P(x)} \\
&= P(z \mid x) P(y \mid x).
\end{aligned}
$$

式**(7.34)**

$$
LL(\Theta|\boldsymbol{X}, \boldsymbol{Z}) = \ln P(\boldsymbol{X}, \boldsymbol{Z}|\Theta)
$$

EM算法这一节建议以李航《统计学习方法》为主，“西瓜书”为辅进行学习

附注

① 贝叶斯估计**[1]**

贝叶斯学派视角下的一类点估计法称为贝叶斯估计，常用的贝叶斯估计有最大后验估计（Maximum A Posteriori Estimation，简称MAP）、后验中位数估计和后验期望值估计这3种参数估计方法，下面给出这3种方法的具体定义.

设总体的概率质量函数（若总体的分布为连续型时则改为概率密度函数，此处以离散型为例）为$P(x|\theta)$，从该总体中抽取出的n个独立同分布的样本构成样本集$D=\{x_1,x_2,\cdots,x_n\}$，则根据贝叶斯式可得，在给定样本集D的条件下，θ的条件概率为

$$P(\theta|D)=\frac{P(D|\theta)P(\theta)}{P(D)}=\frac{P(D|\theta)P(\theta)}{\sum_\theta P(D|\theta)P(\theta)},$$

其中$P(D|\theta)$为似然函数，由于样本集D中的样本是独立同分布的，所以似然函数可以进一步展开，有

$$P(\theta|D)=\frac{P(D|\theta)P(\theta)}{\sum_\theta P(D|\theta)P(\theta)}=\frac{\prod_{i=1}^n P(x_i|\theta)P(\theta)}{\sum_\theta \prod_{i=1}^n P(x_i|\theta)P(\theta)}.$$

根据贝叶斯学派的观点，此条件概率代表了我们在已知样本集D后对θ产生的新的认识，它综合了我们对θ主观预设的先验概率$P(\theta)$和样本集D带来的信息，通常称其为θ的后验概率.

贝叶斯学派认为，在得到$P(\theta|D)$以后，对参数θ的任何统计推断，都只能基于$P(\theta|D)$.至于具体如何去使用它，可以结合某种准则一起去进行，统计学家也有一定的自由度.对于点估计来说， 求使得$P(\theta|D)$达到最大值的$\hat{\theta}_{\text{MAP}}$作为$\theta$的估计称为最大后验估计，求$P(\theta|D)$的中位数$\hat{\theta}_{\text{Median}}$作为$\theta$的估计称为后验中位数估计，求$P(\theta|D)$的期望值（均值）$\hat{\theta}_{\text{Mean}}$作为$\theta$的估计称为后验期望值估计.

② Categorical分布

Categorical分布又称为广义伯努利分布，是将伯努利分布中的随机

变量可取值个数由两个泛化为多个得到的分布.具体地，设离散型随机变量X共有k种可能的取值$\{x_1, x_2, \cdots, x_k\}$，且$X$取到每个值的概率分别为$P(X = x_1) = \theta_1, P(X = x_2) = \theta_2, \cdots, P(X = x_k) = \theta_k$，则称随机变量$X$服从参数为$\theta_1, \theta_2, \cdots, \theta_k$的Categorical分布，其概率质量函数为

$$P(X = x_i) = p(x_i) = \theta_i$$

其中$\mathbb{I}(\cdot)$是指示函数，若.为真则取值1，否则取值0.

③ Dirichlet分布

类似于Categorical分布是伯努利分布的泛化形式，Dirichlet分布是Beta分布的泛化形式.对于一个k维随机变量$\boldsymbol{x} = (x_1, x_2, \cdots, x_k) \in \mathbb{R}^k$，其中$x_i (i = 1, 2, \cdots, k)$满足$0 \leqslant x_i \leqslant 1, \sum_{i=1}^{k} x_i = 1$，若$\boldsymbol{x}$服从参数为$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_k) \in \mathbb{R}^k$的Dirichlet分布，则其概率密度函数为

$$p(\boldsymbol{x}; \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}$$

其中$\Gamma(z) = \int_0^{\infty} x^{z-1} \mathrm{e}^{-x} \mathrm{d}x$为Gamma函数，当$\boldsymbol{\alpha} = (1, 1, \cdots, 1)$时，Dirichlet分布等价于均匀分布.

参考文献

[1] 陈希孺. 概率论与数理统计[M]. 合肥：中国科学技术大学出版社, 2009.

第8章　集成学习

式(8.1)

$$P\left(h_i(\boldsymbol{x}) \neq f(\boldsymbol{x})\right) = \epsilon$$

解析

$h_i(\boldsymbol{x})$是编号为i的基分类器预测的x标记，$f(\boldsymbol{x})$是x的真实标记，它们之间不一致的概率记为ϵ.

式(8.2)

$$H(\boldsymbol{x}) = \operatorname{sgn}\left(\sum_{i=1}^{T} h_i(\boldsymbol{x})\right)$$

“西瓜书”中将符号函数写为$\operatorname{sign}(\cdot)$

解析

当$h_i(\boldsymbol{x})$把$\boldsymbol{x}$分类为1时，有$h_i(\boldsymbol{x}) = 1$，否则$h_i(\boldsymbol{x}) = -1$.各个基分类器h_i的分类结果求和之后数字的正、负或0，代表投票法产生的结果，即“少数服从多数”.符号函数$\operatorname{sgn}(\cdot)$将正数变成1，负数变成$-1$，0仍然是0，所以$H(\boldsymbol{x})$是由投票法产生的分类结果.

式(8.3)

$$P(H(\boldsymbol{x}) \neq f(\boldsymbol{x})) = \sum_{k=0}^{\lfloor T/2 \rfloor} \begin{pmatrix} T \\ k \end{pmatrix} (1-\epsilon)^k \epsilon^{T-k}$$
$$\leqslant \exp\left(-\frac{1}{2}T(1-2\epsilon)^2\right)$$

解析

即X服从二项分布

由于基分类器相互独立，假设随机变量X为T个基分类器分类正确的次数，因此$X \sim \mathcal{B}(T, 1-\epsilon)$，设$x_i$为每一个分类器分类正确的次数，则$x_i \sim \mathcal{B}(1, 1-\epsilon)(i = 1, 2, 3, \cdots, T)$，有

$$X = \sum_{i=1}^{T} x_i,$$

$$\mathbb{E}(X) = \sum_{i=1}^{T} \mathbb{E}(x_i) = (1-\epsilon)T.$$

证明过程如下：

$$P(H(x) \neq f(x)) = P(X \leqslant \lfloor T/2 \rfloor)$$

$$\leqslant P(X \leqslant T/2)$$

$$= P\left(X - (1-\epsilon)T \leqslant \frac{T}{2} - (1-\epsilon)T\right)$$

$$= P\left(X - (1-\epsilon)T \leqslant -\frac{T}{2}(1-2\epsilon)\right)$$

$$= P\left(\sum_{i=1}^{T} x_i - \sum_{i=1}^{T} \mathbb{E}\left(x_i\right) \leqslant -\frac{T}{2}(1-2\epsilon)\right)$$

$$= P\left(\frac{1}{T}\sum_{i=1}^{T} x_i - \frac{1}{T}\sum_{i=1}^{T}\mathbb{E}\left(x_i\right) \leqslant -\frac{1}{2}(1-2\epsilon)\right).$$

根据Hoeffding不等式知

$$P\left(\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left(x_i\right) \leqslant -\delta\right) \leqslant \exp\left(-2m\delta^2\right).$$

令$\delta = \frac{(1-2\epsilon)}{2}, m = T$得

$$\begin{aligned} P(H(\boldsymbol{x}) \neq f(\boldsymbol{x})) &= \sum_{k=0}^{\lfloor T/2 \rfloor} \begin{pmatrix} T \\ k \end{pmatrix} (1-\epsilon)^k \epsilon^{T-k} \\ &\leqslant \exp\left(-\frac{1}{2}T(1-2\epsilon)^2\right). \end{aligned}$$

式**(8.4)**

$$H(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t h_t(\boldsymbol{x})$$

解析

此式是集成学习的加性模型.加性模型不采用梯度下降的思想，而是$H(\boldsymbol{x}) = \sum_{t=1}^{T-1} \alpha_t h_t(\boldsymbol{x}) + \alpha_T h_T(\boldsymbol{x})$每次更新求解一个理论上最优的$h_T$和$\alpha_T$.

参见式(8.18)和式(8.11)

式**(8.5)**

$$\ell_{\exp}(H|\mathcal{D}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H(\boldsymbol{x})}\right]$$

解析

由式(8.4)知

$$H(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t h_t(\boldsymbol{x}),$$

又由式(8.11)可知

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t} \right),$$

由$\ln$函数的单调性可知，该分类器的权重只与分类器的错误率负相关(即错误率越大，权重越低)，指数损失函数的意义如下.

先考虑指数损失函数$\mathrm{e}^{-f(x)H(x)}$的含义.f为真实函数，对于样本x来说，$f(\boldsymbol{x}) \in \{+1, -1\}$只能取$+1$和$-1$，而$H(\boldsymbol{x})$是一个实数.

当$H(\boldsymbol{x})$的符号与$f(x)$一致时，$f(\boldsymbol{x})H(\boldsymbol{x})>0$，因此有$\mathrm{e}^{-f(\boldsymbol{x})H(\boldsymbol{x})} = \mathrm{e}^{-|H(\boldsymbol{x})|} < 1$，且$|H(\boldsymbol{x})|$越大指数损失函数$\mathrm{e}^{-f(\boldsymbol{x})H(\boldsymbol{x})}$越小.这很合理，因为此时$|H(\boldsymbol{x})|$越大意味着分类器本身对预测结果的信心越大，损失应该越小；若$|H(\boldsymbol{x})|$在零附近，虽然预测正确，但表示分类器本身对预测结果信心很小，损失应该较大.

当$H(\boldsymbol{x})$的符号与$f(\boldsymbol{x})$不一致时，$f(\boldsymbol{x})H(\boldsymbol{x}) < 0$，因此$\mathrm{e}^{-f(\boldsymbol{x})H(\boldsymbol{x})} = \mathrm{e}^{|H(\boldsymbol{x})|}>1$，且$|H(\boldsymbol{x})|$越大指数损失函数越大. 此时$|H(\boldsymbol{x})|$越大意味着分类器本身对预测结果的信心越大，但预测结果是错的，因此损失应该越大；若$|H(\boldsymbol{x})|$在零附近，虽然预测错误，但表示分类器本身对预

测结果信心很小，虽然错了，损失应该较小.

接下来考虑符号$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\cdot]$的含义.$\mathcal{D}$为概率分布，可简单理解为在数据集$D$中进行随机抽样时每个样本被取到的概率；$\mathbb{E}[\cdot]$为经典的期望. 因此$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\cdot]$表示在概率分布$\mathcal{D}$上的期望，可理解为对数据集$D$以概率分布$\mathcal{D}$进行加权后的期望. 即

$$\begin{aligned}\ell_{\exp}(H \mid \mathcal{D}) &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H(\boldsymbol{x})}\right] \\ &= \sum_{\boldsymbol{x}\in D}\mathcal{D}(\boldsymbol{x})\mathrm{e}^{-f(\boldsymbol{x})H(\boldsymbol{x})}.\end{aligned}$$

式**(8.6)**

$$\frac{\partial \ell_{\exp}(H \mid \mathcal{D})}{\partial H(\boldsymbol{x})} = -\mathrm{e}^{-H(\boldsymbol{x})}P(f(\boldsymbol{x})=1 \mid \boldsymbol{x}) + \mathrm{e}^{H(\boldsymbol{x})}P(f(\boldsymbol{x})=-1 \mid \boldsymbol{x})$$

解析

由式(8.5)中对于符号$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\cdot]$的解释可知

$$\begin{aligned}&\ell_{\exp}(H \mid \mathcal{D}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H(\boldsymbol{x})}\right] \\ &= \sum_{\boldsymbol{x}\in D}\mathcal{D}(\boldsymbol{x})\mathrm{e}^{-f(\boldsymbol{x})H(\boldsymbol{x})} \\ &= \sum_{i=1}^{|D|}\mathcal{D}\left(\boldsymbol{x}_i\right)\left(\mathrm{e}^{-H(\boldsymbol{x}_i)}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)=1\right)+\mathrm{e}^{H(\boldsymbol{x}_i)}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)=-1\right)\right) \\ &= \mathrm{e}^{-H(\boldsymbol{x}_i)}P\left(f\left(\boldsymbol{x}_i\right)=1 \mid \boldsymbol{x}_i\right)+\mathrm{e}^{H(\boldsymbol{x}_i)}P\left(f\left(\boldsymbol{x}_i\right)=-1 \mid \boldsymbol{x}_i\right),\end{aligned}$$

因此

$$\frac{\partial \ell_{\exp}(H|\mathcal{D})}{\partial H(\boldsymbol{x})} = -\mathrm{e}^{-H(\boldsymbol{x})} P(f(\boldsymbol{x}) = 1|\boldsymbol{x}) + \mathrm{e}^{H(\boldsymbol{x})} P(f(\boldsymbol{x}) = -1|\boldsymbol{x}).$$

式**(8.7)**

$$H(\boldsymbol{x}) = \frac{1}{2} \ln \frac{P(f(x) = 1|\boldsymbol{x})}{P(f(x) = -1|\boldsymbol{x})}$$

解析

令式(8.6)等于0，移项并分离$H(\boldsymbol{x})$，即可得到式(8.7).

式**(8.8)**

$$\operatorname{sgn}(H(\boldsymbol{x})) = \operatorname{sgn}\left(\frac{1}{2} \ln \frac{P(f(\boldsymbol{x}) = 1 \mid \boldsymbol{x})}{P(f(\boldsymbol{x}) = -1 \mid \boldsymbol{x})}\right) \quad ①$$

$$= \begin{cases} 1, & P(f(\boldsymbol{x}) = 1 \mid \boldsymbol{x}) > P(f(\boldsymbol{x}) = -1 \mid \boldsymbol{x}) \\ -1, & P(f(\boldsymbol{x}) = 1 \mid \boldsymbol{x}) < P(f(\boldsymbol{x}) = -1 \mid \boldsymbol{x}) \end{cases} \quad ②$$

$$= \underset{y \in \{-1,1\}}{\arg\max} P(f(\boldsymbol{x}) = y \mid \boldsymbol{x}) \quad ③$$

$\operatorname{sgn}(\cdot)$函数即“西瓜书”中的$\operatorname{sign}(\cdot)$函数

解析

①→②显然成立；②→③利用了$\arg\max$函数的定义：

$\underset{y \in \{-1,1\}}{\arg\max} P(f(x) = y|\boldsymbol{x})$表示使得函数$P(f(x) = y|\boldsymbol{x})$取得最大值的$y$的值，展开则为②.

式**(8.9)**

$$\ell_{\exp}(\alpha_t h_t \mid \mathcal{D}_t) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}\left[\mathrm{e}^{-f(\boldsymbol{x})\alpha_t h_t(\boldsymbol{x})}\right]$$

$$= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}\left[\mathrm{e}^{-\alpha_t}\mathbb{I}(f(\boldsymbol{x}) = h_t(\boldsymbol{x})) + \mathrm{e}^{\alpha_t}\mathbb{I}(f(\boldsymbol{x}) \neq h_t(\boldsymbol{x}))\right]$$

$$= \mathrm{e}^{-\alpha_t} P_{\boldsymbol{x}\sim\mathcal{D}_t}(f(\boldsymbol{x}) = h_t(\boldsymbol{x})) + \mathrm{e}^{\alpha_t} P_{\boldsymbol{x}\sim\mathcal{D}_t}(f(\boldsymbol{x}) \neq h_t(\boldsymbol{x}))$$

$$= \mathrm{e}^{-\alpha_t}(1-\epsilon_t) + \mathrm{e}^{\alpha_t}\epsilon_t$$

ϵ_t与式(8.1)一致，表示$h_t(\boldsymbol{x})$分类错误的概率

式**(8.10**）

$$\frac{\partial \ell_{\exp}(\alpha_t h_t|\mathcal{D}_t)}{\partial \alpha_t} = -\mathrm{e}^{-\alpha_t}(1-\epsilon_t) + \mathrm{e}^{\alpha_t}\epsilon_t$$

指数损失函数对α_t求偏导，以得到使得损失函数取最小值时α_t的值

式**(8.11)**

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

解析

令式(8.10)等于0移项即得到的该式.此时α_t的取值使得该基分类器经α_t加权后的损失函数最小.

式**(8.12)**

$$\begin{aligned}\ell_{\exp}(H_{t-1} + h_t|\mathcal{D}) &= \mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})(H_{t-1}(\boldsymbol{x})+h_t(\boldsymbol{x}))}\right] \\ &= \mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\mathrm{e}^{-f(\boldsymbol{x})h_t(\boldsymbol{x})}\right]\end{aligned}$$

解析

将$H_t(\boldsymbol{x}) = H_{t-1}(\boldsymbol{x}) + h_t(\boldsymbol{x})$代入式(8.5)即可，因为理想的$h_t$可以纠正$H_{t-1}$的全部错误，所以这里指定其权重系数为1.如果权重系数α_t是常数的话，对后续结果也没有影响.

式(8.13)

$$\ell_{\exp}(H_{t-1} + h_t \mid \mathcal{D}) \simeq \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\left(1 - f(\boldsymbol{x})h_t(\boldsymbol{x}) + \frac{f^2(\boldsymbol{x})h_t^2(\boldsymbol{x})}{2}\right)\right]$$

$$= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\left(1 - f(\boldsymbol{x})h_t(\boldsymbol{x}) + \frac{1}{2}\right)\right]$$

解析

由e^x的二阶泰勒展开$1 + x + \frac{x^2}{2} + o(x^2)$得

$$\ell_{\exp}(H_{t-1} + h_t \mid \mathcal{D}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\mathrm{e}^{-f(\boldsymbol{x})h_t(\boldsymbol{x})}\right]$$

$$\simeq \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\left(1 - f(\boldsymbol{x})h_t(\boldsymbol{x}) + \frac{f^2(\boldsymbol{x})h_t^2(\boldsymbol{x})}{2}\right)\right]$$

因为$f(\boldsymbol{x})$与$h_t(\boldsymbol{x})$取值都为1或-1，所以$f^2(\boldsymbol{x}) = h_t^2(\boldsymbol{x}) = 1$，故有

$$\ell_{\exp}(H_{t-1} + h_t|\mathcal{D}) \simeq \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\left(1 - f(\boldsymbol{x})h_t(\boldsymbol{x}) + \frac{1}{2}\right)\right].$$

式(8.14)

$$h_t(\boldsymbol{x}) = \arg\min_h \ell_{\exp}(H_{t-1} + h \mid \mathcal{D}) \quad ①$$

$$= \arg\min_h \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\left(1 - f(\boldsymbol{x})h(\boldsymbol{x}) + \frac{1}{2}\right)\right] \quad ②$$

$$= \arg\max_h \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} \left[e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})} f(\boldsymbol{x})h(\boldsymbol{x}) \right] \quad ③$$

$$= \arg\max_h \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} \left[\frac{e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}}{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\right]} f(\boldsymbol{x})h(\boldsymbol{x}) \right] \quad ④$$

解析

理想的$h_t(\boldsymbol{x})$是使得$H_t(\boldsymbol{x})$的指数损失函数取得最小值时的$h_t(\boldsymbol{x})$，该式将此转化成某个期望的最大值. ②→③是因为$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\frac{3}{2}e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\right]$是一个常数，与$h(\boldsymbol{x})$无关.③→④是因为$\frac{1}{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\frac{3}{2}e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\right]}$也与$h(\boldsymbol{x})$无关，所以可以引入.

式**(8.16)**

$$h_t(\boldsymbol{x}) = \arg\max_h \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} \left[\frac{e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}}{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\right]} f(\boldsymbol{x})h(\boldsymbol{x}) \right]$$

$$= \arg\max_{\boldsymbol{h}} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}[f(\boldsymbol{x})h(\boldsymbol{x})]$$

解析

首先解释下符号$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}$的含义.注意在本章中有两个符号$D$和$\mathcal{D}$，其中$D$表示数据集，而$\mathcal{D}$表示数据集$D$的样本分布，即在数据集$D$上进行一次随机采样时，样本$\boldsymbol{x}$被抽到的概率是$\mathcal{D}(\boldsymbol{x})$. $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}$表示在概率分布$\mathcal{D}$上的期望，可以简单地理解为，对数据及$D$以概率$\mathcal{D}$加权之后的期望，因此有

$$\mathbb{E}(g(\boldsymbol{x})) = \sum_{i=1}^{|D|} f(\boldsymbol{x}_i)g(\boldsymbol{x}_i),$$

故可得

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H(\boldsymbol{x})}\right]=\sum_{i=1}^{|D|}\mathcal{D}\left(\boldsymbol{x}_i\right)\mathrm{e}^{-f(\boldsymbol{x}_i)H(\boldsymbol{x}_i)}.$$

由式(8.15)可知

$$\mathcal{D}_t\left(\boldsymbol{x}_i\right)=\mathcal{D}\left(\boldsymbol{x}_i\right)\frac{\mathrm{e}^{-f(\boldsymbol{x}_i)H_{t-1}(\boldsymbol{x}_i)}}{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\right]},$$

所以有

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\frac{\mathrm{e}^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}}{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\right]}f(\boldsymbol{x})h(\boldsymbol{x})\right]\\
&=\sum_{i=1}^{|D|}\mathcal{D}\left(\boldsymbol{x}_i\right)\frac{\mathrm{e}^{-f(\boldsymbol{x}_i)H_{t-1}(\boldsymbol{x}_i)}}{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\right]}f\left(x_i\right)h\left(x_i\right)\\
&=\sum_{i=1}^{|D|}\mathcal{D}_t\left(\boldsymbol{x}_i\right)f\left(\boldsymbol{x}_i\right)h\left(\boldsymbol{x}_i\right)\\
&=\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}[f(\boldsymbol{x})h(\boldsymbol{x})].
\end{aligned}$$

式**(8.17)**

$$f(\boldsymbol{x})h(\boldsymbol{x})=1-2\mathbb{I}(f(\boldsymbol{x})\neq h(\boldsymbol{x}))$$

当$f(\boldsymbol{x})=h(\boldsymbol{x})$时，$\mathbb{I}(f(\boldsymbol{x})\neq h(\boldsymbol{x}))=0$，$f(\boldsymbol{x})h(\boldsymbol{x})=1$；当$f(\boldsymbol{x})\neq h(\boldsymbol{x})$时，$\mathbb{I}(f(\boldsymbol{x})\neq h(\boldsymbol{x}))=1$，$f(\boldsymbol{x})h(\boldsymbol{x})=-1$

式**(8.18）**

$$h_t(\boldsymbol{x})=\arg\min_{h}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}[\mathbb{I}(f(\boldsymbol{x})\neq h(\boldsymbol{x}))]$$

解析

由式(8.16) 和式(8.17)有

$$h_t(\boldsymbol{x}) = \arg\max_h \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}[f(\boldsymbol{x})h(\boldsymbol{x})]$$

$$= \arg\max_h \left(1 - 2\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}[\mathbb{I}(f(\boldsymbol{x}) \neq h(\boldsymbol{x}))]\right)$$

$$= \arg\max_h \left(-2\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}[\mathbb{I}(f(\boldsymbol{x}) \neq h(\boldsymbol{x}))]\right)$$

$$= \arg\min_h \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}[\mathbb{I}(f(\boldsymbol{x}) \neq h(\boldsymbol{x}))]$$

式**(8.19**）

$$\mathcal{D}_{t+1}(\boldsymbol{x}) = \frac{\mathcal{D}(\boldsymbol{x})\mathrm{e}^{-f(\boldsymbol{x})H_t(\boldsymbol{x})}}{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H_t(\boldsymbol{x})}\right]}$$

$$= \frac{\mathcal{D}(\boldsymbol{x})\mathrm{e}^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\mathrm{e}^{-f(\boldsymbol{x})\alpha_t h_t(\boldsymbol{x})}}{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H_t(\boldsymbol{x})}\right]}$$

$$= \mathcal{D}_t(\boldsymbol{x}) \cdot \mathrm{e}^{-f(\boldsymbol{x})\alpha_t h_t(\boldsymbol{x})} \frac{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\right]}{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{e}^{-f(\boldsymbol{x})H_t(\boldsymbol{x})}\right]}$$

Boosting算法根据调整后的样本训练下一个基分类器.此为“重赋权法”的样本分布的调整式

式**(8.20)**

$$H^{\mathrm{oob}}(\boldsymbol{x}) = \arg\max_{y\in\mathcal{Y}} \sum_{t=1}^{\mathrm{T}} \mathbb{I}\left(h_t(\boldsymbol{x}) = y\right) \cdot \mathbb{I}\left(\boldsymbol{x} \notin D_t\right)$$

解析

$\mathbb{I}(h_t(\boldsymbol{x}) = y)$表示对$T$个基学习器分别判断结果是否与$y$一致，其中$y$的取值一般是$-1$和1.如果基学习器结果与$y$一致，则$\mathbb{I}(h_t(\boldsymbol{x}) = y) = 1$；如果样本不在训练集内，则$\mathbb{I}(\boldsymbol{x} \notin D_t) = 1$.综合起来看，就是对包外的数据，用“投票法”选择包外估计的结果，即1或-1.

式**(8.21**）

$$\epsilon^{\text{oob}} = \frac{1}{|D|} \sum_{(\boldsymbol{x},y)\in D} \mathbb{I}\left(H^{\text{oob}}(\boldsymbol{x}) \neq y\right)$$

由式(8.20)知，$H^{\text{oob}}(\boldsymbol{x})$是对包外的估计.该式表示估计错误的个数除以总的个数，得到泛化误差的包外估计

式**(8.22**）

$$H(\boldsymbol{x}) = \frac{1}{T} \sum_{i=1}^{T} h_i(\boldsymbol{x})$$

此为对基分类器的结果进行简单的平均

式**(8.23**）

$$H(\boldsymbol{x}) = \sum_{i=1}^{T} w_i h_i(\boldsymbol{x})$$

此为对基分类器的结果进行加权平均

式**(8.24**）

$$H(\boldsymbol{x}) = \begin{cases} c_j, & \text{if } \sum_{i=1}^{T} h_i^j(\boldsymbol{x}) > 0.5 \sum_{k=1}^{N} \sum_{i=1}^{T} h_i^k(\boldsymbol{x}) \\ \text{reject}, & \text{otherwise} \end{cases}$$

即当某一个类别j的基分类器的结果之和大于所有结果之和的$\frac{1}{2}$时，选择该类别j为最终结果

式**(8.25**）

$$H(\boldsymbol{x}) = c_{\arg\max_j \sum_{i=1}^{T} h_i^j(\boldsymbol{x})}$$

若类别j的基分类器的结果之和在所有类别中最大，则选择类别j为最终结果

式**(8.26)**

$$H(\boldsymbol{x}) = c_{\arg\max_j \sum_{i=1}^{T} w_i h_i^j(\boldsymbol{x})}$$

此式与式(8.25)的不同在于，在基分类器前面乘上一个权重系数w_i，满足$w_i \geq 0$且$\sum_{i=1}^{T} w_i = 1$

式**(8.27**）

$$A\left(h_i|\boldsymbol{x}\right) = \left(h_i(\boldsymbol{x}) - H(\boldsymbol{x})\right)^2$$

此为个体学习器结果与预测结果的差值的平方，即个体学习器的“分歧”

式**(8.28**）

$$\begin{aligned}\bar{A}(h \mid \boldsymbol{x}) &= \sum_{i=1}^{T} w_i A\left(h_i \mid \boldsymbol{x}\right) \\ &= \sum_{i=1}^{T} w_i \left(h_i(\boldsymbol{x}) - H(\boldsymbol{x})\right)^2\end{aligned}$$

此为对各个个体学习器的“分歧”加权平均的结果，即集成的“分歧”

式**(8.29**）

$$E\left(h_i|\boldsymbol{x}\right) = \left(f(\boldsymbol{x}) - h_i(\boldsymbol{x})\right)^2$$

此为个体学习器与真实值之间差值的平方，即个体学习器的平方误差

式**(8.30)**

$$E(H|\boldsymbol{x}) = \left(f(\boldsymbol{x}) - H(\boldsymbol{x})\right)^2$$

此为集成与真实值之间差值的平方，即集成的平方误差

式**(8.31)**

$$\bar{A}(h|\boldsymbol{x}) = \sum_{i=1}^{T} w_i E\left(h_i|\boldsymbol{x}\right) - E(H|\boldsymbol{x})$$

解析

由式(8.28)知

$$\bar{A}(h \mid \boldsymbol{x}) = \sum_{i=1}^{T} w_i \left(h_i(\boldsymbol{x}) - H(\boldsymbol{x})\right)^2$$

$$= \sum_{i=1}^{T} w_i \left(h_i(\boldsymbol{x})^2 - 2h_i(\boldsymbol{x})H(\boldsymbol{x}) + H(\boldsymbol{x})^2 \right)$$

$$= \sum_{i=1}^{T} w_i h_i(\boldsymbol{x})^2 - H(\boldsymbol{x})^2,$$

又因为

$$\begin{aligned} \sum_{i=1}^{T} w_i E\left(h_i \mid \boldsymbol{x}\right) - E(H \mid \boldsymbol{x}) &= \sum_{i=1}^{T} w_i \left(f(\boldsymbol{x}) - h_i(\boldsymbol{x})\right)^2 - \left(f(\boldsymbol{x}) - H(\boldsymbol{x})\right)^2 \\ &= \sum_{i=1}^{T} w_i h_i(\boldsymbol{x})^2 - H(\boldsymbol{x})^2, \end{aligned}$$

所以

$$\bar{A}(h|\boldsymbol{x}) = \sum_{i=1}^{T} w_i E\left(h_i|\boldsymbol{x}\right) - E(H|\boldsymbol{x}).$$

式**(8.32)**

$$\sum_{i=1}^{T} w_i \int A\left(h_i|\boldsymbol{x}\right) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \sum_{i=1}^{T} w_i \int E\left(h_i|\boldsymbol{x}\right) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int E(H|\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

解析

$\int A\left(h_i|\boldsymbol{x}\right) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ 表示个体学习器在全样本上的“分歧”，

$\sum_{i=1}^{T} w_i \int A\left(h_i|\boldsymbol{x}\right) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ 表示集成在全样本上的“分歧”.根据式(8.31)将

$\sum_{i=1}^{T} w_i \int A(h_i|\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}x$ 拆成误差的形式即有本式.

式**(8.33)**

$$E_i = \int E\left(h_i|\boldsymbol{x}\right) p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

此为个体学习器在全样本上的泛化误差

式**(8.34**）

$$A_i = \int A\left(h_i|\boldsymbol{x}\right) p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

此为个体学习器在全样本上的分歧

式**(8.35**）

$$E = \int E(H|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

此为集成在全样本上的泛化误差

式**(8.36**）

$$E = \bar{E} - \bar{A}$$

$\bar{E}$表示个体学习器泛化误差的加权均值，$\bar{A}$表示个体学习器分歧项的加权均值，该式称为“误差-分歧分解”

第9章　聚类

式(9.5)

$$\mathrm{JC} = \frac{a}{a+b+c}$$

解析

给定两个集合A和B，则Jaccard系数定义为

$$\mathrm{JC} = \frac{|A \bigcap B|}{|A \bigcup B|} = \frac{|A \bigcap B|}{|A| + |B| - |A \bigcap B|}.$$

Jaccard系数可以用来描述两个集合的相似程度.推论：假设全集U共有n个元素，且$A \subseteq U$，$B \subseteq U$，则每一个元素的位置共有4种情况：

(1) 元素同时在集合A和B中，这样的元素个数记为M_{11};

(2) 元素出现在集合A中，但没有出现在集合B中，这样的元素个数记为M_{10};

(3) 元素没有出现在集合A中，但出现在集合B中，这样的元素个数记为M_{01};

(4) 元素既没有出现在集合A中，也没有出现在集合B中，这样的元素个数记为M_{00}.

根据Jaccard系数的定义，此时的Jaccard系数

$$\mathrm{JC} = \frac{M_{11}}{M_{11} + M_{10} + M_{01}}.$$

聚类属于无监督学习.我们并不知道聚类后样本所属类别的类别标记所代表的意义，即便参考模型的类别标记意义是已知的，也无法知道聚类后的类别标记与参考模型的类别标记的对应关系.此外， 聚类后的类别总数与参考模型的类别总数还可能不同.因此无法只用单个样本衡量聚类性能的好坏.

外部指标的基本思想就是以参考模型的类别划分为参照.如果某一个样本对中的两个样本在聚类结果中同属于一个类，在参考模型中也同属于一个类，或者这两个样本在聚类结果中不同属于一个类，在参考模型中也不同属于一个类，那么对于这两个样本，这是一个好的聚类结果.

一般地，样本对中的两个样本共存在4种情况：

(1) 在聚类结果中属于同一个类，在参考模型中也属于同一个类;

(2) 在聚类结果中属于同一个类，但在参考模型中不属于同一个类;

(3) 在聚类结果中不属于同一个类，但在参考模型中属于同一个类;

(4) 在聚类结果中不属于同一个类，在参考模型中也不属于同一个类.

以上4种情况对应“西瓜书”中的式(9.1)~(9.4).

假设集合A中存放着两个样本都同属于聚类结果的同一个类的样本对，即$A = SS \bigcup SD$，集合B中存放着两个样本都同属于参考模型的同一个类的样本对，即$B = SS \bigcup DS$，那么根据Jaccard系数的定义有

$$\mathrm{JC} = \frac{|A \bigcap B|}{|A \bigcup B|} = \frac{|SS|}{|SS \bigcup SD \bigcup DS|} = \frac{a}{a+b+c}.$$

也可直接将“西瓜书”中的式(9.1)~(9.4)的四种情况类比推论，即$M_{11} = a$，$M_{10} = b$，$M_{01} = c$，所以

$$\mathrm{JC} = \frac{M_{11}}{M_{11} + M_{10} + M_{01}} = \frac{a}{a+b+c}.$$

式**(9.6)**

$$\mathrm{FMI} = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

解析

式中$\frac{a}{a+b}$和$\frac{a}{a+c}$为Wallace提出的两个非对称指标，其中a代表两个样本在聚类结果和参考模型中均属于同一类的样本对的数量，$a+b$代表两个样本在聚类结果中属于同一类的样本对的数量，$a+c$代表两个样本在参考模型中属于同一类的样本对的数量.

这两个非对称指标均可理解为样本对中的两个样本在聚类结果和参考模型中均属于同一类的概率.由于指标的非对称性，这两个概率值往往不相等，因此Fowlkes和Mallows提出利用几何平均数将这两个非对称

指标转化为一个对称指标，即FM指数.

式(9.7)

$$\mathrm{RI}=\frac{2(a+d)}{m(m-1)}$$

解析

Rand 指数定义如下：

$$\mathrm{RI}=\frac{a+d}{a+b+c+d}=\frac{a+d}{m(m-1)/2}=\frac{2(a+d)}{m(m-1)},$$

其中a表示聚类结果为同一类别且参考模型给出的划分也属于统一类别的样本对的个数。d表示聚类结果不属于同一类别且参考模型也不划分到同一类别的样本对的个数。分母$m(m-1)/2$是所有样本的总对数，因此RI可以简单理解为聚类结果与参考模型的一致性。

式(9.8）

$$\operatorname{avg}(C)=\frac{2}{|C|(|C|-1)}\sum_{1\leqslant i<j\leqslant|C|}\operatorname{dist}\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)$$

解析

此为簇内距离的定义. $\frac{2}{|C|(|C|-1)}$为$(\boldsymbol{x}_i,\boldsymbol{x}_j)$组合数量的倒数，$\sum_{i\leqslant i<j\leqslant|C|}\operatorname{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j)$为这些组合的距离和.二者相乘即平均距离.

式(9.33）

$$\sum_{j=1}^{m} \frac{\alpha_i \cdot p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\sum_{l=1}^{k} \alpha_l \cdot p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right)} \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right) = 0$$

解析

根据式(9.28)有

$$p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) = \frac{1}{(2\pi)^{\frac{n}{2}} \left|\boldsymbol{\Sigma}_i\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\right),$$

又根据式(9.32)，由 $\frac{\partial LL(D)}{\partial \boldsymbol{\mu}_i} = 0$，有

$$\frac{\partial LL(D)}{\partial p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)} \cdot \frac{\partial p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\partial \boldsymbol{\mu}_i} = 0.$$

为了避免混淆，重写式(9.32)如下：

$$LL(D) = \sum_{j=1}^{m} \ln\left(\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)\right)$$

这里将第 2个求和号的求和变量由式(9.32)的变量i改为l，是为了避免和 $p(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$中的变量$i$混淆,

其中，对于$\frac{\partial LL(D)}{\partial p(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}$，有

$$\frac{\partial LL(D)}{\partial p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)} = \frac{\partial \sum_{j=1}^{m} \ln\left(\sum_{l=1}^{k} \alpha_l \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right)\right)}{\partial p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}$$

$$= \sum_{j=1}^{m} \frac{\partial \ln\left(\sum_{l=1}^{k} \alpha_l \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right)\right)}{\partial p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}$$

$$= \sum_{j=1}^{m} \frac{\alpha_i}{\sum_{l=1}^{k} \alpha_l \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right)}$$

对于 $\frac{\partial p(\boldsymbol{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\partial \boldsymbol{\mu}_i}$ ，有

$$\frac{\partial p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\partial \boldsymbol{\mu}_i}$$

$$= \frac{\partial \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\right)}{\partial \boldsymbol{\mu}_i}$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}\left|\boldsymbol{\Sigma}_i\right|^{\frac{1}{2}}} \frac{\partial \exp\left(-\frac{1}{2}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\right)}{\partial \boldsymbol{\mu}_i}$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}\left|\boldsymbol{\Sigma}_i\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\right)$$

$$\left(-\frac{1}{2} \frac{\partial\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)}{\partial \boldsymbol{\mu}_i}\right)$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}\left|\Sigma_i\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\right) \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)$$

$$= p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right).$$

由矩阵求导的法则 $\frac{\partial \boldsymbol{a}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{a}}{\partial \boldsymbol{a}} = 2\boldsymbol{X}\boldsymbol{a}$ 可得

$$
\begin{aligned}
&-\frac{1}{2}\frac{\partial\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)}{\partial\boldsymbol{\mu}_i}\\
=&-\frac{1}{2}\cdot 2\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\mu}_i-\boldsymbol{x}_j\right)\\
=&\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)
\end{aligned}
$$

因此有

$$
\frac{\partial LL(D)}{\partial\boldsymbol{\mu}_i}=\sum_{j=1}^{m}\frac{\alpha_i}{\sum_{l=1}^{k}\alpha_l\cdot p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l\right)}\cdot p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)\cdot\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)=0.
$$

式**(9.34)**

$$
\mu_i=\frac{\sum_{j=1}^{m}\gamma_{ji}\boldsymbol{X}_j}{\sum_{j=1}^{m}\gamma_{ji}}
$$

解析

由式(9.30)，有

$$
\gamma_{ji}=p_{\mathcal{M}}\left(z_j=i|\boldsymbol{X}_j\right)=\frac{\alpha_i\cdot p\left(\boldsymbol{X}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)}{\sum_{l=1}^{k}\alpha_l\cdot p\left(\boldsymbol{X}_j|\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l\right)},
$$

代入式(9.33)，有

$$
\sum_{j=1}^{m}\gamma_{ji}\left(\boldsymbol{X}_j-\boldsymbol{\mu}_i\right)=0.
$$

因此

$$\mu_i = \frac{\sum_{j=1}^{m} \gamma_{ji} \boldsymbol{X}_j}{\sum_{j=1}^{m} \gamma_{ji}}.$$

式(9.35）

$$\boldsymbol{\Sigma}_i = \frac{\sum_{j=1}^{m} \gamma_{ji} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^{\mathrm{T}}}{\sum_{j=1}^{m} \gamma_{ji}}$$

解析

根据式(9.28)可知

$$p(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)\right),$$

又根据式(9.32)，需令 $\frac{\partial LL(D)}{\partial \boldsymbol{\Sigma}_i} = 0$.考虑等号左侧，有

$$\frac{\partial LL(D)}{\partial \boldsymbol{\Sigma}_i} = \frac{\partial}{\partial \boldsymbol{\Sigma}_i} \left[\sum_{j=1}^{m} \ln \left(\sum_{i=1}^{k} \alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \right) \right]$$

$$= \sum_{j=1}^{m} \frac{\partial}{\partial \boldsymbol{\Sigma}_i} \left[\ln \left(\sum_{i=1}^{k} \alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \right) \right]$$

$$= \sum_{j=1}^{m} \frac{\alpha_i \cdot \frac{\partial}{\partial \boldsymbol{\Sigma}_i} \left(p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)\right)}{\sum_{l=1}^{k} \alpha_l \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right)}$$

其中

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_i}\left(p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)\right)$$

$$= \frac{\partial}{\partial \boldsymbol{\Sigma}_i}\left[\frac{1}{(2\pi)^{\frac{n}{2}}\left|\boldsymbol{\Sigma}_i\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\right)\right]$$

$$= \frac{\partial}{\partial \boldsymbol{\Sigma}_i}\left\{\exp\left[\ln\left(\frac{1}{(2\pi)^{\frac{n}{2}}\left|\boldsymbol{\Sigma}_i\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\right)\right)\right]\right\}$$

$$= p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \cdot \frac{\partial}{\partial \boldsymbol{\Sigma}_i}\left[\ln\left(\frac{1}{(2\pi)^{\frac{n}{2}}\left|\boldsymbol{\Sigma}_i\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\right)\right)\right]$$

$$= p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \cdot \frac{\partial}{\partial \boldsymbol{\Sigma}_i}\left[\ln \frac{1}{(2\pi)^{\frac{n}{2}}} - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| - \frac{1}{2}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\right]$$

$$= p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \cdot \left[-\frac{1}{2}\frac{\partial\left(\ln\left|\boldsymbol{\Sigma}_i\right|\right)}{\partial \boldsymbol{\Sigma}_i} - \frac{1}{2}\frac{\partial\left[\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\right]}{\partial \boldsymbol{\Sigma}_i}\right]$$

$$= p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \cdot \left[-\frac{1}{2}\boldsymbol{\Sigma}_i^{-1} + \frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}\right].$$

矩阵微分式：

$$\frac{\partial|\boldsymbol{X}|}{\partial \boldsymbol{X}} = |\boldsymbol{X}| \cdot \left(\boldsymbol{X}^{-1}\right)^{\mathrm{T}}$$

$$\frac{\partial \boldsymbol{a}^{\mathrm{T}} \boldsymbol{X}^{-1} \boldsymbol{b}}{\partial \boldsymbol{X}} = -\boldsymbol{X}^{-\mathrm{T}} \boldsymbol{a} \boldsymbol{b}^{\mathrm{T}} \boldsymbol{X}^{-\mathrm{T}}$$

将此式代回 $\frac{\partial LL(D)}{\partial \boldsymbol{\Sigma}_i}$ 中可得

$$\frac{\partial LL(D)}{\partial \boldsymbol{\Sigma}_i} = \sum_{j=1}^{m} \frac{\alpha_i \cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \cdot \left[-\frac{1}{2}\boldsymbol{\Sigma}_i^{-1} + \frac{1}{2}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\right],$$

又由式(9.30)可知$\frac{\alpha_i \cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} = \gamma_{ji}$，所以上式可进一步化简为

$$\frac{\partial LL(D)}{\partial \boldsymbol{\Sigma}_i} = \sum_{j=1}^{m} \gamma_{ji} \cdot \left[-\frac{1}{2}\boldsymbol{\Sigma}_i^{-1} + \frac{1}{2}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1} \right].$$

令上式等于0可得

$$\frac{\partial LL(D)}{\partial \boldsymbol{\Sigma}_i} = \sum_{j=1}^{m} \gamma_{ji} \cdot \left(-\frac{1}{2}\boldsymbol{\Sigma}_i^{-1} + \frac{1}{2}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1} \right) = 0.$$

移项推导有

$$\sum_{j=1}^{m} \gamma_{ji} \cdot \left(-\boldsymbol{I} + \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1} \right) = 0,$$

$$\sum_{j=1}^{m} \gamma_{ji}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1} = \sum_{j=1}^{m} \gamma_{ji}\boldsymbol{I},$$

$$\sum_{j=1}^{m} \gamma_{ji}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\mathrm{T}} = \sum_{j=1}^{m} \gamma_{ji}\boldsymbol{\Sigma}_i,$$

$$\boldsymbol{\Sigma}_i^{-1} \cdot \sum_{j=1}^{m} \gamma_{ji}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\mathrm{T}} = \sum_{j=1}^{m} \gamma_{ji},$$

$$\boldsymbol{\Sigma}_i = \frac{\sum_{j=1}^{m} \gamma_{ji}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\mathrm{T}}}{\sum_{j=1}^{m} \gamma_{ji}}.$$

此即式(9.35).

式(9.38)

$$\alpha_i = \frac{1}{m}\sum_{j=1}^{m}\gamma_{ji}$$

解析

式(9.37)两边同时乘以α_i可得

$$\sum_{j=1}^{m}\frac{\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\sum_{l=1}^{k}\alpha_l \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right)} + \lambda\alpha_i = 0,$$

$$\sum_{j=1}^{m}\frac{\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\sum_{l=1}^{k}\alpha_l \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right)} = -\lambda\alpha_i.$$

两边对所有混合成分求和可得

$$\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\sum_{l=1}^{k}\alpha_l \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right)} = -\lambda\sum_{i=1}^{k}\alpha_i,$$

$$\sum_{j=1}^{m}\sum_{i=1}^{k}\frac{\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\sum_{l=1}^{k}\alpha_l \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right)} = -\lambda\sum_{i=1}^{k}\alpha_i.$$

由$m = -\lambda$有

$$\sum_{j=1}^{m}\frac{\alpha_i \cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k}\alpha_l \cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} = -\lambda\alpha_i = m\alpha_i,$$

因此

$$\alpha_i = \frac{1}{m}\sum_{j=1}^{m}\frac{\alpha_i \cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k}\alpha_l \cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}.$$

又由式(9.30)可知$\frac{\alpha_i \cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k}\alpha_l \cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} = \gamma_{ji}$，所以上式可进一步化简

为

$$\alpha_i = \frac{1}{m}\sum_{j=1}^{m}\gamma_{ji}.$$

此即式(9.38).

第10章　降维与度量学习

式(10.1)

$$P(err) = 1 - \sum_{c \in \mathcal{Y}} P(c|\boldsymbol{x}) P(c|\boldsymbol{z})$$

解析

$P(c|\boldsymbol{x})P(c|\boldsymbol{z})$表示$\boldsymbol{x}$和$\boldsymbol{z}$同属类$c$的概率，对所有可能的类别$c \in \mathcal{Y}$求和，则得到$\boldsymbol{x}$和$\boldsymbol{z}$同属相同类别的概率，因此$1 - \sum_{c \in \mathcal{Y}} P(c|\boldsymbol{x}) P(c|\boldsymbol{z})$表示$\boldsymbol{x}$和$\boldsymbol{z}$分属不同类别的概率.

式(10.2)

$$P(err) = 1 - \sum_{c \in \mathcal{Y}} P(c \mid \boldsymbol{x}) P(c \mid \boldsymbol{z}) \quad ①$$

$$\simeq 1 - \sum_{c \in \mathcal{Y}} P^2(c \mid \boldsymbol{x}) \quad ②$$

$$\leqslant 1 - P^2\left(c^* \mid \boldsymbol{x}\right) \quad ③$$

$$= \left(1 + P\left(c^* \mid \boldsymbol{x}\right)\right)\left(1 - P\left(c^* \mid \boldsymbol{x}\right)\right) \quad ④$$

$$\leqslant 2 \times \left(1 - P\left(c^* \mid \boldsymbol{x}\right)\right) \quad ⑤$$

解析

①→②来源于前提假设“假设样本独立同分布，且对任意$\boldsymbol{x}$和任意小正数δ，在x附近δ距离范围内总能找到一个训练样本”，假设所有δ中最小的δ组成和$\boldsymbol{x}$维数相同的向量$\boldsymbol{\delta}$，则$P(c|\boldsymbol{z}) = P(c|\boldsymbol{x} \pm \boldsymbol{\delta}) \simeq P(c|\boldsymbol{x})$.

②→③是因为$c^* \in \mathcal{Y}$，则$P^2(c^*|\boldsymbol{x})$是$\sum_{c\in\mathcal{Y}} P^2(c|\boldsymbol{x})$的一个分量，所以$\sum_{c\in\mathcal{Y}} P^2(c|\boldsymbol{x}) \geqslant P^2(c^*|\boldsymbol{x})$.

③→④是平方差式展开.

④→⑤是因为$1 + P^2(c^*|\boldsymbol{x}) \leqslant 2$.

式**(10.3)**

$$\begin{aligned} dist_{ij}^2 &= \|\boldsymbol{z}_i\|^2 + \|\boldsymbol{z}_j\|^2 - 2\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j \\ &= b_{ii} + b_{jj} - 2b_{ij} \end{aligned}$$

解析

$$dist_{ij}^2 = \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2$$

$$= (\boldsymbol{z}_i - \boldsymbol{z}_j)^{\mathrm{T}}(\boldsymbol{z}_i - \boldsymbol{z}_j)$$

$$= \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i - \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j - \boldsymbol{z}_j^{\mathrm{T}}\boldsymbol{z}_i + \boldsymbol{z}_j^{\mathrm{T}}\boldsymbol{z}_j$$

$$= \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i + \boldsymbol{z}_j^{\mathrm{T}}\boldsymbol{z}_j - 2\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j$$

$$= \|\boldsymbol{z}_i\|^2 + \|\boldsymbol{z}_j\|^2 - 2\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j$$

$$= b_{ii} + b_{jj} - 2b_{ij}$$

式**(10.4）**

$$\sum_{i=1}^{m} dist_{ij}^2 = \mathrm{tr}(\boldsymbol{B}) + mb_{jj}$$

解析

首先根据式(10.3)有

$$\sum_{i=1}^{m} dist_{ij}^2 = \sum_{i=1}^{m} b_{ii} + \sum_{i=1}^{m} b_{jj} - 2\sum_{i=1}^{m} b_{ij}.$$

对于第一项，根据矩阵迹的定义，$\sum_{i=1}^{m} b_{ii} = \mathrm{tr}(\boldsymbol{B})$.对于第二项，由于求和号内元素和$i$无关，因此$\sum_{i=1}^{m} b_{jj} = mb_{jj}$.对于第三项，有

$$\sum_{i=1}^{m} b_{ij} = \sum_{i=1}^{m} \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j = \sum_{i=1}^{m} \boldsymbol{z}_j^{\mathrm{T}}\boldsymbol{z}_i = \boldsymbol{z}_j^{\mathrm{T}}\sum_{i=1}^{m} \boldsymbol{z}_i = \boldsymbol{z}_j^{\mathrm{T}}\boldsymbol{0} = 0.$$

$\sum_{i=1}^{m} \boldsymbol{z}_i = \boldsymbol{0}$是利用了“西瓜书”上的前提条件，即将降维后的样本被中心化

式**(10.5**）

$$\sum_{j=1}^{m} dist_{ij}^2 = \mathrm{tr}(\boldsymbol{B}) + mb_{ii}$$

参见式(10.4)

式**(10.6**）

$$\sum_{i=1}^{m}\sum_{j=1}^{m} dist_{ij}^2 = 2m\,\mathrm{tr}(\boldsymbol{B})$$

解析

$$\sum_{i=1}^{m}\sum_{j=1}^{m} dist_{ij}^2 = \sum_{i=1}^{m}\sum_{j=1}^{m}\left(\|\boldsymbol{z}_i\|^2 + \|\boldsymbol{z}_j\|^2 - 2\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j\right)$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{m}\|\boldsymbol{z}_i\|^2 + \sum_{i=1}^{m}\sum_{j=1}^{m}\|\boldsymbol{z}_j\|^2 - 2\sum_{i=1}^{m}\sum_{j=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j,$$

其中

$$\sum_{i=1}^{m}\sum_{j=1}^{m}\|\boldsymbol{z}_i\|^2 = m\sum_{i=1}^{m}\|\boldsymbol{z}_i\|^2 = m\,\mathrm{tr}(\boldsymbol{B}),$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m}\|\boldsymbol{z}_j\|^2 = m\sum_{j=1}^{m}\|\boldsymbol{z}_j\|^2 = m\,\mathrm{tr}(\boldsymbol{B}),$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j = 0$$

“西瓜书”中假设降维后的样本$\boldsymbol{Z}$被中心化

式**(10.10)**

$$b_{ij} = -\frac{1}{2}(dist_{ij}^2 - dist_{i\cdot}^2 - dist_{\cdot j}^2 + dist_{\cdot\cdot}^2)$$

解析

由式(10.3)可得

$$b_{ij} = -\frac{1}{2}(dist_{ij}^2 - b_{ii} - b_{jj});$$

由式(10.6)和式(10.9)可得

$$\begin{aligned}\mathrm{tr}(\boldsymbol{B}) &= \frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{m} dist_{ij}^2 \\ &= \frac{m}{2} dist_{\cdot\cdot}^2\end{aligned}$$

由式(10.4)和式(10.8)可得

$$\begin{aligned}b_{jj} &= \frac{1}{m}\sum_{i=1}^{m} dist_{ij}^2 - \frac{1}{m}\mathrm{tr}(\boldsymbol{B}) \\ &= dist_{\cdot j}^2 - \frac{1}{2} dist_{\cdot\cdot}^2;\end{aligned}$$

由式(10.5)和式(10.7)可得

$$\begin{aligned}b_{ii} &= \frac{1}{m}\sum_{j=1}^{m} dist_{ij}^2 - \frac{1}{m}\mathrm{tr}(\boldsymbol{B}) \\ &= dist_{i\cdot}^2 - \frac{1}{2} dist_{\cdot\cdot}^2;\end{aligned}$$

综合可得

$$\begin{aligned}b_{ij} &= -\frac{1}{2}\left(dist_{ij}^2 - b_{ii} - b_{jj}\right) \\ &= -\frac{1}{2}\left(dist_{ij}^2 - dist_{i}^2 + \frac{1}{2} dist_{\cdot\cdot}^2 - dist_{\cdot j}^2 + \frac{1}{2} dist_{\cdot\cdot}^2\right) \\ &= -\frac{1}{2}\left(dist_{ij}^2 - dist_{i\cdot}^2 - dist_{\cdot j}^2 + dist_{\cdot\cdot}^2\right).\end{aligned}$$

式**(10.14)**

$$\sum_{i=1}^{m}\left\|\sum_{j=1}^{d'} z_{ij}\boldsymbol{w}_j-\boldsymbol{x}_i\right\|_2^2=\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i-2\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i+\text{const}$$

$$\propto-\operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}}\left(\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right)\boldsymbol{W}\right)$$

解析

已知$\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{I},\boldsymbol{z}_i=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i$，则

$$\sum_{i=1}^{m}\left\|\sum_{j=1}^{d'} z_{ij}\boldsymbol{w}_j-\boldsymbol{x}_i\right\|_2^2=\sum_{i=1}^{m}\|\boldsymbol{W}\boldsymbol{z}_i-\boldsymbol{x}_i\|_2^2$$

$$=\sum_{i=1}^{m}\left(\boldsymbol{W}\boldsymbol{z}_i-\boldsymbol{x}_i\right)^{\mathrm{T}}\left(\boldsymbol{W}\boldsymbol{z}_i-\boldsymbol{x}_i\right)$$

$$=\sum_{i=1}^{m}\left(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{z}_i-\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}\boldsymbol{z}_i+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i\right)$$

$$=\sum_{i=1}^{m}\left(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i-2\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i\right)$$

$$=\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i-2\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i+\sum_{i=1}^{m}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i$$

$$=\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i-2\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_i+\text{const}$$

$$=\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i-2\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i+\text{const}$$

$$= -\sum_{i=1}^{m} \boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{z}_i + \text{const}$$

$$= -\sum_{i=1}^{m} \operatorname{tr}\left(\boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}}\right) + \text{const}$$

$$= -\operatorname{tr}\left(\sum_{i=1}^{m} \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}}\right) + \text{const}$$

$$= -\operatorname{tr}\left(\sum_{i=1}^{m} \boldsymbol{W}^{\mathrm{T}} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}\right) + \text{const}$$

$$\propto -\operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}} \left(\sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right) \boldsymbol{W}\right) + \text{const}$$

$$\propto -\operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}} \left(\sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right) \boldsymbol{W}\right)$$

式**(10.17)**

$$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}_i = \lambda_i \boldsymbol{w}_i$$

解析

由式(10.15)可知，主成分分析的优化目标为

$$\begin{aligned} \min_{\boldsymbol{W}} \quad & -\operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\right) \\ \text{s.t.} \quad & \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} = \boldsymbol{I}, \end{aligned}$$

其中$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m) \in \mathbb{R}^{d \times m}$, $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_{d'}) \in \mathbb{R}^{d \times d'}$, $\boldsymbol{I} \in \mathbb{R}^{d' \times d'}$为单位矩阵.对于带矩阵约束的优化问题，其优化目标的拉格朗日函数为

读者可搜索Lagrangian optimization with matrix constrain s了解更多

$$\begin{aligned} L(\boldsymbol{W}, \boldsymbol{\Theta}) &= -\operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W}\right) + \langle \boldsymbol{\Theta}, \boldsymbol{W}^{\mathrm{T}} \boldsymbol{W} - \mathbf{I} \rangle \\ &= -\operatorname{tr}(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W}) + \operatorname{tr}\left(\boldsymbol{\Theta}^{\mathrm{T}} (\boldsymbol{W}^{\mathrm{T}} \boldsymbol{W} - \mathbf{I})\right) \end{aligned}$$

其中，$\boldsymbol{\Theta} \in \mathbb{R}^{d' \times d'}$为拉格朗日乘子矩阵，其维度恒等于约束条件的维度，且其中的每个元素均为未知的拉格朗日乘子；
$\langle \boldsymbol{\Theta}, \boldsymbol{W}^{\mathrm{T}} \boldsymbol{W} - \boldsymbol{I} \rangle = \operatorname{tr}\left(\boldsymbol{\Theta}^{\mathrm{T}} (\boldsymbol{W}^{\mathrm{T}} \boldsymbol{W} - \boldsymbol{I})\right)$为矩阵的内积.若此时仅考虑约束$\boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{w}_i = 1 (i = 1, 2, \cdots, d')$，则拉格朗日乘子矩阵$\boldsymbol{\Theta}$此时为对角矩阵，令新的拉格朗日乘子矩阵为$\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \cdots, \lambda_{d'}) \in \mathbb{R}^{d' \times d'}$，则新的拉格朗日函数为

读者可搜索Frobenius in ner product 了解更多

$$L(\boldsymbol{W}, \boldsymbol{\Lambda}) = -\operatorname{tr}(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W}) + \operatorname{tr}\left(\boldsymbol{\Lambda}^{\mathrm{T}} (\boldsymbol{W}^{\mathrm{T}} \boldsymbol{W} - \boldsymbol{I})\right).$$

对拉格朗日函数关于$\boldsymbol{W}$求导可得

$$\begin{aligned} &\frac{\partial L(\boldsymbol{W}, \boldsymbol{\Lambda})}{\partial \boldsymbol{W}} = \frac{\partial}{\partial \boldsymbol{W}} \left[-\operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W}\right) + \operatorname{tr}\left(\boldsymbol{\Lambda}^{\mathrm{T}} \left(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{W} - \boldsymbol{I}\right)\right) \right] \\ &= -\frac{\partial}{\partial \boldsymbol{W}} \operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W}\right) + \frac{\partial}{\partial \boldsymbol{W}} \operatorname{tr}\left(\boldsymbol{\Lambda}^{\mathrm{T}} \left(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{W} - \boldsymbol{I}\right)\right) \\ &= -2 \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W} + \boldsymbol{W} \boldsymbol{\Lambda} + \boldsymbol{W} \boldsymbol{\Lambda}^{\mathrm{T}} \\ &= -2 \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W} + \boldsymbol{W} \left(\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^{\mathrm{T}}\right) \\ &= -2 \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W} + 2 \boldsymbol{W} \boldsymbol{\Lambda} \end{aligned}$$

矩阵微分式：

$$\frac{\partial}{\partial \boldsymbol{X}} \operatorname{tr}\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{B} \boldsymbol{X}\right) = \boldsymbol{B} \boldsymbol{X} + \boldsymbol{B}^{\mathrm{T}} \boldsymbol{X},$$

$$\frac{\partial}{\partial \boldsymbol{X}} \operatorname{tr}\left(\boldsymbol{B}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right) = \boldsymbol{X}\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{X}\boldsymbol{B}$$

令$\frac{\partial L(\boldsymbol{W}, \boldsymbol{\Lambda})}{\partial \boldsymbol{W}} = \boldsymbol{0}$可得

$$\begin{aligned} -2\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W} + 2\boldsymbol{W}\boldsymbol{\Lambda} &= \boldsymbol{0}, \\ \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W} &= \boldsymbol{W}\boldsymbol{\Lambda}. \end{aligned}$$

将$\boldsymbol{W}$和$\boldsymbol{\Lambda}$展开可得

$$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}_i = \lambda_i \boldsymbol{w}_i, \quad i = 1, 2, \cdots, d'.$$

显然，此式为矩阵特征值和特征向量的定义式，其中λ_i和$\boldsymbol{w}_i$分别表示矩阵$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$的特征值和单位特征向量.由于以上是仅考虑约束$\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{w}_i = 1$所求得的结果，而$\boldsymbol{w}_i$还需满足约束$\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{w}_j = 0(i \neq j)$. 观察$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$的定义可知，$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$是一个实对称矩阵，实对称矩阵的不同特征值所对应的特征向量之间相互正交，同一特征值的不同特征向量可以通过施密特正交化使其变得正交，所以通过上式求得的$\boldsymbol{w}_i$可以同时满足约束$\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{w}_i = 1$和$\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{w}_j = 0(i \neq j)$.根据拉格朗日乘子法的原理可知，此时求得的结果仅是最优解的必要条件，而且$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$有$d$个相互正交的单位特征向量，所以还需要从这$d$个特征向量里找出$d'$个能使得目标函数达到最优值的特征向量作为最优解.将$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}_i = \lambda_i \boldsymbol{w}_i$代入目标函数可得

$$\begin{aligned} &\min_{\boldsymbol{W}} -\operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\right) = \max_{\boldsymbol{W}} \operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\right) \\ &= \max_{\boldsymbol{W}} \sum_{i=1}^{d'} \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}_i \\ &= \max_{\boldsymbol{W}} \sum_{i=1}^{d'} \boldsymbol{w}_i^{\mathrm{T}}\lambda_i\boldsymbol{w}_i \end{aligned}$$

$$= \max_{\boldsymbol{W}} \sum_{i=1}^{d'} \lambda_i \boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{w}_i$$

$$= \max_{\boldsymbol{W}} \sum_{i=1}^{d'} \lambda_i .$$

显然，此时只需要令$\lambda_1, \lambda_2, \cdots, \lambda_{d'}$和$\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_{d'}$分别为矩阵$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$的前$d'$个最大的特征值和单位特征向量就能使得目标函数达到最优值.

式**(10.24)**

$$\boldsymbol{K}\boldsymbol{\alpha}^j = \lambda_j \boldsymbol{\alpha}^j$$

解析

已知$\boldsymbol{z}_i = \phi(\boldsymbol{x}_i)$，类比$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m\}$可以构造$\boldsymbol{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_m\}$，所以式(10.21)可变换为

$$\left(\sum_{i=1}^{m} \phi(\boldsymbol{x}_i) \phi(\boldsymbol{x}_i)^{\mathrm{T}} \right) \boldsymbol{w}_j = \left(\sum_{i=1}^{m} \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}} \right) \boldsymbol{w}_j = \boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{w}_j = \lambda_j \boldsymbol{w}_j .$$

又由式(10.22)可知

$$\boldsymbol{w}_j = \sum_{i=1}^{m} \phi\left(\boldsymbol{x}_i\right) \alpha_i^j = \sum_{i=1}^{m} \boldsymbol{z}_i \alpha_i^j = \boldsymbol{Z}\boldsymbol{\alpha}^j ,$$

其中，$\boldsymbol{\alpha}^j = (\alpha_1^j; \alpha_2^j; \cdots; \alpha_m^j) \in \mathbb{R}^{m \times 1}$, 所以式(10.21)可以进一步变换为

$$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z}\boldsymbol{\alpha}^j = \lambda_j \boldsymbol{Z}\boldsymbol{\alpha}^j ,$$
$$\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z}\boldsymbol{\alpha}^j = \boldsymbol{Z}\lambda_j \boldsymbol{\alpha}^j .$$

此时的目标是求出$\boldsymbol{w}_j$，等价于求出满足上式的$\boldsymbol{\alpha}^j$.显然，此时满足

$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{\alpha}^j = \lambda_j\boldsymbol{\alpha}^j$的$\boldsymbol{\alpha}^j$一定满足上式，所以问题转化为了求解$\boldsymbol{\alpha}^j$满足

$$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{\alpha}^j = \lambda_j\boldsymbol{\alpha}^j.$$

令$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z} = \boldsymbol{K}$，那么上式可化为

$$\boldsymbol{K}\boldsymbol{\alpha}^j = \lambda_j\boldsymbol{\alpha}^j,$$

即式(10.24)，其中矩阵$\boldsymbol{K}$的第i行第j列的元素
$K_{ij} = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j = \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j) = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

式(10.28)

$$w_{ij} = \frac{\sum\limits_{k\in Q_i} C_{jk}^{-1}}{\sum\limits_{l,s\in Q_i} C_{ls}^{-1}}$$

解析

由“西瓜书”中上下文可知，式(10.28)是如下优化问题的解：

$$\min_{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_m} \sum_{i=1}^{m} \left\| \boldsymbol{x}_i - \sum_{j\in Q_i} w_{ij}\boldsymbol{x}_j \right\|_2^2$$

$$\text{s.t.} \sum_{j\in Q_i} w_{ij} = 1.$$

若令$\boldsymbol{x}_i \in \mathbb{R}^{d\times 1}, Q_i = \{q_i^1, q_i^2, \cdots, q_i^n\}$，则上述优化问题的目标函数可以进行如下恒等变形

$$\sum_{i=1}^{m}\left\|\boldsymbol{x}_i-\sum_{j\in Q_i}w_{ij}\boldsymbol{x}_j\right\|_2^2=\sum_{i=1}^{m}\left\|\sum_{j\in Q_i}w_{ij}\boldsymbol{x}_i-\sum_{j\in Q_i}w_{ij}\boldsymbol{x}_j\right\|_2^2$$

$$=\sum_{i=1}^{m}\left\|\sum_{j\in Q_i}w_{ij}\left(\boldsymbol{x}_i-\boldsymbol{x}_j\right)\right\|_2^2$$

$$=\sum_{i=1}^{m}\left\|\boldsymbol{X}_i\boldsymbol{w}_{\boldsymbol{i}}\right\|_2^2$$

$$=\sum_{i=1}^{m}\boldsymbol{w}_{\boldsymbol{i}}^{\mathrm{T}}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{w}_{\boldsymbol{i}},$$

其中$\boldsymbol{w}_{\boldsymbol{i}}=\left(w_{iq_i^1};w_{iq_i^2};\cdots;w_{iq_i^n}\right)\in\mathbb{R}^{n\times1}$，$\boldsymbol{X}_i=\left(\boldsymbol{x}_i-\boldsymbol{x}_{q_i^1},\boldsymbol{x}_i-\boldsymbol{x}_{q_i^2},\cdots,\boldsymbol{x}_i-\boldsymbol{x}_{q_i^n}\right)\in\mathbb{R}^{d\times n}$. 同理，约束条件也可以进行如下恒等变形：

$$\sum_{j\in Q_i}w_{ij}=\boldsymbol{w}_{\boldsymbol{i}}{}^{\mathrm{T}}\boldsymbol{I}=1,$$

其中$\boldsymbol{I}=(1;1;\cdots;1)\in\mathbb{R}^{n\times1}$为$n$行1列的单位向量. 因此，上述优化问题可以重写为

$$\min_{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_m}\sum_{i=1}^{m}\boldsymbol{w}_{\boldsymbol{i}}^{\mathrm{T}}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{w}_{\boldsymbol{i}}$$
$$\text{s.t. }\boldsymbol{w}_{\boldsymbol{i}}^{\mathrm{T}}\boldsymbol{I}=1.$$

显然，此问题为带约束的优化问题，因此可以考虑使用拉格朗日乘子法来进行求解.由拉格朗日乘子法可得此优化问题的拉格朗日函数

$$L(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_m,\lambda)=\sum_{i=1}^{m}\boldsymbol{w}_{\boldsymbol{i}}{}^{\mathrm{T}}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{w}_{\boldsymbol{i}}+\lambda\left(\boldsymbol{w}_{\boldsymbol{i}}{}^{\mathrm{T}}\boldsymbol{I}-1\right)$$

对拉格朗日函数关于$\boldsymbol{w}_i$求偏导并令其等于**0**有

$$\frac{\partial L\left(\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_m, \lambda\right)}{\partial \boldsymbol{w}_{\boldsymbol{i}}}=\mathbf{0},$$

$$\frac{\partial\left[\sum_{i=1}^m \boldsymbol{w}_{\boldsymbol{i}}^{\mathrm{T}} \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_i \boldsymbol{w}_{\boldsymbol{i}}+\lambda\left(\boldsymbol{w}_{\boldsymbol{i}}^{\mathrm{T}} \boldsymbol{I}-1\right)\right]}{\partial \boldsymbol{w}_i}=\mathbf{0},$$

$$\frac{\partial\left[\boldsymbol{w}_{\boldsymbol{i}}^{\mathrm{T}} \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_i \boldsymbol{w}_{\boldsymbol{i}}+\lambda\left(\boldsymbol{w}_{\boldsymbol{i}}^{\mathrm{T}} \boldsymbol{I}-1\right)\right]}{\partial \boldsymbol{w}_{\boldsymbol{i}}}=\mathbf{0},$$

$$2 \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_i \boldsymbol{w}_{\boldsymbol{i}}+\lambda \boldsymbol{I}=\mathbf{0},$$

$$\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_i \boldsymbol{w}_{\boldsymbol{i}}=-\frac{1}{2} \lambda \boldsymbol{I}.$$

矩阵微分式：

$$\frac{\partial \boldsymbol{x}^{\mathrm{T}} \boldsymbol{B} \boldsymbol{x}}{\partial \boldsymbol{x}}=\left(\boldsymbol{B}+\boldsymbol{B}^{\mathrm{T}}\right) \boldsymbol{x},$$

$$\frac{\partial \boldsymbol{x}^{\mathrm{T}} \boldsymbol{a}}{\partial \boldsymbol{x}}=\boldsymbol{a}$$

若$\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_i$可逆，则

$$\boldsymbol{w}_{\boldsymbol{i}}=-\frac{1}{2} \lambda(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_i)^{-1} \boldsymbol{I}.$$

又因为$\boldsymbol{w}_{\boldsymbol{i}}{}^{\mathrm{T}} \boldsymbol{I}=\boldsymbol{I}^{\mathrm{T}} \boldsymbol{w}_{\boldsymbol{i}}=1$，则上式两边同时左乘$\boldsymbol{I}^{\mathrm{T}}$可得

$$\boldsymbol{I}^{\mathrm{T}} \boldsymbol{w}_{\boldsymbol{i}}=-\frac{1}{2} \lambda \boldsymbol{I}^{\mathrm{T}}(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_i)^{-1} \boldsymbol{I}=1,$$

即

$$-\frac{1}{2} \lambda=\frac{1}{\boldsymbol{I}^{\mathrm{T}}(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_i)^{-1} \boldsymbol{I}}.$$

将其代回$\boldsymbol{w_i} = -\frac{1}{2}\lambda(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i)^{-1}\boldsymbol{I}$即可解得

$$\boldsymbol{w_i} = \frac{(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i)^{-1}\boldsymbol{I}}{\boldsymbol{I}^{\mathrm{T}}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i)^{-1}\boldsymbol{I}}.$$

设矩阵$(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i)^{-1}$第$j$行第$k$列的元素为$C_{jk}^{-1}$，则

$$w_{ij} = w_{iq_i^j} = \frac{\sum\limits_{k\in Q_i} C_{jk}^{-1}}{\sum\limits_{l,s\in Q_i} C_{ls}^{-1}},$$

即式(10.28). 显然，若$\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i$可逆，此优化问题即凸优化问题，且此时用拉格朗日乘子法求得的$\boldsymbol{w}_i$为全局最优解.

式**(10.31)**

$$\begin{aligned} \min_{\boldsymbol{Z}} \quad & \operatorname{tr}\left(\boldsymbol{Z}\boldsymbol{M}\boldsymbol{Z}^{\mathrm{T}}\right) \\ \text{s.t.} \quad & \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z} = \boldsymbol{I} \end{aligned}$$

约束条件$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z} = \boldsymbol{I}$是为了得到标准化（标准正交空间）的低维数据

解析

$$\min_{Z} \sum_{i=1}^{m} \left\| \boldsymbol{z}_i - \sum_{j\in Q_i} w_{ij}\boldsymbol{z}_j \right\|_2^2 = \sum_{i=1}^{m} \|\boldsymbol{Z}\boldsymbol{I}_i - \boldsymbol{Z}\boldsymbol{W}_i\|_2^2$$

$$= \sum_{i=1}^{m} \|\boldsymbol{Z}\left(\boldsymbol{I}_i - \boldsymbol{W}_i\right)\|_2^2$$

$$= \sum_{i=1}^{m} \left(\boldsymbol{Z}\left(\boldsymbol{I}_i - \boldsymbol{W}_i\right)\right)^{\mathrm{T}} \boldsymbol{Z}\left(\boldsymbol{I}_i - \boldsymbol{W}_i\right)$$

$$= \sum_{i=1}^{m} (\boldsymbol{I}_i - \boldsymbol{W}_i)^{\mathrm{T}} \boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z} (\boldsymbol{I}_i - \boldsymbol{W}_i)$$

$$= \mathrm{tr}\left((\boldsymbol{I} - \boldsymbol{W})^{\mathrm{T}} \boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z} (\boldsymbol{I} - \boldsymbol{W})\right)$$

$$= \mathrm{tr}\left(\boldsymbol{Z} (\boldsymbol{I} - \boldsymbol{W})(\boldsymbol{I} - \boldsymbol{W})^{\mathrm{T}} \boldsymbol{Z}^{\mathrm{T}}\right)$$

$$= \mathrm{tr}\left(\boldsymbol{Z} \boldsymbol{M} \boldsymbol{Z}^{\mathrm{T}}\right),$$

其中$\boldsymbol{M} = (\boldsymbol{I} - \boldsymbol{W})(\boldsymbol{I} - \boldsymbol{W})^{\mathrm{T}}$.

第11章　特征选择与稀疏学习

式(11.1)

$$\mathrm{Gain}(A)=\mathrm{Ent}(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\mathrm{Ent}\left(D^v\right)$$

解析

此为信息增益的定义式.对于数据集D和属性子集A，假设根据A的取值将D分为了V个子集$\{D^1,D^2,\cdots,D^V\}$，那么信息增益的定义为划分之前数据集D的信息熵和划分之后每个子数据集D^v的信息熵的差.

熵用来衡量一个系统的混乱程度，因此划分前和划分后熵的差越大，表示划分越有效，划分带来的“信息增益”越大.

式(11.2)

$$\mathrm{Ent}(D)=-\sum_{k=1}^{|\mathcal{Y}|}p_k log_2 p_k$$

解析

此为信息熵的定义式. 其中$p_k(k=1,2,\cdots,|\mathcal{Y}|)$表示$D$中第$k$类样本所占的比例.可以看出，样本越纯，即$p_k\to 0$或$p_k\to 1$时，$\mathrm{Ent}(D)$越小，其最小值 为0.

此处约定$0\log_2 0 = 0$

式**(11.5)**

$$\min_{\boldsymbol{w}} \sum_{i=1}^{m} \left(y_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i\right)^2$$

解析

该式为线性回归的优化目标式y_i表示样本i的真实值，而$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i$表示其预测值，这里使用预测值和真实值差的平方衡量预测值偏离真实值的程度.

式**(11.6)**

$$\min_{\boldsymbol{w}} \sum_{i=1}^{m} \left(y_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i\right)^2 + \lambda||\boldsymbol{w}||_2^2$$

解析

该式为加入了L_2正规化项的优化目标，也叫“岭回归”.λ用来调节误差项和正规化项的相对重要性.引入正规化项的目的是降低因$\boldsymbol{w}$的分量过太而导致过拟合的风险.

式**(11.7)**

$$\min_{\boldsymbol{w}} \sum_{i=1}^{m} \left(y_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i\right)^2 + \lambda||\boldsymbol{w}||_1$$

解析

该式将式(11.6)中的L_2正规化项替换成了L_1正规化项.关于L_2和L_1两个正规化项的区别，“西瓜书”图11.2给出了很形象的解释.具体来说，结合L_1范数优化的模型参数分量更偏向于取0，因此更容易取得稀疏解.

式(11.10)

$$\begin{aligned}\hat{f}(\boldsymbol{x}) &\simeq f\left(\boldsymbol{x}_k\right)+\left\langle\nabla f\left(\boldsymbol{x}_k\right), \boldsymbol{x}-\boldsymbol{x}_k\right\rangle+\frac{L}{2}\left\|\boldsymbol{x}-\boldsymbol{x}_k\right\|^2 \\ &=\frac{L}{2}\left\|\boldsymbol{x}-\left(\boldsymbol{x}_k-\frac{1}{L} \nabla f\left(\boldsymbol{x}_k\right)\right)\right\|_2^2+\text{const}\end{aligned}$$

解析

首先注意优化目标式和式(11.7) LASSO回归的联系和区别，本式中的$\boldsymbol{x}$对应到式(11.7)的$\boldsymbol{w}$，即我们优化的目标.再解释下什么是$L-\text{Lipschitz}$条件：根据维基百科的定义，它是一个比“连续”更强的光滑性条件.直觉上，利普希茨连续函数限制了函数改变的速度，符合利普希茨条件的函数的斜率，必小于一个称为利普希茨常数的实数（该常数依函数而定）.注意这里存在一个笔误，根据维基百科的定义，式(11.9)应该写成

$$\left|\nabla f\left(\boldsymbol{x}'\right)-\nabla f(\boldsymbol{x})\right| \leqslant L\left|\boldsymbol{x}'-\boldsymbol{x}\right| \quad \left(\forall \boldsymbol{x}, \boldsymbol{x}'\right),$$

移项得

$$\frac{\left|\nabla f\left(\boldsymbol{x}'\right)-\nabla f(\boldsymbol{x})\right|}{\left|x'-x\right|} \leqslant L \quad \left(\forall \boldsymbol{x}, \boldsymbol{x}'\right).$$

由于上式对所有的$\boldsymbol{x}$和 $\boldsymbol{x}'$都成立，由导数的定义，上式可以看成是$f(\boldsymbol{x})$的二阶导数恒不大于L, 即 $\nabla^2 f(\boldsymbol{x}) \leqslant L$.

接下来推导式(11.10). 由泰勒式可知，$\boldsymbol{x}_k$附近的$f(\boldsymbol{x})$通过二阶泰勒展开可近似为

$$\hat{f}(\boldsymbol{x}) \simeq f\left(\boldsymbol{x}_k\right)+\left\langle\nabla f\left(\boldsymbol{x}_k\right), \boldsymbol{x}-\boldsymbol{x}_k\right\rangle+\frac{\nabla^2 f\left(\boldsymbol{x}_k\right)}{2}\left\|\boldsymbol{x}-\boldsymbol{x}_k\right\|^2$$

$$\leqslant f\left(\boldsymbol{x}_k\right)+\left\langle\nabla f\left(\boldsymbol{x}_k\right), \boldsymbol{x}-\boldsymbol{x}_k\right\rangle+\frac{L}{2}\left\|\boldsymbol{x}-\boldsymbol{x}_k\right\|^2$$

$$=f\left(\boldsymbol{x}_k\right)+\nabla f\left(\boldsymbol{x}_k\right)^{\mathrm{T}}\left(\boldsymbol{x}-\boldsymbol{x}_k\right)+\frac{L}{2}\left(\boldsymbol{x}-\boldsymbol{x}_k\right)^{\mathrm{T}}\left(\boldsymbol{x}-\boldsymbol{x}_k\right)$$

$$=f\left(\boldsymbol{x}_k\right)+\frac{L}{2}\left(\left(\boldsymbol{x}-\boldsymbol{x}_k\right)^{\mathrm{T}}\left(\boldsymbol{x}-\boldsymbol{x}_k\right)+\frac{2}{L}\nabla f\left(\boldsymbol{x}_k\right)^{\mathrm{T}}\left(\boldsymbol{x}-\boldsymbol{x}_k\right)\right)$$

$$=f\left(\boldsymbol{x}_k\right)+\frac{L}{2}\left(\left(\boldsymbol{x}-\boldsymbol{x}_k\right)^{\mathrm{T}}\left(\boldsymbol{x}-\boldsymbol{x}_k\right)+\frac{2}{L}\nabla f\left(\boldsymbol{x}_k\right)^{\mathrm{T}}\left(\boldsymbol{x}-\boldsymbol{x}_k\right)\right.$$

$$\left.+\frac{1}{L^2}\nabla f\left(x_k\right)^{\mathrm{T}}\nabla f\left(x_k\right)\right)-\frac{1}{2L}\nabla f\left(x_k\right)^{\mathrm{T}}\nabla f\left(x_k\right)$$

$$=f\left(\boldsymbol{x}_k\right)+\frac{L}{2}\left(\left(\boldsymbol{x}-\boldsymbol{x}_k\right)+\frac{1}{L}\nabla f\left(\boldsymbol{x}_k\right)\right)^{\mathrm{T}}\left(\left(\boldsymbol{x}-\boldsymbol{x}_k\right)+\frac{1}{L}\nabla f\left(\boldsymbol{x}_k\right)\right)$$

$$-\frac{1}{2L}\nabla f\left(\boldsymbol{x}_k\right)^{\mathrm{T}}\nabla f\left(\boldsymbol{x}_k\right)$$

$$=\frac{L}{2}\left\|\boldsymbol{x}-\left(\boldsymbol{x}_k-\frac{1}{L}\nabla f\left(\boldsymbol{x}_k\right)\right)\right\|_2^2+\text{const},$$

其中$\text{const}=f(\boldsymbol{x}_k)-\frac{1}{2L}\nabla f\left(\boldsymbol{x}_k\right)^{\mathrm{T}}\nabla f\left(\boldsymbol{x}_k\right)$.

式**(11.11)**

$$\boldsymbol{x}_{k+1}=\boldsymbol{x}_k-\frac{1}{L}\nabla f\left(\boldsymbol{x}_k\right)$$

解析

此式很容易理解. 因为L_2范数的最小值为0，因此当$\boldsymbol{x}_{k+1}=\boldsymbol{x}_k-\frac{1}{L}\nabla f\left(\boldsymbol{x}_k\right)$时， 有$\hat{f}(\boldsymbol{x}_{k+1}) \leqslant \hat{f}(\boldsymbol{x}_k)$恒成立，同理有$\hat{f}(\boldsymbol{x}_{k+2}) \leqslant \hat{f}(\boldsymbol{x}_{k+1})$等.反复迭代能够使$\hat{f}(\boldsymbol{x})$的值不断下降.

式**(11.12）**

$$\boldsymbol{x}_{k+1}=\arg\min_{\boldsymbol{x}}\frac{L}{2}\left\|\boldsymbol{x}-\left(\boldsymbol{x}_k-\frac{1}{L}\nabla f\left(\boldsymbol{x}_k\right)\right)\right\|_2^2+\lambda\|\boldsymbol{x}\|_1$$

解析

式(11.11)的优化目标为$\hat{f}(\boldsymbol{x})$，而式(11.8)优化的函数为$f(\boldsymbol{x})+\lambda\|\boldsymbol{x}\|_1$. 由泰勒展开式，有$f(\boldsymbol{x})+\lambda\|\boldsymbol{x}\|_1 \simeq \hat{f}(\boldsymbol{x})+\lambda\|\boldsymbol{x}\|_1$. $\boldsymbol{x}$的更新由式(11.12)决定.

参见式(11.10)

式**(11.13）**

$$\boldsymbol{x}_{k+1}=\arg\min_{\boldsymbol{x}}\frac{L}{2}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2+\lambda\|\boldsymbol{x}\|_1$$

解析

这里将式(11.12)的优化步骤拆分成了两步. 首先设$\boldsymbol{z}=\boldsymbol{x}_k-\frac{1}{L}\nabla f\left(\boldsymbol{x}_k\right)$并计算$\boldsymbol{z}$，然后再求解式(11.13)，得到的结果是一致的.

式**(11.14)**

$$x_{k+1}^i = \begin{cases} z^i - \lambda/L, & \lambda/L < z^i \\ 0, & |z^i| \leqslant \lambda/L \\ z^i + \lambda/L, & z^i < -\lambda/L \end{cases}$$

解析

设优化函数

$$g(\boldsymbol{x}) = \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{z}\|_2^2 + \lambda\|\boldsymbol{x}\|_1$$

$$= \frac{L}{2}\sum_{i=1}^{d}\left\|x^i - z^i\right\|_2^2 + \lambda\sum_{i=1}^{d}\left\|x^i\right\|_1$$

$$= \sum_{i=1}^{d}\left(\frac{L}{2}\left(x^i - z^i\right)^2 + \lambda\left|x^i\right|\right),$$

这个式子表明优化$g(\boldsymbol{x})$可以被拆解成优化$\boldsymbol{x}$的各个分量的形式，对分量x^i，其优化函数

$$g\left(x^i\right) = \frac{L}{2}\left(x^i - z^i\right)^2 + \lambda\left|x^i\right|,$$

求导得

$$\frac{\mathrm{d}g\left(x^i\right)}{\mathrm{d}x^i} = L\left(x^i - z^i\right) + \lambda\mathrm{sgn}\left(x^i\right),$$

其中

$$\mathrm{sgn}\left(x^i\right) = \begin{cases} 1, & x^i > 0; \\ -1, & x^i < 0. \end{cases}$$

对于$x_i = 0$的特殊情况，由于$|x_i|$在$x_i = 0$点处不光滑，所以其不可

导，需单独讨论. 令$\frac{\mathrm{d}g\left(x^i\right)}{\mathrm{d}x^i}=0$有

$$x^i=z^i-\frac{\lambda}{L}\operatorname{sgn}\left(x^i\right).$$

此式的解即优化目标$g(x^i)$的极值点，因为等式两端均含有未知变量x^i，故分情况讨论.

(1) 当$z^i>\frac{\lambda}{L}$时：

a. 假设$x^i<0$，则$\operatorname{sgn}(x^i)=-1$，那么有$x^i=z^i+\frac{\lambda}{L}>0$与假设矛盾.

b. 假设$x^i>0$，则$\operatorname{sgn}(x^i)=1$，那么有$x^i=z^i-\frac{\lambda}{L}>0$和假设相符合，下面来检验$x^i=z^i-\frac{\lambda}{L}$是否是函数$g(x^i)$的最小值点. 当$x^i>0$时有$\frac{\mathrm{d}g\left(x^i\right)}{\mathrm{d}x^i}=L\left(x^i-z^i\right)+\lambda$在定义域内连续可导，则$g(x^i)$的二阶导数

$$\frac{\mathrm{d}^2g\left(x^i\right)}{\mathrm{d}x^{i2}}=L.$$

由于利普希茨常数L恒大于0，因此$x^i=z^i-\frac{\lambda}{L}$是函数$g(x^i)$的最小值.

(2) 当$z_i<-\frac{\lambda}{L}$时：

a. 假设$x^i>0$，则$\operatorname{sgn}(x^i)=1$，那么有$x^i=z^i-\frac{\lambda}{L}<0$与假设矛盾.

b. 假设$x^i<0$，则$\operatorname{sgn}(x^i)=-1$，那么有$x^i=z^i+\frac{\lambda}{L}<0$与假设相符，

由上述二阶导数恒大于0可知，$x^i = z^i + \frac{\lambda}{L}$是$g(x^i)$的最小值.

(3) 当$-\frac{\lambda}{L} \leqslant z_i \leqslant \frac{\lambda}{L}$时：

a. 假设$x^i>0$，则$\mathrm{sgn}(x^i) = 1$，那么有$x^i = z^i - \frac{\lambda}{L} \leqslant 0$与假设矛盾.

b. 假设$x^i < 0$，则$\mathrm{sgn}(x^i) = -1$，那么有$x^i = z^i + \frac{\lambda}{L} \geqslant 0$与假设矛盾.

(4) 当$x_i = 0$时有$g(x^i) = \frac{L}{2}\left(z^i\right)^2$.

a. 当$|z^i|>\frac{\lambda}{L}$时，由上述推导可知$g(x_i)$的最小值在$x^i = z^i - \frac{\lambda}{L}$处取得.因为

$$\begin{aligned} g\left(x^i\right)\big|_{x^i=0} - g\left(x^i\right)\big|_{x^i=z^i-\frac{\lambda}{L}} &= \frac{L}{2}\left(z^i\right)^2 - \left(\lambda z^i - \frac{\lambda^2}{2L}\right) \\ &= \frac{L}{2}\left(z^i - \frac{\lambda}{L}\right)^2 \\ &>0 \end{aligned}$$

因此当$|z^i|>\frac{\lambda}{L}$时，$x^i = 0$不会是函数$g(x^i)$的最小值.

b. 当$-\frac{\lambda}{L} \leqslant z^i \leqslant \frac{\lambda}{L}$时，对于任何$\Delta x \neq 0$有

$$g(\Delta x) = \frac{L}{2}\left(\Delta x - z^i\right)^2 + \lambda|\Delta x|$$

$$= \frac{L}{2}\left((\Delta x)^2 - 2\Delta x \cdot z^i + \frac{2\lambda}{L}|\Delta x|\right) + \frac{L}{2}\left(z^i\right)^2$$

$$\geqslant \frac{L}{2}\left((\Delta x)^2 - 2\Delta x \cdot z^i + \frac{2\lambda}{L}\Delta x\right) + \frac{L}{2}\left(z^i\right)^2$$

$$\geqslant \frac{L}{2}(\Delta x)^2 + \frac{L}{2}\left(z^i\right)^2$$

$$> g\left(x^i\right)\big|_{x^i=0},$$

因此$x^i = 0$是$g(x^i)$的最小值点.

综上所述，式(11.14)成立.

式**(11.15)**

$$\min_{\boldsymbol{B},\boldsymbol{\alpha}_i} \sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{B}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{i=1}^{m} \|\boldsymbol{\alpha}_i\|_1$$

解析

此式即希望样本$\boldsymbol{x}_i$的稀疏表示$\boldsymbol{\alpha}_i$通过字典$\boldsymbol{B}$重构后和样本$\boldsymbol{x}_i$的原始表示尽量相似，如果满足这个条件，那么稀疏表示$\boldsymbol{\alpha}_i$是比较好的.后面的1范数项是为了使$\boldsymbol{\alpha}_i$更加稀疏.

式**(11.16)**

$$\min_{\boldsymbol{\alpha}_i} \|\boldsymbol{x}_i - \boldsymbol{B}\boldsymbol{\alpha}_i\|_2^2 + \lambda \|\boldsymbol{\alpha}_i\|_1$$

解析

为了优化式(11.15)，我们采用类似EM算法的变量交替优化：首先固定变量$\boldsymbol{B}$，则式(11.15)求解的是m个样本相加的最小值.这是由于式里没有样本之间的交互，即“西瓜书”中所述的$\alpha_i^u \alpha_i^v (u \neq v)$形式，因此可以

对每个变量做分别的优化求出$\boldsymbol{\alpha}_i$，求解方法见式(11.13)和式(11.14).

式**(11.17)**

$$\min_{\boldsymbol{B}} \|\boldsymbol{X}-\boldsymbol{B}\boldsymbol{A}\|_F^2$$

解析

这是优化式(11.15)的第二步，固定住$\boldsymbol{\alpha}_i(i=1,2,\cdots,m)$，此时式(11.15)的第二项为一个常数，优化式(11.15)即优化$\min_{\boldsymbol{B}} \sum_{i=1}^{m} \|\boldsymbol{x}_i-\boldsymbol{B}\boldsymbol{\alpha}_i\|_2^2$,其写成矩阵相乘的形式为$\min_{\boldsymbol{B}} \|\boldsymbol{X}-\boldsymbol{B}\boldsymbol{A}\|_2^2$，将$\mathrm{L}_2$范数扩展到Frobenius范数即得优化目标为$\min_{\boldsymbol{B}} \|\boldsymbol{X}-\boldsymbol{B}\boldsymbol{A}\|_F^2$.

式**(11.18)**

$$\min_{\boldsymbol{B}} \|\boldsymbol{X}-\boldsymbol{B}\boldsymbol{A}\|_F^2 = \min_{\boldsymbol{b}_i} \left\|\boldsymbol{X}-\sum_{j=1}^{k}\boldsymbol{b}_j\boldsymbol{\alpha}^j\right\|_F^2$$

$$= \min_{\boldsymbol{b}_i} \left\|\left(\boldsymbol{X}-\sum_{j\neq i}\boldsymbol{b}_j\boldsymbol{\alpha}^j\right)-\boldsymbol{b}_i\boldsymbol{\alpha}^i\right\|_F^2$$

$$= \min_{\boldsymbol{b}_i} \left\|\boldsymbol{E}_i-\boldsymbol{b}_i\boldsymbol{\alpha}^i\right\|_F^2$$

解析

这个式难点在于推导$\boldsymbol{B}\boldsymbol{A}=\sum_{j=1}^{k}\boldsymbol{b}_j\boldsymbol{\alpha}^j$.大致的思路是$\boldsymbol{b}_j\boldsymbol{\alpha}^j$会生成和矩阵$\boldsymbol{B}\boldsymbol{A}$同样维度的矩阵，这个矩阵对应位置的元素是$\boldsymbol{B}\boldsymbol{A}$中对应位置元素

的一个分量，这样的分量矩阵一共有k个，把所有分量矩阵加起来就得到了最终结果.推导过程如下：

$$
B\boldsymbol{A} = \begin{bmatrix} b_1^1 & b_2^1 & \cdots & b_k^1 \\ b_1^2 & b_2^2 & \cdots & b_k^2 \\ \vdots & \vdots & & \vdots \\ b_1^d & b_2^d & \cdots & b_k^d \end{bmatrix}_{d\times k} \begin{bmatrix} \alpha_1^1 & \alpha_2^1 & \cdots & \alpha_m^1 \\ \alpha_1^2 & \alpha_2^2 & \cdots & \alpha_m^2 \\ \vdots & \vdots & & \vdots \\ \alpha_1^k & \alpha_2^k & \cdots & \alpha_m^k \end{bmatrix}_{k\times m}
$$

$$
= \begin{bmatrix} \sum_{j=1}^k b_j^1\alpha_1^j & \sum_{j=1}^k b_j^1\alpha_2^j & \cdots & \sum_{j=1}^k b_j^1\alpha_m^j \\ \sum_{j=1}^k b_j^2\alpha_1^j & \sum_{j=1}^k b_j^2\alpha_2^j & \cdots & \sum_{j=1}^k b_j^2\alpha_m^j \\ \vdots & \vdots & & \vdots \\ \sum_{j=1}^k b_j^d\alpha_1^j & \sum_{j=1}^k b_j^d\alpha_2^j & \cdots & \sum_{j=1}^k b_j^d\alpha_m^j \end{bmatrix}_{d\times m},
$$

$$
\boldsymbol{b_j\alpha^j} = \begin{bmatrix} b_j^1 \\ b_j^2 \\ \vdots \\ b_j^d \end{bmatrix} \begin{bmatrix} \alpha_1^j & \alpha_2^j & \cdots & \alpha_m^j \end{bmatrix}
$$

$$
= \begin{bmatrix} b_j^1\alpha_1^j & b_j^1\alpha_2^j & \cdots & b_j^1\alpha_m^j \\ b_j^2\alpha_1^j & b_j^2\alpha_2^j & \cdots & b_j^2\alpha_m^j \\ \vdots & \vdots & & \vdots \\ b_j^d\alpha_1^j & b_j^d\alpha_2^j & \cdots & b_j^d\alpha_m^j \end{bmatrix}_{d\times m},
$$

求和可得：

$$
\sum_{j=1}^k \boldsymbol{b_j\alpha^j} = \sum_{j=1}^k \begin{bmatrix} b_1^j \\ b_w^j \\ \vdots \\ b_d^j \end{bmatrix} \begin{bmatrix} \alpha_1^j & \alpha_2^j & \cdots & \alpha_m^j \end{bmatrix}
$$

$$
= \begin{bmatrix} \sum_{j=1}^k b_j^1\alpha_1^j & \sum_{j=1}^k b_j^1\alpha_2^j & \cdots & \sum_{j=1}^k b_j^1\alpha_m^j \\ \sum_{j=1}^k b_j^2\alpha_1^j & \sum_{j=1}^k b_j^2\alpha_2^j & \cdots & \sum_{j=1}^k b_j^2\alpha_m^j \\ \vdots & \vdots & & \vdots \\ \sum_{j=1}^k b_j^d\alpha_1^j & \sum_{j=1}^k b_j^d\alpha_2^j & \cdots & \sum_{j=1}^k b_j^d\alpha_m^j \end{bmatrix}_{d\times m}.
$$

将矩阵$\boldsymbol{B}$分解成矩阵列$\boldsymbol{b}_j(j=1,2,\cdots,k)$带来一个好处，即和式(11.16)的原理相同，矩阵列与列之间无关，因此可以分别优化各个列，即将$\min\limits_{\boldsymbol{B}}\|\cdots\boldsymbol{B}\cdots\|_F^2$转化成了$\min\limits_{\boldsymbol{b}_i}\|\cdots\boldsymbol{b}_i\cdots\|_F^2$，得到第三行的等式之后，再利用文中介绍的KSVD算法求解即可.

第12章　计算学习理论

式(12.1)

$$E(h;\mathcal{D})=P_{\boldsymbol{x}\sim\mathcal{D}}\left(h\left(\boldsymbol{x}\right)\neq y\right)$$

解析

此式为泛化误差的定义式. 所谓泛化误差，是指当样本$\boldsymbol{x}$从真实的样本分布$\mathcal{D}$中采样后其预测值$h(\boldsymbol{x})$不等于真实值y的概率.在现实世界中，我们很难获得样本分布$\mathcal{D}$，而实际获得的数据集可以看作从样本分布$\mathcal{D}$中独立同分布采样得到的.在“西瓜书” 中，我们称实际获得的数据集为样例集D(也叫观测集、样本集，注意与花体$\mathcal{D}$的区别).

式(12.2)

$$\widehat{E}(h;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(h\left(\boldsymbol{x}_i\right)\neq y_i\right)$$

解析

此式为经验误差的定义式. 所谓经验误差，是指观测集D中的样本$\boldsymbol{x}_i(i=1,2,\cdots,m)$的预测值$h(\boldsymbol{x}_i)$和真实值$y_i$的期望误差.

式(12.3)

$$d\left(h_1,h_2\right)=P_{\boldsymbol{x}\sim\mathcal{D}}\left(h_1(\boldsymbol{x})\neq h_2(\boldsymbol{x})\right)$$

解析

假设我们有两个模型h_1和h_2，将它们同时作用于样本$\boldsymbol{x}$上，那么他们的“不合”度定义为这两个模型预测值不相同的概率.

此为Jensen不等式

式**(12.4)**

$$f\left(\mathbb{E}(x)\right) \leqslant \mathbb{E}\left(f(x)\right)$$

解析

此式可以做很直观的理解. 比如在二维空间，凸函数可以想象成开口向上的抛物线. 假设抛物线上有两个点x_1, x_2，那么$f(\mathbb{E}(x))$表示两个点的均值的纵坐标，而$\mathbb{E}(f(x))$表示的是两个点纵坐标的均值.因为两个点的均值落在抛物线的凹处，所以均值的纵坐标会小一些.

式**(12.5)**

$$P\left(\frac{1}{m}\sum_{i=1}^{m}x_i - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left(x_i\right) \geqslant \epsilon\right) \leqslant \exp\left(-2m\epsilon^2\right)$$

此为Hoeffding 不等式

解析

对于独立随机变量$x_1, x_2, \cdots, x_m$来说，它们的观测值x_i的均值$\frac{1}{m}\sum_{i=1}^{m}x_i$总是和它们的期望$\mathbb{E}(x_i)$的均值$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left(x_i\right)$相近.此式从概率的角度

对这个结论进行了描述：事件“观测值的均值和期望的均值之间的差值不小于ϵ”发生的概率不大于$\exp\left(-2m\epsilon^2\right)$.可以看出，当观测到的变量越多，观测值的均值越逼近期望的均值.

式**(12.7)**

$$P\left(f\left(x_1,\cdots,x_m\right)-\mathbb{E}\left(f\left(x_1,\cdots,x_m\right)\right)\geqslant\epsilon\right)\leqslant\exp\left(\frac{-2\epsilon^2}{\sum_i c_i^2}\right)$$

此为McDiarmid不等式

解析

首先解释此式的前提条件

$$\sup_{x_1,\cdots,x_m,x_i'}\left|f\left(x_1,\cdots,x_m\right)-f\left(x_1,\cdots,x_{i-1},x_i',x_{i+1},\cdots,x_m\right)\right|\leqslant c_i.$$

此条件表示当函数f的某个输入由x_i变为x_i'后，其函数值变化的上确界sup仍不大于c_i.所谓上确界sup可以理解成变化的极限最大值，可能被取到也可能被无限逼近.McDiarmid不等式指出，当此条件被满足时，函数值$f(x_1,\cdots,x_m)$和其期望值$\mathbb{E}\left(f(x_1,\cdots,x_m)\right)$也相近.从概率的角度描述，即事件“函数值和其期望之间的差值不小于ϵ”发生的概率不大于$\exp\left(\frac{-2\epsilon^2}{\sum_i c_i^2}\right)$.可以看出，当每次变量改动带来函数值改动的上限越小，函数值和其期望越相近.

式**(12.9**）

$$P\left(E(h)\leqslant\epsilon\right)\geqslant1-\delta$$

解析

此式中$E(h)$表示算法$\mathfrak{L}$在用观测集D训练后输出的假设函数h的泛化误差PAC辨识的定义指出，如果h的泛化误差不大于ϵ的概率不小于$1-\delta$，那么称学习算法$\mathfrak{L}$能从假设空间$\mathcal{H}$中PAC辨识概念类$\mathcal{C}$.

泛化误差见式(12.1)

式**(12.10)**

$$P(h(\boldsymbol{x})=y)=1-P(h(\boldsymbol{x})\neq y) \qquad ①$$

$$=1-E(h) \qquad ②$$

$$<1-\epsilon \qquad ③$$

式(12.10)$\sim$(12.14)是为了回答“西瓜书”中的问题：到底需要多少样例才能学得目标概念c的有效近似？只要训练集D的规模能使学习算法$\mathfrak{L}$以概率$1-\delta$找到目标假设的ϵ近似即可.式(12.10)$\sim$(12.14)用数学式对这个回答进行抽象

解析

①是因为$h(\boldsymbol{x})=y$和$h(\boldsymbol{x})\neq y$是对立事件.①$\rightarrow$②是因为泛化误差的定义[见式(12.1)].由于我们假定了泛化误差$E(h)>\epsilon$，因此有$1-E(h)<1-\epsilon$，即②$\rightarrow$③.

式**(12.11)**

$$P\left(\left(h\left(\boldsymbol{x}_1\right)=y_1\right) \wedge \cdots \wedge\left(h\left(\boldsymbol{x}_m\right)=y_m\right)\right)=\left(1-P(h(\boldsymbol{x}) \neq y)\right)^m$$
$$<(1-\epsilon)^m$$

解析

先解释什么是h与D“表现一致”.“西瓜书”12.2节开头阐述了这样的概念：如果h能将D中所有样本按与真实标记一致的方式完全分开，我们则称问题对学习算法是一致的,即$\left(h\left(\boldsymbol{x}_1\right)=y_1\right) \wedge \cdots \wedge\left(h\left(\boldsymbol{x}_m\right)=y_m\right)$为真.因为每个事件是独立的，所以上式可以写成

$P\left(\left(h\left(\boldsymbol{x}_1\right)=y_1\right) \wedge \cdots \wedge\left(h\left(\boldsymbol{x}_m\right)=y_m\right)\right)=\prod_{i=1}^m P\left(h\left(\boldsymbol{x}_i\right)=y_i\right)$.根据对立事件的定义，有$\prod_{i=1}^m P\left(h\left(\boldsymbol{x}_i\right)=y_i\right)=\prod_{i=1}^m\left(1-P\left(h\left(\boldsymbol{x}_i\right) \neq y_i\right)\right)$，又根据式(12.10)，有

$$\prod_{i=1}^m\left(1-P\left(h\left(\boldsymbol{x}_i\right) \neq y_i\right)\right)<\prod_{i=1}^m(1-\epsilon)=(1-\epsilon)^m .$$

式**(12.12**）

$$P\left(h \in \mathcal{H}: E(h)>\epsilon \wedge \widehat{E}(h)=0\right)<|\mathcal{H}|(1-\epsilon)^m$$
$$<|\mathcal{H}| \mathrm{e}^{-m \epsilon}$$

解析

首先解释为什么“我们事先并不知道学习算法$\mathfrak{L}$会输出$\mathcal{H}$中的哪个假设”.这是因为一些学习算法对用一个观察集D的输出结果是非确 定的.感知机就是个典型的例子：训练样本的顺序也会影响感知机学习到的假设h参数的值.

泛化误差大于ϵ且经验误差为0的假设(即在训练集上表现完美的假设)出现的概率可以表示为$P\left(h \in \mathcal{H} : E(h)>\epsilon \wedge \widehat{E}(h)=0\right)$.根据式(12.11)，每一个这样的假设$h$都满足$P\left(E(h)>\epsilon \wedge \widehat{E}(h)=0\right)<(1-\epsilon)^m$.设这样的假设$h$的数量为$|\mathcal{H}|$.因为每个假设$h$满足$E(h)>\epsilon$且$\widehat{E}(h)=0$是互斥的，因此总的概率$P(h \in \mathcal{H} : E(h)>\epsilon \wedge \widehat{E}(h)=0)$就是这些互斥事件之和，即

$$P\left(h \in \mathcal{H} : E(h)>\epsilon \wedge \widehat{E}(h)=0\right)=\sum_{i}^{|\mathcal{H}|} P\left(E(h_i)>\epsilon \wedge \widehat{E}(h_i)=0\right)$$
$$<|\mathcal{H}|(1-\epsilon)^m.$$

参见式(12.11)

接下来要证明$|\mathcal{H}|(1-\epsilon)^m<|\mathcal{H}|\mathrm{e}^{-m\epsilon}$，即证明$(1-\epsilon)^m<\mathrm{e}^{-m\epsilon}$，其中$\epsilon \in (0,1]$，$m$是正整数.证明如下.

当$\epsilon=1$时，该式显然成立. 当$\epsilon \in (0,1)$时，因为左式和右式的值域均大于0，所以可以左右两边同时取对数；又因为对数函数是单调递增函数，所以即证明$m\ln(1-\epsilon)<-m\epsilon$，即证明$\ln(1-\epsilon)<-\epsilon$.这个式子很容易证明： 设$f(\epsilon)=\ln(1-\epsilon)+\epsilon$，其中$\epsilon \in (0,1)$，则有$f'(\epsilon)=1-\frac{1}{1-\epsilon}=0$，即$\epsilon=0$时$f(\epsilon)$取极大值0，因此有$\ln(1-\epsilon)<-\epsilon$，即$|\mathcal{H}|(1-\epsilon)^m<|\mathcal{H}|\mathrm{e}^{-m\epsilon}$成立.

式(12.13)

$$|\mathcal{H}|\mathrm{e}^{-m\epsilon} \leqslant \delta$$

解析

回到我们要回答的问题：到底需要多少样例才能学得目标概念c的有效近似？只要训练集D的规模能使学习算法$\mathfrak{L}$以概率$1-\delta$找到目标假设的ϵ近似即可.

根据式(12.12)，学习算法$\mathfrak{L}$生成的假设大于目标假设的ϵ近似的概率为$P\left(h \in \mathcal{H}: E(h)>\epsilon \wedge \widehat{E}(h)=0\right)<|\mathcal{H}| \mathrm{e}^{-m \epsilon}$，因此学习算法$\mathfrak{L}$生成的假设落在目标假设的$\epsilon$近似的概率为$1-P\left(h \in \mathcal{H}: E(h)>\epsilon \wedge \widehat{E}(h)=0\right) \geq 1-|\mathcal{H}| \mathrm{e}^{-m \epsilon}$，我们希望这个概率至少是$1-\delta$，因此有$1-\delta \leqslant 1-|\mathcal{H}| \mathrm{e}^{-m \epsilon} \Rightarrow|\mathcal{H}| \mathrm{e}^{-m \epsilon} \leqslant \delta$.

式**(12.14**）

$$m \geqslant \frac{1}{\epsilon}\left(\ln |\mathcal{H}|+\ln \frac{1}{\delta}\right)$$

解析

$$|\mathcal{H}| \mathrm{e}^{-m \epsilon} \leqslant \delta,$$

$$\mathrm{e}^{-m \epsilon} \leqslant \frac{\delta}{|\mathcal{H}|},$$

$$-m \epsilon \leqslant \ln \delta-\ln |\mathcal{H}|,$$

$$m \geqslant \frac{1}{\epsilon}\left(\ln |\mathcal{H}|+\ln \frac{1}{\delta}\right).$$

这个式子告诉我们，若假设空间$\mathcal{H}$是PAC可学习的，输出假设h的泛化误差ϵ随样本数目m增大而收敛到0，收敛速率为$O(\frac{1}{m})$.这也是我们在机器学习中的一个共识，即可供模型训练的观测集样本数量越多，机器

学习模型的泛化性能越好.

式**(12.15)**

$$P\left(\widehat{E}(h)-E(h)\geqslant\epsilon\right)\leqslant\exp\left(-2m\epsilon^2\right)$$

参见式(12.5)

式**(12.16）**

$$P\left(E(h)-\widehat{E}(h)\geqslant\epsilon\right)\leqslant\exp\left(-2m\epsilon^2\right)$$

参见式(12.5)

式**(12.17）**

$$P\left(|E(h)-\widehat{E}(h)|\geqslant\epsilon\right)\leqslant 2\exp\left(-2m\epsilon^2\right)$$

参见式(12.6)

式**(12.18）**

$$\widehat{E}(h)-\sqrt{\frac{\ln(2/\delta)}{2m}}\leqslant E(h)\leqslant\widehat{E}(h)+\sqrt{\frac{\ln(2/\delta)}{2m}}$$

解析

令$\delta=2\mathrm{e}^{-2m\epsilon^2}$，则$\epsilon=\sqrt{\frac{\ln(2/\delta)}{2m}}$，由式(12.17)有

$$P(|E(h)-\widehat{E}(h)|\geqslant\epsilon)\leqslant 2\exp\left(-2m\epsilon^2\right),$$

$$P(|E(h)-\widehat{E}(h)| \geqslant \epsilon) \leqslant \delta,$$

$$P(|E(h)-\widehat{E}(h)| \leqslant \epsilon) \geqslant 1-\delta,$$

$$P(-\epsilon \leqslant E(h)-\widehat{E}(h) \leqslant \epsilon) \geqslant 1-\delta,$$

$$P(\widehat{E}(h)-\epsilon \leqslant E(h) \leqslant \widehat{E}(h)+\epsilon) \geqslant 1-\delta,$$

将$\epsilon=\sqrt{\frac{\ln(2/\delta)}{2m}}$代入即此式得证.

此式进一步阐明了当观测集样本数量足够大的时候，h的经验误差是其泛化误差很好的近似.

式**(12.19)**

$$P\left(\left|E(h)-\widehat{E}(h)\right| \leqslant \sqrt{\frac{\ln|\mathcal{H}|+\ln(2/\delta)}{2m}}\right) \geqslant 1-\delta$$

令$h_1, h_2, \cdots, h_{|\mathcal{H}|}$表示假设空间$\mathcal{H}$中的假设，有

$$\begin{aligned}
&P(\exists h \in \mathcal{H}: |E(h)-\widehat{E}(h)|>\epsilon) \\
&= P\left(\left(\left|E_{h_1}-\widehat{E}_{h_1}\right|>\epsilon\right) \vee \cdots \vee \left(\left|E_{h_{|\mathcal{H}|}}-\widehat{E}_{h_{|\mathcal{H}|}}\right|>\epsilon\right)\right) \\
&\leqslant \sum_{h \in \mathcal{H}} P(|E(h)-\widehat{E}(h)|>\epsilon).
\end{aligned}$$

这一步很好理解：存在一个假设h使得$|E(h)-\widehat{E}(h)|>\epsilon$的概率可以表示为对假设空间内所有的假设$h_i(i \in 1, \cdots, |\mathcal{H}|)$使得$\left|E_{h_i}-\widehat{E}_{h_i}\right|>\epsilon$这个事件成立的“或”事件. 因为$P(A \vee B)=P(A)+P(B)-P(A \wedge B)$，而$P(A \wedge B) \geqslant 0$，所以最后一行的不等式成立.由式(12.17)有

$$P(|E(h)-\widehat{E}(h)| \geqslant \epsilon) \leqslant 2\exp\left(-2m\epsilon^2\right),$$

$$\sum_{h\in\mathcal{H}} P(|E(h)-\widehat{E}(h)|>\epsilon) \leqslant 2|\mathcal{H}|\exp\left(-2m\epsilon^2\right),$$

因此

$$\begin{aligned} P(\exists h\in\mathcal{H}:|E(h)-\widehat{E}(h)|>\epsilon) &\leqslant \sum_{h\in\mathcal{H}} P(|E(h)-\widehat{E}(h)|>\epsilon) \\ &\leqslant 2|\mathcal{H}|\exp\left(-2m\epsilon^2\right), \end{aligned}$$

其对立事件的概率

$$\begin{aligned} P(\forall h\in\mathcal{H}:|E(h)-\widehat{E}(h)| \leqslant \epsilon) &= 1-P(\exists h\in\mathcal{H}:|E(h)-\widehat{E}(h)|>\epsilon) \\ &\geqslant 1-2|\mathcal{H}|\exp\left(-2m\epsilon^2\right). \end{aligned}$$

令$\delta = 2|\mathcal{H}|\mathrm{e}^{-2m\epsilon^2}$，则$\epsilon = \sqrt{\dfrac{\ln|\mathcal{H}|+\ln(2/\delta)}{2m}}$，代入上式中即有

$$P\left(\forall h\in\mathcal{H}:\left|E(h)-\widehat{E}(h)\right| \leqslant \sqrt{\frac{\ln|\mathcal{H}|+\ln(2/\delta)}{2m}}\right) \geqslant 1-\delta,$$

其中前置条件$\forall h\in\mathcal{H}$可以省略.

式(12.20)

$$P\left(E(h)-\min_{h'\in\mathcal{H}} E\left(h'\right) \leqslant \epsilon\right) \geqslant 1-\delta$$

解析

此式是“不可知PAC可学习”的定义式.不可知是指当目标概念c不在

算法$\mathfrak{L}$所能生成的假设空间$\mathcal{H}$里.可学习是指如果$\mathcal{H}$中泛化误差最小的假设是$\arg\min_{h\in\mathcal{H}} E(h)$，且这个假设的泛化误差满足其与目标概念的泛化误差的差值不大于ϵ的概率不小于$1-\delta$.我们称这样的假设空间$\mathcal{H}$是不可知PAC可学习的.

式(12.21）

$$\Pi_{\mathcal{H}}(m)=\max_{\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_m\}\subseteq\mathcal{X}}\left|\left\{\left(h\left(\boldsymbol{x}_1\right),\cdots,h\left(\boldsymbol{x}_m\right)\right)|h\in\mathcal{H}\right\}\right|$$

解析

此式是增长函数的定义式.增长函数$\Pi_{\mathcal{H}}(m)$表示假设空间$\mathcal{H}$对m个样本所能赋予标签的最大可能的结果数.比如对于有2个样本的二分类问题，一共有4种可能的标签组合：$(0,0),(0,1),(1,0),(1,1)$.如果假设空间$\mathcal{H}_1$能赋予这2个样本2种标签组合$(0,0),(1,1)$，则$\Pi_{\mathcal{H}_1}(2)=2$.显然，$\mathcal{H}$对样本所能赋予标签的可能结果数越多，$\mathcal{H}$的表示能力就越强.增长函数可以用来反映假设空间$\mathcal{H}$的复杂度.

式(12.22）

$$P\left(\left|E(h)-\widehat{E}\left(h\right)\right|>\epsilon\right)\leqslant 4\Pi_{\mathcal{H}}(2m)\exp\left(-\frac{m\epsilon^2}{8}\right)$$

解析

截至2018年12月“西瓜书”第1版第30次印刷，式(12.22)的前提假设应当勘误为“对假设空间$\mathcal{H}$，$m\in\mathbb{N}$，$0<\epsilon<1$，存在$h\in\mathcal{H}$”.详细证明可参见“西瓜书”及本书侧栏的原论文.

原论文标题为“On the uniform convergence of relative frequencies of events to their probabilities”[1]

式**(12.23**）

$$\mathrm{VC}(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\}$$

解析

此式是VC维的定义式.VC维的定义是能被$\mathcal{H}$打散的最大示例集的大小.“西瓜书”中例12.1和例12.2 给出了形象的例子.注意，VC维的定义式上的底数2表示这个问题是二分类的问题.如果是n分类的问题，那么定义式中底数需要变为n.

式**(12.24)**

$$\Pi_{\mathcal{H}}(m) \leqslant \sum_{i=0}^{d} \binom{m}{i}$$

解析

首先解释“西瓜书”中数学归纳法的起始条件“当$m = 1, d = 0$或$d = 1$时，定理成立”.

当$m = 1, d = 0$时，由VC维的定义式可知$\Pi_{\mathcal{H}}(1) < 2$，否则d可以取到1. 又因为$\Pi_{\mathcal{H}}(m)$为整数，所以$\Pi_{\mathcal{H}}(1) \in [0, 1]$，式(12.24)右侧为$\sum_{i=0}^{0} \binom{1}{i} = 1$，因此不等式成立.

当$m=1,d=1$时，因为一个样本最多只能有两个类别，所以$\Pi_{\mathcal{H}}(1)=2$，不等式右侧为$\sum_{i=0}^{1}\binom{1}{i}=2$，因此不等式成立.

接下来解释归纳过程. 这里采用的归纳方法是假设式(12.24)对$(m-1,d-1)$和$(m-1,d)$成立，推导出其对(m,d)也成立.证明过程中引入观测集$D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m\}$和观测集$D'=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{m-1}\}$，其中$D$比$D'$多一个样本$x_m$，它们对应的假设空间可以表示为：

$$\mathcal{H}_{|D}=\{(h(\boldsymbol{x}_1),h(\boldsymbol{x}_2),\cdots,h(\boldsymbol{x}_m))\mid h\in\mathcal{H}\};$$
$$\mathcal{H}_{|D'}=\{(h(\boldsymbol{x}_1),h(\boldsymbol{x}_2),\cdots,h(\boldsymbol{x}_{m-1}))\mid h\in\mathcal{H}\}.$$

如果$h\in\mathcal{H}$对$\boldsymbol{x}_m$的分类结果为$+1$或-1，那么任何出现在$\mathcal{H}_{|D'}$中的串都会在$\mathcal{H}_{|D}$中出现一次或者两次.这里举个例子就很容易理解了，假设$m=3$，则有

$$\mathcal{H}_{|D}=\{(+,-,-),(+,+,-),(+,+,+),(-,+,-),(-,-,+)\};$$
$$\mathcal{H}_{|D'}=\{(+,+),(+,-),(-,+),(-,-)\},$$

其中串$(+,+)$在$\mathcal{H}_{|D}$中出现了两次：$(+,+,+),(+,+,-)$，而$\mathcal{H}_{|D'}$中的其他串$(+,-),(-,+),(-,-)$在$\mathcal{H}_{|D}$中均只出现了一次.这里的原因是每个样本是二分类的，所以多出的样本$\boldsymbol{x}_m$要么取$+$，要么取$_$，要么都取到（至少两个假设h对$\boldsymbol{x}_m$做出了不一致的判断）. 记号$\mathcal{H}_{D'|D}$表示在$\mathcal{H}_{|D}$中出现了两次的$\mathcal{H}_{|D'}$组成的集合，有

此时$\mathcal{H}_{D'|D}=\{(+,+)\}$

$$\left|\mathcal{H}_{|D}\right|=\left|\mathcal{H}_{|D'}\right|+\left|\mathcal{H}_{D'|D}\right|.$$

由于$\mathcal{H}_{|D'}$表示限制在样本集D'上的假设空间$\mathcal{H}$的表达能力（即所有假设对样本集D'所能赋予的标记种类数），样本集D'的大小为$m-1$，根据增长函数的定义，假设空间$\mathcal{H}$对包含$(m-1)$个样本的集合所能赋予的最大标记种类数为$\Pi_{\mathcal{H}}(m-1)$，因此$|\mathcal{H}_{|D'}| \leqslant \Pi_{\mathcal{H}}(m-1)$.又根据数学归纳法的前提假设，有

$$\left|\mathcal{H}_{|D'}\right| \leqslant \Pi_{\mathcal{H}}(m-1) \leqslant \sum_{i=0}^{d} \binom{m-1}{i}.$$

由记号$\mathcal{H}_{|D'}$的定义可知$|\mathcal{H}_{|D'}| \geqslant \left\lfloor \frac{|\mathcal{H}_{|D}|}{2} \right\rfloor$，又由于$|\mathcal{H}_{|D'}|$和$|\mathcal{H}_{D'|D}|$均为整数，因此$|\mathcal{H}_{D'|D}| \leqslant \left\lfloor \frac{|\mathcal{H}_{|D}|}{2} \right\rfloor$，由于样本集$D$的大小为$m$，根据增长函数的概念，有$\left|\mathcal{H}_{D'|D}\right| \leqslant \left\lfloor \frac{|\mathcal{H}_{|D}|}{2} \right\rfloor \leqslant \Pi_{\mathcal{H}}(m-1)$.假设$Q$表示能被$\mathcal{H}_{D'|D}$打散的集合.根据$\mathcal{H}_{D'|D}$的定义，$H_D$必对元素$\boldsymbol{x}_m$给定了不一致的判定，因此$Q \cup \{\boldsymbol{x}_m\}$必能被$\mathcal{H}_{|D}$打散，由前提假设$\mathcal{H}$的VC维为$d$，因此$\mathcal{H}_{D'|D}$的VC维最大为$d-1$，综上有

$$\left|\mathcal{H}_{D'|D}\right| \leqslant \Pi_{\mathcal{H}}(m-1) \leqslant \sum_{i=0}^{d-1} \binom{m-1}{i},$$

因此

$$\begin{aligned}
&\left|\mathcal{H}_{|D}\right| = \left|\mathcal{H}_{|D'}\right| + \left|\mathcal{H}_{D'|D}\right| \\
&\leqslant \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d+1} \binom{m-1}{i}
\end{aligned}$$

$$= \sum_{i=0}^{d} \left(\begin{pmatrix} m-1 \\ i \end{pmatrix} + \begin{pmatrix} m-1 \\ i-1 \end{pmatrix} \right)$$

$$= \sum_{i=0}^{d} \begin{pmatrix} m \\ i \end{pmatrix}.$$

其中，最后一步依据组合式. 具体推导如下：

$$\begin{pmatrix} m-1 \\ i \end{pmatrix} + \begin{pmatrix} m-1 \\ i-1 \end{pmatrix} = \frac{(m-1)!}{(m-1-i)!i!} + \frac{(m-1)!}{(m-1-i+1)!(i-1)!}$$

$$= \frac{(m-1)!(m-i)}{(m-i)(m-1-i)!i!} + \frac{(m-1)!i}{(m-i)!(i-1)!i}$$

$$= \frac{(m-1)!(m-i)+(m-1)!i}{(m-i)!i!}$$

$$= \frac{(m-1)!(m-i+i)}{(m-i)!i!}$$

$$= \frac{(m-1)!m}{(m-i)!i!}$$

$$= \frac{m!}{(m-i)!i!}$$

$$= \begin{pmatrix} m \\ i \end{pmatrix}.$$

式**(12.25)**

$$\left|\mathcal{H}_{|D}\right| = \left|\mathcal{H}_{|D'}\right| + \left|\mathcal{H}_{D'|D}\right|$$

参见式(12.24)

式**(12.26**）

$$\left|\mathcal{H}_{|D'}\right| \leqslant \Pi_{\mathcal{H}}(m-1) \leqslant \sum_{i=0}^{d} \binom{m-1}{i}$$

参见式(12.24)

式**(12.27**）

$$\left|\mathcal{H}_{D'|D}\right| \leqslant \Pi_{\mathcal{H}}(m-1) \leqslant \sum_{i=0}^{d-1} \binom{m-1}{i}$$

参见式(12.24)

式**(12.28**）

$$\Pi_{\mathcal{H}}(m) \leqslant \left(\frac{\mathrm{e} \cdot m}{d}\right)^{d}$$

解析

$$\Pi_{\mathcal{H}}(m) \leqslant \sum_{i=0}^{d} \binom{m}{i} \quad ①$$

$$\leqslant \sum_{i=0}^{d} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} \quad ②$$

$$= \left(\frac{m}{d}\right)^{d} \sum_{i=0}^{d} \binom{m}{i} \left(\frac{d}{m}\right)^{i} \quad ③$$

$$\leqslant \left(\frac{m}{d}\right)^{d} \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^{i} \quad ④$$

$$= \left(\frac{m}{d}\right)^d \left(1+\frac{d}{m}\right)^m \qquad ⑤$$

$$< \left(\frac{\mathrm{e}\cdot m}{d}\right)^d. \qquad ⑥$$

①→②和③→④均因为$m \geqslant d$;④→⑤是由于二项式定理

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k,$$

令$k=i, n=m, x=1, y=\frac{d}{m}$得

$$\left(\frac{m}{d}\right)^d \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^i = \left(\frac{m}{d}\right)^d \left(1+\frac{d}{m}\right)^m;$$

⑤→⑥的不等式即需证明$\left(1+\frac{d}{m}\right)^m \leqslant \mathrm{e}^d$，因为$\left(1+\frac{d}{m}\right)^m = \left(1+\frac{d}{m}\right)^{\frac{m}{d}d}$，根据e的定义，有$\left(1+\frac{d}{m}\right)^{\frac{m}{d}d} < \mathrm{e}^d$. 注意“西瓜书”中用的是$\leqslant$，但是由于e的定义是一个极限，所以可用$<$.

式**(12.29)**

$$P\left(\left|E(h)-\widehat{E}(h)\right| \leqslant \sqrt{\frac{8d\ln\frac{2em}{d}+8\ln\frac{4}{\delta}}{m}}\right) \geqslant 1-\delta$$

请注意某些印次的“西瓜书”中漏印了$|E(h)-\hat{E}(h)|$的绝对值符号

解析

将式(12.28)代入式(12.22)得

$$P\left(\left|E(h)-\widehat{E}(h)\right|>\epsilon\right)\leqslant 4\left(\frac{2em}{d}\right)^{d}\exp\left(-\frac{m\epsilon^{2}}{8}\right),$$

令$4\left(\frac{2em}{d}\right)^{d}\exp\left(-\frac{m\epsilon^{2}}{8}\right)=\delta$可解得

$$\delta=\sqrt{\frac{8d\ln\frac{2em}{d}+8\ln\frac{4}{\delta}}{m}},$$

代入式(12.22)，则定理得证.

此式用VC维表示泛化界.可以看出，泛化误差界只与样本数量m有关，收敛速率为$\sqrt{\frac{\ln m}{m}}$ (“西瓜书”上简化为$\frac{1}{\sqrt{m}}$).

式**(12.30**）

$$\widehat{E}(h)=\min_{h'\in\mathcal{H}}\widehat{E}\left(h'\right)$$

解析

此式是经验风险最小化的定义式，从假设空间中找出能使经验风险最小的假设.

式**(12.31**）

$$E(g)=\min_{h\in\mathcal{H}}E(h)$$

解析

首先回忆PAC可学习的概念，见定义12.2，而可知/不可知PAC可学习之间的区别仅仅在于概念类c是否包含于假设空间$\mathcal{H}$中.令

$$\frac{\delta}{2}=\delta'$$

$$\frac{\epsilon}{2}=\sqrt{\frac{(\ln 2/\delta')}{2m}}$$

结合这两个标记的转换，由推论12.1可知,
$\widehat{E}(g)-\frac{\epsilon}{2}\leqslant E(g)\leqslant\widehat{E}(g)+\frac{\epsilon}{2}$至少以$1-\delta/2$的概率成立.写成概率的形式,有

$$P\left(\left|E(g)-\widehat{E}(g)\right|\leqslant\frac{\epsilon}{2}\right)\geqslant 1-\delta/2,$$

即$P\left(\left(E(g)-\widehat{E}(g)\leqslant\frac{\epsilon}{2}\right)\wedge\left(E(g)-\widehat{E}(g)\geqslant-\frac{\epsilon}{2}\right)\right)\geqslant 1-\delta/2$，因此
$P\left(E(g)-\widehat{E}(g)\leqslant\frac{\epsilon}{2}\right)\geqslant 1-\delta/2$且$P\left(E(g)-\widehat{E}(g)\geqslant-\frac{\epsilon}{2}\right)\geqslant 1-\delta/2$成立. 再令

$$\frac{\epsilon}{2}=\sqrt{\frac{8d\ln\frac{2em}{d}+8\ln\frac{4}{\delta'}}{m}},$$

由式(12.29)可知

$$P\left(\left|E(h)-\widehat{E}(h)\right|\leqslant\frac{\epsilon}{2}\right)\geqslant 1-\frac{\delta}{2}.$$

同理，$P\left(E(h)-\widehat{E}(h)\leqslant\frac{\epsilon}{2}\right)\geqslant 1-\delta/2$且
$P\left(E(h)-\widehat{E}(h)\geqslant-\frac{\epsilon}{2}\right)\geqslant 1-\delta/2$成立.由$P\left(E(g)-\widehat{E}(g)\geqslant-\frac{\epsilon}{2}\right)\geqslant 1-\delta/2$和$P\left(E(h)-\widehat{E}(h)\leqslant\frac{\epsilon}{2}\right)\geqslant 1-\delta/2$均成立可知,事件$E(g)-\widehat{E}(g)\geqslant-\frac{\epsilon}{2}$和事

件$E(h)-\widehat{E}(h) \leqslant \frac{\epsilon}{2}$同时成立的概率

$$
\begin{aligned}
&P\left(\left(E(g)-\widehat{E}(g) \geqslant -\frac{\epsilon}{2}\right) \wedge \left(E(h)-\widehat{E}(h) \leqslant \frac{\epsilon}{2}\right)\right) \\
&= P\left(E(g)-\widehat{E}(g) \geqslant -\frac{\epsilon}{2}\right) + P\left(E(h)-\widehat{E}(h) \leqslant \frac{\epsilon}{2}\right) \\
&-P\left(\left(E(g)-\widehat{E}(g) \geqslant -\frac{\epsilon}{2}\right) \vee \left(E(h)-\widehat{E}(h) \leqslant \frac{\epsilon}{2}\right)\right) \\
&\geqslant 1-\delta/2+1-\delta/2-1 \\
&= 1-\delta,
\end{aligned}
$$

即

$$
P\left(\left(E(g)-\widehat{E}(g) \geqslant -\frac{\epsilon}{2}\right) \wedge \left(E(h)-\widehat{E}(h) \leqslant \frac{\epsilon}{2}\right)\right) \geqslant 1-\delta,
$$

因此

$$
\begin{aligned}
&P\left(\widehat{E}(g)-E(g)+E(h)-\widehat{E}(h) \leqslant \frac{\epsilon}{2}+\frac{\epsilon}{2}\right) \\
&= P(E(h)-E(g) \leqslant \widehat{E}(h)-\widehat{E}(g)+\epsilon) \geqslant 1-\delta.
\end{aligned}
$$

再由h和g的定义可知，h表示假设空间中经验误差最小的假设，g表示泛化误差最小的假设.将这两个假设共用作用于样本集D，则一定有$\widehat{E}(h) \leqslant \widehat{E}(g)$，因此上式可以简化为

$$
P\left(E(h)-E(g) \leqslant \epsilon\right) \geqslant 1-\delta.
$$

根据式(12.32)和式(12.34)，可以求出m为关于$(1/\epsilon, 1/\delta, \mathrm{size}(x), \mathrm{size}(c))$的多项式.根据定理12.2和定理12.5可得到结论：任何VC维有限的假设空间$\mathcal{H}$都是（不可知）PAC可学习的.

式**(12.32)**

$$\sqrt{\frac{(\ln 2/\delta')}{2m}} = \frac{\epsilon}{2}$$

参见式(12.31)

式**(12.34**）

$$\sqrt{\frac{8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta'}}{m}} = \frac{\epsilon}{2}$$

参见式(12.31)

式**(12.36**）

$$\widehat{E}(h) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(h\left(\boldsymbol{x}_i\right) \neq y_i\right) \quad ①$$

$$= \frac{1}{m}\sum_{i=1}^{m}\frac{1 - y_i h\left(\boldsymbol{x}_i\right)}{2} \quad ②$$

$$= \frac{1}{2} - \frac{1}{2m}\sum_{i=1}^{m} y_i h\left(\boldsymbol{x}_i\right) \quad ③$$

解析

这里解释①→②的推导.因为前提假设是二分类问题，即 $y_k \in \{-1, +1\}$，因此 $\mathbb{I}\left(h(x_i) \neq y_i\right) \equiv \frac{1 - y_i h(x_i)}{2}$.

具体来说，假如$y_i=+1,h(x_i)=+1$或$y_i=-1,h(x_i)=-1$，有$\mathbb{I}\left(h(x_i)\neq y_i\right)=0=\dfrac{1-y_ih(x_i)}{2}$；反之，假如$y_i=-1,h(x_i)=+1$或$y_i=+1,h(x_i)=-1$，有$\mathbb{I}\left(h(x_i)\neq y_i\right)=1=\dfrac{1-y_ih(x_i)}{2}$.

式**(12.37)**

$$\arg\max_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}y_ih\left(\boldsymbol{x}_i\right)$$

解析

由式(12.36)可知，经验误差$\widehat{E}(h)$和$\dfrac{1}{m}\sum_{i=1}^{m}y_ih\left(\boldsymbol{x}_i\right)$呈反比的关系，因此假设空间中能使经验误差最小的假设h即是使$\dfrac{1}{m}\sum_{i=1}^{m}y_ih\left(\boldsymbol{x}_i\right)$最大的$h$.

上确界sup这个概念的解释见式(12.7)的解析

式**(12.38）**

$$\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_ih\left(\boldsymbol{x}_i\right)$$

解析

由于σ_i是随机变量，因此此式可以理解为求解和随机生成的标签(即$\boldsymbol{\sigma}$)最契合的假设.

当σ_i和$h(\boldsymbol{x}_i)$完全一致时，它们的内积最大

式(12.39）

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h\left(\boldsymbol{x}_i\right)\right]$$

解析

此式可以用来衡量假设空间$\mathcal{H}$的表达能力，对变量$\boldsymbol{\sigma}$求期望可以理解为当变量$\boldsymbol{\sigma}$包含所有可能的结果时，假设空间$\mathcal{H}$中最契合的假设h和变量的平均契合程度.因为前提假设是二分类的问题，因此σ_i一共有2^m种，这些不同的σ_i构成了数据集$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$的“对分”.一个假设空间的表达能力越强，假设空间中就越有可能对于每一种σ_i都存在一个h使得$h(x_i)$和σ_i非常接近甚至相同，对所有可能的σ_i取期望即可衡量假设空间的整体表达能力，这就是这个式子的含义.

参见“西瓜书”12.4节

式(12.40）

$$\widehat{R}_Z(\mathcal{F})=\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f\left(\boldsymbol{z}_i\right)\right]$$

解析

对比式(12.39)，这里使用函数空间$\mathcal{F}$代替假设空间$\mathcal{H}$、函数f代替假设h.这很容易理解，因为假设h即可以看作作用在数据$\boldsymbol{x}_i$上的一个映射，通过这个映射可以得到标签y_i.注意前提假设实值函数空间$\mathcal{F}:\mathcal{Z}\to\mathbb{R}$，即函数$f$将样本$\boldsymbol{z}_i$映射到了实数空间，此时所有的$\sigma_i$将是一个标量，即

$\sigma_i \in \{+1, -1\}$.

式**(12.41)**

$$R_m(\mathcal{F}) = \mathbb{E}_{Z \subseteq \mathcal{Z}:|Z|=m} \left[\widehat{R}_Z(\mathcal{F})\right]$$

解析

这里所要求的是$\mathcal{F}$关于分布$\mathcal{D}$的Rademacher复杂度，因此从$\mathcal{D}$中采出不同的样本Z，计算这些样本对应的Rademacher复杂度的期望.

式**(12.42**）

$$\mathbb{E}[f(\boldsymbol{z})] \leqslant \frac{1}{m}\sum_{i=1}^{m} f\left(\boldsymbol{z}_i\right) + 2R_m(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

解析

首先令记号

$$\widehat{E}_Z(f) = \frac{1}{m}\sum_{i=1}^{m} f\left(\boldsymbol{z}_i\right),$$

$$\Phi(Z) = \sup_{f \in \mathcal{F}} \left(\mathbb{E}[f] - \widehat{E}_Z(f)\right),$$

即$\widehat{E}_Z(f)$表示函数f作为假设下的经验误差，$\Phi(Z)$表示经验误差和泛化误差的上确界.再令Z'为只与Z有一个示例(样本)不同的训练集， 不妨设$\boldsymbol{z}_m \in Z$和$\boldsymbol{z}'_m \in Z'$为不同的示例，那么有

$$\Phi\left(Z'\right) - \Phi(Z) = \sup_{f \in \mathcal{F}} \left(\mathbb{E}[f] - \widehat{E}_{Z'}(f)\right) - \sup_{f \in \mathcal{F}} \left(\mathbb{E}[f] - \widehat{E}_Z(f)\right) \quad ①$$

$$\leqslant \sup_{f\in\mathcal{F}}\left(\widehat{E}_Z(f)-\widehat{E}_{Z'}(f)\right) \qquad ②$$

$$=\sup_{f\in\mathcal{F}}\frac{\sum_{i=1}^{m}f\left(z_i\right)-\sum_{i=1}^{m}f\left(z_i'\right)}{m} \qquad ③$$

$$=\sup_{f\in\mathcal{F}}\frac{f\left(z_m\right)-f\left(z_m'\right)}{m} \qquad ④$$

$$\leqslant\frac{1}{m} \qquad ⑤$$

①→②是因为上确界的差不大于差的上确界，③→④是由于Z'与Z只有$\boldsymbol{z}_m$不相同，④→⑤是因为前提假设$\mathcal{F}:\mathcal{Z}\to[0,1]$，即$f(\boldsymbol{z}_m),f(\boldsymbol{z}_m')\in[0,1]$. 同理有

$$\Phi(Z)-\Phi\left(Z'\right)=\sup_{f\in\mathcal{F}}\frac{f\left(\boldsymbol{z}_m'\right)-f\left(\boldsymbol{z}_m\right)}{m}\leqslant\frac{1}{m}.$$

综上二式有

$$\left|\Phi(Z)-\Phi\left(Z'\right)\right|\leqslant\frac{1}{m}.$$

将Φ看作式(12.7)中的函数f则有

注意区别式(12.7)中的f和Φ定义里的f

$$P\left(\Phi(Z)-\mathbb{E}_Z[\Phi(Z)]\geqslant\epsilon\right)\leqslant\exp\left(\frac{-2\epsilon^2}{\sum_i c_i^2}\right).$$

令$\exp\left(\frac{-2\epsilon^2}{\sum_i c_i^2}\right)=\delta$可以求得$\epsilon=\sqrt{\frac{\ln(1/\delta)}{2m}}$，所以

$$P\left(\Phi(Z)-\mathbb{E}_Z[\Phi(Z)]\geqslant\sqrt{\frac{\ln(1/\delta)}{2m}}\right)\leqslant\delta.$$

由逆事件的概率定义得

$$P\left(\Phi(Z)-\mathbb{E}_Z[\Phi(Z)]\leqslant\sqrt{\frac{\ln(1/\delta)}{2m}}\right)\geqslant 1-\delta.$$

此即式(12.44)的结论

下面来估计$\mathbb{E}_Z[\Phi(Z)]$的上界, 有

$$\mathbb{E}_Z[\Phi(Z)]=\mathbb{E}_Z\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}[f]-\widehat{E}_Z(f)\right)\right] \quad ①$$

$$=\mathbb{E}_Z\left[\sup_{f\in\mathcal{F}}\mathbb{E}_{Z'}\left[\widehat{E}_{Z'}(f)-\widehat{E}_Z(f)\right]\right] \quad ②$$

$$\leqslant\mathbb{E}_{Z,Z'}\left[\sup_{f\in\mathcal{F}}\left(\widehat{E}_{Z'}(f)-\widehat{E}_Z(f)\right)\right] \quad ③$$

$$=\mathbb{E}_{Z,Z'}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\left(f\left(\boldsymbol{z}_i'\right)-f\left(\boldsymbol{z}_i\right)\right)\right] \quad ④$$

$$=\mathbb{E}_{\boldsymbol{\sigma},Z,\boldsymbol{Z}'}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left(f\left(\boldsymbol{z}_i'\right)-f\left(\boldsymbol{z}_i\right)\right)\right] \quad ⑤$$

$$\leqslant\mathbb{E}_{\boldsymbol{\sigma},Z'}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f\left(\boldsymbol{z}_i'\right)\right]+\mathbb{E}_{\boldsymbol{\sigma},Z}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}-\sigma_i f\left(\boldsymbol{z}_i\right)\right] \quad ⑥$$

$$= 2\mathbb{E}_{\boldsymbol{\sigma},Z}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f\left(\boldsymbol{z}_i\right)\right] \qquad ⑦$$

$$= 2R_m(\mathcal{F}). \qquad ⑧$$

Jesen不等式参见式(12.4)

①→②是在外面套了一个对服从分布$\mathcal{D}$的示例集Z'求期望.因为$\mathbb{E}_{Z'\sim\mathcal{D}}[\widehat{E}_{Z'}(f)]=\mathbb{E}(f)$，而采样出来的$Z'$和$Z$相互独立，因此有$\mathbb{E}_{Z'\sim\mathcal{D}}[\widehat{E}_{Z}(f)]=\widehat{E}_{Z}(f)$.

②→③的不等式基于上确界函数$\sup$是凸函数， 根据Jesen不等式，有

$$\mathbb{E}_{Z}\left[\sup_{f\in\mathcal{F}}\mathbb{E}_{Z'}\left[\widehat{E}_{Z'}(f)-\widehat{E}_{Z}(f)\right]\right]$$

$$\leqslant \mathbb{E}_{Z,Z'}\left[\sup_{f\in\mathcal{F}}\left(\widehat{E}_{Z'}(f)-\widehat{E}_{Z}(f)\right)\right],$$

其中$\mathbb{E}_{Z,Z'}[\cdot]$是$\mathbb{E}_{Z}[\mathbb{E}_{Z'}[\cdot]]$的简写形式.

④→⑤引入对Rademacher随机变量的期望，由于函数值空间是标量，而σ_i也是标量，即$\sigma_i\in\{-1,+1\}$，且σ_i总可以以相同概率取到这两个值，因此可以引入$\mathbb{E}_{\sigma}$而不影响最终结果.

⑤→⑥是因为上确界的和不小于和的上确界.同时，因为第一项中只含有变量$\boldsymbol{z}'$，所以可以去掉$\mathbb{E}_{Z}$； 因为第二项中只含有变量$\boldsymbol{z}$，所以可以去掉$\mathbb{E}_{Z'}$.

⑥→⑦利用了$\boldsymbol{\sigma}$的对称性（即$-\boldsymbol{\sigma}$的分布和$\boldsymbol{\sigma}$完全一致）去除第二项中的负号. 又因为Z和Z'均是从$\mathcal{D}$中独立同分布采样得到的数据，因此可以将第一项中的$\boldsymbol{z}_i'$替换成$\boldsymbol{z}$，将Z'替换成Z.最后根据式(12.41)可得$\mathbb{E}_Z[\Phi(Z)] = 2R_m(\mathcal{F})$，式(12.42)得证.

式**(12.43)**

$$\mathbb{E}\left[f(\boldsymbol{z})\right] \leqslant \frac{1}{m}\sum_{i=1}^{m} f\left(\boldsymbol{z}_i\right) + 2\widehat{R}_Z(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

参见式(12.42)

式**(12.44**）

$$\Phi(Z) \leqslant \mathbb{E}_Z\left[\Phi(Z)\right] + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

参见式(12.42)

式**(12.45**）

$$R_m(\mathcal{F}) \leqslant \widehat{R}_Z(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2m}}$$

参见式(12.42)

式**(12.46**）

$$\Phi(Z) \leqslant 2\widehat{R}_Z(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

参见式(12.42)

式**(12.52**）

$$R_m(\mathcal{H}) \leqslant \sqrt{\frac{2\ln \Pi_{\mathcal{H}}(m)}{m}}$$

本式的证明比较烦琐. 同“西瓜书”上所示，参见Mehryar Mohri等人的书籍*Foundations of Machine Learning*(即本章参考文献[2])

式**(12.53**）

$$E(h) \leqslant \widehat{E}(h) + \sqrt{\frac{2d\ln\frac{em}{d}}{m}} + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

解析

根据式(12.28)有$\Pi_{\mathcal{H}}(m) \leqslant \left(\frac{\mathrm{e}\cdot m}{d}\right)^d$. 根据式(12.52)，有$R_m(\mathcal{H}) \leqslant \sqrt{\frac{2\ln \Pi_{\mathcal{H}}(m)}{m}}$，因此$\Pi_{\mathcal{H}}(m) \leqslant \sqrt{\frac{2d\ln\frac{em}{d}}{m}}$，再根据式(12.47)即得证.

式**(12.57)**

$$\begin{aligned}
&\left|\ell\left(\mathfrak{L}_D, \boldsymbol{z}\right) - \ell\left(\mathfrak{L}_{D^i}, \boldsymbol{z}\right)\right| \\
&\leqslant \left|\ell\left(\mathfrak{L}_D, \boldsymbol{z}\right) - \ell\left(\mathfrak{L}_{D\backslash i}, \boldsymbol{z}\right)\right| + \left|\ell\left(\mathfrak{L}_{D^i, \boldsymbol{z}}\right) - \ell\left(\mathfrak{L}_{D\backslash i, \boldsymbol{z}}\right)\right| \qquad ① \\
&\leqslant 2\beta \qquad ②
\end{aligned}$$

解析

根据三角不等式$|a+b| \leq |a|+|b|$，令$a=\ell\left(\mathfrak{L}_D, \boldsymbol{z}\right)-\ell\left(\mathfrak{L}_{D^i}\right)$，$b=\ell\left(\mathfrak{L}_{D^i, \boldsymbol{z}}\right)-\ell\left(\mathfrak{L}_{D^{\backslash i}, \boldsymbol{z}}\right)$，代入即有①的不等关系，由于$D^{\backslash i}$表示移除$D$中第$i$个样本，$D^i$表示替换$D$中第$i$个样本，那么$a$和$b$的变动均为一个样本，根据式(12.57)，有$a \leqslant \beta, b \leqslant \beta$，因此$a+b \leqslant 2\beta$.

式**(12.58**）

$$\ell(\mathfrak{L}, \mathcal{D}) \leqslant \widehat{\ell}(\mathfrak{L}, D)+2\beta+(4m\beta+M)\sqrt{\frac{\ln(1/\delta)}{2m}}$$

证明过程比较烦琐，参见本章参考文献[2]

式**(12.59**）

$$\ell(\mathfrak{L}, \mathcal{D}) \leqslant \ell_{\mathrm{loo}}(\mathfrak{L}, D)+\beta+(4m\beta+M)\sqrt{\frac{\ln(1/\delta)}{2m}}$$

证明过程比较烦琐，参见本章参考文献[2]

式**(12.60**）

$$\ell(\mathfrak{L}, \mathcal{D}) \leqslant \widehat{\ell}(\mathfrak{L}, D)+\frac{2}{m}+(4+M)\sqrt{\frac{\ln(1/\delta)}{2m}}$$

将$\beta=\frac{1}{m}$代入式(12.58)即得证

定理**12.9**

若学习算法$\mathfrak{L}$是ERM且是稳定的，则假设空间$\mathcal{H}$可学习.

解析

首先明确几个概念，ERM表示算法$\mathfrak{L}$满足经验风险最小化(Empirical Risk Min imization).由于算法$\mathfrak{L}$满足经验误差最小化，则可令g表示假设空间中具有最小泛化损失的假设，即

$$\ell(g, \mathcal{D}) = \min_{h \in \mathcal{H}} \ell(h, \mathcal{D}).$$

再令

$$\epsilon' = \frac{\epsilon}{2},$$

$$\frac{\delta}{2} = 2\exp\left(-2m\left(\epsilon'\right)^2\right),$$

将$\epsilon' = \frac{\epsilon}{2}$代入$\frac{\delta}{2} = 2\exp\left(-2m\left(\epsilon'\right)^2\right)$可以解得$m = \frac{2}{\epsilon^2}\ln\frac{4}{\delta}$，由Hoeffding不等式，有

$$P\left(\left|\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left(x_i\right)\right| \geqslant \epsilon\right) \leqslant 2\exp\left(-2m\epsilon^2\right),$$

参见式(12.6)

其中$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left(x_i\right) = \ell(g, \mathcal{D})$，$\frac{1}{m}\sum_{i=1}^{m} x_i = \widehat{\ell}(g, \mathcal{D})$，代入可得

$$P\left(\left|\ell(g, \mathcal{D}) - \widehat{\ell}(g, D)\right| \geqslant \frac{\epsilon}{2}\right) \leqslant \frac{\delta}{2}.$$

根据逆事件的概率可得

$$P\left(\left|\ell(g, \mathcal{D}) - \widehat{\ell}(g, D)\right| \leqslant \frac{\epsilon}{2}\right) \geqslant 1 - \frac{\delta}{2},$$

即文中$|\ell(g,\mathcal{D})-\widehat{\ell}(g,D)|\leqslant\frac{\epsilon}{2}$至少以$1-\delta/2$的概率成立.

由$\frac{2}{m}+(4+M)\sqrt{\frac{\ln(2/\delta)}{2m}}=\frac{\epsilon}{2}$可以求解出

$$\sqrt{m}=\frac{(4+M)\sqrt{\frac{\ln(2/\delta)}{2}}+\sqrt{(4+M)^2\cdot\frac{\ln(2/\delta)}{2}-4\cdot\frac{\epsilon}{2}\cdot(-2)}}{2\cdot\frac{\epsilon}{2}},$$

即$m=O\left(\frac{1}{\epsilon^2}\ln\frac{1}{\delta}\right)$.

由$P\left(\left|\ell(g,\mathcal{D})-\widehat{\ell}(g,D)\right|\leqslant\frac{\epsilon}{2}\right)\geqslant 1-\frac{\delta}{2}$和式(12.31)中介绍的相同的方法，可以推导出

$$P(\ell(\mathfrak{L},\mathcal{D})-\ell(g,\mathcal{D})\leqslant\epsilon)\geqslant 1-\delta.$$

又因为m为与$1/\epsilon,1/\delta,\mathrm{size}(\boldsymbol{x}),\mathrm{size}(c)$相关的多项式的值，根据定理12.2、定理12.5，可得到结论：$\mathcal{H}$是（不可知）PAC可学习的.

参考文献

[1] VLADIMIR V N, CHERVONENKIS A Y. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities[M]//Measures of Complexity. Berlin : Sprin ger,2015:11-30.

[2] MOHRI M, ROSTAMIZADEH A, TALWALKAR A. Foundations of Machin e Learning[M]. Cambridge, M.A.: MIT Press, 2018.

第13章　半监督学习

式(13.1)

$$p(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i \cdot p\left(\boldsymbol{x} | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)$$

解析

此为高斯混合分布的定义式

式(13.2)

$$f(\boldsymbol{x}) = \arg\max_{j \in \mathcal{Y}} p(y = j \mid \boldsymbol{x}) \qquad ①$$

$$= \arg\max_{j \in \mathcal{Y}} \sum_{i=1}^{N} p(y = j, \Theta = i \mid \boldsymbol{x}) \qquad ②$$

$$= \arg\max_{j \in \mathcal{Y}} \sum_{i=1}^{N} p(y = j \mid \Theta = i, \boldsymbol{x}) \cdot p(\Theta = i \mid \boldsymbol{x}) \qquad ③$$

解析

①→②是对概率进行边缘化(margin alization)，引入Θ并对其求和$\sum_{i=1}^{N}$以抵消引入的影响; ②→③推导如下

$$p(y=j,\Theta=i\mid\boldsymbol{x})=\frac{p(y=j,\Theta=i,\boldsymbol{x})}{p(\boldsymbol{x})}$$

$$=\frac{p(y=j,\Theta=i,\boldsymbol{x})}{p(\Theta=i,\boldsymbol{x})}\cdot\frac{p(\Theta=i,\boldsymbol{x})}{p(\boldsymbol{x})}$$

$$=p(y=j\mid\Theta=i,\boldsymbol{x})\cdot p(\Theta=i\mid\boldsymbol{x}).$$

式(13.3)

$$p(\Theta=i|\boldsymbol{x})=\frac{\alpha_i\cdot p\left(\boldsymbol{x}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)}{\sum_{i=1}^{N}\alpha_i\cdot p\left(\boldsymbol{x}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)}$$

解析

根据式(13.1)，有

$$p(\Theta=i\mid\boldsymbol{x})=\frac{p(\Theta=i,\boldsymbol{x})}{P(\boldsymbol{x})}$$

$$=\frac{\alpha_i\cdot p\left(\boldsymbol{x}\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)}{\sum_{i=1}^{N}\alpha_i\cdot p\left(\boldsymbol{x}\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)}.$$

式(13.4)

$$LL\left(D_l\cup D_u\right)=\sum_{(x_j,y_j)\in D_l}\ln\left(\sum_{i=1}^{N}\alpha_i\cdot p\left(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)\cdot p\left(y_j\mid\Theta=i,\boldsymbol{x}_j\right)\right)+$$

$$\sum_{x_j\in D_u}\ln\left(\sum_{i=1}^{N}\alpha_i\cdot p\left(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)\right)$$

解析

首先解释第2项. 当不知道类别信息的时候，样本x_j的概率可以用式

(13.1) 表示. 所有无类别信息样本D_u的似然是所有样本的乘积.因为$\ln$函数是单调的，所以可以将$\ln$函数作用于这个乘积，以消除连乘产生的数值计算问题.

第1项引入了样本的标签信息，由

$$p(y=j \mid \Theta=i, \boldsymbol{x})=\begin{cases} 1, & i=j \\ 0, & i \neq j \end{cases}$$

可知，这项限定了样本x_j只可能来自于y_j所对应的高斯分布.

式(13.5）

$$\gamma_{ji}=\frac{\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\sum_{i=1}^{N} \alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}$$

解析

参见式(13.3)，这项可以理解成样本x_j属于类别标签i（或者说由第i个高斯分布生成）的后验概率.其中$\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$可以通过有标记样本预先计算出来,即

$$\alpha_i=\frac{l_i}{|D_l|}, \text{其中} |D_l|=\sum_{i=1}^{N} l_i;$$

$$\boldsymbol{\mu}_i=\frac{1}{l_i} \sum_{\left(\boldsymbol{x}_j, y_j\right) \in D_l \wedge y_j=i} \boldsymbol{x}_j$$

$$\boldsymbol{\Sigma}_i=\frac{1}{l_i} \sum_{\left(\boldsymbol{x}_j, y_j\right) \in D_l \wedge y_j=i}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}}.$$

式**(13.6)**

$$\boldsymbol{\mu}_i = \frac{1}{\sum\limits_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i} \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \boldsymbol{x}_j + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{x}_j \right)$$

解析

这项可以由 $\frac{\partial LL(D_l \cup D_u)}{\partial \mu_i} = 0$ 而得. 将式(13.4)的两项分别记为

$$LL\left(D_l\right) = \sum_{\left(\boldsymbol{x}_j, y_j \in D_l\right)} \ln \left(\sum_{s=1}^{N} \alpha_s \cdot p\left(\boldsymbol{x_j} \mid \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\right) \cdot p\left(y_i \mid \Theta = s, \boldsymbol{x_j}\right) \right)$$

$$= \sum_{\left(\boldsymbol{x}_j, y_j \in D_l\right)} \ln \left(\alpha_{y_j} \cdot p\left(\boldsymbol{x_j} \mid \boldsymbol{\mu}_{y_j}, \boldsymbol{\Sigma}_{y_j}\right) \right);$$

$$LL\left(D_u\right) = \sum_{\boldsymbol{x}_j \in D_u} \ln \left(\sum_{s=1}^{N} \alpha_s \cdot p\left(\boldsymbol{x_j} \mid \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\right) \right).$$

首先，求$LL(D_l)$对于$\boldsymbol{\mu_i}$的偏导.$LL(D_l)$求和号中只有满足$y_j = i$ 的项能留下来，即

$$\frac{\partial LL\left(D_l\right)}{\partial \boldsymbol{\mu}_i} = \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{\partial \ln\left(\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)\right)}{\partial \boldsymbol{\mu}_i}$$

$$= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{1}{p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)} \cdot \frac{\partial p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\partial \boldsymbol{\mu}_i}$$

$$= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{1}{p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)} \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)$$

$$= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) .$$

然后求$LL(D_u)$对于$\boldsymbol{\mu_i}$的偏导，参考式(9.33)的推导，有

$$\frac{\partial LL(D_u)}{\partial \boldsymbol{\mu}_i} = \sum_{\boldsymbol{x}_j \in D_u} \frac{\alpha_i}{\sum_{s=1}^{N} \alpha_s \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)} \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)$$

$$= \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) .$$

综上，

$$\frac{\partial LL(D_l \cup D_u)}{\partial \boldsymbol{\mu}_i}$$

$$= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)$$

$$= \boldsymbol{\Sigma}_i^{-1} \left(\sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot (\boldsymbol{x}_j - \boldsymbol{\mu}_i) \right)$$

$$= \boldsymbol{\Sigma}_i^{-1} \left(\sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{x}_j + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{x}_j - \right.$$

$$\left. \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{\mu}_i - \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{\mu}_i \right) .$$

令$\frac{\partial LL(D_l \cup D_u)}{\partial \boldsymbol{\mu}_i} = 0$，两边同时左乘$\boldsymbol{\Sigma}_i$并移项有

$$\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{\mu}_i + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{\mu}_i = \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{x}_j + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{x}_j .$$

上式中，$\boldsymbol{\mu_i}$作为常量可以移到求和号外，而$\sum\limits_{(x_j,y_j)\in D_l \wedge y_j=i} 1 = l_i$，即第$i$类样本的有标记样本数目，因此有

$$\left(\sum_{x_j\in D_u}\gamma_{ji}+\sum_{(x_j,y_j)\in D_l\wedge y_j=i}1\right)\boldsymbol{\mu}_i=\sum_{x_j\in D_u}\gamma_{ji}\cdot\boldsymbol{x}_j+\sum_{(x_j,y_j)\in D_l\wedge y_j=i}\boldsymbol{x}_j,$$

即得式(13.6).

式**(13.7)**

$$\boldsymbol{\Sigma}_i=\frac{1}{\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}+l_i}\left(\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}\cdot\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}}+\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}}\right)$$

解析

首先求$LL(D_l)$关于$\boldsymbol{\Sigma_i}$的偏导，类似于式(13.6)，有

$$\frac{\partial LL\left(D_l\right)}{\partial\boldsymbol{\Sigma}_i}=\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\frac{\partial\ln\left(\alpha_i\cdot p\left(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)\right)}{\partial\boldsymbol{\Sigma}_i}$$

$$=\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\frac{1}{p\left(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)}\cdot\frac{\partial p\left(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)}{\partial\boldsymbol{\Sigma}_i}$$

$$=\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\frac{1}{p\left(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)}\cdot p\left(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)$$

.

$$\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}}-\boldsymbol{I}\right)\left(\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\right)$$

$$=\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}}-\boldsymbol{I}\right)\left(\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\right).$$

然后求$LL(D_u)$对于$\boldsymbol{\Sigma}_i$的偏导，有

参见式 (9.35)

$$\frac{\partial LL\left(D_u\right)}{\partial\boldsymbol{\Sigma}_i}=\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}\cdot\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}}-\boldsymbol{I}\right)\left(\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\right).$$

综合可得

$$\frac{\partial LL\left(D_l\cup D_u\right)}{\partial\boldsymbol{\Sigma}_i}=\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}\cdot\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}}-\boldsymbol{I}\right)\left(\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\right)+$$

$$\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}}-\boldsymbol{I}\right)\left(\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\right)$$

$$=\left(\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}\cdot\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}}-\boldsymbol{I}\right)+\right.$$

$$\left.\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}}-\boldsymbol{I}\right)\right)\left(\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\right).$$

令$\dfrac{\partial LL\left(D_l\cup D_u\right)}{\partial\boldsymbol{\Sigma}_i}=0$，两边同时右乘$2\boldsymbol{\Sigma}_i$并移项得

$$\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}\cdot\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}}$$

$$+\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}}$$

$$=\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}\cdot\boldsymbol{I}+\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\boldsymbol{I}$$

$$=\left(\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}+l_i\right)\boldsymbol{I},$$

两边同时左乘以$\boldsymbol{\Sigma}_i$, 有

$$\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}\cdot\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}}+\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\mathrm{T}}$$

$$=\left(\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}+l_i\right)\boldsymbol{\Sigma}_i,$$

即得式(13.7).

式**(13.8)**

$$\alpha_i=\frac{1}{m}\left(\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}+l_i\right)$$

解析

写出$LL(D_l\cup D_u)$的拉格朗日形式,有

参见式(9.36)

$$L\left(D_l \cup D_u, \lambda\right) = LL\left(D_l \cup D_u\right) + \lambda\left(\sum_{s=1}^{N} \alpha_s - 1\right)$$
$$= LL\left(D_l\right) + LL\left(D_u\right) + \lambda\left(\sum_{s=1}^{N} \alpha_s - 1\right).$$

求上式关于α_i的偏导. 对于$LL(D_u)$，求导结果与式(9.37)的推导过程相同，即

参见式(9.37)

$$\frac{\partial LL\left(D_u\right)}{\partial \alpha_i} = \sum_{\boldsymbol{x}_j \in D_u} \frac{1}{\sum_{s=1}^{N} \alpha_s \cdot p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\right)} \cdot p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right);$$

对于$LL(D_l)$，则有类似于式(13.6)和式(13.7)的推导过程，即

$$\frac{\partial LL\left(D_l\right)}{\partial \alpha_i} = \sum_{\left(x_j, y_j\right) \in D_l \wedge y_j = i} \frac{\partial \ln\left(\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)\right)}{\partial \alpha_i}$$

$$= \sum_{\left(\boldsymbol{x}_j, y_j\right) \in D_l \wedge y_j = i} \frac{1}{\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)} \cdot \frac{\partial\left(\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)\right)}{\partial \alpha_i}$$

$$= \sum_{\left(\boldsymbol{x}_j, y_j\right) \in D_l \wedge y_j = i} \frac{1}{\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)} \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)$$

$$= \sum_{\left(\boldsymbol{x}_j, y_j\right) \in D_l \wedge y_j = i} \frac{1}{\alpha_i} = \frac{1}{\alpha_i} \cdot \sum_{\left(\boldsymbol{x}_j, y_j\right) \in D_l \wedge y_j = i} 1 = \frac{l_i}{\alpha_i}.$$

上式推导过程中需重点注意，α_i是变量，$p(x_j|\mu_i, \boldsymbol{\Sigma}_i)$是常量.最后一行中的$\alpha_i$相对于求和变量为常量，因此作为公因子移到求和号外.l_i 为第i类样本的有标记样本数目. 综合两项结果，有

$$\frac{\partial L\left(D_l \cup D_u, \lambda\right)}{\partial \alpha_i}=\frac{l_i}{\alpha_i}+\sum_{\boldsymbol{x}_j \in D_u} \frac{p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\sum_{s=1}^N \alpha_s \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\right)}+\lambda;$$

令 $\frac{\partial LL(D_l \cup D_u)}{\partial \alpha_i}=0$ 并且两边同乘以α_i，得

$$\alpha_i \cdot \frac{l_i}{\alpha_i}+\sum_{\boldsymbol{x}_j \in D_u} \frac{\alpha_i \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\sum_{s=1}^N \alpha_s \cdot p\left(\boldsymbol{x}_j \mid \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\right)}+\lambda \cdot \alpha_i=0;$$

结合式(9.30)发现，求和号内即后验概率γ_{ji},即

$$l_i+\sum_{x_i \in D_u} \gamma_{ji}+\lambda \alpha_i=0;$$

对所有混合成分求和，得

$$\sum_{i=1}^N l_i+\sum_{i=1}^N \sum_{x_i \in D_u} \gamma_{ji}+\sum_{i=1}^N \lambda \alpha_i=0;$$

又有 $\sum_{i=1}^N \alpha_i=1$ ，因此 $\sum_{i=1}^N \lambda \alpha_i=\lambda \sum_{i=1}^N \alpha_i=\lambda$. 根据式(9.30)对γ_{ji}的定义可知

$$\sum_{i=1}^N \gamma_{ji}=\sum_{i=1}^N \frac{\alpha_i \cdot p(x_j \mid \mu_i, \boldsymbol{\Sigma}_i)}{\Sigma_{s=1}^N \alpha_s \cdot p(x_j \mid \mu_s, \boldsymbol{\Sigma}_s)}=\frac{\sum_{i=1}^N \alpha_i \cdot p(x_j \mid \mu_i, \boldsymbol{\Sigma}_i)}{\sum_{s=1}^N \alpha_s \cdot p(x_j \mid \mu_s, \boldsymbol{\Sigma}_s)}=1;$$

再结合加法满足交换律，有

$$\sum_{i=1}^{N}\sum_{x_i\in D_u}\gamma_{ji}=\sum_{x_i\in D_u}\sum_{i=1}^{N}\gamma_{ji}=\sum_{x_i\in D_u}1=u.$$

以上分析过程中，$\sum_{x_j\in D_u}$形式与$\sum_{j=1}^{u}$等价，其中u为未标记样本集的样本个数. 又有$\sum_{i=1}^{N}l_i=l$，其中l为有标记样本集的样本个数.将这些结果代入$\sum_{i=1}^{N}l_i+\sum_{i=1}^{N}\sum_{x_i\in D_u}\gamma_{ji}+\sum_{i=1}^{N}\lambda\alpha_i=0$，解出

$$l+u+\lambda=0,$$

又$l+u=m$，其中m为样本总个数，移项即得$\lambda=-m$. 最后代入整理解得

$$l_i+\sum_{x_j\in D_u}\gamma_{ji}-\lambda\alpha_i=0,$$

整理即得式(13.8).

式**(13.9)**

$$\min_{\boldsymbol{w},\boldsymbol{b},\boldsymbol{y},\boldsymbol{\xi}}\quad\frac{1}{2}\|\boldsymbol{w}\|_2^2+C_l\sum_{i=1}^{l}\xi_i+C_u\sum_{i=l+1}^{m}\xi_i$$

$$\text{s.t. } y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)\geqslant 1-\xi_i,\quad i=1,2,\cdots,l;$$

$$\hat{y}_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)\geqslant 1-\xi_i,\quad i=l+1,l+2,\cdots,m;$$

$$\xi_i\geqslant 0,\quad i=1,2,\cdots,m.$$

解析

此式与式(6.35)基本一致，只不过引入了无标记样本的松弛变量$\xi_i(i=l+1,\cdots,m)$对应的权重系数C_u和无标记样本的标记指派$\hat{y}_i$.

式**(13.12)**

$$E(f)=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\boldsymbol{W})_{ij}\left(f\left(\boldsymbol{x}_i\right)-f\left(\boldsymbol{x}_j\right)\right)^2$$

$$=\frac{1}{2}\left(\sum_{i=1}^{m}d_i f^2\left(\boldsymbol{x}_i\right)+\sum_{j=1}^{m}d_j f^2\left(\boldsymbol{x}_j\right)-2\sum_{i=1}^{m}\sum_{j=1}^{m}(\boldsymbol{W})_{ij}f\left(\boldsymbol{x}_i\right)f\left(\boldsymbol{x}_j\right)\right)$$

$$=\sum_{i=1}^{m}d_i f^2\left(\boldsymbol{x}_i\right)-\sum_{i=1}^{m}\sum_{j=1}^{m}(\boldsymbol{W})_{ij}f\left(\boldsymbol{x}_i\right)f\left(\boldsymbol{x}_j\right)$$

$$=f^{\mathrm{T}}(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{f}$$

解析

首先解释下这个能量函数的定义.原则上，我们希望能量函数$E(f)$越小越好，对于节点i和j，如果它们不相邻，则$(\boldsymbol{W})_{ij}=0$；如果它们相邻，则最小化能量函数要求$f(x_i)$和$f(x_j)$尽量相似，和逻辑相符.

下面进行式的推导. 首先由二项展开可得

$$E(f)=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\boldsymbol{W})_{ij}\left(f\left(\boldsymbol{x}_i\right)-f\left(\boldsymbol{x}_j\right)\right)^2$$

$$=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\boldsymbol{W})_{ij}\left(f^2\left(\boldsymbol{x}_i\right)-2f\left(\boldsymbol{x}_i\right)f\left(\boldsymbol{x}_j\right)+f^2\left(\boldsymbol{x}_j\right)\right)$$

$$=\frac{1}{2}\left(\sum_{i=1}^{m}\sum_{j=1}^{m}(\boldsymbol{W})_{ij}f^2\left(\boldsymbol{x}_i\right)+\sum_{i=1}^{m}\sum_{j=1}^{m}(\boldsymbol{W})_{ij}f^2\left(\boldsymbol{x}_j\right)-2\sum_{i=1}^{m}\sum_{j=1}^{m}(\boldsymbol{W})_{ij}f\left(\boldsymbol{x}_i\right)f\left(\boldsymbol{x}_j\right)\right)$$

.

由于$\boldsymbol{W}$是对称矩阵，替换变量可以得到

$$\sum_{i=1}^{m}\sum_{j=1}^{m}(\boldsymbol{W})_{ij}f^2(\boldsymbol{x}_j)=\sum_{j=1}^{m}\sum_{i=1}^{m}(\boldsymbol{W})_{ji}f^2(\boldsymbol{x}_i)$$

$$=\sum_{i=1}^{m}\sum_{j=1}^{m}(\boldsymbol{W})_{ij}f^2(\boldsymbol{x}_i)$$

$$=\sum_{i=1}^{m}\sum_{j=1}^{m}(\boldsymbol{W})_{ij}f^2(\boldsymbol{x}_j),$$

因此可化简$E(f)$，有

$$E(f)=\sum_{i=1}^{m}\sum_{j=1}^{m}(\boldsymbol{W})_{ij}f^2(\boldsymbol{x}_i)-\sum_{i=1}^{m}\sum_{j=1}^{m}(\boldsymbol{W})_{ij}f(\boldsymbol{x}_i)f(\boldsymbol{x}_j).$$

根据定义$d_i=\sum_{j=1}^{l+u}(\boldsymbol{W})_{ij}$，又$m=l+u$，则有

$$E(f)=\sum_{i=1}^{m}d_if^2(\boldsymbol{x}_i)-\sum_{i=1}^{m}\sum_{j=1}^{m}(\boldsymbol{W})_{ij}f(\boldsymbol{x}_i)f(\boldsymbol{x}_j)$$

$$=\boldsymbol{f}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{f}-\boldsymbol{f}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{f}$$

$$=\boldsymbol{f}^{\mathrm{T}}(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{f}.$$

式**(13.13)**

$$E(f)=\begin{pmatrix}\boldsymbol{f}_l^{\mathrm{T}} & \boldsymbol{f}_u^{\mathrm{T}}\end{pmatrix}\left(\begin{bmatrix}\boldsymbol{D}_{ll} & \boldsymbol{0}_{lu}\\ \boldsymbol{0}_{ul} & \boldsymbol{D}_{uu}\end{bmatrix}-\begin{bmatrix}\boldsymbol{W}_{ll} & \boldsymbol{W}_{lu}\\ \boldsymbol{W}_{ul} & \boldsymbol{W}_{uu}\end{bmatrix}\right)\begin{bmatrix}\boldsymbol{f}_l\\ \boldsymbol{f}_u\end{bmatrix}$$

$$=\boldsymbol{f}_l^{\mathrm{T}}(\boldsymbol{D}_{ll}-\boldsymbol{W}_{ll})\boldsymbol{f}_l-2\boldsymbol{f}_u^{\mathrm{T}}\boldsymbol{W}_{ul}\boldsymbol{f}_l+\boldsymbol{f}_u^{\mathrm{T}}(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu})\boldsymbol{f}_u$$

解析

这里第一项“西瓜书”中的符号有歧义，表示成$\left[\begin{array}{cc} \boldsymbol{f}_l^{\mathrm{T}} & \boldsymbol{f}_u^{\mathrm{T}} \end{array}\right]$, 即一个$\mathbb{R}^{1\times(l+u)}$的行向量更佳. 根据矩阵乘法的定义，有

$$
\begin{aligned}
E(f) &= \left[\begin{array}{cc} \boldsymbol{f}_l^{\mathrm{T}} & \boldsymbol{f}_u^{\mathrm{T}} \end{array}\right] \left[\begin{array}{cc} \boldsymbol{D}_{ll}-\boldsymbol{W}_{ll} & -\boldsymbol{W}_{lu} \\ -\boldsymbol{W}_{ul} & \boldsymbol{D}_{uu}-\boldsymbol{W}_{uu} \end{array}\right] \left[\begin{array}{c} \boldsymbol{f}_l \\ \boldsymbol{f}_u \end{array}\right] \\
&= \left[\boldsymbol{f}_l^{\mathrm{T}}\left(\boldsymbol{D}_{ll}-\boldsymbol{W}_{ll}\right)-\boldsymbol{f}_u^{\mathrm{T}}\boldsymbol{W}_{ul} \quad -\boldsymbol{f}_l^{\mathrm{T}}\boldsymbol{W}_{lu}+\boldsymbol{f}_u^{\mathrm{T}}\left(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu}\right)\right] \left[\begin{array}{c} \boldsymbol{f}_l \\ \boldsymbol{f}_u \end{array}\right] \\
&= \left(\boldsymbol{f}_l^{\mathrm{T}}\left(\boldsymbol{D}_{ll}-\boldsymbol{W}_{ll}\right)-\boldsymbol{f}_u^{\mathrm{T}}\boldsymbol{W}_{ul}\right)\boldsymbol{f}_l+\left(-\boldsymbol{f}_l^{\mathrm{T}}\boldsymbol{W}_{lu}+\boldsymbol{f}_u^{\mathrm{T}}\left(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu}\right)\right)\boldsymbol{f}_u \\
&= \boldsymbol{f}_l^{\mathrm{T}}\left(\boldsymbol{D}_{ll}-\boldsymbol{W}_{ll}\right)\boldsymbol{f}_l-\boldsymbol{f}_u^{\mathrm{T}}\boldsymbol{W}_{ul}\boldsymbol{f}_l-\boldsymbol{f}_l^{\mathrm{T}}\boldsymbol{W}_{lu}\boldsymbol{f}_u+\boldsymbol{f}_u^{\mathrm{T}}\left(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu}\right)\boldsymbol{f}_u \\
&= \boldsymbol{f}_l^{\mathrm{T}}\left(\boldsymbol{D}_{ll}-\boldsymbol{W}_{ll}\right)\boldsymbol{f}_l-2\boldsymbol{f}_u^{\mathrm{T}}\boldsymbol{W}_{ul}\boldsymbol{f}_l+\boldsymbol{f}_u^{\mathrm{T}}\left(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu}\right)\boldsymbol{f}_u,
\end{aligned}
$$

其中最后一步之所以有$\boldsymbol{f}_l^{\mathrm{T}}\boldsymbol{W}_{lu}\boldsymbol{f}_u=\left(\boldsymbol{f}_l^{\mathrm{T}}\boldsymbol{W}_{lu}\boldsymbol{f}_u\right)^{\mathrm{T}}=\boldsymbol{f}_u^{\mathrm{T}}\boldsymbol{W}_{ul}\boldsymbol{f}_l$，是因为$\boldsymbol{f}_i^{\mathrm{T}}\boldsymbol{W}_{lu}\boldsymbol{f}_u$的结果是一个标量.

式**(13.14)**

$$
\begin{aligned}
E(f) &= \left(\begin{array}{cc} \boldsymbol{f}_l^{\mathrm{T}} & \boldsymbol{f}_u^{\mathrm{T}} \end{array}\right)\left(\left[\begin{array}{cc} \boldsymbol{D}_{ll} & \boldsymbol{0}_{lu} \\ \boldsymbol{0}_{ul} & \boldsymbol{D}_{uu} \end{array}\right]-\left[\begin{array}{cc} \boldsymbol{W}_{ll} & \boldsymbol{W}_{lu} \\ \boldsymbol{W}_{ul} & \boldsymbol{W}_{uu} \end{array}\right]\right)\left[\begin{array}{c} \boldsymbol{f}_l \\ \boldsymbol{f}_u \end{array}\right] \\
&= \boldsymbol{f}_l^{\mathrm{T}}\left(\boldsymbol{D}_{ll}-\boldsymbol{W}_{ll}\right)\boldsymbol{f}_l-2\boldsymbol{f}_u^{\mathrm{T}}\boldsymbol{W}_{ul}\boldsymbol{f}_l+\boldsymbol{f}_u^{\mathrm{T}}\left(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu}\right)\boldsymbol{f}_u
\end{aligned}
$$

参见式(13.13)

式**(13.15**）

$$
\boldsymbol{f}_u=\left(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu}\right)^{-1}\boldsymbol{W}_{ul}\boldsymbol{f}_l
$$

解析

由式(13.13)，有

$$\frac{\partial E(f)}{\partial \boldsymbol{f}_u}=\frac{\partial \boldsymbol{f}_l^{\mathrm{T}}\left(\boldsymbol{D}_{ll}-\boldsymbol{W}_{ll}\right)\boldsymbol{f}_l-2\boldsymbol{f}_u^{\mathrm{T}}\boldsymbol{W}_{ul}\boldsymbol{f}_l+\boldsymbol{f}_u^{\mathrm{T}}\left(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu}\right)\boldsymbol{f}_u}{\partial \boldsymbol{f}_u}$$

$$=-2\boldsymbol{W}_{ul}\boldsymbol{f}_l+2\left(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu}\right)\boldsymbol{f}_u,$$

令结果等于 **0** 即得式(13.15).

式**(13.16**）

$$\boldsymbol{P}=\boldsymbol{D}^{-1}\boldsymbol{W}=\begin{bmatrix}\boldsymbol{D}_{ll}^{-1} & \boldsymbol{0}_{lu}\\ \boldsymbol{0}_{ul} & \boldsymbol{D}_{uu}^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{W}_{ll} & \boldsymbol{W}_{lu}\\ \boldsymbol{W}_{ul} & \boldsymbol{W}_{uu}\end{bmatrix}$$

$$=\begin{bmatrix}\boldsymbol{D}_{ll}^{-1}\boldsymbol{W}_{ll} & \boldsymbol{D}_{ll}^{-1}\boldsymbol{W}_{lu}\\ \boldsymbol{D}_{uu}^{-1}\boldsymbol{W}_{ul} & \boldsymbol{D}_{uu}^{-1}\boldsymbol{W}_{uu}\end{bmatrix}$$

解析

根据矩阵乘法的定义计算可得. 其中需要注意的是，求对角矩阵$\boldsymbol{D}$的逆等于将其各个对角元素取倒数.

式**(13.17)**

$$\boldsymbol{f}_u=\left(\boldsymbol{D}_{uu}\left(\boldsymbol{I}-\boldsymbol{D}_{uu}^{-1}\boldsymbol{W}_{uu}\right)\right)^{-1}\boldsymbol{W}_{ul}\boldsymbol{f}_l \qquad ①$$

$$=\left(\boldsymbol{I}-\boldsymbol{D}_{uu}^{-1}\boldsymbol{W}_{uu}\right)^{-1}\boldsymbol{D}_{uu}^{-1}\boldsymbol{W}_{ul}\boldsymbol{f}_l \qquad ②$$

$$=\left(\boldsymbol{I}-\boldsymbol{P}_{uu}\right)^{-1}\boldsymbol{P}_{ul}\boldsymbol{f}_l \qquad ③$$

解析

①→②是根据矩阵乘法逆的定义：$(\boldsymbol{AB})^{-1}=\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$. 此式中$\boldsymbol{P}_{uu}=\boldsymbol{D}_{uu}^{-1}\boldsymbol{W}_{uu}$ 和$\boldsymbol{P}_{ul}=\boldsymbol{D}_{uu}^{-1}\boldsymbol{W}_{ul}$均可以根据$\boldsymbol{W}_{ij}$计算得到，因此可以通过标记$\boldsymbol{f}_l$计算未标记数据的标签$\boldsymbol{f}_u$.

式**(13.20**）

$$\boldsymbol{F}^*=\lim_{t\to\infty}\boldsymbol{F}(t)=(1-\alpha)(\boldsymbol{I}-\alpha\boldsymbol{S})^{-1}\boldsymbol{Y}$$

解析

在式(13.19)

$$\boldsymbol{F}(t+1)=\alpha\boldsymbol{S}\boldsymbol{F}(t)+(1-\alpha)\boldsymbol{Y}$$

中，尝试将t取不同的值，有以下情况.

(1) $t=0$时: $\boldsymbol{F}(1)=\alpha\boldsymbol{S}\boldsymbol{F}(0)+(1-\alpha)\boldsymbol{Y}$

$$=\alpha\boldsymbol{S}\boldsymbol{Y}+(1-\alpha)\boldsymbol{Y};$$

(2) $t=1$时:

$$\boldsymbol{F}(2)=\alpha\boldsymbol{S}\boldsymbol{F}(1)+(1-\alpha)\boldsymbol{Y}=\alpha\boldsymbol{S}(\alpha\boldsymbol{S}\boldsymbol{Y}+(1-\alpha)\boldsymbol{Y})+(1-\alpha)\boldsymbol{Y}$$

$$=(\alpha\boldsymbol{S})^2\boldsymbol{Y}+(1-\alpha)\left(\sum_{i=0}^{1}(\alpha\boldsymbol{S})^i\right)\boldsymbol{Y};$$

(3) $t=2$时: $\boldsymbol{F}(3)=\alpha\boldsymbol{S}\boldsymbol{F}(2)+(1-\alpha)\boldsymbol{Y}$

$$=\alpha\boldsymbol{S}\left((\alpha\boldsymbol{S})^2\boldsymbol{Y}+(1-\alpha)\left(\sum_{i=0}^{1}(\alpha\boldsymbol{S})^i\right)\boldsymbol{Y}\right)+(1-\alpha)\boldsymbol{Y}$$

$$= (\alpha \boldsymbol{S})^3 \boldsymbol{Y} + (1-\alpha)\left(\sum_{i=0}^{2}(\alpha \boldsymbol{S})^i\right)\boldsymbol{Y}.$$

此处将i个相同矩阵连乘记为$(\cdot)^i$

可以观察到规律

$$\boldsymbol{F}(t) = (\alpha \boldsymbol{S})^t \boldsymbol{Y} + (1-\alpha)\left(\sum_{i=0}^{t-1}(\alpha \boldsymbol{S})^i\right)\boldsymbol{Y},$$

则

$$\boldsymbol{F}^* = \lim_{t\to\infty}\boldsymbol{F}(t) = \lim_{t\to\infty}(\alpha \boldsymbol{S})^t \boldsymbol{Y} + \lim_{t\to\infty}(1-\alpha)\left(\sum_{i=0}^{t-1}(\alpha \boldsymbol{S})^i\right)\boldsymbol{Y},$$

注意区分t个相同矩阵连乘$(\cdot)^t$和矩阵转置$(\cdot)^{\mathrm{T}}$

其中第一项是由于$\boldsymbol{S} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}$的特征值介于[-1,1]之间，而$\alpha \in (0,1)$，所以$\lim_{t\to\infty}(\alpha \boldsymbol{S})^t = 0$. 第二项由等比数列式可推出，即

$$\lim_{t\to\infty}\sum_{i=0}^{t-1}(\alpha \boldsymbol{S})^i = \frac{\boldsymbol{I} - \lim_{t\to\infty}(\alpha \boldsymbol{S})^t}{\boldsymbol{I} - \alpha \boldsymbol{S}} = \frac{\boldsymbol{I}}{\boldsymbol{I} - \alpha \boldsymbol{S}} = (\boldsymbol{I} - \alpha \boldsymbol{S})^{-1}.$$

综合可得式(13.20).

第14章　概率图模型

式(14.1)

$$P\left(x_1, y_1, \cdots, x_n, y_n\right)=P\left(y_1\right) P\left(x_1 | y_1\right) \prod_{i=2}^{n} P\left(y_i | y_{i-1}\right) P\left(x_i | y_i\right)$$

解析

所有的相乘关系都表示概率所对应事件相互独立的关系.三种概率$P\left(y_i\right), P\left(x_i | y_i\right), P\left(y_i | y_{i-1}\right)$分别表示初始状态概率、输出观测概率、条件转移概率.

式(14.2)

$$P(\mathbf{x})=\frac{1}{Z} \prod_{Q \in C} \psi_Q\left(\mathbf{x}_Q\right)$$

各个团之间概率分布相互独立，故可以用速乘表示最终的概率

式(14.3)

$$P(\mathbf{x})=\frac{1}{Z^*} \prod_{Q \in \mathcal{C}^*} \psi_Q\left(\mathbf{x}_Q\right)$$

意义同式(14.2), 区别在于此处的团为极大团

式(14.4)

$$P\left(x_A, x_B, x_C\right)=\frac{1}{Z} \psi_{AC}\left(x_A, x_C\right) \psi_{BC}\left(x_B, x_C\right)$$

将图14.3分解成x_A, x_C 和 x_B, x_C 两个团

式**(14.7)**

$$P\left(x_A, x_B | x_C\right)=P\left(x_A | x_C\right) P\left(x_B | x_C\right)$$

由式(14.5)、式(14.6)联立可得此式

式**(14.8)**

$$\psi_Q\left(\mathbf{x}_Q\right)=\mathrm{e}^{-H_Q\left(\mathbf{x}_Q\right)}$$

解析

此式为势函数的定义式，其中势函数写作指数函数的形式.指数函数具有非负性，且便于求导，因此在机器学习中具有广泛应用，例如式(8.5)和式(13.11).

式**(14.9)**

$$H_Q\left(\mathbf{x}_Q\right)=\sum_{u, v \in Q, u \neq v} \alpha_{uv} x_u x_v+\sum_{v \in Q} \beta_v x_v$$

解析

此式为定义在变量$\mathbf{x}_Q$上的函数$H_Q\left(\cdot\right)$的定义式. 其中第一项考虑每一对节点之间的关系，第二项考虑单节点.

式**(14.10）**

$$P\left(y_v|\mathbf{x},\mathbf{y}_{V\setminus\{v\}}\right)=P\left(y_v|\mathbf{x},\mathbf{y}_{n(v)}\right)$$

解析

根据局部马尔科夫性，给定某变量的邻接变量，则该变量独立于其他变量，即该变量只与其邻接变量有关. 所以此式中给定变量v以外的所有变量与仅给定变量v的邻接变量是等价的.

式(14.14）

$$P\left(x_5\right)=\sum_{x_4}\sum_{x_3}\sum_{x_2}\sum_{x_1}P\left(x_1,x_2,x_3,x_4,x_5\right)$$

$$=\sum_{x_4}\sum_{x_3}\sum_{x_2}\sum_{x_1}P\left(x_1\right)P\left(x_2\mid x_1\right)P\left(x_3\mid x_2\right)P\left(x_4\mid x_3\right)P\left(x_5\mid x_3\right)$$

解析

在消去变量的过程中，在消去每一个变量时都需要保证其依赖的变量已经消去，因此消去顺序应该是有向概率图中的一条以目标节点为终点的拓扑序列.

式(14.15）

$$P\left(x_5\right)=\sum_{x_3}P\left(x_5\mid x_3\right)\sum_{x_4}P\left(x_4\mid x_3\right)\sum_{x_2}P\left(x_3\mid x_2\right)\sum_{x_1}P\left(x_1\right)P\left(x_2\mid x_1\right)$$

$$=\sum_{x_3}P\left(x_5\mid x_3\right)\sum_{x_4}P\left(x_4\mid x_3\right)\sum_{x_2}P\left(x_3\mid x_2\right)m_{12}\left(x_2\right)$$

解析

变量按从右至左求和号下标的顺序消去. 应当注意x_4与x_5相互独立，

因此可与x_3互换消去顺序，对最终结果无影响.

式**(14.16)**

$$P(x_5) = \sum_{x_3} P(x_5 \mid x_3) \sum_{x_4} P(x_4 \mid x_3) m_{23}(x_3)$$

$$= \sum_{x_3} P(x_5 \mid x_3) m_{23}(x_3) \sum_{x_4} P(x_4 \mid x_3)$$

$$= \sum_{x_3} P(x_5 \mid x_3) m_{23}(x_3)$$

$$= m_{35}(x_5)$$

此处有$\sum_{x_4} P(x_4|x_3) = 1$

式**(14.17)**

$$P(x_1, x_2, x_3, x_4, x_5) = \frac{1}{Z}\psi_{12}(x_1, x_2)\psi_{23}(x_2, x_3)\psi_{34}(x_3, x_4)\psi_{35}(x_3, x_5)$$

解析

首先忽略图14.7(a)中的箭头，然后把无向图中的每条边的两个端点作为一个团，将其分解为四个团因子的乘积.Z为确保所有可能性的概率之和为1的规范化因子.

式**(14.18)**

$$P(x_5) = \frac{1}{Z}\sum_{x_3}\psi_{35}(x_3, x_5)\sum_{x_4}\psi_{34}(x_3, x_4)\sum_{x_2}\psi_{23}(x_2, x_3)\sum_{x_1}\psi_{12}(x_1, x_2)$$

$$= \frac{1}{Z} \sum_{x_3} \psi_{35}(x_3, x_5) \sum_{x_4} \psi_{34}(x_3, x_4) \sum_{x_2} \psi_{23}(x_2, x_3) m_{12}(x_2)$$

$$= \cdots$$

$$= \frac{1}{Z} m_{35}(x_5)$$

区别在于此式把条件概率替换为势函数

式**(14.19）**

$$m_{ij}(x_j) = \sum_{x_i} \psi(x_i, x_j) \prod_{k \in n(i) \backslash j} m_{ki}(x_i)$$

解析

此式表示从节点i传递到节点j的过程.其中求和号表示要考虑节点i的所有可能取值，连乘号解释见式(14.20).应当注意此式中连乘号的下标不包括节点j，节点i只需要把自己知道的关于j以外的消息传递给节点j即可.

式**(14.20）**

$$P(x_i) \propto \prod_{k \in n(i)} m_{ki}(x_i)$$

解析

应当注意，此式中是正比于而不是等于. 这涉及概率的规范化，可以这样解释：每个变量都可以看作一位有一些邻居的居民，每个邻居根据其自己的见闻告诉你一些事情 （即传递消息），任何一条消息的可信度应当与所有邻居都有相关性，此式中用乘积来表达这种相关性.

式(14.22)

$$\hat{f} = \frac{1}{N}\sum_{i=1}^{N} f\left(x_i\right)$$

解析

假设x有M种不同的取值，x_i的采样数量为m_i（连续取值可以采用微积分的方法分割为离散的取值），则

$$\begin{aligned}\hat{f} &= \frac{1}{N}\sum_{j=1}^{M} f\left(x_j\right)\cdot m_j \\ &= \sum_{j=1}^{M} f\left(x_j\right)\cdot \frac{m_j}{N} \\ &\simeq \sum_{j=1}^{M} f\left(x_j\right)\cdot p\left(x_j\right) \\ &\simeq \int f(x)p(x)\mathrm{d}x\end{aligned}.$$

式(14.26）

$$p(x^t)T(x^{t-1}|x^t) = p(x^{t-1})T(x^t|x^{t-1})$$

解析

假设变量$\mathbf{x}$所在的空间有n个状态$(s_1, s_2, \cdots, s_n)$,定义在该空间上的一个转移矩阵$\boldsymbol{T} \in \mathbb{R}^{n\times n}$满足一定的条件则该马尔可夫过程存在一个稳态分布$\boldsymbol{\pi}$,使得

$$\boldsymbol{\pi}\boldsymbol{T}=\boldsymbol{\pi},$$

其中,$\boldsymbol{\pi}$是一个n维向量，代表$s_1,s_2,\cdots,s_n$对应的概率.反过来,如果我们希望采样得到符合某个分布$\boldsymbol{\pi}$的一系列变量$\mathbf{x}^1,\mathbf{x}^2,\cdots,\mathbf{x}^t$,应当采用哪一个转移矩阵$\boldsymbol{T}\in\mathbb{R}^{n\times n}$呢？

事实上，转移矩阵只需要满足马尔可夫细致平稳条件

$$\pi_i T_{ij}=\pi_j T_{ji},$$

即式(14.26)，证明如下：

$$(\boldsymbol{\pi}\boldsymbol{T})_j=\sum_i \pi_i T_{ij}=\sum_i \pi_j T_{ji}=\pi_j.$$

这里采用的符号与“西瓜书”略有区别以便于理解

假设采样得到的序列为$\mathbf{x}^1,\mathbf{x}^2,\cdots,\mathbf{x}^{t-1},\mathbf{x}^t$，则可以使用Metropolis-Hastings算法来使得$\mathbf{x}^{t-1}$（假设为状态s_i）转移到$\mathbf{x}^t$（假设为状态s_j）的概率满足式.

式(14.27)

$$p\left(\mathbf{x}^{t-1}\right)Q\left(\mathbf{x}^{*}|\mathbf{x}^{t-1}\right)A\left(\mathbf{x}^{*}|\mathbf{x}^{t-1}\right)=p\left(\mathbf{x}^{*}\right)Q\left(\mathbf{x}^{t-1}|\mathbf{x}^{*}\right)A\left(\mathbf{x}^{t-1}|\mathbf{x}^{*}\right)$$

解析

这里把式(14.26)中的函数T拆分为两个函数(即先验概率和接受概率)之积，便于实际算法的实现.

式(14.28）

$$A(\mathbf{x}^*|\mathbf{x}^{t-1}) = \min\left(1, \frac{p(\mathbf{x}^*)Q(\mathbf{x}^{t-1}|\mathbf{x}^*)}{p(\mathbf{x}^{t-1})Q(\mathbf{x}^*|\mathbf{x}^{t-1})}\right)$$

解析

此式其实是拒绝采样的一个技巧. 因为根据式(14.27)，只需要

$$A(\mathbf{x}^*|\mathbf{x}^{t-1}) = p(\mathbf{x}^*)Q(\mathbf{x}^{t-1}|\mathbf{x}^*),$$
$$A(\mathbf{x}^{t-1}|\mathbf{x}^*) = p(\mathbf{x}^{t-1})Q(\mathbf{x}^*|\mathbf{x}^{t-1}),$$

即可满足式(14.26).但是实际上等号右侧的数值可能比较小，比如各为0.1和0.2，那么好不容易才到的样本只有百分之十几得到利用，所以不妨将接受率设为0.5和1，则细致平稳分布条件依然满足，样本利用率大大提高,所以可以改进为

$$A\left(\mathbf{x}^* \mid \mathbf{x}^{t-1}\right) = \frac{p(\mathbf{x}^*)Q\left(\mathbf{x}^{t-1}|\mathbf{x}^*\right)}{norm},$$
$$A\left(\mathbf{x}^{t-1} \mid \mathbf{x}^*\right) = \frac{p\left(\mathbf{x}^{t-1}\right)Q\left(\mathbf{x}^*|\mathbf{x}^{t-1}\right)}{norm},$$

其中

$$norm = \max\left(p(\mathbf{x}^{t-1})Q(\mathbf{x}^*|\mathbf{x}^{t-1}), p(\mathbf{x}^*)Q(\mathbf{x}^{t-1}|\mathbf{x}^*)\right),$$

即可得到此式.

式**(14.29)**

$$p(\mathbf{x}|\Theta) = \prod_{i=1}^{N}\sum_{\mathbf{z}} p\left(x_i, \mathbf{z}|\Theta\right)$$

解析

此式中，连乘号是因为N个变量的生成过程相互独立，求和号是因为每个变量的生成过程需要考虑中间隐变量的所有可能性.

类似于边际分布的计算方式

式**(14.30**）

$$\ln p(\mathbf{x}|\Theta) = \sum_{i=1}^{N} \ln \left\{ \sum_{\mathbf{z}} p\left(x_i, \mathbf{z}|\Theta\right) \right\}$$

对式(14.29)两侧取对数即有此式

式**(14.31**）

$$\Theta^{t+1} = \arg\max_{\Theta} \mathcal{Q}\left(\Theta; \Theta^{t}\right)$$

$$= \arg\max_{\Theta} \sum_{\mathbf{z}} p\left(\mathbf{z} \mid \mathbf{x}, \Theta^{t}\right) \ln p(\mathbf{x}, \mathbf{z} \mid \Theta)$$

解析

此式为EM算法中的M步.

参见“西瓜书”7.6节

式**(14.32**）

$$\ln p(x) = \mathcal{L}(q) + \mathrm{KL}(q \parallel p)$$

解析

根据条件概率式$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z}|\mathbf{x})p(\mathbf{x})$，可以得到

$$p(\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{z}|\mathbf{x})};$$

然后两边同时取自然对数，可得

$$\ln p(\mathbf{x}) = \ln \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{z}|\mathbf{x})};$$

因为$q(\cdot)$是概率密度函数，所以有

$$1 = \int q(\mathbf{z}) \mathrm{d}\mathbf{z};$$

等式两边同时乘以$\ln p(\mathbf{x})$，因为$\ln p(\mathbf{x})$不是关于变量$\mathbf{z}$的函数，所以可以将$\ln p(\mathbf{x})$拿进积分里面，得到$\ln p(\mathbf{x}) = \int q(\mathbf{z}) \ln p(\mathbf{x}) \mathrm{d}\mathbf{z}$，即有

$$\begin{aligned}
\ln p(\mathbf{x}) &= \int q(\mathbf{z}) \ln p(\mathbf{x}) \mathrm{d}\mathbf{z} \\
&= \int q(\mathbf{z}) \ln \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})} \\
&= \int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \cdot \frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})} \right\} \\
&= \int q(\mathbf{z}) \left(\ln \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} - \ln \frac{p(\mathbf{z} \mid \mathbf{x})}{q(\mathbf{z})} \right) \\
&= \int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right\} - \int q(z) \ln \frac{p(\mathbf{z} \mid \mathbf{x})}{q(\mathbf{z})} \\
&= \mathcal{L}(q) + \mathrm{KL}(q||p).
\end{aligned}$$

最后一行可根据$\mathcal{L}$和KL的定义得到.

式**(14.33)**

$$\mathcal{L}(q) = \int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right\} \mathrm{d}\mathbf{z}$$

参见式(14.32)

式**(14.34）**

$$\mathrm{KL}(q \,\|\, p) = -\int q(\mathbf{z}) \ln \frac{p(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})} \mathrm{d}\mathbf{z}$$

参见式(14.32)

式**(14.35）**

$$q(\mathbf{z}) = \prod_{i=1}^{M} q_i\left(\mathbf{z}_i\right)$$

解析

再一次，此式需满足条件独立的假设. 可以看到，当问题复杂时，往往可尝试将问题简化为最简单、最容易计算的形式.实际上，这样做往往可以获得不错的结果.

式**(14.36)**

$$\mathcal{L}(q) = \int \prod_i q_i \left\{ \ln p(\mathbf{x}, \mathbf{z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{z}$$

$$= \int q_j \left\{ \int p(x, z) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{z}_i \right\} \mathrm{d}\mathbf{z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{z}_j + \text{const}$$

$$= \int q_j \ln \tilde{p}\left(\mathbf{x}, \mathbf{z}_j\right) \mathrm{d}\mathbf{z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{z}_j + \text{const}$$

解析

$$\mathcal{L}(q) = \int \prod_i q_i \left\{ \ln p(\mathbf{x}, \mathbf{z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{z}$$

$$= \int \prod_i q_i \ln p(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z} - \int \prod_i q_i \sum_i \ln q_i \, \mathrm{d}\mathbf{z},$$

其中

$$\int \prod_i q_i \ln p(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z} = \int q_j \prod_{i \neq j} q_i \ln p(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z}$$

$$= \int q_j \left\{ \int \ln p(\mathbf{x}, \mathbf{z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{z}_i \right\} \mathrm{d}\mathbf{z}_j,$$

即先对$\mathbf{z}_j$求积分，再对$\mathbf{z}_i$求积分，这就是式(14.36)左侧的积分部分.

接下来推导右侧的积分部分. 首先计算$\int \prod_i q_i \ln q_k \mathrm{d}\mathbf{z}$, 有

$$\int \prod_i q_i \ln q_k \, \mathrm{d}\mathbf{z} = \int q_{i'} \prod_{i \neq i'} q_i \ln q_k \, \mathrm{d}\mathbf{z}$$

$$= \int q_{i'} \left\{ \int \prod_{i \neq i'} q_i \ln q_k \, \mathrm{d}\mathbf{z}_i \right\} \mathrm{d}\mathbf{z}_{i'}$$

其中第一步是一个展开项，选取了一个变量$q_{i'}(i' \neq k)$. 由于$\left\{ \int \prod_{i \neq i'} q_i \ln q_k \mathrm{d}\mathbf{z}_i \right\}$部分与变量$q_{i'}$无关，所以可以将$q_{i'}$拿到积分外面. 又因

为$\int q_{i'}\,\mathrm{d}\mathbf{z}_{i'}=1$，所以

$$\int\prod_i q_i\ln q_k\ \mathrm{d}\mathbf{z}=\int\prod_{i\neq i'}q_i\ln q_k\ \mathrm{d}\mathbf{z}_i$$

$$=\int q_k\ln q_k\ \mathrm{d}\mathbf{z}_k,$$

即所有k以外的变量都可以通过上面的方式消除. 因此有

$$\int\prod_i q_i\sum_i\ln q_i\ \mathrm{d}\mathbf{z}=\int\prod_i q_i\ln q_j\ \mathrm{d}\mathbf{z}+\sum_{k\neq j}\int\prod_i q_i\ln q_k\ \mathrm{d}\mathbf{z}$$

$$=\int q_j\ln q_j\ \mathrm{d}\mathbf{z}_j+\sum_{z\neq j}\int q_k\ln q_k\ \mathrm{d}\mathbf{z}_k$$

$$=\int q_j\ln q_j\ \mathrm{d}\mathbf{z}_j+\mathrm{const},$$

即式(14.36)右侧的积分部分.

将与q_j无关的部分写作const

式**(14.37)**

$$\ln\tilde{p}\left(\mathbf{x},\mathbf{z}_j\right)=\mathbb{E}_{i\neq j}[\ln p(\mathbf{x},\mathbf{z})]+\ \mathrm{const}$$

参见式(14.36)

式**(14.38**）

$$\mathbb{E}_{i\neq j}\left[\ln p(\mathbf{x},\mathbf{z})\right]=\int\ln p(\mathbf{x},\mathbf{z})\prod_{i\neq j}q_i\mathrm{d}\mathbf{z}_i$$

参见式(14.36)

式**(14.39**）

$$\ln q_j^*(\mathbf{z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{x}, \mathbf{z})] + \text{const}$$

散度取得极值的条件是两个概率分布相同，见“西瓜书”附注C.3

式**(14.40**）

$$q_j^*(\mathbf{z}_j) = \frac{\exp\left(\mathbb{E}_{i \neq j}[\ln p(\mathbf{x}, \mathbf{z})]\right)}{\int \exp\left(\mathbb{E}_{i \neq j}[\ln p(\mathbf{x}, \mathbf{z})]\right) \mathrm{d}\mathbf{z}_j}$$

解析

由式(14.39)取对数并求积分, 有

$$\int q_j^*(\mathbf{z}_j)\,\mathrm{d}\mathbf{z}_j = \int \exp\left(\mathbb{E}_{i \neq j}[\ln p(\mathbf{x}, \mathbf{z})]\right) \cdot \exp(\text{const})\mathrm{d}\mathbf{z}_j$$

$$= \exp(\text{const}) \int \exp\left(\mathbb{E}_{i \neq j}[\ln p(\mathbf{x}, \mathbf{z})]\right) \mathrm{d}\mathbf{z}_j$$

$$= 1,$$

所以

$$\exp(\text{const}) = \frac{1}{\int \exp\left(\mathbb{E}_{i \neq j}[\ln p(\mathbf{x}, \mathbf{z})]\right)\, \mathrm{d}\mathbf{z}_j},$$

因此有

$$q_j^*(\mathbf{z}_j) = \exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{x},\mathbf{z})]\right)\cdot \exp(\text{ const })$$

$$= \frac{\exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{x},\mathbf{z})]\right)}{\int \exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{x},\mathbf{z})]\right) \mathrm{d}\mathbf{z}_j}.$$

式**(14.41**）

$$p(\mathbf{W},\mathbf{z},\boldsymbol{\beta},\Theta|\boldsymbol{\alpha},\boldsymbol{\eta}) = \prod_{t=1}^{T} p(\Theta_t|\boldsymbol{\alpha}) \prod_{k=1}^{K} p(\boldsymbol{\beta}_k|\boldsymbol{\eta}) \left(\prod_{n=1}^{N} p(w_{t,n}|z_{t,n},\boldsymbol{\beta}_k) p(z_{t,n}|\Theta_t)\right)$$

解析

此式表示LDA模型下根据参数$\boldsymbol{\alpha},\boldsymbol{\eta}$生成文档$\boldsymbol{W}$的概率. 其中$\mathbf{z},\boldsymbol{\beta},\Theta$是生成过程的中间变量.具体的生成步骤可见“西瓜书”图14.12，图中的箭头和式(14.41)中的条件概率中的因果项目一一对应.这里共有3个连乘号，表示3个相互独立的概率关系：第1个连乘表示T个文档的话题分布都是相互独立的；第2个连乘表示K个话题的单词分布是相互独立的；第3个连乘表示每篇文档中的所有单词的生成是相互独立的.

式**(14.42)**

$$p(\Theta_t|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \prod_k \Theta_{t,k}^{\alpha_k - 1}$$

参见“西瓜书”附录C.1.6

式**(14.43**）

$$LL(\boldsymbol{\alpha}, \boldsymbol{\eta}) = \sum_{t=1}^{T} \ln p\left(\boldsymbol{w}_t | \boldsymbol{\alpha}, \boldsymbol{\eta}\right)$$

解析

此为对数似然函数.

参见西瓜7.2节

式**(14.44**）

$$p(\mathbf{z}, \boldsymbol{\beta}, \Theta | \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{p(\mathbf{W}, \mathbf{z}, \boldsymbol{\beta}, \Theta | \boldsymbol{\alpha}, \boldsymbol{\eta})}{p(\mathbf{W} | \boldsymbol{\alpha}, \boldsymbol{\eta})}$$

解析

此式中分母为边际分布，需要对变量$\mathbf{z}, \boldsymbol{\beta}, \Theta$ 求积分或者求和，所以往往难以直接求解.

第15章　规则学习

式(15.2)

$$\mathrm{LRS} = 2 \cdot \left(\hat{m}_+ \log_2 \frac{\left(\frac{\hat{m}_+}{\hat{m}_+ + \hat{m}_-}\right)}{\left(\frac{m_+}{m_+ + m_-}\right)} + \hat{m}_- \log_2 \frac{\left(\frac{\hat{m}_-}{\hat{m}_+ + \hat{m}_-}\right)}{\left(\frac{m_-}{m_+ + m_-}\right)} \right)$$

解析

此为似然率统计量(Likelihood Ratio Statistics, LRS)的定义式.

式(15.3)

$$\mathrm{F\_Gain} = \hat{m}_+ \times \left(\log_2 \frac{\hat{m}_+}{\hat{m}_+ + \hat{m}_-} - \log_2 \frac{m_+}{m_+ + m_-} \right)$$

解析

此为FOIL增益(FOIL gain)的定义式.

式(15.6)

$$(A \vee B) - \{B\} = A$$

解析

此为析合范式的删除操作定义式，表示在A和B的析合式中删除成分B，得到成分A.

式(15.7）

$$C = (C_1 - \{L\}) \vee (C_2 - \{\neg L\})$$

解析

$C = A \vee B$，把$A = C_1 - \{L\}$和$B = C_2 - \{\neg L\}$代入即得.

式(15.9）

$$C_2 = (C - (C_1 - \{L\})) \vee \{\neg L\}$$

解析

由式(15.7)可知

$$C_2 - \{\neg L\} = C - (C_1 - \{L\}),$$

由式(15.6)移项即证得.

式(15.10)

$$\frac{p \leftarrow A \wedge B \quad q \leftarrow A}{p \leftarrow q \wedge B \quad q \leftarrow A}$$

解析

此为吸收(absorption)操作的定义式.

式(15.11）

$$\frac{p \leftarrow A \wedge B \quad p \leftarrow A \wedge q}{q \leftarrow B \quad p \leftarrow A \wedge q}$$

解析

此为辨识(identification)操作的定义式.

式**(15.12**）

$$\frac{p \leftarrow A \wedge B \quad p \leftarrow A \wedge C}{q \leftarrow B \quad p \leftarrow A \wedge q \quad q \leftarrow C}$$

解析

此为内构(in tra-construction)操作的定义式.

式**(15.13**）

$$\frac{p \leftarrow A \wedge B \quad q \leftarrow A \wedge C}{p \leftarrow r \wedge B \quad r \leftarrow A \quad q \leftarrow r \wedge C}$$

解析

此为互构(in ter-construction)操作的定义式.

式**(15.14**）

$$C = (C_1 - \{L_1\})\,\theta \vee (C_2 - \{L_2\})\,\theta$$

解析

由式(15.7)分别对析合的两个子项进行归结即得证.

式**(15.16)**

$$C_2 = (C - (C_1 - \{L_1\})\,\theta_1 \vee \{\neg L_1\theta_1\})\,\theta_2^{-1}$$

这里θ_2^{-1}应该放在括号里. 由式(15.9)有

$$C_2 = (C - (C_1 - \{L_1\})) \vee \{L_2\},$$

其中 $L_2 = (\neg L_1\theta_1)\theta_2^{-1}$, 代入即得证.

第16章　强化学习

式(16.2）

$$Q_n(k) = \frac{1}{n}\left((n-1) \times Q_{n-1}(k) + v_n\right)$$

解析

$$Q_n(k) = \frac{1}{n}\sum_{i=1}^{n} v_i$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n-1} v_i + v_n\right)$$

$$= \frac{1}{n}\left((n-1)Q_{n-1}(k) + v_n\right)$$

$$= Q_{n-1}(k) + \frac{1}{n}\left(v_n - Q_{n-1}(k)\right).$$

式(16.3）

$$Q_n(k) = \frac{1}{n}\left((n-1) \times Q_{n-1}(k) + v_n\right)$$

$$= Q_{n-1}(k) + \frac{1}{n}\left(v_n - Q_{n-1}(k)\right)$$

参见式 (16.2)

式(16.4）

$$P(k)=\frac{\mathrm{e}^{\frac{Q(k)}{\tau}}}{\sum_{i=1}^{K}\mathrm{e}^{\frac{Q(i)}{\tau}}}$$

解析

$$P(k)=\frac{\mathrm{e}^{\frac{Q(k)}{\tau}}}{\sum_{i=1}^{K}\mathrm{e}^{\frac{Q(i)}{\tau}}}\propto \mathrm{e}^{\frac{Q(k)}{\tau}}\propto \frac{Q(k)}{\tau}\propto \frac{1}{\tau}.$$

式**(16.7)**

$$V_T^{\pi}(x)=\mathbb{E}_{\pi}\left[\frac{1}{T}\sum_{t=1}^{T}r_t \mid x_0=x\right] \quad ①$$

$$=\mathbb{E}_{\pi}\left[\frac{1}{T}r_1+\frac{T-1}{T}\frac{1}{T-1}\sum_{t=2}^{T}r_t \mid x_0=x\right] \quad ②$$

$$=\sum_{a\in A}\pi(x,a)\sum_{x'\in X}P_{x\to x'}^{a}\left(\frac{1}{T}R_{x\to x'}^{a}+\frac{T-1}{T}\mathbb{E}_{\pi}\left[\frac{1}{T-1}\sum_{t=1}^{T-1}r_t \mid x_0=x'\right]\right) \quad ③$$

$$=\sum_{a\in A}\pi(x,a)\sum_{x'\in X}P_{x\to x'}^{a}\left(\frac{1}{T}R_{x\to x'}^{a}+\frac{T-1}{T}V_{T-1}^{\pi}\left(x'\right)\right) \quad ④$$

解析

此处主要解释②→③. 因为 $\pi(x,a)$表示在状态x下选择动作a的概率，且动作事件之间两两互斥且和为动作空间，由全概率展开式

$$P(A)=\sum_{i=1}^{\in fty}P(B_i)P(A \mid B_i)$$

可得

$$\mathbb{E}_\pi\left[\frac{1}{T}r_1+\frac{T-1}{T}\frac{1}{T-1}\sum_{t=2}^{T}r_t\mid x_0=x\right]$$

$$=\sum_{a\in A}\pi(x,a)\sum_{x'\in X}P^a_{x\to x'}\left(\frac{1}{T}R^a_{x\to x'}+\frac{T-1}{T}\mathbb{E}_\pi\left[\frac{1}{T-1}\sum_{t=1}^{T-1}r_t\mid x_0=x'\right]\right),$$

其中

$$r_1=\pi(x,a)P^a_{x\to x'}R^a_{x\to x'}.$$

③→④为递归.

式(16.8）

$$V^\pi_\gamma(x)=\sum_{a\in A}\pi(x,a)\sum_{x'\in X}P^a_{x\to x'}(R^a_{x\to x'}+\gamma V^\pi_\gamma(x'))$$

解析

$$V^\pi_\gamma(x)=\mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r_{t+1}\mid x_0=x\right]$$

$$=\mathbb{E}_\pi\left[r_1+\sum_{t=1}^{\infty}\gamma^t r_{t+1}\mid x_0=x\right]$$

$$=\mathbb{E}_\pi\left[r_1+\gamma\sum_{t=1}^{\infty}\gamma^{t-1} r_{t+1}\mid x_0=x\right]$$

$$=\sum_{a\in A}\pi(x,a)\sum_{x'\in X}P^a_{x\to x'}\left(R^a_{x\to x'}+\gamma\mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r_{t+1}\mid x_0=x'\right]\right)$$

$$= \sum_{a \in A} \pi(x,a) \sum_{x' \in X} P_{x \to x'}^{a} \left(R_{x \to x'}^{a} + \gamma V_{\gamma}^{\pi}(x') \right)$$

.

式**(16.10)**

$$\begin{cases} Q_{T}^{\pi}(x,a) = \sum_{x' \in X} P_{x \to x'}^{a} \left(\frac{1}{T} R_{x \to x'}^{a} + \frac{T-1}{T} V_{T-1}^{\pi}(x') \right) \\ Q_{\gamma}^{\pi}(x,a) = \sum_{x' \in X} P_{x \to x'}^{a} \left(R_{x \to x'}^{a} + \gamma V_{\gamma}^{\pi}(x') \right) \end{cases}$$

参见式(16.7)和式(16.8)

式**(16.14）**

$$V^{*}(x) = \max_{a \in A} Q^{\pi^{*}}(x,a)$$

解析

此式的目的是获得最优的状态值函数V.这里取了两层最优，分别是采用最优策略π^{*}和选取使得状态动作值函数Q最大的动作$\max_{a \in A}$.

式**(16.16）**

$$V^{\pi}(x) \leqslant V^{\pi'(x)}$$

解析

$$V^{\pi}(x) \leqslant Q^{\pi}(x, \pi'(x))$$

$$= \sum_{x' \in X} P_{x \to x'}^{\pi'(x)} \left(R_{x \to x'}^{\pi'(x)} + \gamma V^{\pi}(x') \right)$$

$$\leqslant \sum_{x' \in X} P_{x \to x'}^{\pi'(x)} \left(R_{x \to x'}^{\pi'(x)} + \gamma Q^{\pi}(x', \pi'(x')) \right)$$

$$= \sum_{x' \in X} P_{x \to x'}^{\pi'(x)} \left(R_{x \to x'}^{\pi'(x)} + \sum_{x'' \in X} P_{x' \to x''}^{\pi'(x')} \left(\gamma R_{x' \to x''}^{\pi'(x')} + \gamma^2 V^{\pi} \left(x'' \right) \right) \right)$$

$$\leqslant \sum_{x' \in X} P_{x \to x'}^{\pi'(x)} \left(R_{x \to x'}^{\pi'(x)} + \sum_{x'' \in X} P_{x' \to x''}^{\pi'(x')} \left(\gamma R_{x' \to x''}^{\pi'(x')} + \gamma^2 Q^{\pi} \left(x'', \pi' \left(x'' \right) \right) \right) \right)$$

$$\leqslant \cdots$$

$$\leqslant \sum_{x' \in X} P_{x \to x'}^{\pi'(x)} \left(R_{x \to x'}^{\pi'(x)} + \sum_{x'' \in X} P_{x' \to x''}^{\pi'(x')'} \right.$$

$$\left. \left(\gamma R_{x' \to x''}^{\pi'(x')} + \sum_{x'' \in X} P_{x'' \to x'''}^{\pi'(x'')} \left(\gamma^2 R_{x'' \to x'''}^{\pi'(x'')} + \cdots \right) \right) \right)$$

$$= V^{\pi'}(x)$$

其中，使用了动作改变条件

$$Q^{\pi}(x, \pi'(x)) \geqslant V^{\pi}(x)$$

以及状态-动作值函数

$$Q^{\pi}(x', \pi'(x')) = \sum_{x' \in X} P_{x' \to x'}^{\pi'(x')} \left(R_{x' \to x'}^{\pi'(x')} + \gamma V^{\pi}(x') \right).$$

于是，当前状态的最优值函数

$$V^{*}(x) = V^{\pi'}(x) \geqslant V^{\pi}(x).$$

式**(16.31)**

$$Q_{t+1}^{\pi}(x, a) = Q_t^{\pi}(x, a) + \alpha \left(R_{x \to x'}^{a} + \gamma Q_t^{\pi}(x', a') - Q_t^{\pi}(x, a) \right)$$

解析

对比式(16.29)

$$Q_{t+1}^{\pi}(x,a) = Q_t^{\pi}(x,a) + \frac{1}{t+1}\left(r_{t+1} - Q_t^{\pi}(x,a)\right).$$

由 $\frac{1}{t+1} = \alpha$ 可知, 若下式成立，则式(16.31)成立

$$r_{t+1} = R_{x\to x'}^{a} + \gamma Q_t^{\pi}(x',a'),$$

而r_{t+1}表示$t+1$步的奖赏，即状态x变化到x'的奖赏加上前面t步奖赏总和$Q_t^{\pi}(x',a')$的γ折扣，因此这个式子成立.