南瓜书

PUMPKIN BOOK

Datawhale

版本号:1.1.0
发布日期:2022.09

前言

“周志华老师的《机器学习》（西瓜书）是机器学习领域的经典入门教材之一，周老师为了使尽可能多的读者通过西瓜书对机器学习有所了解，所以在书中对部分公式的推导细节没有详述，但是这对那些想深究公式推导细节的读者来说可能“不太友好”，本书旨在对西瓜书里比较难理解的公式加以解析，以及对部分公式补充具体的推导细节。”

读到这里，大家可能会疑问为啥前面这段话加了引号，因为这只是我们最初的遐想，后来我们了解到，周老师之所以省去这些推导细节的真实原因是，他本尊认为“理工科数学基础扎实点的大二下学生应该对西瓜书中的推导细节无困难吧，要点在书里都有了，略去的细节应能脑补或做练习”。所以..... 本南瓜书只能算是我等数学渣渣在自学的时候记下来的笔记，希望能够帮助大家都成为一名合格的“理工科数学基础扎实点的大二下学生”。

使用说明

- 南瓜书的所有内容都是以西瓜书的内容为前置知识进行表述的，所以南瓜书的最佳使用方法是以西瓜书为主线，遇到自己推导不出来或者看不懂的公式时再来查阅南瓜书；
- 对于初学机器学习的小白，西瓜书第 1 章和第 2 章的公式**强烈不建议深究**，简单过一下即可，等你学得有点飘的时候再回来啃都来得及；
- 每个公式的解析和推导我们都力 (zhi) 争 (neng) 以本科数学基础的视角进行讲解，所以超纲的数学知识我们通常都会以附录和参考文献的形式给出，感兴趣的同学可以继续沿着我们给的资料进行深入学习；
- 若南瓜书里没有你想要查阅的公式，或者你发现南瓜书哪个地方有错误，请毫不犹豫地去我们 GitHub 的 Issues（地址：`https://github.com/datawhalechina/pumpkin-book/issues`）进行反馈，在对应版块提交你希望补充的公式编号或者勘误信息，我们通常会在 24 小时以内给您回复，超过 24 小时未回复的话可以邮件联系我们（微信号：at-Sm1les）；

配套视频教程： `https://www.bilibili.com/video/BV1Mh411e7VU`
在线阅读地址： `https://datawhalechina.github.io/pumpkin-book`（仅供第 1 版）
最新版 PDF 获取地址： `https://github.com/datawhalechina/pumpkin-book/releases`

编委会

主编： Sm1les、archwalker、jbb0523
编委： juxiao、Majingmin、MrBigFan、shanry、Ye980226

致谢

特别感谢 awyd234、feijuan、Ggmatch、Heitao5200、huaqing89、LongJH、LilRachel、LeoLRH、Nono17、spareribs、sunchaothu、StevenLzq 在最早期的时候对南瓜书所做的贡献。

扫描下方二维码，然后回复关键词“**南瓜书**”，即可加入“南瓜书读者交流群”

Datawhale
一个专注于 AI 领域的开源组织

目录

第 1 章 绪论

本章作为“西瓜书”的开篇，主要讲解什么是机器学习以及机器学习的相关数学符号，为后续内容作铺垫，并未涉及复杂的算法理论，因此阅读本章时只需耐心梳理清楚所有概念和数学符号即可。此外，在阅读本章前建议先阅读西瓜书目录前页的《主要符号表》，它能解答在阅读“西瓜书”过程中产生的大部分对数学符号的疑惑。

本章也作为本书的开篇，笔者在此赘述一下本书的撰写初衷，本书旨在以“过来人”的视角陪读者一起阅读“西瓜书”，尽力帮读者消除阅读过程中的“数学恐惧”，只要读者学习过《高等数学》、《线性代数》和《概率论与数理统计》这三门大学必修的数学课，均能看懂本书对西瓜书中的公式所做的解释和推导，同时也能体会到这三门数学课在机器学习上碰撞产生的“数学之美”。

1.1 引言

本节以概念理解为主，在此对“算法”和“模型”作补充说明。“算法”是指从数据中学得“模型”的具体方法，例如后续章节中将会讲述的线性回归、对数几率回归、决策树等。“算法”产出的结果称为“模型”，通常是具体的函数或者可抽象地看作为函数，例如一元线性回归算法产出的模型即为形如 $f(x) = wx + b$ 的一元一次函数。

1.2 基本术语

本节涉及的术语较多且很多术语都有多个称呼，下面梳理各个术语，并将最常用的称呼加粗标注。

样本：也称为“示例”，是关于一个事件或对象的描述。因为要想让计算机能对现实生活中的事物进行机器学习，必须先将其抽象为计算机能理解的形式，计算机最擅长做的就是进行数学运算，因此考虑如何将其抽象为某种数学形式。显然，线性代数中的向量就很适合，因为任何事物都可以由若干“特征”（或称为“属性”）唯一刻画出来，而向量的各个维度即可用来描述各个特征。例如，如果用色泽、根蒂和敲声这 3 个特征来刻画西瓜，那么一个“色泽青绿，根蒂蜷缩，敲声清脆”的西瓜用向量来表示即为 $\boldsymbol{x} =$ (青绿; 蜷缩; 清脆)（向量中的元素用分号“;”分隔时表示此向量为列向量，用逗号“,”分隔时表示为行向量），其中青绿、蜷缩和清脆分别对应为相应特征的取值，也称为“属性值”。显然，用中文书写向量的方式不够“数学”，因此需要将属性值进一步数值化，具体例子参见“西瓜书”第 3 章 3.2。此外，仅靠以上 3 个特征来刻画西瓜显然不够全面细致，因此还需要扩展更多维度的特征，一般称此类与特征处理相关的工作为“特征工程”。

样本空间：也称为“输入空间”或“属性空间”。由于样本采用的是标明各个特征取值的“特征向量”来进行表示，根据线性代数的知识可知，有向量便会有向量所在的空间，因此称表示样本的特征向量所在的空间为样本空间，通常用花式大写的 $\mathcal{X}$ 表示。

数据集：数据集通常用集合来表示，令集合 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m\}$ 表示包含 m 个样本的数据集，一般同一份数据集中的每个样本都含有相同个数的特征，假设此数据集中的每个样本都含有 d 个特征，则第 i 个样本的数学表示为 d 维向量：$\boldsymbol{x}_i = (x_{i1}; x_{i2}; ...; x_{id})$，其中 x_{ij} 表示样本 $\boldsymbol{x}_i$ 在第 j 个属性上的取值。

模型：机器学习的一般流程如下：首先收集若干样本（假设此时有 100 个），然后将其分为训练样本（80 个）和测试样本（20 个），其中 80 个训练样本构成的集合称为“训练集”，20 个测试样本构成的集合称为“测试集”，接着选用某个机器学习算法，让其在训练集上进行“学习”（或称为“训练”），然后产出得到“模型”（或称为“学习器”），最后用测试集来测试模型的效果。执行以上流程时，表示我们已经默认样本的背后是存在某种潜在的规律，我们称这种潜在的规律为“真相”或者“真实”，例如样本是一堆好西瓜和坏西瓜时，我们默认的便是好西瓜和坏西瓜背后必然存在某种规律能将其区分开。当我们应用某个机器学习算法来学习时，产出得到的模型便是该算法所找到的它自己认为的规律，由于该规律通常并不一定就是所谓的真相，所以也将其称为“假设”。通常机器学习算法都有可配置的参数，同一个机器学习算法，使用不同的参数配置或者不同的训练集，训练得到的模型通常都不同。

标记：上文提到机器学习的本质就是在学习样本在某个方面的表现是否存在潜在的规律，我们称该方面的信息为“标记”。例如在学习西瓜的好坏时，“好瓜”和“坏瓜”便是样本的标记。一般第 i 个样本的标记的数学表示为 y_i，标记所在的空间称为“标记空间”或“输出空间”，数学表示为花式大写的 $\mathcal{Y}$。标记通常也看作为样本的一部分，因此，一个完整的样本通常表示为 $(\boldsymbol{x}, y)$。

根据标记的取值类型不同，可将机器学习任务分为以下两类：

- 当标记取值为离散型时，称此类任务为“分类”，例如学习西瓜是好瓜还是坏瓜、学习猫的图片是白猫还是黑猫等。当分类的类别只有两个时，称此类任务为“二分类”，通常称其中一个为“正类”，另一个为“反类”或“负类”；当分类的类别超过两个时，称此类任务为“多分类”。由于标记也属于样本的一部分，通常也需要参与运算，因此也需要将其数值化，例如对于二分类任务，通常将正类记为 1，反类记为 0，即 $\mathcal{Y}=\{0,1\}$。这只是一般默认的做法，具体标记该如何数值化可根据具体机器学习算法进行相应地调整，例如第 6 章的支持向量机算法则采用的是 $\mathcal{Y}=\{-1,+1\}$；
- 当标记取值为连续型时，称此类任务为“回归”，例如学习预测西瓜的成熟度、学习预测未来的房价等。由于是连续型，因此标记的所有可能取值无法直接罗列，通常只有取值范围，回归任务的标记取值范围通常是整个实数域 $\mathbb{R}$，即 $\mathcal{Y}=\mathbb{R}$。

无论是分类还是回归，机器学习算法最终学得的模型都可以抽象地看作为以样本 $\boldsymbol{x}$ 为自变量，标记 y 为因变量的函数 $y=f(\boldsymbol{x})$，即一个从输入空间 $\mathcal{X}$ 到输出空间 $\mathcal{Y}$ 的映射。例如在学习西瓜的好坏时，机器学习算法学得的模型可看作为一个函数 $f(\boldsymbol{x})$，给定任意一个西瓜样本 $\boldsymbol{x}_i=$(青绿;蜷缩;清脆)，将其输入进函数即可计算得到一个输出 $y_i=f(\boldsymbol{x}_i)$，此时得到的 y_i 便是模型给出的预测结果，当 y_i 取值为 1 时表明模型认为西瓜 $\boldsymbol{x}_i$ 是好瓜，当 y_i 取值为 0 时表明模型认为西瓜 $\boldsymbol{x}_i$ 是坏瓜。

根据是否有用到标记信息，可将机器学习任务分为以下两类：

- 在模型训练阶段有用到标记信息时，称此类任务为“监督学习”，例如第 3 章的线性模型；
- 在模型训练阶段没用到标记信息时，称此类任务为“无监督学习”，例如第 9 章的聚类。

泛化：由于机器学习的目标是根据已知来对未知做出尽可能准确的判断，因此对未知事物判断的准确与否才是衡量一个模型好坏的关键，我们称此为“泛化”能力。例如学习西瓜好坏时，假设训练集中共有 3 个样本：$\{(\boldsymbol{x}_1=$(青绿;蜷缩),$y_1=$好瓜),($\boldsymbol{x}_2=$(乌黑;蜷缩),$y_2=$好瓜),($\boldsymbol{x}_3=$(浅白;蜷缩),$y_3=$好瓜)$\}$，同时假设判断西瓜好坏的真相是“只要根蒂蜷缩就是好瓜”，如果应用算法 A 在此训练集上训练得到模型 $f_a(\boldsymbol{x})$，模型 a 学到的规律是“色泽等于青绿、乌黑或者浅白时，同时根蒂蜷缩即为好瓜，否则便是坏瓜”，再应用算法 B 在此训练集上训练得到模型 $f_b(\boldsymbol{x})$，模型 $f_b(\boldsymbol{x})$ 学到的规律是“只要根蒂蜷缩就是好瓜”，因此对于一个未见过的西瓜样本 $\boldsymbol{x}=$(金黄;蜷缩) 来说，模型 $f_a(\boldsymbol{x})$ 给出的预测结果为“坏瓜”，模型 $f_b(\boldsymbol{x})$ 给出的预测结果为“好瓜”，此时我们称模型 $f_b(\boldsymbol{x})$ 的泛化能力优于模型 $f_a(\boldsymbol{x})$。

通过以上举例可知，尽管模型 $f_a(\boldsymbol{x})$ 和模型 $f_b(\boldsymbol{x})$ 对训练集学得一样好，即两个模型对训练集中每个样本的判断都对，但是其所学到的规律是不同的。导致此现象最直接的原因是算法的不同，但是算法通常是有限的，可穷举的，尤其是在特定任务场景下可使用的算法更是有限，因此，数据便是导致此现象的另一重要原因，这也就是机器学习领域常说的“数据决定模型的上限，而算法则是让模型无限逼近上限”，下面详细解释此话的含义。

先解释“数据决定模型效果的上限”，其中数据是指从数据量和特征工程两个角度考虑。从数据量的角度来说，通常数据量越大模型效果越好，因为数据量大即表示累计的经验多，因此模型学习到的经验也多，自然表现效果越好。例如以上举例中如果训练集中含有相同颜色但根蒂不蜷缩的坏瓜，模型 a 学到真相的概率则也会增大；从特征工程的角度来说，通常对特征数值化越合理，特征收集越全越细致，模型效果通常越好，因为此时模型更易学得样本之间潜在的规律。例如学习区分亚洲人和非洲人时，此时样本即为人，在进行特征工程时，如果收集到每个样本的肤色特征，则其他特征例如年龄、身高和体重等便可省略，因为只需靠肤色这一个特征就足以区分亚洲人和非洲人。

而“算法则是让模型无限逼近上限”是指当数据相关的工作已准备充分时，接下来便可用各种可适用的算法从数据中学习其潜在的规律进而得到模型，不同的算法学习得到的模型效果自然有高低之分，效果越好则越逼近上限，即逼近真相。

分布：此处的“分布”指的是概率论中的概率分布，通常假设样本空间服从一个未知“分布”$\mathcal{D}$，而我们收集到的每个样本都是独立地从该分布中采样得到，即“独立同分布”。通常收集到的样本越多，越能从样本中反推出 $\mathcal{D}$ 的信息，即越接近真相。此假设属于机器学习中的经典假设，在后续学习机器学习算法过程中会经常用到。

1.3 假设空间

本节的重点是理解“假设空间”和“版本空间”，下面以“房价预测”举例说明。假设现已收集到某地区近几年的房价和学校数量数据，希望利用收集到的数据训练出能通过学校数量预测房价的模型，具体收集到的数据如下。

表 1-1 房价预测

年份	学校数量	房价
2020	1 所	1 万/m^2
2021	2 所	4 万/m^2

基于对以上数据的观察以及日常生活经验，不难得出“房价与学校数量成正比”的假设，若将学校数量设为 x，房价设为 y，则该假设等价表示学校数量和房价呈 $y = wx + b$ 的一元一次函数关系，此时房价预测问题的假设空间即为“一元一次函数”。确定假设空间以后便可以采用机器学习算法从假设空间中学得模型，即从一元一次函数空间中学得能满足表1-1中数值关系的某个一元一次函数。学完第 3 章的线性回归可知当前问题属于一元线性回归问题，根据一元线性回归算法可学得模型为 $y = 3x - 2$。

除此之外，也可以将问题复杂化，假设学校数量和房价呈 $y = wx^2 + b$ 一元二次函数关系，此时问题变为了线性回归中的多项式回归问题，按照多项式回归算法可学得模型为 $y = x^2$。因此，以表1-1中数据作为训练集可以有多个假设空间，且在不同的假设空间中都有可能学得能够拟合训练集的模型，我们将所有能够拟合训练集的模型构成的集合称为“版本空间”。

1.4 归纳偏好

在上一节“房价预测”的例子中，当选用一元线性回归算法时，学得的模型是一元一次函数，当选用多项式回归算法时，学得的模型是一元二次函数，所以不同的机器学习算法有不同的偏好，我们称为“归纳偏好”。对于当前房价预测这个例子来说，这两个算法学得的模型哪个更好呢？著名的“奥卡姆剃刀”原则认为“若有多个假设与观察一致，则选最简单的那个”，但是何为“简单”便见仁见智了，如果认为函数的幂次越低越简单，则此时一元线性回归算法更好，如果认为幂次越高越简单，则此时多项式回归算法更好，因此该方法其实并不“简单”，所以并不常用，而最常用的方法则是基于模型在测试集上的表现来评判模型之间的优劣。测试集是指由训练集之外的样本构成的集合，例如在当前房价预测问题中，通常会额外留有部分未参与模型训练的数据来对模型进行测试。假设此时额外留有 1 条数据：(年份：2022年;学校数量：3所;房价：7万/m^2) 用于测试，模型 $y = 3x - 2$ 的预测结果为 $3 * 3 - 2 = 7$，预测正确，模型 $y = x^2$ 的预测结果为 $3^2 = 9$，预测错误，因此，在当前房价预测问题上，我们认为一元线性回归算法优于多项式回归算法。

机器学习算法之间没有绝对的优劣之分，只有是否适合当前待解决的问题之分，例如上述测试集中的数据如果改为 (年份：2022年;学校数量：3所;房价：9万/m^2) 则结论便逆转为多项式回归算法优于一元线性回归算法。

1.4.1 式 (1.1) 和式 (1.2) 的解释

$$\begin{aligned}
\sum_f E_{ote}(\mathfrak{L}_a|X,f) &= \sum_f \sum_h \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x})\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))P(h|X,\mathfrak{L}_a) && ① \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h|X,\mathfrak{L}_a) \sum_f \mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x})) && ② \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h|X,\mathfrak{L}_a) \frac{1}{2}2^{|\mathcal{X}|} && ③ \\
&= \frac{1}{2}2^{|\mathcal{X}|} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h|X,\mathfrak{L}_a) && ④ \\
&= 2^{|\mathcal{X}|-1} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \cdot 1 && ⑤
\end{aligned}$$

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 $\sum_f \mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))=\frac{1}{2}2^{|\mathcal{X}|}$。

需要注意的是，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。套用到上述“房价预测”的例子中，我们认为只有能正确拟合测试集的函数才是真实目标函数，也就是我们希望学得的模型。

第 2 章 模型评估与选择

如“西瓜书”前言所述，本章仍属于机器学习基础知识，如果说第 1 章介绍了什么是机器学习及机器学习的相关数学符号，那么本章则进一步介绍机器学习的相关概念。具体来说，介绍内容正如本章名称“模型评估与选择”所述，讲述的是如何评估模型的优劣和选择最适合自己业务场景的模型。

由于“模型评估与选择”是在模型产出以后进行的下游工作，要想完全吸收本章内容需要读者对模型有一些基本的认知，因此零基础的读者直接看本章会很吃力，实属正常，在此建议零基础的读者可以简单泛读本章，仅看能看懂的部分即可，或者直接跳过本章从第 3 章开始看，直至看完第 6 章以后再回头来看本章便会轻松许多。

2.1 经验误差与过拟合

梳理本节的几个概念。

错误率：$E=\frac{a}{m}$，其中 m 为样本个数，a 为分类错误样本个数。

精度：精度 =1-错误率。

误差：学习器的实际预测输出与样本的真实输出之间的差异。

经验误差：学习器在训练集上的误差，又称为“训练误差”。

泛化误差：学习器在新样本上的误差。

经验误差和泛化误差用于分类问题的定义式可参见“西瓜书”第 12 章的式 (12.1) 和式 (12.2)，接下来辨析一下以上几个概念。

错误率和精度很容易理解，而且很明显是针对分类问题的。误差的概念更适用于回归问题，但是，根据“西瓜书”第 12 章的式 (12.1) 和式 (12.2) 的定义可以看出，在分类问题中也会使用误差的概念，此时的“差异”指的是学习器的实际预测输出的类别与样本真实的类别是否一致，若一致则“差异”为 0，若不一致则“差异”为 1，训练误差是在训练集上差异的平均值，而泛化误差则是在新样本（训练集中未出现过的样本）上差异的平均值。

过拟合是由于模型的学习能力相对于数据来说过于强大，反过来说，**欠拟合**是因为模型的学习能力相对于数据来说过于低下。暂且抛开“没有免费的午餐”定理不谈，例如对于“西瓜书”第 1 章图 1.4 中的训练样本（黑点）来说，用类似于抛物线的曲线 A 去拟合则较为合理，而比较崎岖的曲线 B 相对于训练样本来说学习能力过于强大，但若仅用一条直线去训练则相对于训练样本来说直线的学习能力过于低下。

2.2 评估方法

本节介绍了 3 种模型评估方法：留出法、交叉验证法、自助法。留出法由于操作简单，因此最常用；交叉验证法常用于对比同一算法的不同参数配置之间的效果，以及对比不同算法之间的效果；自助法常用于集成学习（详见“西瓜书”第 8 章的 8.2 节和 8.3 节）产生基分类器。留出法和自助法简单易懂，在此不再赘述，下面举例说明交叉验证法的常用方式。

对比同一算法的不同参数配置之间的效果：假设现有数据集 D，且有一个被评估认为适合用于数据集 D 的算法 $\mathfrak{L}$，该算法有可配置的参数，假设备选的参数配置方案有两套：方案 a，方案 b。下面通过交叉验证法为算法 $\mathfrak{L}$ 筛选出在数据集 D 上效果最好的参数配置方案。以 3 折交叉验证为例，首先按照“西瓜书”中所说的方法，通过分层采样将数据集 D 划分为 3 个大小相似的互斥子集：D_1, D_2, D_3，然后分别用其中 1 个子集作为测试集，其他子集作为训练集，这样就可获得 3 组训练集和测试集：

$$\text{训练集 1：}D_1 \cup D_2\text{，测试集 1:}D_3$$
$$\text{训练集 2：}D_1 \cup D_3\text{，测试集 2:}D_2$$
$$\text{训练集 3：}D_2 \cup D_3\text{，测试集 3:}D_1$$

接下来用算法 $\mathfrak{L}$ 搭配方案 a 在训练集 1 上进行训练，训练结束后将训练得到的模型在测试集 1 上进行测试，得到测试结果 1，依此方法再分别通过训练集 2 和测试集 2、训练集 3 和测试集 3 得到测试结果

2 和测试结果 3，最后将 3 次测试结果求平均即可得到算法 $\mathfrak{L}$ 搭配方案 a 在数据集 D 上的最终效果，记为 $Score_a$。同理，按照以上方法也可得到算法 $\mathfrak{L}$ 搭配方案 b 在数据集 D 上的最终效果 $Score_b$，最后通过比较 $Score_a$ 和 $Score_b$ 之间的优劣来确定算法 $\mathfrak{L}$ 在数据集 D 上效果最好的参数配置方案。

对比不同算法之间的效果：同上述“对比同一算法的不同参数配置之间的效果”中所讲的方法一样，只需将其中的“算法 $\mathfrak{L}$ 搭配方案 a”和“算法 $\mathfrak{L}$ 搭配方案 b”分别换成需要对比的算法 α 和算法 β 即可。

从以上的举例可以看出，交叉验证法本质上是在进行多次留出法，且每次都换不同的子集做测试集，最终让所有样本均至少做 1 次测试样本。这样做的理由其实很简单，因为一般的留出法只会划分出 1 组训练集和测试集，仅依靠 1 组训练集和测试集去对比不同算法之间的效果显然不够置信，偶然性太强，因此要想基于固定的数据集产生多组不同的训练集和测试集，则只有进行多次划分，每次采用不同的子集作为测试集，也即为交叉验证法。

2.2.1 算法参数（超参数）与模型参数

算法参数是指算法本身的一些参数（也称超参数），例如 k 近邻的近邻个数 k、支持向量机的参数 C（详见“西瓜书”第 6 章式 (6.29)）。算法配置好相应参数后进行训练，训练结束会得到一个模型，例如支持向量机最终会得到 $\boldsymbol{w}$ 和 b 的具体数值（此处不考虑核函数），这就是模型参数，模型配置好相应模型参数后即可对新样本做预测。

2.2.2 验证集

带有参数的算法一般需要从候选参数配置方案中选择相对于当前数据集的最优参数配置方案，例如支持向量机的参数 C，一般采用的是前面讲到的交叉验证法，但是交叉验证法操作起来较为复杂，实际中更多采用的是：先用留出法将数据集划分出训练集和测试集，然后再对训练集采用留出法划分出训练集和新的测试集，称新的测试集为验证集，接着基于验证集的测试结果来调参选出最优参数配置方案，最后将验证集合并进训练集（训练集数据量够的话也可不合并），用选出的最优参数配置在合并后的训练集上重新训练，再用测试集来评估训练得到的模型的性能。

2.3 性能度量

本节性能度量指标较多，但是一般常用的只有错误率、精度、查准率、查全率、F1、ROC 和 AUC。

2.3.1 式 (2.2) 到式 (2.7) 的解释

这几个公式简单易懂，几乎不需要额外解释，但是需要补充说明的是式 (2.2)、式 (2.4) 和式 (2.5) 假设了数据分布为均匀分布，即每个样本出现的概率相同，而式 (2.3)、式 (2.6) 和式 (2.7) 则为更一般的表达式。此外，在无特别说明的情况下，2.3 节所有公式中的“样例集 D”均默认为非训练集（测试集、验证集或其他未用于训练的样例集）。

2.3.2 式 (2.8) 和式 (2.9) 的解释

查准率 P：被学习器预测为**正例**的样例中有多大比例是**真正例**。

查全率 R：所有**正例**当中有多大比例被学习器预测为**正例**。

2.3.3 图 2.3 的解释

P-R 曲线的画法与 ROC 曲线的画法类似，也是通过依次改变模型阈值，然后计算出查准率和查全率并画出相应坐标点，具体参见“式 (2.20) 的推导”部分的讲解。这里需要说明的是，“西瓜书”中的图 2.3 仅仅是示意图，除了图左侧提到的“现实任务中的 P-R 曲线常是非单调、不平滑的，在很多局部有上

下波动”以外，通常也不会取到 $(1,0)$ 点。因为当取到 $(1,0)$ 点时，此时是将所有样本均判为正例，因此 $FN=0$，根据式 (2.8) 可算得查全率为 1，但是此时 $TP+FP$ 为样本总数，根据式 (2.9) 可算得查准率此时为正例在全体样本中的占比，显然在现实任务中正例的占比通常不为 0，因此 P-R 曲线在现实任务中通常不会取到 $(1,0)$ 点。

2.3.4 式 (2.10) 的推导

将式 (2.8) 和式 (2.9) 代入式 (2.10)，得

$$\begin{aligned}
F1 &= \frac{2\times P\times R}{P+R} = \frac{2\times \frac{TP}{TP+FP}\times\frac{TP}{TP+FN}}{\frac{TP}{TP+FP}+\frac{TP}{TP+FN}} \\
&= \frac{2\times TP\times TP}{TP(TP+FN)+TP(TP+FP)} \\
&= \frac{2\times TP}{(TP+FN)+(TP+FP)} \\
&= \frac{2\times TP}{(TP+FN+FP+TN)+TP-TN} \\
&= \frac{2\times TP}{\text{样例总数}+TP-TN}
\end{aligned}$$

若现有数据集 $D=\{(\boldsymbol{x}_i,y_i)\mid 1\leqslant i\leqslant m\}$，其中标记 $y_i\in\{0,1\}$（1 表示正例，0 表示反例），假设模型 $f(\boldsymbol{x})$ 对 $\boldsymbol{x}_i$ 的预测结果为 $h_i\in\{0,1\}$，则模型 $f(\boldsymbol{x})$ 在数据集 D 上的 $F1$ 为

$$F1=\frac{2\sum_{i=1}^{m}y_ih_i}{\sum_{i=1}^{m}y_i+\sum_{i=1}^{m}h_i}$$

不难看出上式的本质为

$$F1=\frac{2\times TP}{(TP+FN)+(TP+FP)}$$

2.3.5 式 (2.11) 的解释

“西瓜书”在式 (2.11) 左侧提到 F_β 本质是加权调和平均，且和常用的算数平均相比，其更重视较小值，在此举例说明。例如 a 同学有两门课的成绩分别为 100 分和 60 分，b 同学相应的成绩为 80 分和 80 分，此时若计算 a 同学和 b 同学的算数平均分则均为 80 分，无法判断两位同学成绩的优劣，但是若计算加权调和平均，当 $\beta=1$ 时，a 同学的加权调和平均为 $\frac{2\times100\times60}{100+60}=75$，b 同学的加权调和平均为 $\frac{2\times80\times80}{80+80}=80$，此时 b 同学的平均成绩更优，原因是 a 同学由于偏科导致其中一门成绩过低，而调和平均更重视较小值，所以 a 同学的偏科便被凸显出来。

式 (2.11) 下方有提到“$\beta>1$ 时查全率有更大影响；$\beta<1$ 时查准率有更大影响”，下面解释其原因。将式 (2.11) 恒等变形为如下形式

$$F_\beta=\frac{1}{\frac{1}{1+\beta^2}\cdot\frac{1}{P}+\frac{\beta^2}{1+\beta^2}\cdot\frac{1}{R}}$$

从上式可以看出，当 $\beta>1$ 时 $\frac{\beta^2}{1+\beta^2}>\frac{1}{1+\beta^2}$，所以 $\frac{1}{R}$ 的权重比 $\frac{1}{P}$ 的权重高，因此查全率 R 对 F_β 的影响更大，反之查准率 P 对 F_β 的影响更大。

2.3.6 式 (2.12) 到式 (2.17) 的解释

式 (2.12) 的 macro-P 和式 (2.13) 的 macro-R 是基于各个二分类问题的 P 和 R 计算而得的；式 (2.15) 的 micro-P 和式 (2.16) 的 micro-R 是基于各个二分类问题的 TP、FP、TN、FN 计算而得的；“宏”可以认为是只关注宏观而不看具体细节，而“微”可以认为是要从具体细节做起，因为相比于 P 和 R 指标来说，TP、FP、TN、FN 更微观，毕竟 P 和 R 是基于 TP、FP、TN、FN 计算而得。

从“宏”和“微”的计算方式可以看出，“宏”没有考虑每个类别下的的样本数量，所以平等看待每个类别，因此会受到高 P 和高 R 类别的影响，而“微”则考虑到了每个类别的样本数量，因为样本数量多的类相应的 TP、FP、TN、FN 也会占比更多，所以在各类别样本数量极度不平衡的情况下，数量较多的类别会主导最终结果。

式 (2.14) 的 macro-$F1$ 是将 macro-P 和 macro-R 代入式 (2.10) 所得；式 (2.17) 的 macro-$F1$ 是将 macro-P 和 macro-R 代入式 (2.10) 所得。值得一提的是，以上只是 macro-$F1$ 和 micro-$F1$ 的常用计算方式之一，如若在查阅资料的过程中看到其他的计算方式也属正常。

2.3.7 式 (2.18) 和式 (2.19) 的解释

式 (2.18) 定义了真正例率 TPR。先解释公式中出现的真正例和假反例，真正例即实际为正例预测结果也为正例，假反例即实际为正例但预测结果为反例，式 (2.18) 分子为真正例，分母为真正例和假反例之和（即实际的正例个数），因此式 (2.18) 的含义是所有**正例**当中有多大比例被预测为**正例**（即查全率 Recall）。

式 (2.19) 定义了假正例率 FPR。先解释式子中出现的假正例和真反例，假正例即实际为反例但预测结果为正例，真反例即实际为反例预测结果也为反例，式 (2.19) 分子为假正例，分母为真反例和假正例之和（即实际的反例个数），因此式 (2.19) 的含义是所有**反例**当中有多大比例被预测为**正例**。

除了真正例率 TPR 和假正例率 FPR，还有真反例率 TNR 和假反例率 FNR：

$$\text{TNR} = \frac{TN}{FP + TN}$$
$$\text{FNR} = \frac{FN}{TP + FN}$$

2.3.8 式 (2.20) 的推导

在推导式 (2.20) 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（此处用 $\boldsymbol{s}$ 表示样本，以和坐标 (x, y) 作出区分）：

$$(\boldsymbol{s}_1, 0.77, +), (\boldsymbol{s}_2, 0.62, -), (\boldsymbol{s}_3, 0.58, +), (\boldsymbol{s}_4, 0.47, +),$$
$$(\boldsymbol{s}_5, 0.47, -), (\boldsymbol{s}_6, 0.33, -), (\boldsymbol{s}_7, 0.23, +), (\boldsymbol{s}_8, 0.15, -)$$

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。

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（上面给出的预测结果已经按照预测值从大到小排序），接着将分类阈值设为一个不可能取到的超大值，例如设为 1。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。

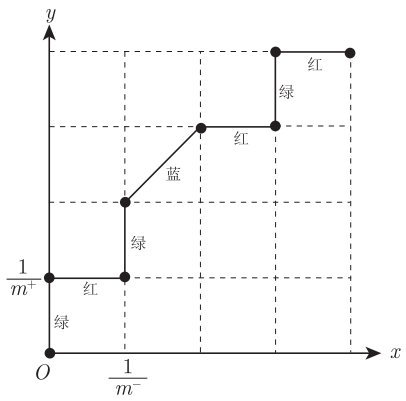

图 2-1 ROC 曲线示意

在这里，为了能在解释式 (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

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

此即为 AUC。

通过以上 ROC 曲线的绘制流程可以看出，ROC 曲线上每一个点都表示学习器 $f(s)$ 在特定阈值下构成的一个二分类器，越好的二分类器其假正例率（反例被预测错误的概率，横轴）越小，真正例率（正例被预测正确的概率，纵轴）越大，所以这个点越靠左上角（即点 $(0,1)$）越好。因此，越好的学习器，其 ROC 曲线上的点越靠左上角，相应的 ROC 曲线下的面积也越大，即 AUC 也越大。

2.3.9 式 (2.21) 和式 (2.22) 的推导

下面针对“西瓜书”上所说的“ℓ_{rank} 对应的是 ROC 曲线之上的面积”进行推导。按照我们上述对式 (2.20) 的推导思路，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：

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

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：

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

与式 (2.20) 中的推导思路相同，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

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

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

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

到此，推导完毕。

若不考虑 $f(\boldsymbol{x}^+)=f(\boldsymbol{x}^-)$，从直观上理解 ℓ_{rank}，其表示的是：对于待测试的模型 $f(\boldsymbol{x})$，从测试集中随机抽取一个正反例对儿 $\{\boldsymbol{x}^+,\boldsymbol{x}^-\}$，模型 $f(\boldsymbol{x})$ 对正例的打分 $f(\boldsymbol{x}^+)$ 小于对反例的打分 $f(\boldsymbol{x}^-)$ 的概率，即“排序错误”的概率。推导思路如下：采用频率近似概率的思路，组合出测试集中的所有正反例对儿，假设组合出来的正反例对儿的个数为 m，用模型 $f(\boldsymbol{x})$ 对所有正反例对儿打分并统计“排序错误”的正反例对儿个数 n，然后计算出 $\frac{n}{m}$ 即为模型 $f(\boldsymbol{x})$“排序错误”的正反例对儿的占比，其可近似看作为 $f(\boldsymbol{x})$ 在测试集上“排序错误”的概率。具体推导过程如下：测试集中的所有正反例对儿的个数为

$$m^+\times m^-$$

“排序错误”的正反例对儿个数为

$$\sum_{\boldsymbol{x}^+\in D^+}\sum_{\boldsymbol{x}^-\in D^-}\big(\mathbb{I}\big(f(\boldsymbol{x}^+)<f(\boldsymbol{x}^-)\big)\big)$$

因此，“排序错误”的概率为

$$\frac{\sum_{\boldsymbol{x}^+\in D^+}\sum_{\boldsymbol{x}^-\in D^-}\big(\mathbb{I}\big(f(\boldsymbol{x}^+)<f(\boldsymbol{x}^-)\big)\big)}{m^+\times m^-}$$

若再考虑 $f(\boldsymbol{x}^+)=f(\boldsymbol{x}^-)$ 时算半个“排序错误”，则上式可进一步扩展为

$$\frac{\sum_{\boldsymbol{x}^+\in D^+}\sum_{\boldsymbol{x}^-\in D^-}\big(\mathbb{I}\big(f(\boldsymbol{x}^+)<f(\boldsymbol{x}^-)+\frac{1}{2}\mathbb{I}\big(f(\boldsymbol{x}^+)=f(\boldsymbol{x}^-)\big)\big)\big)}{m^+\times m^-}$$

此即为 ℓ_{rank}。

如果说 ℓ_{rank} 指的是从测试集中随机抽取正反例对儿，模型 $f(\boldsymbol{x})$“排序错误”的概率，那么根据式 (2.22) 可知，AUC 则指的是从测试集中随机抽取正反例对儿，模型 $f(\boldsymbol{x})$“排序正确”的概率。显然，此概率越大越好。

2.3.10 式 (2.23) 的解释

本公式很容易理解，只是需要注意该公式上方交代了“若将表 2.2 中的第 0 类作为正类、第 1 类作为反类”，若不注意此条件，按习惯（0 为反类、1 为正类）会产生误解。为避免产生误解，在接下来的解释

中将 $cost_{01}$ 记为 $cost_{+-}$，$cost_{10}$ 记为 $cost_{-+}$。本公式还可以作如下恒等变形

$$\begin{aligned}E(f;D;cost) &= \frac{1}{m}\left(m^+ \times \frac{1}{m^+}\sum_{\boldsymbol{x}_i\in D^+}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i))\times cost_{+-} + m^- \times \frac{1}{m^-}\sum_{\boldsymbol{x}_i\in D^-}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i))\times cost_{-+}\right)\\ &= \frac{m^+}{m}\times\frac{1}{m^+}\sum_{\boldsymbol{x}_i\in D^+}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i))\times cost_{+-} + \frac{m^-}{m}\times\frac{1}{m^-}\sum_{\boldsymbol{x}_i\in D^-}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i))\times cost_{-+}\end{aligned}$$

其中 m^+ 和 m^- 分别表示正例子集 D^+ 和反例子集 D^- 的样本个数。

$\frac{1}{m^+}\sum_{\boldsymbol{x}_i\in D^+}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i))$ 表示正例子集 D^+ 预测错误样本所占比例，即假反例率 FNR。

$\frac{1}{m^-}\sum_{\boldsymbol{x}_i\in D^-}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i))$ 表示反例子集 D^- 预测错误样本所占比例，即假反例率 FPR。

$\frac{m^+}{m}$ 表示样例集 D 中正例所占比例，或理解为随机从 D 中取一个样例取到正例的概率。

$\frac{m^-}{m}$ 表示样例集 D 中反例所占比例，或理解为随机从 D 中取一个样例取到反例的概率。

因此，若将样例为正例的概率 $\frac{m^+}{m}$ 记为 p，则样例为 f 反例的概率 $\frac{m^-}{m}$ 为 $1-p$，上式可进一步写为

$$E(f;D;cost) = p\times \text{FNR}\times cost_{+-} + (1-p)\times \text{FPR}\times cost_{-+}$$

此公式在接下来式 (2.25) 的解释中会用到。

2.3.11 式 (2.24) 的解释

当 $cost_{+-} = cost_{-+}$ 时，本公式可化简为

$$P(+)cost = \frac{p}{p+(1-p)} = p$$

其中 p 是样例为正例的概率（一般用正例在样例集中所占的比例近似代替）。因此，当代价不敏感时（也即 $cost_{+-} = cost_{-+}$），$P(+)cost$ 就是正例在样例集中的占比。那么，当代价敏感时（也即 $cost_{+-} \neq cost_{-+}$），$P(+)cost$ 即为正例在样例集中的加权占比。具体来说，对于样例集

$$D = \left\{\boldsymbol{x}_1^+, \boldsymbol{x}_2^+, \boldsymbol{x}_3^-, \boldsymbol{x}_4^-, \boldsymbol{x}_5^-, \boldsymbol{x}_6^-, \boldsymbol{x}_7^-, \boldsymbol{x}_8^-, \boldsymbol{x}_9^-, \boldsymbol{x}_{10}^-\right\}$$

其中 $\boldsymbol{x}^+$ 表示正例，$\boldsymbol{x}^-$ 表示反例。可以看出 $p=0.2$，若想让正例得到更多重视，考虑代价敏感 $cost_{+-}=4$ 和 $cost_{-+}=1$，这实际等价于在以下样例集上进行代价不敏感的正例概率代价计算

$$D' = \left\{\boldsymbol{x}_1^+, \boldsymbol{x}_1^+, \boldsymbol{x}_1^+, \boldsymbol{x}_1^+, \boldsymbol{x}_2^+, \boldsymbol{x}_2^+, \boldsymbol{x}_2^+, \boldsymbol{x}_2^+, \boldsymbol{x}_3^-, \boldsymbol{x}_4^-, \boldsymbol{x}_5^-, \boldsymbol{x}_6^-, \boldsymbol{x}_7^-, \boldsymbol{x}_8^-, \boldsymbol{x}_9^-, \boldsymbol{x}_{10}^-\right\}$$

即将每个正例样本复制 4 份，若有 1 个出错，则有 4 个一起出错，代价为 4。此时可计算出

$$\begin{aligned}P(+)cost &= \frac{p\times cost_{+-}}{p\times cost_{+-} + (1-p)\times cost_{-+}}\\ &= \frac{0.2\times 4}{0.2\times 4 + (1-0.2)\times 1} = 0.5\end{aligned}$$

也就是正例在等价的样例集 D' 中的占比。所以，无论代价敏感还是不敏感，$P(+)cost$ 本质上表示的都是样例集中正例的占比。在实际应用过程中，如果由于某种原因无法将 $cost_{+-}$ 和 $cost_{-+}$ 设为不同取值，可以采用上述“复制样本”的方法间接实现将 $cost_{+-}$ 和 $cost_{-+}$ 设为不同取值。

对于不同的 $cost_{+-}$ 和 $cost_{-+}$ 取值，若二者的比值保持相同，则 $P(+)cost$ 不变。例如，对于上面的例子，若设 $cost_{+-}=40$ 和 $cost_{-+}=10$，所得 $P(+)cost$ 仍为 0.5。

此外，根据此式还可以相应地推导出反例概率代价

$$P(-)cost = 1 - P(+)cost = \frac{(1-p)\times cost_{-+}}{p\times cost_{+-} + (1-p)\times cost_{-+}}$$

2.3.12 式 (2.25) 的解释

对于包含 m 个样本的样例集 D，可以算出学习器 $f(\boldsymbol{x})$ 总的代价是

$$\begin{aligned} cost_{se} = &m \times p \times \text{FNR} \times cost_{+-} + m \times (1-p) \times \text{FPR} \times cost_{-+} \\ &+ m \times p \times \text{TPR} \times cost_{++} + m \times (1-p) \times \text{TNR} \times cost_{--} \end{aligned}$$

其中 p 是正例在样例集中所占的比例（或严格地称为样例为正例的概率），$cost_{se}$ 下标中的“se”表示 sensitive，即代价敏感，根据前面讲述的 FNR、FPR、TPR、TNR 的定义可知：

$m \times p \times \text{FNR}$ 表示正例被预测为反例（正例预测错误）的样本个数；

$m \times (1-p) \times \text{FPR}$ 表示反例被预测为正例（反例预测错误）的样本个数；

$m \times p \times \text{TPR}$ 表示正例被预测为正例（正例预测正确）的样本个数；

$m \times (1-p) \times \text{TNR}$ 表示反例预测为反例（反例预测正确）的样本个数。

以上各种样本个数乘以相应的代价则得到总的代价 $cost_{se}$。但是，按照此公式计算出的代价与样本个数 m 呈正比，显然不具有一般性，因此需要除以样本个数 m，而且一般来说，预测出错才会产生代价，预测正确则没有代价，也即 $cost_{++} = cost_{--} = 0$，所以 $cost_{se}$ 更为一般化的表达式为

$$cost_{se} = p \times \text{FNR} \times cost_{+-} + (1-p) \times \text{FPR} \times cost_{-+}$$

回顾式 (2.23) 的解释可知，此式即为式 (2.23) 的恒等变形，所以此式可以同式 (2.23) 一样理解为学习器 $f(\boldsymbol{x})$ 在样例集 D 上的“代价敏感错误率”。显然，$cost_{se}$ 的取值范围并不在 0 到 1 之间，且 $cost_{se}$ 在 $\text{FNR} = \text{FPR} = 1$ 时取到最大值，因为 $\text{FNR} = \text{FPR} = 1$ 时表示所有正例均被预测为反例，反例均被预测为正例，代价达到最大，即

$$max(cost_{se}) = p \times cost_{+-} + (1-p) \times cost_{-+}$$

所以，如果要将 $cost_{se}$ 的取值范围归一化到 0 到 1 之间，则只需将其除以其所能取到的最大值即可，也即

$$\frac{cost_{se}}{max(cost_{se})} = \frac{p \times \text{FNR} \times cost_{+-} + (1-p) \times \text{FPR} \times cost_{-+}}{p \times cost_{+-} + (1-p) \times cost_{-+}}$$

此即为式 (2.25)，也即为 $cost_{norm}$，其中下标“norm”表示 normalization。

进一步地，根据式 (2.24) 中 $P(+)cost$ 的定义可知，式 (2.25) 可以恒等变形为

$$cost_{norm} = \text{FNR} \times P(+)cost + \text{FPR} \times (1 - P(+)cost)$$

对于二维直角坐标系中的两个点 $(0, B)$ 和 $(1, A)$ 以及实数 $p \in [0, 1]$，$(p, pA + (1-p)B)$ 一定是线段 $A - B$ 上的点，且当 p 从 0 变到 1 时，点 $(p, pA + (1-p)B)$ 的轨迹为从 $(0, B)$ 到 $(1, A)$，基于此，结合上述 $cost_{norm}$ 的表达式可知：$(P(+)cost, cost_{norm})$ 即为线段 $\text{FPR} - \text{FNR}$ 上的点，当 $(P(+)cost$ 从 0 变到 1 时，$(P(+)cost, cost_{norm})$ 的轨迹为从 $(0, \text{FPR})$ 到 $(1, \text{FNR})$，也即图 2.5 中的各条线段。需要注意的是，以上只是从数学逻辑自洽的角度对图 2.5 中的各条线段进行解释，实际中各条线段并非按照上述方法绘制。理由如下：

$P(+)cost$ 表示的是样例集中正例的占比，而在进行学习器的比较时，变动的只是训练学习器的算法或者算法的超参数，用来评估学习器性能的样例集是固定的（单一变量原则），所以 $P(+)cost$ 是一个固定值，因此图 2.5 中的各条线段并不是通过变动 $P(+)cost$ 然后计算 $cost_{norm}$ 画出来的，而是按照“西瓜书”上式 (2.25) 下方所说对 ROC 曲线上每一点计算 FPR 和 FNR，然后将点 $(0, \text{FPR})$ 和点 $(1, \text{FNR})$ 直接连成线段。

虽然图 2.5 中的各条线段并不是通过变动横轴表示的 $P(+)cost$ 来进行绘制，但是横轴仍然有其他用处，例如用来找使学习器的归一化代价 $cost_{norm}$ 达到最小的阈值（暂且称其为最佳阈值）。具体地，首先计算当前样例集的 $P(+)cost$ 值，然后根据计算出来的值在横轴上标记出具体的点，再基于该点作一条垂

直于横轴的垂线，与该垂线最先相交（从下往上看）的线段所对应的阈值（因为每条线段都对应 ROC 曲线上的点，ROC 曲线上的点又对应着具体的阈值）即为最佳阈值。原因是与该垂线最先相交的线段必然最靠下，因此其交点的纵坐标最小，而纵轴表示的便是归一化代价 $cost_{norm}$，所以此时归一化代价 $cost_{norm}$ 达到最小。特别地，当 $P(+)cost = 0$ 时，即样例集中没有正例，全是负例，因此最佳阈值应该是学习器不可能取到的最大值，且按照此阈值计算出来出来的 $\text{FPR} = 0, \text{FNR} = 1, cost_{norm} = 0$。那么按照上述作垂线的方法去图 2.5 中进行实验，也即在横轴 0 刻度处作垂线，显然与该垂线最先相交的线段是点 $(0, 0)$ 和点 $(1, 1)$ 连成的线段，交点为 $(0, 0)$，此时对应的也为 $\text{FPR} = 0, \text{FNR} = 1, cost_{norm} = 0$，且该条线段所对应的阈值也确实为“学习器不可能取到的最大值”（因为该线段对应的是 ROC 曲线中的起始点）。

2.4 比较检验

为什么要做比较检验？“西瓜书”在本节开篇的两段话已经交代原由。简单来说，从统计学的角度，取得的性能度量的值本质上仍是一个随机变量，因此并不能简单用比较大小来直接判定算法（或者模型）之间的优劣，而需要更置信的方法来进行判定。

在此说明一下，如果不做算法理论研究，也不需要对算法（或模型）之间的优劣给出严谨的数学分析，本节可以暂时跳过。本节主要使用的数学知识是“统计假设检验”，该知识点在各个高校的概率论与数理统计教材（例如参考文献 [1]）上均有讲解。此外，有关检验变量的公式，例如式 (2.30) 至式 (2.36)，并不需要清楚是怎么来的（这是统计学家要做的事情），只需要会用即可。

2.4.1 式 (2.26) 的解释

理解本公式时需要明确的是：ϵ 是未知的，是当前希望估算出来的，$\hat{\epsilon}$ 是已知的，是已经用 m 个测试样本对学习器进行测试得到的。因此，本公式也可理解为：当学习器的泛化错误率为 ϵ 时，被测得测试错误率为 $\hat{\epsilon}$ 的条件概率。所以本公式可以改写为

$$P(\hat{\epsilon}|\epsilon) = \begin{pmatrix} m \\ \hat{\epsilon} \times m \end{pmatrix} \epsilon^{\hat{\epsilon} \times m} (1-\epsilon)^{m - \hat{\epsilon} \times m}$$

其中

$$\begin{pmatrix} m \\ \hat{\epsilon} \times m \end{pmatrix} = \frac{m!}{(\hat{\epsilon} \times m)!(m - \hat{\epsilon} \times m)!}$$

为中学时学的组合数，即 $C_m^{\hat{\epsilon} \times m}$。

在已知 $\hat{\varepsilon}$ 时，求使得条件概率 $P(\hat{\epsilon}|\epsilon)$ 达到最大的 ϵ 是概率论与数理统计中经典的极大似然估计问题。从极大似然估计的角度可知，由于 $\hat{\epsilon}, m$ 均为已知量，所以 $P(\hat{\epsilon}|\epsilon)$ 可以看作为一个关于 ϵ 的函数，称为似然函数，于是问题转化为求使得似然函数取到最大值的 ϵ，即

$$\epsilon = \arg\max_{\epsilon} P(\hat{\epsilon}|\epsilon)$$

首先对 ϵ 求一阶导数

$$\begin{aligned}
\frac{\partial P(\hat{\epsilon} \mid \epsilon)}{\partial \epsilon} &= \begin{pmatrix} m \\ \hat{\epsilon} \times m \end{pmatrix} \frac{\partial \epsilon^{\hat{\epsilon} \times m} (1-\epsilon)^{m - \hat{\epsilon} \times m}}{\partial \epsilon} \\
&= \begin{pmatrix} m \\ \hat{\epsilon} \times m \end{pmatrix} \left(\hat{\epsilon} \times m \times \epsilon^{\hat{\epsilon} \times m - 1} (1-\epsilon)^{m - \hat{\epsilon} \times m} + \epsilon^{\hat{\epsilon} \times m} \times (m - \hat{\epsilon} \times m) \times (1-\epsilon)^{m - \hat{\epsilon} \times m - 1} \times (-1) \right) \\
&= \begin{pmatrix} m \\ \hat{\epsilon} \times m \end{pmatrix} \epsilon^{\hat{\epsilon} \times m - 1} (1-\epsilon)^{m - \hat{\epsilon} \times m - 1} (\hat{\epsilon} \times m \times (1-\epsilon) - \epsilon \times (m - \hat{\epsilon} \times m)) \\
&= \begin{pmatrix} m \\ \hat{\epsilon} \times m \end{pmatrix} \epsilon^{\hat{\epsilon} \times m - 1} (1-\epsilon)^{m - \hat{\epsilon} \times m - 1} (\hat{\epsilon} \times m - \epsilon \times m)
\end{aligned}$$

分析上式可知，其中 $\begin{pmatrix} m \\ \hat{\epsilon} \times m \end{pmatrix}$ 为常数，由于 $\epsilon \in [0,1]$，所以 $\epsilon^{\hat{\epsilon}\times m-1}(1-\epsilon)^{m-\hat{\epsilon}\times m-1}$ 恒大于 0，$(\hat{\epsilon}\times m-\epsilon\times m)$ 在 $0 \leqslant \epsilon < \hat{\epsilon}$ 时大于 0，在 $\epsilon = \hat{\epsilon}$ 时等于 0，在 $\hat{\epsilon} \leqslant \epsilon < 1$ 时小于 0，因此 $P(\hat{\epsilon} \mid \epsilon)$ 是关于 ϵ 开口向下的凹函数（此处采用的是最优化中对凹凸函数的定义，“西瓜书”第 3 章 3.2 节左侧边注对凹凸函数的定义也是如此）。所以，当且仅当一阶导数 $\frac{\partial P(\hat{\epsilon}|\epsilon)}{\partial \epsilon} = 0$ 时 $P(\hat{\epsilon} \mid \epsilon)$ 取到最大值，此时 $\epsilon = \hat{\epsilon}$。

2.4.2 式 (2.27) 的推导

截至 2021 年 5 月，“西瓜书”第 1 版第 36 次印刷，式 (2.27) 应当勘误为

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

在推导此公式之前，先铺垫讲解一下“二项分布参数 p 的假设检验”[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

$$\varphi : \text{当} X \leqslant C \text{时接受} H_0, \text{否则就拒绝} H_0$$

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

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

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，$\beta_\varphi(p_0)$ 即为 $\beta_\varphi(p)$ 的上确界。（更为严格的数学证明参见参考文献 [1] 中第 2 章习题 7）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：

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

显然，当 $p \leqslant p_0$ 时有

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

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 $\overline{C}$ 使得

$$\sum_{i=\overline{C}+1}^{m} \begin{pmatrix} m \\ i \end{pmatrix} p_0^i (1-p_0)^{m-i} < \alpha$$

$$\sum_{i=\overline{C}}^{m} \begin{pmatrix} m \\ i \end{pmatrix} p_0^i (1-p_0)^{m-i} > \alpha$$

此时，C 只能取 $\overline{C}$ 或者 $\overline{C}+1$。若 C 取 $\overline{C}$，则相当于升高了检验水平 α；若 C 取 $\overline{C}+1$ 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 $\overline{C}+1$。

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

$$\overline{C}=\min C \quad \text{s.t.} \sum_{i=C+1}^{m}\left(\begin{array}{c} m \\ i \end{array}\right) p_0^i(1-p_0)^{m-i}<\alpha$$

由“西瓜书”中的上下文可知，对 $\epsilon \leqslant \epsilon_0$ 进行假设检验，等价于“二项分布参数 p 的假设检验”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 $\bar{\epsilon}$ 等价于在“二项分布参数 p 的假设检验”中求解事件最大发生频率 $\frac{\overline{C}}{m}$。由上述“二项分布参数 p 的假设检验”中的推导可知

$$\overline{C}=\min C \quad \text{s.t.} \sum_{i=C+1}^{m}\left(\begin{array}{c} m \\ i \end{array}\right) p_0^i(1-p_0)^{m-i}<\alpha$$

所以

$$\frac{\overline{C}}{m}=\min \frac{C}{m} \quad \text{s.t.} \sum_{i=C+1}^{m}\left(\begin{array}{c} m \\ i \end{array}\right) p_0^i(1-p_0)^{m-i}<\alpha$$

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

$$\bar{\epsilon}=\min \epsilon \quad \text{s.t.} \sum_{i=\epsilon \times m+1}^{m}\left(\begin{array}{c} m \\ i \end{array}\right) \epsilon_0^i(1-\epsilon_0)^{m-i}<\alpha$$

2.5 偏差与方差

2.5.1 式 (2.37) 到式 (2.42) 的推导

首先，梳理一下“西瓜书”中的符号，书中称 $\boldsymbol{x}$ 为测试样本，但是书中又提到“令 y_D 为 $\boldsymbol{x}$ 在数据集中的标记”，那么 $\boldsymbol{x}$ 究竟是测试集中的样本还是训练集中的样本呢？这里暂且理解为 $\boldsymbol{x}$ 为从训练集中抽取出来用于测试的样本。此外，“西瓜书”中左侧边注中提到“有可能出现噪声使得 $y_D \neq y$”，其中所说的“噪声”通常是指人工标注数据时带来的误差，例如标注“身高”时，由于测量工具的精度等问题，测出来的数值必然与真实的“身高”之间存在一定误差，此即为“噪声”。

为了进一步解释式 (2.37)、(2.38) 和 (2.39)，在这里设有 n 个训练集 $D_1, ..., D_n$，这 n 个训练集都是以独立同分布的方式从样本空间中采样而得，并且恰好都包含测试样本 $\boldsymbol{x}$，该样本在这 n 个训练集的标记分别为 $y_{D_1}, ..., y_{D_n}$。书中已明确，此处以回归任务为例，也即 $y_D, y, f(\boldsymbol{x}; D)$ 均为实值。

式 (2.37) 可理解为：

$$\bar{f}(\boldsymbol{x})=\mathbb{E}_D[f(\boldsymbol{x}; D)]=\frac{1}{n}\left(f\left(\boldsymbol{x}; D_1\right)+\ldots+f\left(\boldsymbol{x}; D_n\right)\right)$$

式 (2.38) 可理解为：

$$\begin{aligned}\operatorname{var}(\boldsymbol{x}) &= \mathbb{E}_D\left[(f(\boldsymbol{x}; D)-\bar{f}(\boldsymbol{x}))^2\right] \\ &= \frac{1}{n}\left(\left(f\left(\boldsymbol{x}; D_1\right)-\bar{f}(\boldsymbol{x})\right)^2+\ldots+\left(f\left(\boldsymbol{x}; D_n\right)-\bar{f}(\boldsymbol{x})\right)^2\right)\end{aligned}$$

式 (2.39) 可理解为：

$$\varepsilon^2=\mathbb{E}_D\left[\left(y_D-y\right)^2\right]=\frac{1}{n}\left(\left(y_{D_1}-y\right)^2+\ldots+\left(y_{D_n}-y\right)^2\right)$$

最后，推导一下式 (2.41) 和式 (2.42)，由于推导完式 (2.41) 自然就会得到式 (2.42)，因此下面仅推导式 (2.41) 即可。

$$
\begin{aligned}
E(f;D) =& \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-y_D\right)^2\right] && ①\\
=& \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})+\bar{f}(\boldsymbol{x})-y_D\right)^2\right] && ②\\
=& \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y_D\right)^2\right]+\\
& \mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\left(\bar{f}(\boldsymbol{x})-y_D\right)\right] && ③\\
=& \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y_D\right)^2\right] && ④\\
=& \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y+y-y_D\right)^2\right] && ⑤\\
=& \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y\right)^2\right]+\mathbb{E}_D\left[\left(y-y_D\right)^2\right]+\\
& 2\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x})-y\right)\left(y-y_D\right)\right] && ⑥\\
=& \mathbb{E}_D\left[\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)^2\right]+\left(\bar{f}(\boldsymbol{x})-y\right)^2+\mathbb{E}_D\left[\left(y_D-y\right)^2\right] && ⑦
\end{aligned}
$$

上式即为式 (2.41)，下面给出每一步的推导过程：

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。

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

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

③ → ④：再次利用期望的运算性质将 ③ 的最后一项展开，有

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

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 2 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

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

所以

$$\begin{aligned}\mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\left(\bar{f}(\boldsymbol{x})-y_D\right)\right] &= \mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\cdot\bar{f}(\boldsymbol{x})\right]-\mathbb{E}_D\left[2\left(f(\boldsymbol{x};D)-\bar{f}(\boldsymbol{x})\right)\cdot y_D\right]\\ &= 0+0\\ &= 0\end{aligned}$$

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。

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

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

同理，⑥ 中的最后一项可化为

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

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

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

参考文献

[1] 陈希孺. 概率论与数理统计. 中国科学技术大学出版社, 2009.

第 3 章 线性模型

如“西瓜书”前言所述，本章仍属于第 1 部分机器学习基础知识。作为“西瓜书”介绍机器学习模型的开篇，线性模型也是机器学习中最为基础的模型，很多复杂模型均可认为由线性模型衍生而得。

3.1 基本形式

第 1 章的 1.2 基本术语中讲述样本的定义时，我们说明了“西瓜书”和本书中向量的写法，当向量中的元素用分号“;”分隔时表示此向量为列向量，用逗号“,”分隔时表示为行向量。因此，式 (3.2) 中 $\boldsymbol{w}=(w_1;w_2;...;w_d)$ 和 $\boldsymbol{x}=(x_1;x_2;...;x_d)$ 均为 d 行 1 列的列向量。

3.2 线性回归

3.2.1 属性数值化

为了能进行数学运算，样本中的非数值类属性都需要进行数值化。对于存在“序”关系的属性，可通过连续化将其转化为带有相对大小关系的连续值；对于不存在“序”关系的属性，可根据属性取值将其拆解为多个属性，例如“西瓜书”中所说的“瓜类”属性，可将其拆解为“是否是西瓜”、“是否是南瓜”、“是否是黄瓜”3 个属性，其中每个属性的取值为 1 或 0，1 表示“是”，0 表示“否”。具体地，假如现有 3 个瓜类样本：$\boldsymbol{x}_1=$(甜度 = 高; 瓜类 = 西瓜), $\boldsymbol{x}_2=$(甜度 = 中; 瓜类 = 南瓜), $\boldsymbol{x}_3=$(甜度 = 低; 瓜类 = 黄瓜)，其中“甜度”属性存在序关系，因此可将“高”、“中”、“低”转化为 $\{1.0,0.5,0.0\}$，“瓜类”属性不存在序关系，则按照上述方法进行拆解，3 个瓜类样本数值化后的结果为：$\boldsymbol{x}_1=(1.0;1;0;0), \boldsymbol{x}_1=(0.5;0;1;0), \boldsymbol{x}_1=(0.0;0;0;1)$。

以上针对样本属性所进行的处理工作便是第 1 章 1.2 基本术语中提到的“特征工程”范畴，完成属性数值化以后通常还会进行缺失值处理、规范化、降维等一系列处理工作。由于特征工程属于算法实践过程中需要掌握的内容，待学完机器学习算法以后，再进一步学习特征工程相关知识即可，在此先不展开。

3.2.2 式 (3.4) 的解释

下面仅针对式 (3.4) 中的数学符号进行解释。首先解释一下符号“$\arg\min$”，其中“arg”是“argument”（参数）的前三个字母，“min”是“minimum”（最小值）的前三个字母，该符号表示求使目标函数达到最小值的参数取值。例如式 (3.4) 表示求出使目标函数 $\sum_{i=1}^{m}(y_i-wx_i-b)^2$ 达到最小值的参数取值 (w^*,b^*)，注意目标函数是以 (w,b) 为自变量的函数，(x_i,y_i) 均是已知常量，即训练集中的样本数据。

类似的符号还有“min”，例如将式 (3.4) 改为

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

则表示求目标函数的最小值。对比知道，“min”和“$\arg\min$”的区别在于，前者输出目标函数的最小值，而后者输出使得目标函数达到最小值时的参数取值。

若进一步修改式 (3.4) 为

$$\begin{aligned}&\min_{(w,b)}\sum_{i=1}^{m}(y_i-wx_i-b)^2\\&\text{s.t.}w>0,\\&b<0.\end{aligned}$$

则表示在 $w>0, b<0$ 范围内寻找目标函数的最小值，“s.t.”是“subject to”的简写，意思是“受约束于”，即为约束条件。

以上介绍的符号都是应用数学领域的一个分支——“最优化”中的内容，若想进一步了解可找一本最优化的教材（例如参考文献 [3]）进行系统性地学习。

3.2.3 式 (3.5) 的推导

“西瓜书” 在式 (3.5) 左侧给出的凸函数的定义是最优化中的定义，与高等数学中的定义不同，本书也默认采用此种定义。由于一元线性回归可以看作是多元线性回归中元的个数为 1 时的情形，所以此处暂不给出 $E_{(w,b)}$ 是关于 w 和 b 的凸函数的证明，在推导式 (3.11) 时一并给出，下面开始推导式 (3.5)。

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

$$\begin{aligned}
\frac{\partial E_{(w,b)}}{\partial w} &= \frac{\partial}{\partial w}\left[\sum_{i=1}^{m} (y_i - wx_i - b)^2\right] \\
&= \sum_{i=1}^{m} \frac{\partial}{\partial w}\left[(y_i - wx_i - b)^2\right] \\
&= \sum_{i=1}^{m} [2 \cdot (y_i - wx_i - b) \cdot (-x_i)] \\
&= \sum_{i=1}^{m} \left[2 \cdot \left(wx_i^2 - y_i x_i + b x_i\right)\right] \\
&= 2 \cdot \left(w \sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} y_i x_i + b \sum_{i=1}^{m} x_i\right) \\
&= 2 \left(w \sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} (y_i - b)\, x_i\right)
\end{aligned}$$

3.2.4 式 (3.6) 的推导

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

$$\begin{aligned}
\frac{\partial E_{(w,b)}}{\partial b} &= \frac{\partial}{\partial b}\left[\sum_{i=1}^{m} (y_i - wx_i - b)^2\right] \\
&= \sum_{i=1}^{m} \frac{\partial}{\partial b}\left[(y_i - wx_i - b)^2\right] \\
&= \sum_{i=1}^{m} [2 \cdot (y_i - wx_i - b) \cdot (-1)] \\
&= \sum_{i=1}^{m} [2 \cdot (b - y_i + wx_i)] \\
&= 2 \cdot \left[\sum_{i=1}^{m} b - \sum_{i=1}^{m} y_i + \sum_{i=1}^{m} wx_i\right] \\
&= 2 \left(mb - \sum_{i=1}^{m} (y_i - wx_i)\right)
\end{aligned}$$

3.2.5 式 (3.7) 的推导

推导之前先重点说明一下 “闭式解” 或称为 “解析解”。闭式解是指可以通过具体的表达式解出待解参数，例如可根据式 (3.7) 直接解得 w。机器学习算法很少有闭式解，线性回归是一个特例，接下来推导式 (3.7)。

令式 (3.5) 等于 0

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

$$w\sum_{i=1}^{m}x_i^2=\sum_{i=1}^{m}y_ix_i-\sum_{i=1}^{m}bx_i$$

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 $\frac{1}{m}\sum_{i=1}^{m}y_i=\bar{y}$，$\frac{1}{m}\sum_{i=1}^{m}x_i=\bar{x}$，则 $b=\bar{y}-w\bar{x}$，代入上式可得

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

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

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

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

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

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

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

若令 $\boldsymbol{x}=(x_1;x_2;...;x_m)$，$\boldsymbol{x}_d=(x_1-\bar{x};x_2-\bar{x};...;x_m-\bar{x})$ 为去均值后的 $\boldsymbol{x}$；$\boldsymbol{y}=(y_1;y_2;...;y_m)$，$\boldsymbol{y}_d=(y_1-\bar{y};y_2-\bar{y};...;y_m-\bar{y})$ 为去均值后的 $\boldsymbol{y}$，（$\boldsymbol{x}$、$\boldsymbol{x}_d$、$\boldsymbol{y}$、$\boldsymbol{y}_d$ 均为 m 行 1 列的列向量）代入上式可得

$$w=\frac{\boldsymbol{x}_d^{\mathrm{T}}\boldsymbol{y}_d}{\boldsymbol{x}_d^{\mathrm{T}}\boldsymbol{x}_d}$$

3.2.6 式 (3.9) 的推导

式 (3.4) 是最小二乘法运用在一元线性回归上的情形，那么对于多元线性回归来说，我们可以类似得到

$$\begin{aligned}(\boldsymbol{w}^*,b^*)&=\arg\min_{(\boldsymbol{w},b)}\sum_{i=1}^{m}\left(f\left(\boldsymbol{x}_i\right)-y_i\right)^2\\&=\arg\min_{(\boldsymbol{w},b)}\sum_{i=1}^{m}\left(y_i-f\left(\boldsymbol{x}_i\right)\right)^2\\&=\arg\min_{(\boldsymbol{w},b)}\sum_{i=1}^{m}\left(y_i-\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)\right)^2\end{aligned}$$

为便于讨论，我们令 $\hat{\boldsymbol{w}}=(\boldsymbol{w};b)=(w_1;...;w_d;b)\in\mathbb{R}^{(d+1)\times 1}$，$\hat{\boldsymbol{x}}_i=(x_{i1};...;x_{id};1)\in\mathbb{R}^{(d+1)\times 1}$，那么上式可以简化为

$$\begin{aligned}\hat{\boldsymbol{w}}^*&=\arg\min_{\hat{\boldsymbol{w}}}\sum_{i=1}^{m}\left(y_i-\hat{\boldsymbol{w}}^{\mathrm{T}}\hat{\boldsymbol{x}}_i\right)^2\\&=\arg\min_{\hat{\boldsymbol{w}}}\sum_{i=1}^{m}\left(y_i-\hat{\boldsymbol{x}}_i^{\mathrm{T}}\hat{\boldsymbol{w}}\right)^2\end{aligned}$$

根据向量内积的定义可知，上式可以写成如下向量内积的形式

$$\hat{\boldsymbol{w}}^*=\arg\min_{\hat{\boldsymbol{w}}}\begin{bmatrix}y_1-\hat{\boldsymbol{x}}_1^{\mathrm{T}}\hat{\boldsymbol{w}} & \cdots & y_m-\hat{\boldsymbol{x}}_m^{\mathrm{T}}\hat{\boldsymbol{w}}\end{bmatrix}\begin{bmatrix}y_1-\hat{\boldsymbol{x}}_1^{\mathrm{T}}\hat{\boldsymbol{w}}\\ \vdots\\ y_m-\hat{\boldsymbol{x}}_m^{\mathrm{T}}\hat{\boldsymbol{w}}\end{bmatrix}$$

其中

$$\begin{aligned}\begin{bmatrix}y_1-\hat{\boldsymbol{x}}_1^{\mathrm{T}}\hat{\boldsymbol{w}}\\ \vdots\\ y_m-\hat{\boldsymbol{x}}_m^{\mathrm{T}}\hat{\boldsymbol{w}}\end{bmatrix}&=\begin{bmatrix}y_1\\ \vdots\\ y_m\end{bmatrix}-\begin{bmatrix}\hat{\boldsymbol{x}}_1^{\mathrm{T}}\hat{\boldsymbol{w}}\\ \vdots\\ \hat{\boldsymbol{x}}_m^{\mathrm{T}}\hat{\boldsymbol{w}}\end{bmatrix}\\&=\boldsymbol{y}-\begin{bmatrix}\hat{\boldsymbol{x}}_1^{\mathrm{T}}\\ \vdots\\ \hat{\boldsymbol{x}}_m^{\mathrm{T}}\end{bmatrix}\cdot\hat{\boldsymbol{w}}\\&=\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{w}}\end{aligned}$$

所以

$$\hat{\boldsymbol{w}}^*=\arg\min_{\hat{\boldsymbol{w}}}(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{w}})$$

3.2.7 式 (3.10) 的推导

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

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

对 $\hat{\boldsymbol{w}}$ 求导可得

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

由矩阵微分公式 $\frac{\partial\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}}{\partial\boldsymbol{x}}=\frac{\partial\boldsymbol{x}^{\mathrm{T}}\boldsymbol{a}}{\partial\boldsymbol{x}}=\boldsymbol{a}$, $\frac{\partial\boldsymbol{x}^{\mathrm{T}}\mathbf{A}\boldsymbol{x}}{\partial\boldsymbol{x}}=(\mathbf{A}+\mathbf{A}^{\mathrm{T}})\boldsymbol{x}$（更多矩阵微分公式可查阅 [1]，矩阵微分原理可查阅 [2]）可得

$$\begin{aligned}\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial\hat{\boldsymbol{w}}}&=0-\mathbf{X}^{\mathrm{T}}\boldsymbol{y}-\mathbf{X}^{\mathrm{T}}\boldsymbol{y}+(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\boldsymbol{w}}\\&=2\mathbf{X}^{\mathrm{T}}(\mathbf{X}\hat{\boldsymbol{w}}-\boldsymbol{y})\end{aligned}$$

3.2.8 式 (3.11) 的推导

首先铺垫讲解接下来以及后续内容将会用到的多元函数相关基础知识[3]。

n **元实值函数**：含 n 个自变量，值域为实数域 $\mathbb{R}$ 的函数称为 n 元实值函数，记为 $f(\boldsymbol{x})$，其中 $\boldsymbol{x}=(x_1;x_2;...;x_n)$ 为 n 维向量。“西瓜书”和本书中的多元函数未加特殊说明均为实值函数。

凸集：设集合 $D\subset\mathbb{R}^n$ 为 n 维欧式空间中的子集，如果对 D 中任意的 n 维向量 $\boldsymbol{x}\in D$ 和 $\boldsymbol{y}\in D$ 与任意的 $\alpha\in[0,1]$，有

$$\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y}\in D$$

则称集合 D 是凸集。凸集的几何意义是：若两个点属于此集合，则这两点连线上的任意一点均属于此集合。常见的凸集有空集 $\varnothing$，整个 n 维欧式空间 $\mathbb{R}^n$。

凸函数：设 $D \subset \mathbb{R}^n$ 是非空凸集，f 是定义在 D 上的函数，如果对任意的 $\boldsymbol{x}^1, \boldsymbol{x}^2 \in D, \alpha \in (0,1)$，均有

$$f\left(\alpha \boldsymbol{x}^1 + (1-\alpha)\boldsymbol{x}^2\right) \leqslant \alpha f(\boldsymbol{x}^1) + (1-\alpha) f(\boldsymbol{x}^2)$$

则称 f 为 D 上的凸函数。若其中的 $\leqslant$ 改为 $<$ 也恒成立，则称 f 为 D 上的严格凸函数。

梯度：若 n 元函数 $f(\boldsymbol{x})$ 对 $\boldsymbol{x} = (x_1; x_2; ...; x_n)$ 中各分量 x_i 的偏导数 $\frac{\partial f(\boldsymbol{x})}{\partial x_i} (i = 1, 2, ..., n)$ 都存在，则称函数 $f(\boldsymbol{x})$ 在 $\boldsymbol{x}$ 处一阶可导，并称以下列向量

$$\nabla f(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial f(\boldsymbol{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix}$$

为函数 $f(\boldsymbol{x})$ 在 $\boldsymbol{x}$ 处的一阶导数或梯度，易证梯度指向的方向是函数值增大速度最快的方向。$\nabla f(\boldsymbol{x})$ 也可写成行向量形式

$$\nabla f(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}^{\mathrm{T}}} = \left[\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \frac{\partial f(\boldsymbol{x})}{\partial x_2}, \cdots, \frac{\partial f(\boldsymbol{x})}{\partial x_n}\right]$$

我们称列向量形式为“分母布局”，行向量形式为“分子布局”，由于在最优化中习惯采用分母布局，因此“西瓜书”以及本书中也采用分母布局。为了便于区分当前采用何种布局，通常在采用分母布局时偏导符号 ∂ 后接的是 $\boldsymbol{x}$，采用分子布局时后接的是 $\boldsymbol{x}^{\mathrm{T}}$。

Hessian 矩阵：若 n 元函数 $f(\boldsymbol{x})$ 对 $\boldsymbol{x} = (x_1; x_2; ...; x_n)$ 中各分量 x_i 的二阶偏导数 $\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j} (i = 1, 2, ..., n; j = 1, 2, ..., n)$ 都存在，则称函数 $f(\boldsymbol{x})$ 在 $\boldsymbol{x}$ 处二阶阶可导，并称以下矩阵

$$\nabla^2 f(\boldsymbol{x}) = \frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}} = \begin{bmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1^2} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n^2} \end{bmatrix}$$

为函数 $f(\boldsymbol{x})$ 在 $\boldsymbol{x}$ 处的二阶导数或 Hessian 矩阵。若其中的二阶偏导数均连续，则

$$\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j} = \frac{\partial^2 f(\boldsymbol{x})}{\partial x_j \partial x_i}$$

此时 Hessian 矩阵为对称矩阵。

定理 3.1：设 $D \subset \mathbb{R}^n$ 是非空开凸集，$f(\boldsymbol{x})$ 是定义在 D 上的实值函数，且 $f(\boldsymbol{x})$ 在 D 上二阶连续可微，如果 $f(\boldsymbol{x})$ 的 Hessian 矩阵 $\nabla^2 f(\boldsymbol{x})$ 在 D 上是半正定的，则 $f(\boldsymbol{x})$ 是 D 上的凸函数；如果 $\nabla^2 f(\boldsymbol{x})$ 在 D 上是正定的，则 $f(\boldsymbol{x})$ 是 D 上的严格凸函数。

定理 3.2：若 $f(\boldsymbol{x})$ 是凸函数，且 $f(\boldsymbol{x})$ 一阶连续可微，则 $\boldsymbol{x}^*$ 是全局解的充分必要条件是其梯度等于零向量，即 $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$。

式 (3.11) 的推导思路如下：首先根据定理 3.1 推导出 $E_{\hat{\boldsymbol{w}}}$ 是 $\hat{\boldsymbol{w}}$ 的凸函数，接着根据定理 3.2 推导出式 (3.11)。下面按照此思路进行推导。

由于式 (3.10) 已推导出 $E_{\hat{\boldsymbol{w}}}$ 关于 $\hat{\boldsymbol{w}}$ 的一阶导数，接着基于此进一步推导出二阶导数，即 Hessian 矩阵。推导过程如下：

$$\begin{aligned} \nabla^2 E_{\hat{\boldsymbol{w}}} &= \frac{\partial}{\partial \hat{\boldsymbol{w}}^{\mathrm{T}}} \left(\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}}\right) \\ &= \frac{\partial}{\partial \hat{\boldsymbol{w}}^{\mathrm{T}}} \left[2\mathbf{X}^{\mathrm{T}}(\mathbf{X}\hat{\boldsymbol{w}} - \boldsymbol{y})\right] \\ &= \frac{\partial}{\partial \hat{\boldsymbol{w}}^{\mathrm{T}}} \left(2\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{w}} - 2\mathbf{X}^{\mathrm{T}}\boldsymbol{y}\right) \end{aligned}$$

由矩阵微分公式 $\frac{\partial \mathbf{A}\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}}=\mathbf{A}$ 可得

$$\nabla^2 E_{\hat{\boldsymbol{w}}}=2\mathbf{X}^{\mathrm{T}}\mathbf{X}$$

如“西瓜书”中式 (3.11) 上方的一段话所说，假定 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 为正定矩阵，根据定理 3.1 可知此时 $E_{\hat{\boldsymbol{w}}}$ 是 $\hat{\boldsymbol{w}}$ 的严格凸函数，接着根据定理 3.2 可知只需令 $E_{\hat{\boldsymbol{w}}}$ 关于 $\hat{\boldsymbol{w}}$ 的一阶导数等于零向量，即令式 (3.10) 等于零向量即可求得全局最优解 $\hat{\boldsymbol{w}}^*$，具体求解过程如下：

$$\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}}=2\mathbf{X}^{\mathrm{T}}(\mathbf{X}\hat{\boldsymbol{w}}-\boldsymbol{y})=\mathbf{0}$$

$$2\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{w}}-2\mathbf{X}^{\mathrm{T}}\boldsymbol{y}=\mathbf{0}$$

$$2\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{w}}=2\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$$

$$\hat{\boldsymbol{w}}=(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$$

令其为 $\hat{\boldsymbol{w}}^*$ 即为式 (3.11)。

由于 $\mathbf{X}$ 是由样本构成的矩阵，而样本是千变万化的，因此无法保证 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 一定是正定矩阵，极易出现非正定的情形。当 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 非正定矩阵时，除了“西瓜书”中所说的引入正则化外，也可用 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 的伪逆矩阵代入式 (3.11) 求解出 $\hat{\boldsymbol{w}}^*$，只是此时并不保证求解得到的 $\hat{\boldsymbol{w}}^*$ 一定是全局最优解。除此之外，也可用下一节将会讲到的“梯度下降法”求解，同样也不保证求得全局最优解。

3.3 对数几率回归

对数几率回归的一般使用流程如下：首先在训练集上学得模型

$$y=\frac{1}{1+e^{-(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)}}$$

然后对于新的测试样本 $\boldsymbol{x}_i$，将其代入模型得到预测结果 y_i，接着自行设定阈值 θ，通常设为 $\theta=0.5$，如果 $y_i\geqslant\theta$ 则判 $\boldsymbol{x}_i$ 为正例，反之判为反例。

3.3.1 式 (3.27) 的推导

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

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

由于 y_i=0 或 1，则

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

两式综合可得

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

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

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

显然，此种方式更易推导出式 (3.27)。

“西瓜书” 在式 (3.27) 下方有提到式 (3.27) 是关于 $\boldsymbol{\beta}$ 的凸函数，其证明过程如下：由于若干半正定矩阵的加和仍为半正定矩阵，则根据定理 3.1 可知，若干凸函数的加和仍为凸函数。因此，只需证明式 (3.27) 求和符号后的式子 $-y_i\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i+\ln(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i})$（记为 $f(\boldsymbol{\beta})$）为凸函数即可。根据式 (3.31) 可知，$f(\boldsymbol{\beta})$ 的二阶导数，即 Hessian 矩阵为

$$\hat{\boldsymbol{x}}_i\hat{\boldsymbol{x}}_i^{\mathrm{T}}p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\left(1-p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\right)$$

对于任意非零向量 $\boldsymbol{y}\in\mathbb{R}^{d+1}$，恒有

$$\boldsymbol{y}^{\mathrm{T}}\cdot\hat{\boldsymbol{x}}_i\hat{\boldsymbol{x}}_i^{\mathrm{T}}p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\left(1-p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\right)\cdot\boldsymbol{y}$$

$$\boldsymbol{y}^{\mathrm{T}}\hat{\boldsymbol{x}}_i\hat{\boldsymbol{x}}_i^{\mathrm{T}}\boldsymbol{y}p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\left(1-p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\right)$$

$$\left(\boldsymbol{y}^{\mathrm{T}}\hat{\boldsymbol{x}}_i\right)^2p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\left(1-p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\right)$$

由于 $p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)>0$，因此上式恒大于等于 0，根据半正定矩阵的定义可知此时 $f(\boldsymbol{\beta})$ 的 Hessian 矩阵为半正定矩阵，所以 $f(\boldsymbol{\beta})$ 是关于 $\boldsymbol{\beta}$ 的凸函数。

3.3.2 梯度下降法

不同于式 (3.7) 可求得闭式解，式 (3.27) 中的 $\boldsymbol{\beta}$ 没有闭式解，因此需要借助其他工具进行求解。求解使得式 (3.27) 取到最小值的 $\boldsymbol{\beta}$ 属于最优化中的 “无约束优化问题”，在无约束优化问题中最常用的求解算法有 “梯度下降法” 和 “牛顿法” [3]，下面分别展开讲解。

梯度下降法是一种迭代求解算法，其基本思路如下：先在定义域中随机选取一个点 $\boldsymbol{x}^0$，将其代入函数 $f(\boldsymbol{x})$ 并判断此时 $f(\boldsymbol{x}^0)$ 是否是最小值，如果不是的话，则找下一个点 $\boldsymbol{x}^1$，且保证 $f(\boldsymbol{x}^1)<f(\boldsymbol{x}^0)$，然后接着判断 $f(\boldsymbol{x}_1)$ 是否是最小值，如果不是的话则重复上述步骤继续迭代寻找 $\boldsymbol{x}^2$、$\boldsymbol{x}^3$、..... 直到找到使得 $f(\boldsymbol{x})$ 取到最小值的 $\boldsymbol{x}^*$。

显然，此算法要想行得通就必须解决在找到第 t 个点 $\boldsymbol{x}^t$ 时，能进一步找到第 $t+1$ 个点 $\boldsymbol{x}^{t+1}$，且保证 $f(\boldsymbol{x}^{t+1})<f(\boldsymbol{x}^t)$。梯度下降法利用 “梯度指向的方向是函数值增大速度最快的方向” 这一特性，每次迭代时朝着梯度的反方向进行，进而实现函数值越迭代越小，下面给出完整的数学推导过程。

根据泰勒公式可知，当函数 $f(\boldsymbol{x})$ 在 $\boldsymbol{x}^t$ 处一阶可导时，在其邻域内进行一阶泰勒展开恒有

$$f(\boldsymbol{x})=f\left(\boldsymbol{x}^t\right)+\nabla f\left(\boldsymbol{x}^t\right)^{\mathrm{T}}\left(\boldsymbol{x}-\boldsymbol{x}^t\right)+o\left(\left\|\boldsymbol{x}-\boldsymbol{x}^t\right\|\right)$$

其中 $\nabla f(\boldsymbol{x}^t)$ 是函数 $f(\boldsymbol{x})$ 在点 $\boldsymbol{x}^t$ 处的梯度，$\|\boldsymbol{x}-\boldsymbol{x}^t\|$ 是指向量 $\boldsymbol{x}-\boldsymbol{x}^t$ 的模。若令 $\boldsymbol{x}-\boldsymbol{x}^t=a\boldsymbol{d}^t$，其中 $a>0$，$\boldsymbol{d}^t$ 是模长为 1 的单位向量，则上式可改写为

$$f(\boldsymbol{x}^t+a\boldsymbol{d}^t)=f\left(\boldsymbol{x}^t\right)+a\nabla f\left(\boldsymbol{x}^t\right)^{\mathrm{T}}\boldsymbol{d}^t+o\left(\left\|\boldsymbol{d}^t\right\|\right)$$

$$f(\boldsymbol{x}^t+a\boldsymbol{d}^t)-f\left(\boldsymbol{x}^t\right)=a\nabla f\left(\boldsymbol{x}^t\right)^{\mathrm{T}}\boldsymbol{d}^t+o\left(\left\|\boldsymbol{d}^t\right\|\right)$$

观察上式可知，如果能保证 $a\nabla f(\boldsymbol{x}^t)^{\mathrm{T}}\boldsymbol{d}^t<0$，则一定能保证 $f(\boldsymbol{x}^t+a\boldsymbol{d}^t)<f(\boldsymbol{x}^t)$，此时再令 $\boldsymbol{x}^{t+1}=\boldsymbol{x}^t+a\boldsymbol{d}^t$，即可推得我们想要的 $f(\boldsymbol{x}^{t+1})<f(\boldsymbol{x}^t)$。所以，此时问题转化为了求解能使得 $a\nabla f(\boldsymbol{x}^t)^{\mathrm{T}}\boldsymbol{d}^t<0$ 的 $\boldsymbol{d}^t$，且 $a\nabla f(\boldsymbol{x}^t)^{\mathrm{T}}\boldsymbol{d}^t$ 比 0 越小，相应地 $f(\boldsymbol{x}^{t+1})$ 也会比 $f(\boldsymbol{x}^t)$ 越小，也更接近最小值。

根据向量的内积公式可知

$$a\nabla f\left(\boldsymbol{x}^t\right)^{\mathrm{T}}\boldsymbol{d}^t=a\times\|\nabla f\left(\boldsymbol{x}^t\right)\|\times\|\boldsymbol{d}^t\|\times\cos\theta^t$$

其中 θ^t 是向量 $\nabla f(\boldsymbol{x}^t)$ 与向量 $\boldsymbol{d}^t$ 之间的夹角。观察上式易知，此时 $\|\nabla f(\boldsymbol{x}^t)\|$ 是固定常量，$\|\boldsymbol{d}^t\|=1$，所以当 a 也固定时，取 $\theta^t=\pi$，即向量 $\boldsymbol{d}^t$ 与向量 $\nabla f(\boldsymbol{x}^t)$ 的方向刚好相反时，上式取到最小值。通常为了精简计算步骤，可直接令 $\boldsymbol{d}^t=-\nabla f(\boldsymbol{x}^t)$，因此便得到了第 $t+1$ 个点 $\boldsymbol{x}^{t+1}$ 的迭代公式

$$\boldsymbol{x}^{t+1}=\boldsymbol{x}^t-a\nabla f\left(\boldsymbol{x}^t\right)$$

其中 a 也称为“步长”或“学习率”，是需要自行设定的参数，且每次迭代时可取不同值。

除了需要解决如何找到 $\boldsymbol{x}^{t+1}$ 以外，梯度下降法通常还需要解决如何判断当前点是否使得函数取到了最小值，否则的话迭代过程便可能会无休止进行。常用的做法是预先设定一个极小的阈值 ϵ，当某次迭代造成的函数值波动已经小于 ϵ 时，即 $|f(\boldsymbol{x}^{t+1})-f(\boldsymbol{x}^t)|<\epsilon$，我们便近似地认为此时 $f(\boldsymbol{x}^{t+1})$ 取到了最小值。

3.3.3 牛顿法

同梯度下降法，牛顿法也是一种迭代求解算法，其基本思路和梯度下降法一致，只是在选取第 $t+1$ 个点 $\boldsymbol{x}^{t+1}$ 时所采用的策略有所不同，即迭代公式不同。梯度下降法每次选取 $\boldsymbol{x}^{t+1}$ 时，只要求通过泰勒公式在 $\boldsymbol{x}^t$ 的邻域内找到一个函数值比其更小的点即可，而牛顿法则期望在此基础之上，$\boldsymbol{x}^{t+1}$ 还必须是 $\boldsymbol{x}^t$ 的邻域内的极小值点。

类似一元函数取到极值点的必要条件是一阶导数等于 0，多元函数取到极值点的必要条件是其梯度等于零向量 $\mathbf{0}$，为了能求解出 $\boldsymbol{x}^t$ 的邻域内梯度等于 $\mathbf{0}$ 的点，需要进行二阶泰勒展开，其展开式如下

$$f(\boldsymbol{x})=f\left(\boldsymbol{x}^t\right)+\nabla f\left(\boldsymbol{x}^t\right)^{\mathrm{T}}\left(\boldsymbol{x}-\boldsymbol{x}^t\right)+\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{x}^t\right)^{\mathrm{T}}\nabla^2 f\left(\boldsymbol{x}^t\right)\left(\boldsymbol{x}-\boldsymbol{x}^t\right)+o\left(\left\|\boldsymbol{x}-\boldsymbol{x}^t\right\|\right)$$

为了后续计算方便，我们取其近似形式

$$f(\boldsymbol{x})\approx f\left(\boldsymbol{x}^t\right)+\nabla f\left(\boldsymbol{x}^t\right)^{\mathrm{T}}\left(\boldsymbol{x}-\boldsymbol{x}^t\right)+\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{x}^t\right)^{\mathrm{T}}\nabla^2 f\left(\boldsymbol{x}^t\right)\left(\boldsymbol{x}-\boldsymbol{x}^t\right)$$

首先对上式求导

$$\begin{aligned}\frac{\partial f(\boldsymbol{x})}{\partial\boldsymbol{x}}&=\frac{\partial f\left(\boldsymbol{x}^t\right)}{\partial\boldsymbol{x}}+\frac{\partial\nabla f\left(\boldsymbol{x}^t\right)^{\mathrm{T}}\left(\boldsymbol{x}-\boldsymbol{x}^t\right)}{\partial\boldsymbol{x}}+\frac{1}{2}\frac{\partial\left(\boldsymbol{x}-\boldsymbol{x}^t\right)^{\mathrm{T}}\nabla^2 f\left(\boldsymbol{x}^t\right)\left(\boldsymbol{x}-\boldsymbol{x}^t\right)}{\partial\boldsymbol{x}}\\&=0+\nabla f\left(\boldsymbol{x}^t\right)+\frac{1}{2}\left(\nabla^2 f\left(\boldsymbol{x}^t\right)+\nabla^2 f\left(\boldsymbol{x}^t\right)^{\mathrm{T}}\right)\left(\boldsymbol{x}-\boldsymbol{x}^t\right)\end{aligned}$$

假设函数 $f(\boldsymbol{x})$ 在 $\boldsymbol{x}^t$ 处二阶可导，且偏导数连续，则 $\nabla^2 f(\boldsymbol{x}^t)$ 是对称矩阵，上式可写为

$$\begin{aligned}\frac{\partial f(\boldsymbol{x})}{\partial\boldsymbol{x}}&=0+\nabla f\left(\boldsymbol{x}^t\right)+\frac{1}{2}\times 2\times\nabla^2 f\left(\boldsymbol{x}^t\right)\left(\boldsymbol{x}-\boldsymbol{x}^t\right)\\&=\nabla f\left(\boldsymbol{x}^t\right)+\nabla^2 f\left(\boldsymbol{x}^t\right)\left(\boldsymbol{x}-\boldsymbol{x}^t\right)\end{aligned}$$

令上式等于 $\mathbf{0}$

$$\nabla f\left(\boldsymbol{x}^t\right)+\nabla^2 f\left(\boldsymbol{x}^t\right)\left(\boldsymbol{x}-\boldsymbol{x}^t\right)=\mathbf{0}$$

当 $\nabla^2 f\left(\boldsymbol{x}^t\right)$ 是可逆矩阵时，解得

$$\boldsymbol{x}=\boldsymbol{x}^t-\left[\nabla^2 f\left(\boldsymbol{x}^t\right)\right]^{-1}\nabla f\left(\boldsymbol{x}^t\right)$$

令上式为 $\boldsymbol{x}^{t+1}$ 即可得到牛顿法的迭代公式

$$\boldsymbol{x}^{t+1}=\boldsymbol{x}^t-\left[\nabla^2 f\left(\boldsymbol{x}^t\right)\right]^{-1}\nabla f\left(\boldsymbol{x}^t\right)$$

通过上述推导可知，牛顿法每次迭代时需要求解 Hessian 矩阵的逆矩阵，该步骤计算量通常较大，因此有人基于牛顿法，将其中求 Hessian 矩阵的逆矩阵改为求计算量更低的近似逆矩阵，我们称此类算法为“拟牛顿法”。

牛顿法虽然期望在每次迭代时能取到极小值点，但是通过上述推导可知，迭代公式是根据极值点的必要条件推导而得，因此并不保证一定是极小值点。

无论是梯度下降法还是牛顿法，根据其终止迭代的条件可知，其都是近似求解算法，即使 $f(\boldsymbol{x})$ 是凸函数，也并不一定保证最终求得的是全局最优解，仅能保证其接近全局最优解。不过在解决实际问题时，并不一定苛求解得全局最优解，在能接近全局最优甚至局部最优时通常也能很好地解决问题。

3.3.4 式 (3.29) 的解释

根据上述牛顿法的迭代公式可知，此式为式 (3.27) 应用牛顿法时的迭代公式。

3.3.5 式 (3.30) 的推导

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

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

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

其中 $\hat{\boldsymbol{y}}=(\hat{y}_1;\hat{y}_2;...;\hat{y}_m),\boldsymbol{y}=(y_1;y_2;...;y_m)$。

3.3.6 式 (3.31) 的推导

继续对上述式 (3.30) 中倒数第二个等号的结果求导

$$\begin{aligned}\frac{\partial^2\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}} &= -\frac{\partial\sum_{i=1}^{m}\hat{\boldsymbol{x}}_i\left(y_i-\frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}{1+e^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\\ &= -\sum_{i=1}^{m}\hat{\boldsymbol{x}}_i\frac{\partial\left(y_i-\frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\\ &= -\sum_{i=1}^{m}\hat{\boldsymbol{x}}_i\left(\frac{\partial y_i}{\partial\boldsymbol{\beta}^{\mathrm{T}}}-\frac{\partial\left(\frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\right)\\ &= \sum_{i=1}^{m}\hat{\boldsymbol{x}}_i\cdot\frac{\partial\left(\frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\end{aligned}$$

根据矩阵微分公式 $\frac{\partial\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}}{\partial\boldsymbol{x}^{\mathrm{T}}}=\frac{\partial\boldsymbol{x}^{\mathrm{T}}\boldsymbol{a}}{\partial\boldsymbol{x}^{\mathrm{T}}}=\boldsymbol{a}^{\mathrm{T}}$，其中

$$\begin{aligned}\frac{\partial\left(\frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}} &= \frac{\frac{\partial e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\cdot\left(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)-e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\cdot\frac{\partial\left(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}}{\left(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)^2}\\ &= \frac{\hat{\boldsymbol{x}}_i^{\mathrm{T}}e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\cdot\left(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)-e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\cdot\hat{\boldsymbol{x}}_i^{\mathrm{T}}e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}{\left(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)^2}\\ &= \hat{\boldsymbol{x}}_i^{\mathrm{T}}e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\cdot\frac{\left(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)-e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}{\left(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)^2}\\ &= \hat{\boldsymbol{x}}_i^{\mathrm{T}}e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\cdot\frac{1}{\left(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)^2}\\ &= \hat{\boldsymbol{x}}_i^{\mathrm{T}}\cdot\frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}\cdot\frac{1}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}\end{aligned}$$

所以

$$\begin{aligned}\frac{\partial^2\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}} &= \sum_{i=1}^{m}\hat{\boldsymbol{x}}_i\cdot\hat{\boldsymbol{x}}_i^{\mathrm{T}}\cdot\frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}\cdot\frac{1}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}\\ &= \sum_{i=1}^{m}\hat{\boldsymbol{x}}_i\hat{\boldsymbol{x}}_i^{\mathrm{T}}p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\left(1-p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\right)\end{aligned}$$

3.4 线性判别分析

线性判别分析的一般使用流程如下：首先在训练集上学得模型

$$y=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$$

由向量内积的几何意义可知，y 可以看作是 $\boldsymbol{x}$ 在 $\boldsymbol{w}$ 上的投影，因此在训练集上学得的模型能够保证训练集中的同类样本在 $\boldsymbol{w}$ 上的投影 y 很相近，而异类样本在 $\boldsymbol{w}$ 上的投影 y 很疏远。然后对于新的测试样本 $\boldsymbol{x}_i$，将其代入模型得到它在 $\boldsymbol{w}$ 上的投影 y_i，然后判别这个投影 y_i 与哪一类投影更近，则将其判为该类。

最后，线性判别分析也是一种降维方法，但不同于第 10 章介绍的无监督降维方法，线性判别分析是一种监督降维方法，即降维过程中需要用到样本类别标记信息。

3.4.1 式 (3.32) 的推导

式 (3.32) 中 $\|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1\|_2^2$ 左下角的 “2” 表示求 “2 范数”，向量的 2 范数即为模，右上角的 “2” 表示求平方数，基于此，下面推导式 (3.32)。

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

3.4.2 式 (3.37) 到式 (3.39) 的推导

由式 (3.36)，可定义拉格朗日函数为

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

对 $\boldsymbol{w}$ 求偏导可得

$$\begin{aligned}\frac{\partial L(\boldsymbol{w},\lambda)}{\partial\boldsymbol{w}}&=-\frac{\partial(\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w})}{\partial\boldsymbol{w}}+\lambda\frac{\partial(\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}-1)}{\partial\boldsymbol{w}}\\&=-(\mathbf{S}_b+\mathbf{S}_b^{\mathrm{T}})\boldsymbol{w}+\lambda(\mathbf{S}_w+\mathbf{S}_w^{\mathrm{T}})\boldsymbol{w}\end{aligned}$$

由于 $\mathbf{S}_b=\mathbf{S}_b^{\mathrm{T}},\mathbf{S}_w=\mathbf{S}_w^{\mathrm{T}}$，所以

$$\frac{\partial L(\boldsymbol{w},\lambda)}{\partial\boldsymbol{w}}=-2\mathbf{S}_b\boldsymbol{w}+2\lambda\mathbf{S}_w\boldsymbol{w}$$

令上式等于 0 即可得

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

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

$$(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}=\lambda\mathbf{S}_w\boldsymbol{w}$$

若令 $(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{w}=\gamma$，则有

$$\gamma(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)=\lambda\mathbf{S}_w\boldsymbol{w}$$

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

由于最终要求解的 $\boldsymbol{w}$ 不关心其大小，只关心其方向，所以其大小可以任意取值。又因为 $\boldsymbol{\mu}_0$ 和 $\boldsymbol{\mu}_1$ 的大小是固定的，所以 γ 的大小只受 $\boldsymbol{w}$ 的大小影响，因此可以通过调整 $\boldsymbol{w}$ 的大小使得 $\gamma=\lambda$，西瓜书中所说的 “不妨令 $\mathbf{S}_b\boldsymbol{w}=\lambda(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)$” 也可等价理解为令 $\gamma=\lambda$，因此，此时 $\frac{\gamma}{\lambda}=1$，求解出的 $\boldsymbol{w}$ 即为式 (3.39)。

3.4.3 式 (3.43) 的推导

由式 (3.40)、式 (3.41)、式 (3.42) 可得

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

3.4.4 式 (3.44) 的推导

此式是式 (3.35) 的推广形式，证明如下。

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

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

所以式 (3.44) 可变形为

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

对比式 (3.35) 易知，上式即式 (3.35) 的推广形式。

除了式 (3.35) 以外，还有一种常见的优化目标形式如下

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

无论是采用何种优化目标形式，其优化目标只要满足“同类样例的投影点尽可能接近，异类样例的投影点尽可能远离”即可。

3.4.5 式 (3.45) 的推导

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

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

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

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

其中，$\mathbf{I} \in \mathbb{R}^{N-1\times N-1}$ 为单位矩阵，$= diag(\lambda_1,\lambda_2,...,\lambda_{N-1}) \in \mathbb{R}^{N-1\times N-1}$ 是由 $N-1$ 个拉格朗日乘子构成的对角矩阵矩阵，根据矩阵微分公式 $\frac{\partial}{\partial \mathbf{X}}\ \operatorname{tr}\ (\mathbf{X}^{\mathrm{T}}\mathbf{B}\mathbf{X}) = (\mathbf{B}+\mathbf{B}^{\mathrm{T}})\mathbf{X}$ 对上式关于 $\mathbf{W}$ 求偏导可得

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

由于 $\mathbf{S}_b = \mathbf{S}_b^{\mathrm{T}}, \mathbf{S}_w = \mathbf{S}_w^{\mathrm{T}}$，所以

$$\frac{\partial L(\mathbf{W},\lambda)}{\partial \mathbf{W}} = -2\mathbf{S}_b\mathbf{W} + 2\lambda\mathbf{S}_w\mathbf{W}$$

令上式等于 $\mathbf{0}$ 即可得

$$-2\mathbf{S}_b\mathbf{W} + 2\lambda\mathbf{S}_w\mathbf{W} = \mathbf{0}$$

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

此即为式 (3.45)，但是此式在解释为何要取 $N-1$ 个最大广义特征值所对应的特征向量来构成 $\mathbf{W}$ 时不够直观。因此，我们换一种更为直观的方式求解式 (3.44)，只需换一种方式构造拉格朗日函数即可。

重新定义上述优化问题的拉格朗日函数

$$L(\mathbf{W},\mathbf{\Lambda}) = -\operatorname{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{S}_b\mathbf{W}) + \operatorname{tr}\left(\mathbf{\Lambda}(\mathbf{W}^{\mathrm{T}}\mathbf{S}_w\mathbf{W} - \mathbf{I})\right)$$

其中，$\mathbf{I} \in \mathbb{R}^{(N-1)\times(N-1)}$ 为单位矩阵，$\mathbf{\Lambda} = \operatorname{diag}(\lambda_1,\lambda_2,...,\lambda_{N-1}) \in \mathbb{R}^{(N-1)\times(N-1)}$ 是由 $N-1$ 个拉格朗日乘子构成的对角矩阵。根据矩阵微分公式 $\frac{\partial}{\partial \mathbf{X}}\operatorname{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{A}\mathbf{X}) = (\mathbf{A}+\mathbf{A}^{\mathrm{T}})\mathbf{X}, \frac{\partial}{\partial \mathbf{X}}\operatorname{tr}(\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}\mathbf{B}) = \frac{\partial}{\partial \mathbf{X}}\operatorname{tr}(\mathbf{A}\mathbf{X}^{\mathrm{T}}\mathbf{B}\mathbf{X}) = \mathbf{B}^{\mathrm{T}}\mathbf{X}\mathbf{A}^{\mathrm{T}} + \mathbf{B}\mathbf{X}\mathbf{A}$，对上式关于 $\mathbf{W}$ 求偏导可得

$$\begin{aligned}\frac{\partial L(\mathbf{W},\mathbf{\Lambda})}{\partial \mathbf{W}} &= -\frac{\partial\left(\operatorname{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{S}_b\mathbf{W})\right)}{\partial \mathbf{W}} + \frac{\partial\left(\operatorname{tr}\left(\mathbf{\Lambda}\mathbf{W}^{\mathrm{T}}\mathbf{S}_w\mathbf{W} - \mathbf{\Lambda}\mathbf{I}\right)\right)}{\partial \mathbf{W}} \\ &= -(\mathbf{S}_b+\mathbf{S}_b^{\mathrm{T}})\mathbf{W} + (\mathbf{S}_w^{\mathrm{T}}\mathbf{W}\mathbf{\Lambda}^{\mathrm{T}} + \mathbf{S}_w\mathbf{W}\mathbf{\Lambda})\end{aligned}$$

由于 $\mathbf{S}_b = \mathbf{S}_b^{\mathrm{T}}, \mathbf{S}_w = \mathbf{S}_w^{\mathrm{T}}, \mathbf{\Lambda}^{\mathrm{T}} = \mathbf{\Lambda}$，所以

$$\frac{\partial L(\mathbf{W},\mathbf{\Lambda})}{\partial \mathbf{W}} = -2\mathbf{S}_b\mathbf{W} + 2\mathbf{S}_w\mathbf{W}\mathbf{\Lambda}$$

令上式等于 $\mathbf{0}$ 即可得

$$-2\mathbf{S}_b\mathbf{W} + 2\mathbf{S}_w\mathbf{W}\mathbf{\Lambda} = \mathbf{0}$$

$$\mathbf{S}_b\mathbf{W} = \mathbf{S}_w\mathbf{W}\mathbf{\Lambda}$$

将 $\mathbf{W}$ 和 $\mathbf{\Lambda}$ 展开可得

$$\mathbf{S}_b\boldsymbol{w}_i = \lambda_i\mathbf{S}_w\boldsymbol{w}_i, \quad i = 1,2,...,N-1$$

此时便得到了 $N-1$ 个广义征值问题。进一步地，将其代入优化问题的目标函数可得

$$\begin{aligned}\min_{\mathbf{W}} -\operatorname{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{S}_b\mathbf{W}) &= \max_{\mathbf{W}} \operatorname{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{S}_b\mathbf{W}) \\ &= \max_{\mathbf{W}} \sum_{i=1}^{N-1} \boldsymbol{w}_i^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w}_i \\ &= \max_{\mathbf{W}} \sum_{i=1}^{N-1} \lambda_i \boldsymbol{w}_i^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}_i\end{aligned}$$

由于存在约束 $\operatorname{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{S}_w\mathbf{W}) = \sum\limits_{i=1}^{N-1} \boldsymbol{w}_i^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}_i = 1$，所以欲使上式取到最大值，只需取 $N-1$ 个最大的 λ_i 即可。根据 $\mathbf{S}_b\boldsymbol{w}_i = \lambda_i\mathbf{S}_w\boldsymbol{w}_i$ 可知，λ_i 对应的便是广义特征值，$\boldsymbol{w}_i$ 是 λ_i 所对应的特征向量。

（广义特征值的定义和常用求解方法可查阅 [2]）

对于 N 分类问题，一定要求出 $N-1$ 个 $\boldsymbol{w}_i$ 吗？其实不然。之所以将 $\mathbf{W}$ 定义为 $d\times(N-1)$ 维的矩阵是因为当 $d>(N-1)$ 时，实对称矩阵 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 的秩至多为 $N-1$，所以理论上至多能解出 $N-1$ 个非零特征值 λ_i 及其对应的特征向量 $\boldsymbol{w}_i$。但是 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 的秩是受当前训练集中的数据分布所影响的，因此并不一定为 $N-1$。此外，当数据分布本身就足够理想时，即使能求解出多个 $\boldsymbol{w}_i$，但是实际可能只需要求解出 1 个 $\boldsymbol{w}_i$ 便可将同类样本聚集，异类样本完全分离。

当 $d>(N-1)$ 时，实对称矩阵 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 的秩至多为 $N-1$ 的证明过程如下：由于 $\boldsymbol{\mu} = \frac{1}{N}\sum\limits_{i=1}^{N} m_i\boldsymbol{\mu}_i$，所以 $\boldsymbol{\mu}_1-\boldsymbol{\mu}$ 一定可以由 $\boldsymbol{\mu}$ 和 $\boldsymbol{\mu}_2,...,\boldsymbol{\mu}_N$ 线性表示，因此矩阵 $\mathbf{S}_b$ 中至多有 $\boldsymbol{\mu}_2-\boldsymbol{\mu},...,\boldsymbol{\mu}_N-\boldsymbol{\mu}$ 共 $N-1$ 个线性无关的向量，由于此时 $d>(N-1)$，所以 $\mathbf{S}_b$ 的秩 $r(\mathbf{S}_b)$ 至多为 $N-1$。同时假设矩阵 $\mathbf{S}_w$ 满秩，即 $r(\mathbf{S}_w) = r(\mathbf{S}_w^{-1}) = d$，则根据矩阵秩的性质 $r(\mathbf{AB}) \leqslant \min\{r(\mathbf{A}), r(\mathbf{B})\}$ 可知，$\mathbf{S}_w^{-1}\mathbf{S}_b$ 的秩也至多为 $N-1$。

3.5 多分类学习

3.5.1 图 3.5 的解释

图 3.5 中所说的“海明距离”是指两个码对应位置不相同的个数，“欧式距离”则是指两个向量之间的欧氏距离，例如图 3.5(a) 中第 1 行的编码可以视作为向量 $(-1,+1,-1,+1,+1)$，测试示例的编码则为 $(-1,-1,+1,-1,+1)$，其中第 2 个、第 3 个、第 4 个元素不相同，所以它们的海明距离为 3，欧氏距离为 $\sqrt{(-1-(-1))^2+(1-(-1))^2+(-1-1)^2+(1-(-1))^2+(1-1)^2} = \sqrt{0+4+4+4+0} = 2\sqrt{3}$。

3.6 类别不平衡问题

对于类别平衡问题，“西瓜书”2.3.1 节中的“精度”通常无法满足该特殊任务的需求，例如“西瓜书”在本节第一段的举例：有 998 个反例和 2 个正例，若机器学习算法返回一个永远将新样本预测为反例的学习器则能达到 99.8% 的精度，显然虚高，因此在类别不平衡时常采用 2.3.2 节中的查准率、查全率和 F1 来度量学习器的性能。

参考文献

[1] Wikipedia contributors. Matrix calculus, 2022.

[2] 张贤达. *矩阵分析与应用. 第 2 版*. 清华大学出版社, 2013.

[3] 王燕军. *最优化基础理论与方法*. 复旦大学出版社, 2011.

第 4 章 决策树

4.1 公式 (4.1)

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

[解析]：证明 $0 \leq \mathrm{Ent}(D) \leq \log_2 |\mathcal{Y}|$：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，且 $0 \leq p_k \leq 1, \sum_{k=1}^{n} p_k = 1$。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，也即

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

其中，$0 \leq x_k \leq 1, \sum_{k=1}^{n} x_k = 1$，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 $0 \leq x_k \leq 1$，仅考虑 $\sum_{k=1}^{n} x_k = 1$ 的话，对 $f(x_1, ..., x_n)$ 求最大值等价于如下最小化问题

$$\begin{aligned} \min \quad & \sum_{k=1}^{n} x_k \log_2 x_k \\ \text{s.t.} \quad & \sum_{k=1}^{n} x_k = 1 \end{aligned}$$

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

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

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

$$\begin{aligned} \frac{\partial L(x_1, ..., x_n, \lambda)}{\partial x_1} &= \frac{\partial}{\partial x_1}\left[\sum_{k=1}^{n} x_k \log_2 x_k + \lambda(\sum_{k=1}^{n} x_k - 1)\right] = 0 \\ &= \log_2 x_1 + x_1 \cdot \frac{1}{x_1 \ln 2} + \lambda = 0 \\ &= \log_2 x_1 + \frac{1}{\ln 2} + \lambda = 0 \\ &\Rightarrow \lambda = -\log_2 x_1 - \frac{1}{\ln 2} \end{aligned}$$

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

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

$$\vdots$$

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

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

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

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

整理一下可得

$$\begin{cases}\lambda=-\log_2 x_1-\dfrac{1}{\ln 2}=-\log_2 x_2-\dfrac{1}{\ln 2}=...=-\log_2 x_n-\dfrac{1}{\ln 2}\\ \sum\limits_{k=1}^{n}x_k=1\end{cases}$$

由以上两个方程可以解得

$$x_1=x_2=...=x_n=\frac{1}{n}$$

又因为 x_k 还需满足约束 $0\le x_k\le 1$，显然 $0\le\frac{1}{n}\le 1$，所以 $x_1=x_2=...=x_n=\frac{1}{n}$ 是满足所有约束的最优解，也即为当前最小化问题的最小值点，同时也是 $f(x_1,...,x_n)$ 的最大值点。将 $x_1=x_2=...=x_n=\frac{1}{n}$ 代入 $f(x_1,...,x_n)$ 中可得

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

所以 $f(x_1,...,x_n)$ 在满足约束 $0\le x_k\le 1,\sum_{k=1}^{n}x_k=1$ 时的最大值为 $\log_2 n$。求完最大值后下面我们再来求最小值，如果不考虑约束 $\sum_{k=1}^{n}x_k=1$，仅考虑 $0\le x_k\le 1$ 的话，$f(x_1,...,x_n)$ 可以看做是 n 个互不相关的一元函数的加和，也即

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

其中，$g(x_k)=-x_k\log_2 x_k, 0\le x_k\le 1$。那么当 $g(x_1),g(x_2),...,g(x_n)$ 分别取到其最小值时，$f(x_1,...,x_n)$ 也就取到了最小值。所以接下来考虑分别求 $g(x_1),g(x_2),...,g(x_n)$ 各自的最小值，由于 $g(x_1),g(x_2),...,g(x_n)$ 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 $g(x_2),...,g(x_n)$ 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

$$g'(x_1)=\frac{d(-x_1\log_2 x_1)}{dx_1}=-\log_2 x_1-x_1\cdot\frac{1}{x_1\ln 2}=-\log_2 x_1-\frac{1}{\ln 2}$$

$$g''(x_1)=\frac{d\left(g'(x_1)\right)}{dx_1}=\frac{d\left(-\log_2 x_1-\dfrac{1}{\ln 2}\right)}{dx_1}=-\frac{1}{x_1\ln 2}$$

显然，当 $0 \leq x_k \leq 1$ 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，于是分别取 $x_1 = 0$ 和 $x_1 = 1$，代入 $g(x_1)$ 可得

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

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

所以，$g(x_1)$ 的最小值为 0，同理可得 $g(x_2), ..., g(x_n)$ 的最小值也为 0，那么 $f(x_1, ..., x_n)$ 的最小值此时也为 0。但是，此时是不考虑约束 $\sum_{k=1}^{n} x_k = 1$，仅考虑 $0 \leq x_k \leq 1$ 时取到的最小值，若考虑约束 $\sum_{k=1}^{n} x_k = 1$ 的话，那么 $f(x_1, ..., x_n)$ 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 $x_1 = x_2 = ... = x_{k-1} = x_{k+1} = ... = x_n = 0$，将其代入 $f(x_1, ..., x_n)$ 可得

$$f(0, 0, ..., 0, 1, 0, ..., 0) = -0 \log_2 0 - 0 \log_2 0 ... - 0 \log_2 0 - 1 \log_2 1 - 0 \log_2 0 ... - 0 \log_2 0 = 0$$

所以 $x_k = 1, x_1 = x_2 = ... = x_{k-1} = x_{k+1} = ... = x_n = 0$ 一定是 $f(x_1, ..., x_n)$ 在满足约束 $\sum_{k=1}^{n} x_k = 1$ 和 $0 \leq x_k \leq 1$ 的条件下的最小值点，其最小值为 0。
 综上可知，当 $f(x_1, ..., x_n)$ 取到最大值时：$x_1 = x_2 = ... = x_n = \dfrac{1}{n}$，此时样本集合纯度最低；当 $f(x_1, ..., x_n)$ 取到最小值时：$x_k = 1, x_1 = x_2 = ... = x_{k-1} = x_{k+1} = ... = x_n = 0$，此时样本集合纯度最高。

4.2 公式 (4.2)

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

[解析]：这个是信息增益的定义公式，在信息论中信息增益也称为互信息（参见附录①），其表示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。所以在这里，这个公式可以理解为在属性 a 的取值已知后，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，也即为书上所说的“纯度提升”越大。

4.3 公式 (4.6)

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

[解析]：这个是数据集 D 中属性 a 的基尼指数的定义，它表示在属性 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 分类树时并不会严格按照此公式来选择最优划分属性，主要是因为 CART 分类树是一颗二叉树，如果用上面的公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。CART 分类树的构造算法如下：

- 首先，对每个属性 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

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

- 然后，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；
- 最后，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 分类树，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 $\neq$ 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 $p_2=\frac{6}{11}$，根据公式（4.5）可计算出用“色泽 $=$ 青绿”划分之后得到基尼指数为

$$\text{Gini\_index}(D,\text{色泽}=\text{青绿})=\frac{6}{17}\times\left(1-(\frac{3}{6})^2-(\frac{3}{6})^2\right)+\frac{11}{17}\times\left(1-(\frac{5}{11})^2-(\frac{6}{11})^2\right)=0.497$$

类似的，可以计算出以下不同属性取不同值的基尼指数

$$\text{Gini\_index}(D,\text{色泽}=\text{乌黑})=\frac{6}{17}\times\left(1-(\frac{4}{6})^2-(\frac{2}{6})^2\right)+\frac{11}{17}\times\left(1-(\frac{4}{11})^2-(\frac{7}{11})^2\right)=0.456$$

$$\text{Gini\_index}(D,\text{色泽}=\text{浅白})=\frac{5}{17}\times\left(1-(\frac{1}{5})^2-(\frac{4}{5})^2\right)+\frac{12}{17}\times\left(1-(\frac{7}{12})^2-(\frac{5}{12})^2\right)=0.426$$

$$\text{Gini\_index}(D,\text{根蒂}=\text{蜷缩})=0.456$$
$$\text{Gini\_index}(D,\text{根蒂}=\text{稍蜷})=0.496$$
$$\text{Gini\_index}(D,\text{根蒂}=\text{硬挺})=0.439$$
$$\text{Gini\_index}(D,\text{敲声}=\text{浊响})=0.450$$
$$\text{Gini\_index}(D,\text{敲声}=\text{沉闷})=0.494$$
$$\text{Gini\_index}(D,\text{敲声}=\text{清脆})=0.439$$
$$\text{Gini\_index}(D,\text{纹理}=\text{清晰})=0.286$$
$$\text{Gini\_index}(D,\text{纹理}=\text{稍稀})=0.437$$
$$\text{Gini\_index}(D,\text{纹理}=\text{模糊})=0.403$$
$$\text{Gini\_index}(D,\text{脐部}=\text{凹陷})=0.415$$
$$\text{Gini\_index}(D,\text{脐部}=\text{稍凹})=0.497$$
$$\text{Gini\_index}(D,\text{脐部}=\text{平坦})=0.362$$
$$\text{Gini\_index}(D,\text{触感}=\text{硬挺})=0.494$$
$$\text{Gini\_index}(D,\text{触感}=\text{软粘})=0.494$$

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 $\neq$ 清晰) 两个子节点，对于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 分类树的构建过程，从构建过程中可以看出，CART 分类树最终构造出来的是一颗二叉树。CART 决策树除了能处理分类问题以外，它还可以处理回归问题，附录②中给出了 CART 回归树的构造算法。

4.4 公式 (4.7)

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

[解析]：这个公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 为例来说明此公式的用法。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，根据公式（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 $T_a = \{\frac{(0.243+0.245)}{2}, \frac{(0.245+0.343)}{2}, \frac{(0.343+0.360)}{2}, \frac{(0.360+0.403)}{2}, \frac{(0.403+0.437)}{2}, \frac{(0.437+0.481)}{2}, \frac{(0.481+0.556)}{2}, \frac{(0.556+0.593)}{2}, \frac{(0.593+0.608)}{2}, \frac{(0.608+0.634)}{2}, \frac{(0.634+0.639)}{2}, \frac{(0.639+0.657)}{2}, \frac{(0.657+0.666)}{2}, \frac{(0.666+0.697)}{2}, \frac{(0.697+0.719)}{2}, \frac{(0.719+0.774)}{2}\}$

4.5 公式 (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]：此公式是公式（4.2）用于离散化后的连续属性的版本，其中 T_a 由公式（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，也即当 $\lambda = -$ 时：$D_t^\lambda = D_t^{a\leq t}$，当 $\lambda = +$ 时：$D_t^\lambda = D_t^{a>t}$。

4.6 附录

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

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

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，$p_i = P(X=x_i)\ i=1,2,...,n$。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

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

此即为互信息的数学定义。

②CART 回归树 [1]

假设给定数据集

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

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 $\boldsymbol{x}_i$，他们所对应的输出值 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：

- 任意选择一个属性 a，遍历其所有可能取值，根据如下公式找出属性 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，$R_1(a,v)=\{\boldsymbol{x}|\boldsymbol{x}\in D^{a\leq v}\}, R_2(a,v)=\{\boldsymbol{x}|\boldsymbol{x}\in D^{a>v}\}$，$c_1$ 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，也即

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

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

- 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。这样就生成了一颗 CART 回归树，假设最终将特征空间被划分为了 M 个子空间 $R_1,R_2,...,R_M$，那么 CART 回归树的模型公式可以表示为

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

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。此公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. 清华大学出版社, 2012.

第 5 章　神经网络

5.1　公式 (5.2)

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

[解析]：此公式是感知机学习算法中的参数更新公式，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，故感知机模型的公式可表示为

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

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，θ 为阈值。上式中的 f 通常设为符号函数，那么感知机模型的公式可进一步表示为

$$y = \mathrm{sign}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta) = \begin{cases} 1, & \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta \geq 0 \\ 0, & \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta < 0 \end{cases}$$

由于 n 维空间中的超平面方程为

$$w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0$$

所以此时感知机模型公式中的 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta$ 可以看作是 n 维空间中的一个超平面，通过它将 n 维空间划分为 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta \geq 0$ 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，以此来实现分类功能。

感知机学习策略

给定一个线性可分的数据集 T（参见附录①），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 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} - \theta \geq 0$ 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，以下公式恒成立

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

所以，给定数据集 T，其损失函数可以定义为：

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

显然，此损失函数是非负的。如果没有误分类点，损失函数值是 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

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

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 $L(\boldsymbol{w})$ 的梯度为：

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

感知机的学习算法具体采用的是随机梯度下降法，也就是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使其梯度下降。所以权重 $\boldsymbol{w}$ 的更新公式为

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

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

相应地，$\boldsymbol{w}$ 中的某个分量 w_i 的更新公式即为公式 (5.2)。

5.2 公式 (5.10)

$$\begin{aligned} g_j &= -\frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \\ &= -(\hat{y}_j^k - y_j^k) f'(\beta_j - \theta_j) \\ &= \hat{y}_j^k (1 - \hat{y}_j^k)(y_j^k - \hat{y}_j^k) \end{aligned}$$

[推导]：参见公式 (5.12)

5.3 公式 (5.12)

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

[推导]：因为

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

又

$$\begin{aligned}
\frac{\partial E_k}{\partial\theta_j} &= \frac{\partial E_k}{\partial\hat{y}_j^k}\cdot\frac{\partial\hat{y}_j^k}{\partial\theta_j} \\
&= \frac{\partial E_k}{\partial\hat{y}_j^k}\cdot\frac{\partial[f(\beta_j-\theta_j)]}{\partial\theta_j} \\
&= \frac{\partial E_k}{\partial\hat{y}_j^k}\cdot f'(\beta_j-\theta_j)\times(-1) \\
&= \frac{\partial E_k}{\partial\hat{y}_j^k}\cdot f\left(\beta_j-\theta_j\right)\times\left[1-f\left(\beta_j-\theta_j\right)\right]\times(-1) \\
&= \frac{\partial E_k}{\partial\hat{y}_j^k}\cdot\hat{y}_j^k\left(1-\hat{y}_j^k\right)\times(-1) \\
&= \frac{\partial\left[\frac{1}{2}\sum_{j=1}^{l}\left(\hat{y}_j^k-y_j^k\right)^2\right]}{\partial\hat{y}_j^k}\cdot\hat{y}_j^k\left(1-\hat{y}_j^k\right)\times(-1) \\
&= \frac{1}{2}\times 2(\hat{y}_j^k-y_j^k)\times 1\cdot\hat{y}_j^k\left(1-\hat{y}_j^k\right)\times(-1) \\
&= (y_j^k-\hat{y}_j^k)\hat{y}_j^k\left(1-\hat{y}_j^k\right) \\
&= g_j
\end{aligned}$$

所以

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

5.4 公式 (5.13)

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

[推导]：因为

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

又

$$\begin{aligned}\frac{\partial E_k}{\partial v_{ih}} &= \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \cdot \frac{\partial \beta_j}{\partial b_h} \cdot \frac{\partial b_h}{\partial \alpha_h} \cdot \frac{\partial \alpha_h}{\partial v_{ih}} \\ &= \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \cdot \frac{\partial \beta_j}{\partial b_h} \cdot \frac{\partial b_h}{\partial \alpha_h} \cdot x_i \\ &= \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \cdot \frac{\partial \beta_j}{\partial b_h} \cdot f'(\alpha_h - \gamma_h) \cdot x_i \\ &= \sum_{j=1}^{l} \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \cdot w_{hj} \cdot f'(\alpha_h - \gamma_h) \cdot x_i \\ &= \sum_{j=1}^{l} (-g_j) \cdot w_{hj} \cdot f'(\alpha_h - \gamma_h) \cdot x_i \\ &= -f'(\alpha_h - \gamma_h) \cdot \sum_{j=1}^{l} g_j \cdot w_{hj} \cdot x_i \\ &= -b_h(1 - b_h) \cdot \sum_{j=1}^{l} g_j \cdot w_{hj} \cdot x_i \\ &= -e_h \cdot x_i\end{aligned}$$

所以

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

5.5 公式 (5.14)

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

[推导]：因为

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

又

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

所以

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

5.6 公式 (5.15)

$$\begin{aligned} e_h &= -\frac{\partial E_k}{\partial b_h} \cdot \frac{\partial b_h}{\partial \alpha_h} \\ &= -\sum_{j=1}^{l} \frac{\partial E_k}{\partial \beta_j} \cdot \frac{\partial \beta_j}{\partial b_h} f'(\alpha_h - \gamma_h) \\ &= \sum_{j=1}^{l} w_{hj} g_j f'(\alpha_h - \gamma_h) \\ &= b_h(1-b_h) \sum_{j=1}^{l} w_{hj} g_j \end{aligned}$$

[推导]：参见公式 (5.13)

5.7 公式 (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 机本质上是一个引入了隐变量的无向图模型，无向图的能量可理解为

$$E_{graph} = E_{edges} + E_{nodes}$$

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：$E_{edge_{ij}} = -w_{ij} s_i s_j$，结点能量由结点的值及其阈值的乘积确定：$E_{node_i} = -\theta_i s_i$；图中边的能量为图中所有边能量之和

$$E_{edges} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E_{edge_{ij}} = -\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij} s_i s_j$$

图中结点的能量为图中所有结点能量之和

$$E_{nodes} = \sum_{i=1}^{n} E_{node_i} = -\sum_{i=1}^{n} \theta_i s_i$$

故状态向量 $\boldsymbol{s}$ 所对应的 Boltzmann 机能量为

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

5.8 公式 (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，显层的状态向量为 $\boldsymbol{v}$，隐层的状态向量为 $\boldsymbol{h}$。

$$\boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{bmatrix} \qquad \boldsymbol{h} = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_q \end{bmatrix}$$

对于显层状态向量 $\boldsymbol{v}$ 中的变量 v_i，其仅与隐层状态向量 $\boldsymbol{h}$ 有关，所以给定隐层状态向量 $\boldsymbol{h}$，$v_1, v_2, ..., v_d$ 相互独立。

5.9 公式 (5.23)

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

[解析]：由公式 5.22 的解析同理可得：给定显层状态向量 $\boldsymbol{v}$，$h_1, h_2, \cdots, h_q$ 相互独立。

5.10 公式 (5.24)

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

[推导]：由公式 (5.20) 可推导出受限 Boltzmann 机（以下简称 RBM）的能量函数为：

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

其中

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

再由公式 (5.21) 可知，RBM 的联合概率分布为

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

其中 Z 为规范化因子

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

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 $\boldsymbol{\theta} = \{\mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$，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

$$\begin{aligned} L(\boldsymbol{\theta}) &= \ln\left(\prod_{k=1}^{m} P(\boldsymbol{v}_k)\right) \\ &= \sum_{k=1}^{m} \ln P(\boldsymbol{v}_k) \\ &= \sum_{k=1}^{m} L_k(\boldsymbol{\theta}) \end{aligned}$$

具体采用的是梯度上升法来求解参数 $\boldsymbol{\theta}$，因此，下面来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 $\boldsymbol{v}_k$ 来说，其 $L_k(\boldsymbol{\theta})$ 的具体形式为

$$\begin{aligned}
L_k(\boldsymbol{\theta}) &= \ln P(\boldsymbol{v}_k) \\
&= \ln\left(\sum_{\boldsymbol{h}} P(\boldsymbol{v}_k, \boldsymbol{h})\right) \\
&= \ln\left(\sum_{\boldsymbol{h}} \frac{1}{Z} e^{-E(\boldsymbol{v}_k, \boldsymbol{h})}\right) \\
&= \ln\left(\sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v}_k, \boldsymbol{h})}\right) - \ln Z \\
&= \ln\left(\sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v}_k, \boldsymbol{h})}\right) - \ln\left(\sum_{\boldsymbol{v},\boldsymbol{h}} e^{-E(\boldsymbol{v}, \boldsymbol{h})}\right)
\end{aligned}$$

对 $L_k(\boldsymbol{\theta})$ 进行求导

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

由于

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

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

故

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

由于 $\boldsymbol{\theta} = \{\mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$ 包含三个参数，在这里我们仅以 $\mathbf{W}$ 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

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

根据公式 (5.23) 可知

$$\begin{aligned}
&\sum_{\boldsymbol{h}} P(\boldsymbol{h}|\boldsymbol{v})\frac{\partial E(\boldsymbol{v},\boldsymbol{h})}{\partial w_{ij}} \\
=&-\sum_{\boldsymbol{h}} P(\boldsymbol{h}|\boldsymbol{v})h_i v_j \\
=&-\sum_{\boldsymbol{h}}\prod_{l=1}^{q} P(h_l|\boldsymbol{v})h_i v_j \\
=&-\sum_{\boldsymbol{h}} P(h_i|\boldsymbol{v})\prod_{l=1,l\neq i}^{q} P(h_l|\boldsymbol{v})h_i v_j \\
=&-\sum_{\boldsymbol{h}} P(h_i|\boldsymbol{v})P(h_1,...,h_{i-1},h_{i+1},...,h_q|\boldsymbol{v})h_i v_j \\
=&-\sum_{h_i} P(h_i|\boldsymbol{v})h_i v_j \sum_{h_1,...,h_{i-1},h_{i+1},...,h_q} P(h_1,...,h_{i-1},h_{i+1},...,h_q|\boldsymbol{v}) \\
=&-\sum_{h_i} P(h_i|\boldsymbol{v})h_i v_j \cdot 1 \\
=&-[P(h_i=0|\boldsymbol{v})\cdot 0\cdot v_j + P(h_i=1|\boldsymbol{v})\cdot 1\cdot v_j] \\
=&-P(h_i=1|\boldsymbol{v})v_j
\end{aligned}$$

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 [2] 是：用步长为 s（通常设为 1）的 CD 算法

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

近似代替

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

由此可知对于 w_{ij} 来说，就是用

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

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 $\Delta w_{ij} := \frac{\partial L_k(\boldsymbol{\theta})}{\partial w_{ij}}$，$RBM(\boldsymbol{\theta})$ 表示参数为 $\boldsymbol{\theta}$ 的 RBM 网络，则 $CD_s(w_{ij},\boldsymbol{v})$ 的具体算法为 Algorithm 1。其中函数 $h\_given\_v(\boldsymbol{v},RBM(\boldsymbol{\theta}))$ 表示在给定 $\boldsymbol{v}$ 的条件下，从 $RBM(\boldsymbol{\theta})$ 中采样生成 $\boldsymbol{h}$，同理，函数 $v\_given\_h(\boldsymbol{h},RBM(\boldsymbol{\theta}))$ 表示在给定 $\boldsymbol{h}$ 的条件下，从 $RBM(\boldsymbol{\theta})$ 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，下面仅给出函数 $h\_given\_v(\boldsymbol{v},RBM(\boldsymbol{\theta}))$ 的具体算法 Algorithm 2。综上可知，公式 (5.24) 其实就是带有学习率为 η 的 Δw_{ij} 的一种形式化的表示。

Algorithm 1 $CD_s(w_{ij}, \boldsymbol{v})$

输入: $s, V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_m\}, RBM(\boldsymbol{\theta})$

过程:

初始化：$\Delta w_{ij} = 0$
for $\boldsymbol{v} \in V$ **do**
  $\boldsymbol{v}^{(0)} := \boldsymbol{v}$
  for $t = 1, 2, ..., s-1$ **do**
    $\boldsymbol{h}^{(t)} = h\_given\_v(\boldsymbol{v}^{(t)}, RBM(\boldsymbol{\theta}))$
    $\boldsymbol{v}^{(t+1)} = v\_given\_h(\boldsymbol{h}^{(t)}, RBM(\boldsymbol{\theta}))$
  for $i = 1, 2, ..., q; j = 1, 2, ..., d$ **do**
    $\Delta w_{ij} = \Delta w_{ij} + \left[P(h_i = 1|\boldsymbol{v}^{(0)})v_j^{(0)} - P(h_i = 1|\boldsymbol{v}^{(s)})v_j^{(s)}\right]$

输出: Δw_{ij}

Algorithm 2 $h\_given\_v(\boldsymbol{v}, RBM(\boldsymbol{\theta}))$

输入: $\boldsymbol{v}, RBM(\boldsymbol{\theta})$

过程:

for $i = 1, 2, ..., q$ **do**
  随机生成$0 \leq \alpha_i \leq 1$
  $h_j = \begin{cases} 1, & \text{if } \alpha_i < P(h_i = 1|\boldsymbol{v}) \\ 0, & \text{otherwise} \end{cases}$

输出: $\boldsymbol{h} = (h_1, h_2, ..., h_q)^{\mathrm{T}}$

5.11 附录

①数据集的线性可分 [1]

给定一个数据集

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

其中，$\boldsymbol{x}_i \in \mathbb{R}^n, y_i \in \{0, 1\}, i = 1, 2, ..., N$，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. 清华大学出版社, 2012.

[2] 皮果提. 受限玻尔兹曼机（rbm）学习笔记（六）对比散度算法, 2014. URL: https://blog.csdn.net/itplus/article/details/19408143.

第 6 章　支持向量机

6.1　公式 (6.9)

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

[推导]：公式 (6.8) 可作如下展开

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

对 $\boldsymbol{w}$ 和 b 分别求偏导数 并令其等于 0

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

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

值得一提的是，上述求解过程遵循的是西瓜书附录 B 中公式 (B.7) 左边的那段话 “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啥呢？在这里我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 $\boldsymbol{x}^*$ 一定满足附录①给出的 KKT 条件（证明参见参考文献 [2] 的 § 5.5），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；
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可知，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。

6.2　公式 (6.10)

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

[解析]：参见公式 (6.9)

6.3　公式 (6.11)

$$\begin{aligned}&\max_{\boldsymbol{\alpha}}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\boldsymbol{x}_i^T\boldsymbol{x}_j\\ &\text{s.t.}\ \sum_{i=1}^{m}\alpha_i y_i=0\\ &\quad\alpha_i\geq 0\quad i=1,2,\ldots,m\end{aligned}$$

[推导]：将公式 (6.9) 和公式 (6.10) 代入公式 (6.8) 即可将 $L(\boldsymbol{w},b,\boldsymbol{\alpha})$ 中的 $\boldsymbol{w}$ 和 b 消去，再考虑公式 (6.10) 的约束，就得到了公式 (6.6) 的对偶问题

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

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

$$\begin{aligned}\inf_{\boldsymbol{w},b} L(\boldsymbol{w},b,\boldsymbol{\alpha}) &= -\frac{1}{2}\boldsymbol{w}^T\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i + \sum_{i=1}^{m}\alpha_i \\ &= -\frac{1}{2}(\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i)^T(\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i) + \sum_{i=1}^{m}\alpha_i \\ &= -\frac{1}{2}\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i^T\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i + \sum_{i=1}^{m}\alpha_i \\ &= \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j\end{aligned}$$

所以

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

6.4 公式 (6.13)

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

[解析]：参见公式 (6.9) 中给出的第 1 点理由

6.5 公式 (6.35)

$$\begin{aligned}&\min_{\boldsymbol{w},b,\xi_i}\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\xi_i \\ &\text{s.t. } y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) \geqslant 1 - \xi_i \\ &\xi_i \geqslant 0, i = 1,2,\ldots,m\end{aligned}$$

[解析]：令

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

显然 $\xi_i \geq 0$，而且当 $1 - y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) > 0$ 时

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

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

$$\xi_i = 0$$

所以综上可得

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

6.6 公式 (6.37)

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

[解析]：参见公式 (6.9)

6.7 公式 (6.38)

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

[解析]：参见公式 (6.10)

6.8 公式 (6.39)

$$C=\alpha_i+\mu_i$$

[推导]：公式 (6.36) 关于 ξ_i 求偏导并令其等于 0 可得：

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

6.9 公式 (6.40)

$$\begin{aligned}&\max_{\boldsymbol{\alpha}}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\boldsymbol{x}_i^T\boldsymbol{x}_j\\&s.t.\sum_{i=1}^{m}\alpha_i y_i=0\\&0\leq\alpha_i\leq C\quad i=1,2,\ldots,m\end{aligned}$$

将公式 (6.37)-(6.39) 代入公式 (6.36) 可以得到公式 (6.35) 的对偶问题：

$$\begin{aligned}&\frac{1}{2}||\boldsymbol{w}||^2+C\sum_{i=1}^{m}\xi_i+\sum_{i=1}^{m}\alpha_i(1-\xi_i-y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b))-\sum_{i=1}^{m}\mu_i\xi_i\\=&\frac{1}{2}||\boldsymbol{w}||^2+\sum_{i=1}^{m}\alpha_i(1-y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b))+C\sum_{i=1}^{m}\xi_i-\sum_{i=1}^{m}\alpha_i\xi_i-\sum_{i=1}^{m}\mu_i\xi_i\\=&-\frac{1}{2}\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{x}_i^T\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{x}_i+\sum_{i=1}^{m}\alpha_i+\sum_{i=1}^{m}C\xi_i-\sum_{i=1}^{m}\alpha_i\xi_i-\sum_{i=1}^{m}\mu_i\xi_i\\=&-\frac{1}{2}\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{x}_i^T\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{x}_i+\sum_{i=1}^{m}\alpha_i+\sum_{i=1}^{m}(C-\alpha_i-\mu_i)\xi_i\\=&\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\boldsymbol{x}_i^T\boldsymbol{x}_j\\=&\min_{\boldsymbol{w},b,\boldsymbol{\xi}}L(\boldsymbol{w},b,\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\mu})\end{aligned}$$

所以

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

又

$$\begin{aligned}&\alpha_i \geq 0 \\ &\mu_i \geq 0 \\ &C = \alpha_i + \mu_i\end{aligned}$$

消去 μ_i 可得等价约束条件为：

$$0 \leq \alpha_i \leq C \quad i = 1,2,\ldots,m$$

6.10 公式 (6.41)

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

[解析]：参见公式 (6.13)

6.11 公式 (6.52)

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

[推导]：将公式 (6.45) 的约束条件全部恒等变形为小于等于 0 的形式可得：

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

由于以上四个约束条件的拉格朗日乘子分别为 $\alpha_i, \hat{\alpha}_i, \mu_i, \hat{\mu}_i$，所以由附录①可知，以上四个约束条件可相应转化为以下 KKT 条件：

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

由公式 (6.49) 和公式 (6.50) 可知：

$$\begin{aligned}&\mu_i = C - \alpha_i \\ &\hat{\mu}_i = C - \hat{\alpha}_i\end{aligned}$$

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，那么约束条件 $f\left(\boldsymbol{x}_i\right)-y_i-\epsilon-\xi_i=0$ 和 $y_i-f\left(\boldsymbol{x}_i\right)-\epsilon-\hat{\xi}_i=0$ 不可能同时成立，所以 α_i 和 $\hat{\alpha}_i$ 中至少有一个为 0，也即 $\alpha_i\hat{\alpha}_i=0$。在此基础上再进一步分析可知，如果 $\alpha_i=0$ 的话，那么根据约束 $(C-\alpha_i)\xi_i=0$ 可知此时 $\xi_i=0$，同理，如果 $\hat{\alpha}_i=0$ 的话，那么根据约束 $(C-\hat{\alpha}_i)\hat{\xi}_i=0$ 可知此时 $\hat{\xi}_i=0$，所以 ξ_i 和 $\hat{\xi}_i$ 中也是至少有一个为 0，也即 $\xi_i\hat{\xi}_i=0$。将 $\alpha_i\hat{\alpha}_i=0, \xi_i\hat{\xi}_i=0$ 整合进上述 KKT 条件中即可得到公式 (6.52)。

6.12 公式 (6.60)

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

[解析]：类似于第 3 章的公式 (3.35)。

6.13 公式 (6.62)

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

[解析]：类似于第 3 章的公式 (3.34)。

6.14 公式 (6.63)

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

[解析]：类似于第 3 章的公式 (3.33)。

6.15 公式 (6.65)

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

[推导]：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 $\kappa(\boldsymbol{x}, \boldsymbol{x}_i) = \phi(\boldsymbol{x})^{\mathrm{T}}\phi(\boldsymbol{x}_i)$ 代入可得

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

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

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

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

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

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

6.16 公式 (6.66)

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

[解析]：为了详细地说明此公式的计算原理，下面首先先举例说明，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：

$$m = 4$$

$$m_0 = 2, m_1 = 2$$

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

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

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

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

所以

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

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

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

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

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

6.17 公式 (6.67)

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

[解析]：参见公式 (6.66) 的解析。

6.18 公式 (6.70)

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

[推导]：此公式是将公式 (6.65) 代入公式 (6.60) 后推得而来的，下面给出详细地推导过程。首先将公式 (6.65) 代入公式 (6.60) 的分子可得：

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

其中

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

将其代入上式可得

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

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,也即 $\kappa\left(\boldsymbol{x}_i,\boldsymbol{x}\right)=\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x})=\left(\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x})\right)^{\mathrm{T}}=\phi(\boldsymbol{x})^{\mathrm{T}}\phi(\boldsymbol{x}_i)=\kappa\left(\boldsymbol{x}_i,\boldsymbol{x}\right)^{\mathrm{T}}$，将其代入上式可得

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

令 $\boldsymbol{\alpha}=(\alpha_1;\alpha_2;...;\alpha_m)^{\mathrm{T}}\in\mathbb{R}^{m\times 1}$，同时结合公式 (6.66) 的解析中得到的 $\hat{\boldsymbol{\mu}}_0,\hat{\boldsymbol{\mu}}_1$ 的一般形式，上式可以化简为

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

以上便是公式 (6.70) 分子部分的推导，下面继续推导公式 (6.70) 的分母部分。将公式 (6.65) 代入公式 (6.60) 的分母可得:

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

其中

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

由于

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

所以

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

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

$$\begin{aligned}\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w^{\phi}\boldsymbol{w} &= \sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\cdot\mathbf{S}_w^{\phi}\cdot\sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right) \\ &= \sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\cdot\left(\sum_{\boldsymbol{x}\in D}\phi(\boldsymbol{x})\phi(\boldsymbol{x})^{\mathrm{T}} - m_0\boldsymbol{\mu}_0^{\phi}\left(\boldsymbol{\mu}_0^{\phi}\right)^{\mathrm{T}} - m_1\boldsymbol{\mu}_1^{\phi}\left(\boldsymbol{\mu}_1^{\phi}\right)^{\mathrm{T}}\right)\cdot\sum_{i=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{\boldsymbol{x}\in D}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\phi(\boldsymbol{x})\phi(\boldsymbol{x})^{\mathrm{T}}\alpha_j\phi\left(\boldsymbol{x}_j\right) - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}m_0\boldsymbol{\mu}_0^{\phi}\left(\boldsymbol{\mu}_0^{\phi}\right)^{\mathrm{T}}\alpha_j\phi\left(\boldsymbol{x}_j\right) \\ &\quad - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}m_1\boldsymbol{\mu}_1^{\phi}\left(\boldsymbol{\mu}_1^{\phi}\right)^{\mathrm{T}}\alpha_j\phi\left(\boldsymbol{x}_j\right)\end{aligned}$$

其中，第 1 项可化简为

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

第 2 项可化简为

$$\begin{aligned}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}m_0\boldsymbol{\mu}_0^{\phi}\left(\boldsymbol{\mu}_0^{\phi}\right)^{\mathrm{T}}\alpha_j\phi\left(\boldsymbol{x}_j\right) &= m_0\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\boldsymbol{\mu}_0^{\phi}\left(\boldsymbol{\mu}_0^{\phi}\right)^{\mathrm{T}}\phi\left(\boldsymbol{x}_j\right) \\ &= m_0\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\left[\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\phi(\boldsymbol{x})\right]\left[\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\phi(\boldsymbol{x})\right]^{\mathrm{T}}\phi\left(\boldsymbol{x}_j\right) \\ &= m_0\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\left[\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\phi(\boldsymbol{x})\right]\left[\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\phi(\boldsymbol{x})^{\mathrm{T}}\phi\left(\boldsymbol{x}_j\right)\right] \\ &= m_0\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\left[\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\kappa\left(\boldsymbol{x}_i,\boldsymbol{x}\right)\right]\left[\frac{1}{m_0}\sum_{\boldsymbol{x}\in X_0}\kappa\left(\boldsymbol{x}_j,\boldsymbol{x}\right)\right] \\ &= m_0\boldsymbol{\alpha}^{\mathrm{T}}\hat{\boldsymbol{\mu}}_0\hat{\boldsymbol{\mu}}_0^{\mathrm{T}}\boldsymbol{\alpha}\end{aligned}$$

同理可得，第 3 项可化简为

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

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

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

6.19 附录

①KKT 条件 [3]

对于一般地约束优化问题

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

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（constraint qualifications or regularity conditions）[1]，则一定存在 $\boldsymbol{\mu}^* = (\mu_1^*, \mu_2^*, ..., \mu_m^*)^T, \boldsymbol{\lambda}^* = (\lambda_1^*, \lambda_2^*, ..., \lambda_n^*)^T$，使得

$$\begin{cases}\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) = \nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m}\mu_i^*\nabla g_i(\boldsymbol{x}^*) + \sum_{j=1}^{n}\lambda_j^*\nabla h_j(\boldsymbol{x}^*) = 0 & (1) \\ h_j(\boldsymbol{x}^*) = 0 & (2) \\ g_i(\boldsymbol{x}^*) \leq 0 & (3) \\ \mu_i^* \geq 0 & (4) \\ \mu_i^* g_i(\boldsymbol{x}^*) = 0 & (5)\end{cases}$$

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 条即为 KKT 条件，严格数学证明参见参考文献 [3] 的 § 4.2.1。

参考文献

[1] Wikipedia contributors. Karush–kuhn–tucker conditions, 2020. URL: https://en.wikipedia.org/w/index.php?title=Karush%E2%80%93Kuhn%E2%80%93Tucker_conditions&oldid=936587706.

[2] 王书宁. 凸优化. 清华大学出版社, 2013.

[3] 王燕军. 最优化基础理论与方法. 复旦大学出版社, 2011.

第 7 章　贝叶斯分类器

7.1　公式 (7.5)

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

[推导]：由公式 (7.1) 和公式 (7.4) 可得：

$$R(c_i|\boldsymbol{x}) = 1 * P(c_1|\boldsymbol{x}) + ... + 1 * P(c_{i-1}|\boldsymbol{x}) + 0 * P(c_i|\boldsymbol{x}) + 1 * P(c_{i+1}|\boldsymbol{x}) + ... + 1 * P(c_N|\boldsymbol{x})$$

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

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

此即为公式 (7.5)

7.2　公式 (7.6)

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

[推导]：将公式 (7.5) 带入公式 (7.3) 即可推得此式。

7.3　公式 (7.12)

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

[推导]：参见公式 (7.13)

7.4　公式 (7.13)

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

[推导]：根据公式 (7.11) 和公式 (7.10) 可知参数求解公式为

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

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

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

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。将其代入参数求解公式可得

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

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

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

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

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

所以

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

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，上式最后一项取值的大小此时仅与 $\boldsymbol{\mu}_c - \bar{\boldsymbol{x}}$ 相关，而且当且仅当 $\boldsymbol{\mu}_c - \bar{\boldsymbol{x}} = \mathbf{0}$ 时，上式最后一项取到最小值 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 $\hat{\boldsymbol{\mu}}_c$ 代回参数求解公式可得新的参数求解公式为

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

此时的参数求解公式是仅与 $\boldsymbol{\Sigma}_c$ 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（具体证明参见参考文献 [6]）：设 $\mathbf{B}$ 为 p 阶正定矩阵，$n > 0$ 为实数，在对所有 p 阶正定矩阵 $\boldsymbol{\Sigma}$ 有

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

当且仅当 $\boldsymbol{\Sigma} = \frac{1}{n}\mathbf{B}$ 时等号成立。所以根据此引理可知，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，上述参数求解公式中 $\arg\min$ 后面的式子取到最小值，那么此时的 $\boldsymbol{\Sigma}_c$ 即为我们想要求解的 $\hat{\boldsymbol{\Sigma}}_c$。

7.5 公式 (7.19)

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

[推导]：从贝叶斯估计（参见附录①）的角度来说，拉普拉斯修正就等价于先验概率为 Dirichlet 分布（参见附录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。设包含 m 个独立同分布样本的训练集为 D，D 中可能的类别数为 k，其类别的具体取值范围为 $\{c_1, c_2, ..., c_k\}$。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 $P(C = c_1) = \theta_1, P(C = c_2) = \theta_2, ..., P(C = c_k) = \theta_k$，那么显然 C 服从参数为 $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k) \in \mathbb{R}^k$ 的 Categorical 分布（参见附录②），其概率质量函数为

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

其中 $P(c_i) = \theta_i$ 就是公式 (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 [7] 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 [3] 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，所以此时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。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

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

那么后验概率 $P(D|\boldsymbol{\theta})$ 为

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

假设此时先验概率 $P(\boldsymbol{\theta})$ 是参数为 $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,...,\alpha_k)\in\mathbb{R}^k$ 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

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

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

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

将此结论代入 $P(D|\boldsymbol{\theta})$ 可得

$$\begin{aligned}
P(\boldsymbol{\theta}|D)&=\frac{\prod_{i=1}^k\theta_i^{\alpha_i+y_i-1}}{\sum_{\boldsymbol{\theta}}\left[\prod_{i=1}^k\theta_i^{\alpha_i+y_i-1}\right]}\\
&=\frac{\Gamma\left(\sum_{i=1}^k(\alpha_i+y_i)\right)}{\prod_{i=1}^k\Gamma(\alpha_i+y_i)}\prod_{i=1}^k\theta_i^{\alpha_i+y_i-1}\\
&=P(\boldsymbol{\theta};\boldsymbol{\alpha}+\boldsymbol{y})
\end{aligned}$$

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 [3]。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

$$\begin{aligned}\theta_i &= \mathbb{E}_{P(\boldsymbol{\theta}|D)}[\theta_i] \\ &= \mathbb{E}_{P(\boldsymbol{\theta};\boldsymbol{\alpha}+\boldsymbol{y})}[\theta_i] \\ &= \frac{\alpha_i + y_i}{\sum_{j=1}^{k}(\alpha_j + y_j)} \\ &= \frac{\alpha_i + y_i}{\sum_{j=1}^{k}\alpha_j + \sum_{j=1}^{k} y_j} \\ &= \frac{\alpha_i + y_i}{\sum_{j=1}^{k}\alpha_j + m}\end{aligned}$$

显然，公式 (7.9) 是当 $\boldsymbol{\alpha} = (1, 1, ..., 1)$ 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，此即为拉普拉斯修正。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。

7.6 公式 (7.20)

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

[推导]：参见公式 (7.19)

7.7 公式 (7.24)

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

[推导]：参见公式 (7.19)

7.8 公式 (7.25)

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

[推导]：参见公式 (7.20)

7.9 公式 (7.27)

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

[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

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

顺序结构：在给定节点 x 的条件下 y,z 独立

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

7.10 公式 (7.34)

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

[解析]：EM 算法这一节建议以李航老师的《统计学习方法》为主，西瓜书为辅进行学习。

7.11 附录

①贝叶斯估计 [8]

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的样本集为 $D=\{x_1,x_2,...,x_n\}$，则根据贝叶斯公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

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

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 $\hat{\theta}_{MAP}$ 作为 θ 的估计称为最大后验估计；求 $P(\theta|D)$ 的中位数 $\hat{\theta}_{Median}$ 作为 θ 的估计称为后验中位数估计；求 $P(\theta|D)$ 的期望值（均值）$\hat{\theta}_{Mean}$ 作为 θ 的估计称为后验期望值估计。

②Categorical 分布 [1]

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 $\{x_1,x_2,...,x_k\}$，且 X 取到每个值的概率分别为 $P(X=x_1)=\theta_1,P(X=x_2)=\theta_2,...,P(X=x_k)=\theta_k$，则称随机变量 X 服从参数为 $\theta_1,\theta_2,...,\theta_k$ 的 Categorical 分布，其概率质量函数为

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

③Dirichlet 分布 [4]

类似于 Categorical 分布是伯努利分布的泛化形式，Dirichlet 分布是 Beta 分布 [2] 的泛化形式。对于一个 k 维随机变量 $\boldsymbol{x}=(x_1,x_2,...,x_k)\in\mathbb{R}^k$，其中 $x_i(i=1,2,...,k)$ 满足 $0\leqslant x_i\leqslant 1,\sum_{i=1}^k x_i=1$，若 $\boldsymbol{x}$ 服从参数为 $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,...,\alpha_k)\in\mathbb{R}^k$ 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 $\Gamma(z)=\int_0^\infty x^{z-1}e^{-x}dx$ 为 Gamma 函数 [5]，当 $\boldsymbol{\alpha}=(1,1,...,1)$ 时，Dirichlet 分布等价于均匀分布。

参考文献

[1] Wikipedia contributors. Categorical distribution, 2019. URL: https://en.wikipedia.org/w/index.php?title=Categorical_distribution&oldid=905316786.

[2] Wikipedia contributors. Beta distribution, 2020. URL: https://en.wikipedia.org/w/index.php?title=Beta_distribution&oldid=953406542.

[3] Wikipedia contributors. Conjugate prior, 2020. URL: https://en.wikipedia.org/w/index.php?title=Conjugate_prior&oldid=946918786.

[4] Wikipedia contributors. Dirichlet distribution, 2020. URL: https://en.wikipedia.org/w/index.php?title=Dirichlet_distribution&oldid=954542610.

[5] Wikipedia contributors. Gamma function, 2020. URL: https://en.wikipedia.org/w/index.php?title=Gamma_function&oldid=953059892.

[6] 张伟平. 多元正态分布参数的估计和数据的清洁与变换, 2020. URL: http://staff.ustc.edu.cn/~zwp/teach/MVA/Lec5_slides.pdf.

[7] 百度百科. 共轭先验分布, 2020. URL: https://baike.baidu.com/item/%E5%85%B1%E8%BD%AD%E5%85%88%E9%AA%8C%E5%88%86%E5%B8%83.

[8] 陈希孺. 概率论与数理统计. 中国科学技术大学出版社, 2009.

第 8 章　集成学习

8.1　公式 (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 的基分类器给 x 的预测标记，$f(\boldsymbol{x})$ 是 x 的真实标记，它们之间不一致的概率记为 ϵ。

8.2　公式 (8.2)

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

[解析]：$h_i(\boldsymbol{x})$ 当把 $\boldsymbol{x}$ 分成 1 时，$h_i(\boldsymbol{x}) = 1$，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 sign，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。

8.3　公式 (8.3)

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

[推导]：由基分类器相互独立，假设随机变量 X 为 T 个基分类器分类正确的次数，因此随机变量 X 服从二项分布：$\mathrm{X} \sim \mathcal{B}(\mathrm{T}, 1-\epsilon)$，设 x_i 为每一个分类器分类正确的次数，则 $x_i \sim \mathcal{B}(1, 1-\epsilon)$ $i = 1\ 2\ 3\ ...\ \mathrm{T}$，那么有

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

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

证明过程如下：

$$\begin{aligned} P(H(x) \neq f(x)) =& P(X \leq \lfloor T/2 \rfloor) \\ \leqslant& P(X \leq T/2) \\ =& P\left[X - (1-\epsilon)T \leqslant \frac{T}{2} - (1-\epsilon)T\right] \\ =& P\left[X - (1-\epsilon)T \leqslant -\frac{T}{2}(1-2\epsilon)]\right] \\ =& P\left[\sum_{i=1}^{\mathrm{T}} x_i - \sum_{i=1}^{\mathrm{T}} \mathbb{E}(x_i) \leqslant -\frac{T}{2}(1-2\epsilon)]\right] \\ =& P\left[\frac{1}{\mathrm{T}} \sum_{i=1}^{\mathrm{T}} x_i - \frac{1}{\mathrm{T}} \sum_{i=1}^{\mathrm{T}} \mathbb{E}(x_i) \leqslant -\frac{1}{2}(1-2\epsilon)]\right] \end{aligned}$$

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

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

8.4 公式 (8.4)

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

[解析]: 这个式子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（见式 8.18）和 α_T（见式 8.11）

8.5 公式 (8.5)

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

[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)，下面解释指数损失函数的意义:

1. 先考虑指数损失函数 $e^{-f(x)H(x)}$ 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 $e^{-f(\boldsymbol{x})H(\boldsymbol{x})}=e^{-|H(\boldsymbol{x})|}<1$，且 $|H(\boldsymbol{x})|$ 越大指数损失函数 $e^{-f(\boldsymbol{x})H(\boldsymbol{x})}$ 越小（这很合理: 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 $e^{-f(\boldsymbol{x})H(\boldsymbol{x})}=e^{|H(\boldsymbol{x})|}>1$，且 $|H(\boldsymbol{x})|$ 越大指数损失函数越大（这很合理: 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）

2. 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 中进行一次随机抽样，每个样本被取到的概率；$\mathbb{E}[\cdot]$ 为经典的期望，则综合起来 $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\cdot]$ 表示在概率分布 $\mathcal{D}$ 上的期望，可简单理解为对数据集 D 以概率 $\mathcal{D}$ 进行加权后的期望。即

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

8.6 公式 (8.6)

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

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

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

其中 $\mathcal{D}\left(\boldsymbol{x}_i\right)\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)=1\right)=P\left(f\left(\boldsymbol{x}_i\right)=1\mid\boldsymbol{x}_i\right)$ 可以这样理解：$\mathcal{D}(x_i)$ 表示在数据集 D 中进行一次随机抽样，样本 x_i 被取到的概率，$\mathcal{D}\left(\boldsymbol{x}_i\right)\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)=1\right)$ 表示在数据集 D 中进行一次随机抽样，使得 $f(x_i)=1$ 的样本 x_i 被抽到的概率，即为 $P\left(f\left(\boldsymbol{x}_i\right)=1\mid\boldsymbol{x}_i\right)$。

当对 $H(x_i)$ 求导时，求和号中只有含 x_i 项不为 0，由求导公式

$$\frac{\partial e^{-H(\boldsymbol{x})}}{\partial H(\boldsymbol{x})} = -e^{-H(\boldsymbol{x})} \qquad \frac{\partial e^{H(\boldsymbol{x})}}{\partial H(\boldsymbol{x})} = e^{H(\boldsymbol{x})}$$

有

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

8.7 公式 (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)。

8.8 公式 (8.8)

$$\begin{aligned}
\operatorname{sign}(H(\boldsymbol{x})) &= \operatorname{sign}\left(\frac{1}{2}\ln\frac{P(f(x)=1|\boldsymbol{x})}{P(f(x)=-1|\boldsymbol{x})}\right) \\
&= \begin{cases} 1, & P(f(x)=1|\boldsymbol{x}) > P(f(x)=-1|\boldsymbol{x}) \\ -1, & P(f(x)=1|\boldsymbol{x}) < P(f(x)=-1|\boldsymbol{x}) \end{cases} \\
&= \underset{y\in\{-1,1\}}{\arg\max} P(f(x)=y|\boldsymbol{x})
\end{aligned}$$

[解析]：第一行到第二行显然成立，第二行到第三行是利用了 $\arg\max$ 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，展开刚好是第二行的式子。

8.9 公式 (8.9)

$$\begin{aligned}\ell_{\exp}\left(\alpha_t h_t|\mathcal{D}_t\right) &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}\left[e^{-f(\boldsymbol{x})\alpha_t h_t(\boldsymbol{x})}\right] \\ &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}\left[e^{-\alpha_t}\mathbb{I}\left(f(\boldsymbol{x})=h_t(\boldsymbol{x})\right)+e^{\alpha_t}\mathbb{I}\left(f(\boldsymbol{x})\neq h_t(\boldsymbol{x})\right)\right] \\ &= e^{-\alpha_t}P_{\boldsymbol{x}\sim\mathcal{D}_t}\left(f(\boldsymbol{x})=h_t(\boldsymbol{x})\right)+e^{\alpha_t}P_{\boldsymbol{x}\sim\mathcal{D}_t}\left(f(\boldsymbol{x})\neq h_t(\boldsymbol{x})\right) \\ &= e^{-\alpha_t}\left(1-\epsilon_t\right)+e^{\alpha_t}\epsilon_t\end{aligned}$$

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

8.10 公式 (8.10)

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

[解析]：指数损失函数对 α_t 求偏导，为了得到使得损失函数取最小值时 α_t 的值。

8.11 公式 (8.11)

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

[解析]：令公式 (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。

8.12 公式 (8.12)

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

[解析]：将 $H_t(\boldsymbol{x})=H_{t-1}(\boldsymbol{x})+h_t(\boldsymbol{x})$ 带入公式 (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 是个常数的话，对后续结果也没有影响。

8.13 公式 (8.13)

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

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

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

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

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

8.14 公式 (8.14)

$$\begin{aligned}
h_t(\boldsymbol{x}) &= \arg\min_h \ell_{\exp}\left(H_{t-1}+h|\mathcal{D}\right) \\
&= \arg\min_h \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\left(1-f(\boldsymbol{x})h(\boldsymbol{x})+\frac{1}{2}\right)\right] \\
&= \arg\max_h \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}f(\boldsymbol{x})h(\boldsymbol{x})\right] \\
&= \arg\max_h \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\frac{e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}}{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\right]}f(\boldsymbol{x})h(\boldsymbol{x})\right]
\end{aligned}$$

[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。第二个式子到第三个式子是因为 $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\frac{3}{2}e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\right]$ 与 $h(\boldsymbol{x})$ 无关，是一个常数。第三个式子到最后一个式子是因为 $\frac{1}{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\frac{3}{2}e^{-f(\boldsymbol{x})H_{t-1}(\boldsymbol{x})}\right]}$ 与 $h(\boldsymbol{x})$ 无关因此可以引入进来。

8.15 公式 (8.16)

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

[推导]：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，可以理解为在数据集 D 上进行一次随机采样，样本 x 被抽到的概率是 $\mathcal{D}(x)$，那么符号 $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}$ 表示的是在概率分布 $\mathcal{D}$ 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

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

由式 (8.15) 可知

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

所以式 (8.16) 可以表示为

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

8.16 公式 (8.17)

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

[解析]：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，$\mathbb{I}(f(\boldsymbol{x}) \neq h(\boldsymbol{x})) = 1$，$f(\boldsymbol{x})h(\boldsymbol{x}) = -1$。

8.17 公式 (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]：由公式 (8.16) 和公式 (8.17) 有：

$$\begin{aligned} h_t(\boldsymbol{x}) &= \arg\max_h \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}[f(\boldsymbol{x})h(\boldsymbol{x})] \\ &= \arg\max_h (1 - 2\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}[\mathbb{I}(f(\boldsymbol{x}) \neq h(\boldsymbol{x}))]) \\ &= \arg\max_h (-2\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}[\mathbb{I}(f(\boldsymbol{x}) \neq h(\boldsymbol{x}))]) \\ &= \arg\min \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_t}[\mathbb{I}(f(\boldsymbol{x}) \neq h(\boldsymbol{x}))] \end{aligned}$$

8.18 公式 (8.19)

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

[解析]：boosting 算法是根据调整后的样本再去训练下一个基分类器，这就是“重赋权法”的样本分布的调整公式。

8.19 公式 (8.20)

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

[解析]：$\mathbb{I}\left(h_t(\boldsymbol{x}) = y\right)$ 表示对 T 个基学习器，每一个都判断结果是否与 y 一致，y 的取值一般是 -1 和 1，如果基学习器结果与 y 一致，则 $\mathbb{I}\left(h_t(\boldsymbol{x}) = y\right) = 1$，如果样本不在训练集内，则 $\mathbb{I}\left(\boldsymbol{x} \notin D_t\right) = 1$，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。

8.20 公式 (8.21)

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

[解析]：由 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。

8.21 公式 (8.22)

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

[解析]：对基分类器的结果进行简单的平均。

8.22 公式 (8.23)

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

[解析]：对基分类器的结果进行加权平均。

8.23 公式 (8.24)

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

[解析]：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 $\frac{1}{2}$，则选择该类别 j 为最终结果。

8.24 公式 (8.25)

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

[解析]：相比于其他类别，该类别 j 的基分类器的结果之和最大，则选择类别 j 为最终结果。

8.25 公式 (8.26)

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

[解析]：相比于其他类别，该类别 j 的基分类器的结果之和最大，则选择类别 j 为最终结果，与式 (8.25) 不同的是，该式在基分类器前面乘上一个权重系数，该系数大于等于 0，且 T 个权重之和为 1。

8.26 公式 (8.27)

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

[解析]：该式表示个体学习器结果与预测结果的差值的平方，即为个体学习器的“分歧”。

8.27 公式 (8.28)

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

[解析]：该式表示对各个个体学习器的“分歧”加权平均的结果，即集成的“分歧”。

8.28 公式 (8.29)

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

[解析]：该式表示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。

8.29 公式 (8.30)

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

[解析]：该式表示集成与真实值之间差值的平方，即集成的平方误差。

8.30 公式 (8.31)

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

[推导]：由 (8.28) 知

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

又因为

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

所以

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

8.31 公式 (8.32)

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

[解析]：$\int A\left(h_i|\boldsymbol{x}\right) p(\boldsymbol{x}) d\boldsymbol{x}$ 表示个体学习器在全样本上的“分歧”，$\sum_{i=1}^{T} w_i \int A\left(h_i|\boldsymbol{x}\right) p(\boldsymbol{x}) d\boldsymbol{x}$ 表示集成在全样本上的“分歧”，然后根据式 (8.31) 拆成误差的形式。

8.32 公式 (8.33)

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

[解析]：表示个体学习器在全样本上的泛化误差。

8.33 公式 (8.34)

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

[解析]：表示个体学习器在全样本上的分歧。

8.34 公式 (8.35)

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

[解析]：表示集成在全样本上的泛化误差。

8.35 公式 (8.36)

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

[解析]：$\bar{E}$ 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

9.1 公式 (9.5)

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

[解析]：给定两个集合 A 和 B，则 Jaccard 系数定义为如下公式

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

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四种情况：

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 系数为如下公式

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

由于聚类属于无监督学习，事先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，我们也无法知道聚类后的类别标记与参考模型的类别标记是如何对应的，况且聚类后的类别总数与参考模型的类别总数还可能不一样，因此只用单个样本无法衡量聚类性能的好坏。

由于外部指标的基本思想就是以参考模型的类别划分为参照，因此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来说这是一个好的聚类结果。

总的来说所有样本对中的两个样本共存在四种情况：

1. 样本对中的两个样本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. 样本对中的两个样本在聚类结果中属于同一个类，在参考模型中不属于同一个类
3. 样本对中的两个样本在聚类结果中不属于同一个类，在参考模型中属于同一个类
4. 样本对中的两个样本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

综上所述，即所有样本对存在着书中公式 (9.1)-(9.4) 的四种情况，现在假设集合 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：

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

也可直接将书中公式 (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

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

9.2 公式 (9.6)

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

[解析]：其中 $\frac{a}{a+b}$ 和 $\frac{a}{a+c}$ 为 Wallace 提出的两个非对称指标，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个数，$a+b$ 代表两个样本在聚类结果中属于同一类的样本对的个数，$a+c$ 代表两个样本在参考模型中属于同一类的样本对的个数，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一样，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 Fowlkes and Mallows Index, FMI。

9.3 公式 (9.7)

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

[解析]：Rand Index 定义如下：

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

其可以理解为两个样本都属于聚类结果和参考模型中的同一类的样本对的个数与两个样本都分别不属于聚类结果和参考模型中的同一类的样本对的个数的总和在所有样本对中出现的频率，可以简单理解为聚类结果与参考模型的一致性。

9.4 公式 (9.8)

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

[解析]：簇内距离的定义式：求和号左边是 (x_i, x_j) 组合个数的倒数，求和号右边是这些组合的距离和，所以两者相乘定义为平均距离。

9.5 公式 (9.33)

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

[推导]：根据公式 (9.28) 可知：

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

又根据公式 (9.32)，由

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

其中：

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

$$\begin{aligned}\frac{\partial p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\partial \boldsymbol{\mu}_i} &= \frac{\partial \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\top} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\right)}{\partial \boldsymbol{\mu}_i} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}\left|\boldsymbol{\Sigma}_i\right|^{\frac{1}{2}}} \cdot \frac{\partial \exp\left(-\frac{1}{2}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\top} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\right)}{\partial \boldsymbol{\mu}_i} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}\left|\boldsymbol{\Sigma}_i\right|^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\top} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\right) \cdot -\frac{1}{2}\frac{\partial\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\top} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)}{\partial \boldsymbol{\mu}_i} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}\left|\boldsymbol{\Sigma}_i\right|^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\top} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\right) \cdot \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right) \\ &= p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \cdot \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\end{aligned}$$

其中，由矩阵求导的法则 $\frac{\partial \mathbf{a}^T \mathbf{X}\mathbf{a}}{\partial \mathbf{a}} = 2\mathbf{X}\mathbf{a}$ 可得：

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

因此有：

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

9.6 公式 (9.34)

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

[推导]：由式 9.30

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

带入 9.33

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

因此有

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

9.7 公式 (9.35)

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

[推导]：根据公式 (9.28) 可知：

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

又根据公式 (9.32)，由

$$\frac{\partial LL(D)}{\partial\boldsymbol{\Sigma}_i}=0$$

可得

$$\begin{aligned}\frac{\partial LL(D)}{\partial\boldsymbol{\Sigma}_i}&=\frac{\partial}{\partial\boldsymbol{\Sigma}_i}\left[\sum_{j=1}^{m}\ln\left(\sum_{i=1}^{k}\alpha_i\cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\right)\right]\\&=\sum_{j=1}^{m}\frac{\partial}{\partial\boldsymbol{\Sigma}_i}\left[\ln\left(\sum_{i=1}^{k}\alpha_i\cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\right)\right]\\&=\sum_{j=1}^{m}\frac{\alpha_i\cdot\dfrac{\partial}{\partial\boldsymbol{\Sigma}_i}(p(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i))}{\sum_{l=1}^{k}\alpha_l\cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)}\end{aligned}$$

其中

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

由矩阵微分公式 $\dfrac{\partial|\mathbf{X}|}{\partial\mathbf{X}}=|\mathbf{X}|\cdot(\mathbf{X}^{-1})^T$, $\dfrac{\partial\boldsymbol{a}^T\mathbf{X}^{-1}\mathbf{B}}{\partial\mathbf{X}}=-\mathbf{X}^{-T}\boldsymbol{a}\boldsymbol{b}^T\mathbf{X}^{-T}$ 可得

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

将此式代回 $\dfrac{\partial LL(D)}{\partial\boldsymbol{\Sigma}_i}$ 中可得

$$\frac{\partial LL(D)}{\partial\boldsymbol{\Sigma}_i}=\sum_{j=1}^{m}\frac{\alpha_i\cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k}\alpha_l\cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)}\cdot\left[-\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}+\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}\right]$$

又由公式 (9.30) 可知 $\dfrac{\alpha_i\cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k}\alpha_l\cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)}=\gamma_{ji}$，所以上式可进一步化简为

$$\frac{\partial LL(D)}{\partial\boldsymbol{\Sigma}_i}=\sum_{j=1}^{m}\gamma_{ji}\cdot\left[-\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}+\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}\right]$$

令上式等于 0 可得

$$\frac{\partial LL(D)}{\partial \boldsymbol{\Sigma}_i} = \sum_{j=1}^{m} \gamma_{ji} \cdot \left[-\frac{1}{2}\boldsymbol{\Sigma}_i^{-1} + \frac{1}{2}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} \right] = 0$$

移项推导有：

$$\begin{aligned}
\sum_{j=1}^{m} \gamma_{ji} \cdot \left[-\boldsymbol{I} + (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} \right] &= 0 \\
\sum_{j=1}^{m} \gamma_{ji} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} &= \sum_{j=1}^{m} \gamma_{ji} \boldsymbol{I} \\
\sum_{j=1}^{m} \gamma_{ji} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T &= \sum_{j=1}^{m} \gamma_{ji} \boldsymbol{\Sigma}_i \\
\boldsymbol{\Sigma}_i^{-1} \cdot \sum_{j=1}^{m} \gamma_{ji} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T &= \sum_{j=1}^{m} \gamma_{ji} \\
\boldsymbol{\Sigma}_i &= \frac{\sum_{j=1}^{m} \gamma_{ji} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T}{\sum_{j=1}^{m} \gamma_{ji}}
\end{aligned}$$

此即为公式 (9.35)。

9.8 公式 (9.38)

$$\alpha_i = \frac{1}{m} \sum_{j=1}^{m} \gamma_{ji}$$

[推导]：对公式 (9.37) 两边同时乘以 α_i 可得

$$\begin{aligned}
\sum_{j=1}^{m} \frac{\alpha_i \cdot p(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} + \lambda \alpha_i &= 0 \\
\sum_{j=1}^{m} \frac{\alpha_i \cdot p(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} &= -\lambda \alpha_i
\end{aligned}$$

两边对所有混合成分求和可得

$$\begin{aligned}
\sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\alpha_i \cdot p(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} &= -\lambda \sum_{i=1}^{k} \alpha_i \\
\sum_{j=1}^{m} \sum_{i=1}^{k} \frac{\alpha_i \cdot p(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} &= -\lambda \sum_{i=1}^{k} \alpha_i
\end{aligned}$$

因为

$$\sum_{i=1}^{k} \frac{\alpha_i \cdot p(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \ _i)}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j | \boldsymbol{\mu}_l, \ _l)} = \frac{\sum_{i=1}^{k} \alpha_i \cdot p(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \ _i)}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j | \boldsymbol{\mu}_l, \ _l)} = 1$$

且 $\sum_{i=1}^{k} \alpha_i = 1$，所以有 $m = -\lambda$，因此

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

因此

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

又由公式 (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

$$\alpha_i = \frac{1}{m}\sum_{j=1}^{m} \gamma_{ji}$$

此即为公式 (9.38)。

第 10 章　降维与度量学习

10.1　公式 (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 x 和 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 x 和 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 x 和 z 分属不同类别的概率。

10.2　公式 (10.2)

$$\begin{aligned} P(err) &= 1 - \sum_{c \in \mathcal{Y}} P(c|\boldsymbol{x}) P(c|\boldsymbol{z}) \\ &\simeq 1 - \sum_{c \in \mathcal{Y}} P^2(c|\boldsymbol{x}) \\ &\leqslant 1 - P^2\left(c^*|\boldsymbol{x}\right) \\ &= \left(1 + P\left(c^*|\boldsymbol{x}\right)\right)\left(1 - P\left(c^*|\boldsymbol{x}\right)\right) \\ &\leqslant 2 \times \left(1 - P\left(c^*|\boldsymbol{x}\right)\right) \end{aligned}$$

[解析]：第二个式子是来源于前提假设” 假设样本独立同分布，且对任意 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 $\boldsymbol{x}$ 同一维度的向量 $\boldsymbol{\delta}$ 则 $P(c|\boldsymbol{z}) = P(c|\boldsymbol{x} \pm \boldsymbol{\delta}) \simeq P(c|\boldsymbol{x})$。第三个式子是应为 $c^* \in \mathcal{Y}$，因此 $P^2\left(c^*|\boldsymbol{x}\right)$ 是 $\sum_{c \in \mathcal{Y}} P^2(c|\boldsymbol{x})$ 的一个分量，所以 $\sum_{c \in \mathcal{Y}} P^2(c|\boldsymbol{x}) \geqslant P^2\left(c^*|\boldsymbol{x}\right)$。第四个式子是平方差公式展开，最后一个式子因为 $1 + P\left(c^*|\boldsymbol{x}\right) \leqslant 2$。

10.3　公式 (10.3)

$$\begin{aligned} \mathrm{dist}_{ij}^2 &= \left\|\boldsymbol{z}_i\right\|^2 + \left\|\boldsymbol{z}_j\right\|^2 - 2\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j \\ &= b_{ii} + b_{jj} - 2b_{ij} \end{aligned}$$

[推导]：

$$\begin{aligned} \mathrm{dist}_{ij}^2 &= \left\|\boldsymbol{z}_i - \boldsymbol{z}_j\right\|^2 = \left(\boldsymbol{z}_i - \boldsymbol{z}_j\right)^{\top}\left(\boldsymbol{z}_i - \boldsymbol{z}_j\right) \\ &= \boldsymbol{z}_i^{\top}\boldsymbol{z}_i - \boldsymbol{z}_i^{\top}\boldsymbol{z}_j - \boldsymbol{z}_j^{\top}\boldsymbol{z}_i + \boldsymbol{z}_j^{\top}\boldsymbol{z}_j \\ &= \boldsymbol{z}_i^{\top}\boldsymbol{z}_i + \boldsymbol{z}_j^{\top}\boldsymbol{z}_j - 2\boldsymbol{z}_i^{\top}\boldsymbol{z}_j \\ &= \left\|\boldsymbol{z}_i\right\|^2 + \left\|\boldsymbol{z}_j\right\|^2 - 2\boldsymbol{z}_i^{\top}\boldsymbol{z}_j \\ &= b_{ii} + b_{jj} - 2b_{ij} \end{aligned}$$

10.4　公式 (10.4)

$$\sum_{i=1}^{m} dist_{ij}^2 = tr(\boldsymbol{B}) + mb_{jj}$$

[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，$\sum_{i=1}^{m} b_{ii} = tr(\boldsymbol{B})$，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，

$$\sum_{i=1}^{m} b_{ij} = \sum_{i=1}^{m} \boldsymbol{z}_i^\top \boldsymbol{z}_j = \sum_{i=1}^{m} \boldsymbol{z}_j^\top \boldsymbol{z}_i = \boldsymbol{z}_j^\top \sum_{i=1}^{m} \boldsymbol{z}_i = \boldsymbol{z}_j^\top \cdot \boldsymbol{0} = 0$$

其中 $\sum_{i=1}^{m} \boldsymbol{z}_i = \boldsymbol{0}$ 是利用了书上的前提条件，即将降维后的样本被中心化。

10.5 公式 (10.5)

$$\sum_{j=1}^{m} \text{dist}_{ij}^2 = \text{tr}(\mathbf{B}) + mb_{ii}$$

[解析]：参考 10.4

10.6 公式 (10.6)

$$\sum_{i=1}^{m}\sum_{j=1}^{m} \text{dist}_{ij}^2 = 2m\,\text{tr}(\mathbf{B})$$

[推导]：

$$\begin{aligned}\sum_{i=1}^{m}\sum_{j=1}^{m} \text{dist}_{ij}^2 &= \sum_{i=1}^{m}\sum_{j=1}^{m} \left(\|z_i\|^2 + \|\boldsymbol{z}_j\|^2 - 2\boldsymbol{z}_i^\top \boldsymbol{z}_j\right) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} \|\boldsymbol{z}_i\|^2 + \sum_{i=1}^{m}\sum_{j=1}^{m} \|\boldsymbol{z}_j\|^2 - 2\sum_{i=1}^{m}\sum_{j=1}^{m} \boldsymbol{z}_i^\top \boldsymbol{z}_j\end{aligned}$$

其中

$$\sum_{i=1}^{m}\sum_{j=1}^{m} \|\boldsymbol{z}_i\|^2 = m\sum_{i=1}^{m} \|\boldsymbol{z}_i\|^2 = m\,\text{tr}(\mathbf{B})$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} \|\boldsymbol{z}_j\|^2 = m\sum_{j=1}^{m} \|\boldsymbol{z}_j\|^2 = m\,\text{tr}(\mathbf{B})$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} \boldsymbol{z}_i^\top \boldsymbol{z}_j = 0$$

最后一个式子是来自于书中的假设，假设降维后的样本 $\mathbf{Z}$ 被中心化。

10.7 公式 (10.10)

$$b_{ij} = -\frac{1}{2}(dist_{ij}^2 - dist_{i\cdot}^2 - dist_{\cdot j}^2 + dist_{\cdot\cdot}^2)$$

[推导]：由公式（10.3）可得

$$b_{ij} = -\frac{1}{2}(dist_{ij}^2 - b_{ii} - b_{jj})$$

由公式（10.6）和（10.9）可得

$$\begin{aligned}tr(\boldsymbol{B}) &= \frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{m} dist_{ij}^2 \\ &= \frac{m}{2} dist_{\cdot}^2\end{aligned}$$

由公式（10.4）和（10.8）可得

$$\begin{aligned} b_{jj} &= \frac{1}{m}\sum_{i=1}^{m} dist_{ij}^2 - \frac{1}{m}tr(\boldsymbol{B}) \\ &= dist_{\cdot j}^2 - \frac{1}{2}dist_{\cdot}^2 \end{aligned}$$

由公式（10.5）和（10.7）可得

$$\begin{aligned} b_{ii} &= \frac{1}{m}\sum_{j=1}^{m} dist_{ij}^2 - \frac{1}{m}tr(\boldsymbol{B}) \\ &= dist_{i\cdot}^2 - \frac{1}{2}dist_{\cdot}^2 \end{aligned}$$

综合可得

$$\begin{aligned} b_{ij} &= -\frac{1}{2}(dist_{ij}^2 - b_{ii} - b_{jj}) \\ &= -\frac{1}{2}(dist_{ij}^2 - dist_{i\cdot}^2 + \frac{1}{2}dist_{\cdot\cdot}^2 - dist_{\cdot j}^2 + \frac{1}{2}dist_{\cdot\cdot}^2) \\ &= -\frac{1}{2}(dist_{ij}^2 - dist_{i\cdot}^2 - dist_{\cdot j}^2 + dist_{\cdot\cdot}^2) \end{aligned}$$

10.8 公式 (10.11)

$$\mathbf{Z} = \boldsymbol{\Lambda}_*^{1/2}\mathbf{V}_*^{\mathrm{T}} \in \mathbb{R}^{d^* \times m}$$

[解析]：由题设知，d^* 为 $\mathbf{V}$ 的非零特征值，因此 $\mathbf{B} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$ 可以写成 $\mathbf{B} = \mathbf{V}_*\boldsymbol{\Lambda}_*\mathbf{V}_*^\top$，其中 $\boldsymbol{\Lambda}_* \in \mathbb{R}^{d\times d}$ 为 d 个非零特征值构成的特征值对角矩阵，而 $\mathbf{V}_* \in \mathbb{R}^{m\times d}$ 为 $\boldsymbol{\Lambda}_* \in \mathbb{R}^{d\times d}$ 对应的特征值向量矩阵，因此有

$$\mathbf{B} = \left(\mathbf{V}_*\boldsymbol{\Lambda}_*^{1/2}\right)\left(\boldsymbol{\Lambda}_*^{1/2}\mathbf{V}_*^\top\right)$$

故而 $\mathbf{Z} = \boldsymbol{\Lambda}_*^{1/2}\mathbf{V}_*^\top \in \mathbb{R}^{d\times m}$

10.9 公式 (10.14)

$$\begin{aligned} \sum_{i=1}^{m}\left\|\sum_{j=1}^{d'} z_{ij}\boldsymbol{w}_j - \boldsymbol{x}_i\right\|_2^2 &= \sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i - 2\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i + \text{ const} \\ &\propto -\operatorname{tr}(\mathbf{W}^{\mathrm{T}}(\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}})\mathbf{W}) \end{aligned}$$

[推导]：已知 $\mathbf{W}^{\mathrm{T}}\mathbf{W}=\mathbf{I}, \boldsymbol{z}_i=\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i$，则

$$\begin{aligned}
\sum_{i=1}^{m}\left\|\sum_{j=1}^{d'} z_{ij}\boldsymbol{w}_j-\boldsymbol{x}_i\right\|_2^2 &= \sum_{i=1}^{m}\|\mathbf{W}\boldsymbol{z}_i-\boldsymbol{x}_i\|_2^2 \\
&= \sum_{i=1}^{m}\left(\mathbf{W}\boldsymbol{z}_i-\boldsymbol{x}_i\right)^{\mathrm{T}}\left(\mathbf{W}\boldsymbol{z}_i-\boldsymbol{x}_i\right) \\
&= \sum_{i=1}^{m}\left(\boldsymbol{z}_i^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\boldsymbol{z}_i-\boldsymbol{z}_i^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{x}_i^{\mathrm{T}}\mathbf{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}}\mathbf{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}}\mathbf{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}}\mathbf{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}\mathrm{tr}\left(\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}\right)+\ \text{const} \\
&= -\mathrm{tr}\left(\sum_{i=1}^{m}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}\right)+\ \text{const} \\
&= -\mathrm{tr}\left(\sum_{i=1}^{m}\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\mathbf{W}\right)+\ \text{const} \\
&= -\mathrm{tr}\left(\mathbf{W}^{\mathrm{T}}\left(\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right)\mathbf{W}\right)+\ \text{const} \\
&\propto -\mathrm{tr}\left(\mathbf{W}^{\mathrm{T}}\left(\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right)\mathbf{W}\right)
\end{aligned}$$

10.10 公式 (10.17)

$$\mathbf{X}\mathbf{X}^{\mathrm{T}}\boldsymbol{w}_i=\lambda_i\boldsymbol{w}_i$$

[推导]：由式（10.15）可知，主成分分析的优化目标为

$$\begin{aligned}
&\min_{\mathbf{W}} \quad -\mathrm{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}\right) \\
&s.t. \quad \mathbf{W}^{\mathrm{T}}\mathbf{W}=\mathbf{I}
\end{aligned}$$

其中，$\mathbf{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m)\in\mathbb{R}^{d\times m}, \mathbf{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_{d'})\in\mathbb{R}^{d\times d'}$，$\mathbf{I}\in\mathbb{R}^{d'\times d'}$ 为单位矩阵。对于带矩阵约束的优化问题，根据 [1] 中讲述的方法可得此优化目标的拉格朗日函数为

$$\begin{aligned}
L(\mathbf{W},\Theta) &= -\mathrm{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}\right)+\langle\Theta,\mathbf{W}^{\mathrm{T}}\mathbf{W}-\mathbf{I}\rangle \\
&= -\mathrm{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}\right)+\mathrm{tr}\left(\Theta^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\mathbf{W}-\mathbf{I})\right)
\end{aligned}$$

其中，$\Theta\in\mathbb{R}^{d'\times d'}$ 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，$\langle\Theta,\mathbf{W}^{\mathrm{T}}\mathbf{W}-\mathbf{I}\rangle=\mathrm{tr}\left(\Theta^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\mathbf{W}-\mathbf{I})\right)$ 为矩阵的内积 [2]。若此时仅考虑约束 $\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{w}_i=1(i=$

$1,2,...,d'$)，则拉格朗日乘子矩阵 Θ 此时为对角矩阵，令新的拉格朗日乘子矩阵为 $\mathbf{\Lambda}=\mathrm{diag}(\lambda_1,\lambda_2,...,\lambda_{d'})\in\mathbb{R}^{d'\times d'}$，则新的拉格朗日函数为

$$L(\mathbf{W},\mathbf{\Lambda})=-\ \mathrm{tr}\ (\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W})+\ \mathrm{tr}\ \left(\mathbf{\Lambda}^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\mathbf{W}-\mathbf{I})\right)$$

对拉格朗日函数关于 $\mathbf{W}$ 求导可得

$$\begin{aligned}\frac{\partial L(\mathbf{W},\mathbf{\Lambda})}{\partial\mathbf{W}}&=\frac{\partial}{\partial\mathbf{W}}\left[-\ \mathrm{tr}\ (\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W})+\ \mathrm{tr}\ \left(\mathbf{\Lambda}^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\mathbf{W}-\mathbf{I})\right)\right]\\&=-\frac{\partial}{\partial\mathbf{W}}\ \mathrm{tr}\ (\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W})+\frac{\partial}{\partial\mathbf{W}}\ \mathrm{tr}\ \left(\mathbf{\Lambda}^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\mathbf{W}-\mathbf{I})\right)\end{aligned}$$

由矩阵微分公式 $\frac{\partial}{\partial\mathbf{X}}\ \mathrm{tr}\ (\mathbf{X}^{\mathrm{T}}\mathbf{B}\mathbf{X})=\mathbf{B}\mathbf{X}+\mathbf{B}^{\mathrm{T}}\mathbf{X},\frac{\partial}{\partial\mathbf{X}}\ \mathrm{tr}\ \left(\mathbf{B}\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)=\mathbf{X}\mathbf{B}^{\mathrm{T}}+\mathbf{X}\mathbf{B}$ 可得

$$\begin{aligned}\frac{\partial L(\mathbf{W},\mathbf{\Lambda})}{\partial\mathbf{W}}&=-2\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}+\mathbf{W}\mathbf{\Lambda}+\mathbf{W}\mathbf{\Lambda}^{\mathrm{T}}\\&=-2\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}+\mathbf{W}(\mathbf{\Lambda}+\mathbf{\Lambda}^{\mathrm{T}})\\&=-2\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}+2\mathbf{W}\mathbf{\Lambda}\end{aligned}$$

令 $\frac{\partial L(\mathbf{W},\mathbf{\Lambda})}{\partial\mathbf{W}}=\mathbf{0}$ 可得

$$\begin{aligned}-2\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}+2\mathbf{W}\mathbf{\Lambda}&=\mathbf{0}\\\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}&=\mathbf{W}\mathbf{\Lambda}\end{aligned}$$

将 $\mathbf{W}$ 和 $\mathbf{\Lambda}$ 展开可得

$$\mathbf{X}\mathbf{X}^{\mathrm{T}}\boldsymbol{w}_i=\lambda_i\boldsymbol{w}_i,\quad i=1,2,...,d'$$

显然，此式为矩阵特征值和特征向量的定义式，其中 $\lambda_i,\boldsymbol{w}_i$ 分别表示矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ 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 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ 的定义可知，$\mathbf{X}\mathbf{X}^{\mathrm{T}}$ 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 $\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{w}_i=1,\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{w}_j=0(i\neq j)$。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 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ 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 $\mathbf{X}\mathbf{X}^{\mathrm{T}}\boldsymbol{w}_i=\lambda_i\boldsymbol{w}_i$ 代入目标函数可得

$$\begin{aligned}\min_{\mathbf{W}}-\ \mathrm{tr}\ (\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W})&=\max_{\mathbf{W}}\ \mathrm{tr}\ (\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W})\\&=\max_{\mathbf{W}}\sum_{i=1}^{d'}\boldsymbol{w}_i^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\boldsymbol{w}_i\\&=\max_{\mathbf{W}}\sum_{i=1}^{d'}\boldsymbol{w}_i^{\mathrm{T}}\cdot\lambda_i\boldsymbol{w}_i\\&=\max_{\mathbf{W}}\sum_{i=1}^{d'}\lambda_i\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{w}_i\\&=\max_{\mathbf{W}}\sum_{i=1}^{d'}\lambda_i\end{aligned}$$

显然，此时只需要令 $\lambda_1,\lambda_2,...,\lambda_{d'}$ 和 $\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_{d'}$ 分别为矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ 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。

10.11 公式 (10.24)

$$\mathbf{K}\boldsymbol{\alpha}^j = \lambda_j \boldsymbol{\alpha}^j$$

[推导]：已知 $\boldsymbol{z}_i = \phi(\boldsymbol{x}_i)$，类比 $\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m\}$ 可以构造 $\mathbf{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_m\}$，所以公式 (10.21) 可变换为

$$\left(\sum_{i=1}^{m} \phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^{\mathrm{T}}\right)\boldsymbol{w}_j = \left(\sum_{i=1}^{m} \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}}\right)\boldsymbol{w}_j = \mathbf{Z}\mathbf{Z}^{\mathrm{T}}\boldsymbol{w}_j = \lambda_j \boldsymbol{w}_j$$

又由公式 (10.22) 可知

$$\boldsymbol{w}_j = \sum_{i=1}^{m} \phi\left(\boldsymbol{x}_i\right)\alpha_i^j = \sum_{i=1}^{m} \boldsymbol{z}_i \alpha_i^j = \mathbf{Z}\boldsymbol{\alpha}^j$$

其中，$\boldsymbol{\alpha}^j = (\alpha_1^j; \alpha_2^j; ...; \alpha_m^j) \in \mathbb{R}^{m\times 1}$。所以公式 (10.21) 可以进一步变换为

$$\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\boldsymbol{\alpha}^j = \lambda_j \mathbf{Z}\boldsymbol{\alpha}^j$$
$$\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\boldsymbol{\alpha}^j = \mathbf{Z}\lambda_j \boldsymbol{\alpha}^j$$

由于此时的目标是要求出 $\boldsymbol{w}_j$，也就等价于要求出满足上式的 $\boldsymbol{\alpha}^j$，显然，此时满足 $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\boldsymbol{\alpha}^j = \lambda_j \boldsymbol{\alpha}^j$ 的 $\boldsymbol{\alpha}^j$ 一定满足上式，所以问题转化为了求解满足下式的 $\boldsymbol{\alpha}^j$：

$$\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\boldsymbol{\alpha}^j = \lambda_j \boldsymbol{\alpha}^j$$

令 $\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \mathbf{K}$，那么上式可化为

$$\mathbf{K}\boldsymbol{\alpha}^j = \lambda_j \boldsymbol{\alpha}^j$$

此式即为公式 (10.24)，其中矩阵 $\mathbf{K}$ 的第 i 行第 j 列的元素 $(\mathbf{K})_{ij} = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j = \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j) = \kappa\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$

10.12 公式 (10.28)

$$w_{ij} = \frac{\sum\limits_{k\in Q_i} C_{jk}^{-1}}{\sum\limits_{l,s\in Q_i} C_{ls}^{-1}}$$

[推导]：由书中上下文可知，式 (10.28) 是如下优化问题的解。

$$\min_{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_m} \sum_{i=1}^{m} \left\| \boldsymbol{x}_i - \sum_{j\in Q_i} w_{ij}\boldsymbol{x}_j \right\|_2^2$$
$$\text{s.t.} \sum_{j\in Q_i} w_{ij} = 1$$

若令 $\boldsymbol{x}_i \in \mathbb{R}^{d\times 1}, Q_i = \{q_i^1, q_i^2, ..., q_i^n\}$，则上述优化问题的目标函数可以进行如下恒等变形

$$\begin{aligned}
\sum_{i=1}^{m} \left\| \boldsymbol{x}_i - \sum_{j\in Q_i} w_{ij}\boldsymbol{x}_j \right\|_2^2 &= \sum_{i=1}^{m} \left\| \sum_{j\in Q_i} w_{ij}\boldsymbol{x}_i - \sum_{j\in Q_i} w_{ij}\boldsymbol{x}_j \right\|_2^2 \\
&= \sum_{i=1}^{m} \left\| \sum_{j\in Q_i} w_{ij}(\boldsymbol{x}_i - \boldsymbol{x}_j) \right\|_2^2 \\
&= \sum_{i=1}^{m} \left\| \mathbf{X}_i \boldsymbol{w_i} \right\|_2^2 \\
&= \sum_{i=1}^{m} \boldsymbol{w_i}^{\mathrm{T}} \mathbf{X}_i^{\mathrm{T}} \mathbf{X}_i \boldsymbol{w_i}
\end{aligned}$$

其中 $\boldsymbol{w_i}=(w_{iq_i^1},w_{iq_i^2},...,w_{iq_i^n})\in\mathbb{R}^{n\times 1}$，$\mathbf{X}_i=\left(\boldsymbol{x}_i-\boldsymbol{x}_{q_i^1},\boldsymbol{x}_i-\boldsymbol{x}_{q_i^2},...,\boldsymbol{x}_i-\boldsymbol{x}_{q_i^n}\right)\in\mathbb{R}^{d\times n}$。同理，约束条件也可以进行如下恒等变形

$$\sum_{j\in Q_i}w_{ij}=\boldsymbol{w_i}^{\mathrm{T}}\boldsymbol{I}=1$$

其中 $\boldsymbol{I}=(1,1,...,1)\in\mathbb{R}^{n\times 1}$ 为 n 行 1 列的元素值全为 1 的向量。因此，上述优化问题可以重写为

$$\min_{\boldsymbol{w}_1,\boldsymbol{w}_2,...,\boldsymbol{w}_m}\sum_{i=1}^{m}\boldsymbol{w_i}^{\mathrm{T}}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\boldsymbol{w_i}$$
$$\text{s.t. }\boldsymbol{w_i}^{\mathrm{T}}\boldsymbol{I}=1$$

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为

$$L(\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_m,\lambda)=\sum_{i=1}^{m}\boldsymbol{w_i}^{\mathrm{T}}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\boldsymbol{w_i}+\lambda\left(\boldsymbol{w_i}^{\mathrm{T}}\boldsymbol{I}-1\right)$$

对拉格朗日函数关于 $\boldsymbol{w_i}$ 求偏导并令其等于 0 可得

$$\begin{aligned}\frac{\partial L(\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_m,\lambda)}{\partial\boldsymbol{w_i}}&=\frac{\partial\left[\sum_{i=1}^{m}\boldsymbol{w_i}^{\mathrm{T}}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\boldsymbol{w_i}+\lambda\left(\boldsymbol{w_i}^{\mathrm{T}}\boldsymbol{I}-1\right)\right]}{\partial\boldsymbol{w_i}}=0\\&=\frac{\partial\left[\boldsymbol{w_i}^{\mathrm{T}}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\boldsymbol{w_i}+\lambda\left(\boldsymbol{w_i}^{\mathrm{T}}\boldsymbol{I}-1\right)\right]}{\partial\boldsymbol{w_i}}=0\end{aligned}$$

又由矩阵微分公式 $\frac{\partial\boldsymbol{x}^T\mathbf{B}\boldsymbol{x}}{\partial\boldsymbol{x}}=\left(\mathbf{B}+\mathbf{B}^{\mathrm{T}}\right)\boldsymbol{x},\frac{\partial\boldsymbol{x}^T\boldsymbol{a}}{\partial\boldsymbol{x}}=\boldsymbol{a}$ 可得

$$\frac{\partial\left[\boldsymbol{w_i}^{\mathrm{T}}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\boldsymbol{w_i}+\lambda\left(\boldsymbol{w_i}^{\mathrm{T}}\boldsymbol{I}-1\right)\right]}{\partial\boldsymbol{w_i}}=2\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\boldsymbol{w_i}+\lambda\boldsymbol{I}=0$$
$$\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\boldsymbol{w_i}=-\frac{1}{2}\lambda\boldsymbol{I}$$

若 $\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i$ 可逆，则

$$\boldsymbol{w_i}=-\frac{1}{2}\lambda(\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i)^{-1}\boldsymbol{I}$$

又因为 $\boldsymbol{w_i}^{\mathrm{T}}\boldsymbol{I}=\boldsymbol{I}^{\mathrm{T}}\boldsymbol{w_i}=1$，则上式两边同时左乘 $\boldsymbol{I}^{\mathrm{T}}$ 可得

$$\boldsymbol{I}^{\mathrm{T}}\boldsymbol{w_i}=-\frac{1}{2}\lambda\boldsymbol{I}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i)^{-1}\boldsymbol{I}=1$$
$$-\frac{1}{2}\lambda=\frac{1}{\boldsymbol{I}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i)^{-1}\boldsymbol{I}}$$

将其代回 $\boldsymbol{w_i}=-\frac{1}{2}\lambda(\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i)^{-1}\boldsymbol{I}$ 即可解得

$$\boldsymbol{w_i}=\frac{(\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i)^{-1}\boldsymbol{I}}{\boldsymbol{I}^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i)^{-1}\boldsymbol{I}}$$

若令矩阵 $(\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i)^{-1}$ 第 j 行第 k 列的元素为 C_{jk}^{-1}，则

$$w_{ij}=w_{iq_i^j}=\frac{\sum\limits_{k\in Q_i}C_{jk}^{-1}}{\sum\limits_{l,s\in Q_i}C_{ls}^{-1}}$$

此即为公式 (10.28)。显然，若 $\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i$ 可逆，此优化问题即为凸优化问题，且此时用拉格朗日乘子法求得的 $\boldsymbol{w_i}$ 为全局最优解。

10.13 公式 (10.31)

$$\min_{\boldsymbol{Z}} tr(\boldsymbol{Z}\boldsymbol{M}\boldsymbol{Z}^T)$$
$$s.t. \boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{I}.$$

[推导]:

$$\begin{aligned}
\min_{\boldsymbol{Z}} \sum_{i=1}^{m} \|\boldsymbol{z}_i - \sum_{j\in Q_i} w_{ij}\boldsymbol{z}_j\|_2^2 &= \sum_{i=1}^{m} \|\boldsymbol{Z}\boldsymbol{I}_i - \boldsymbol{Z}\boldsymbol{W}_i\|_2^2 \\
&= \sum_{i=1}^{m} \|\boldsymbol{Z}(\boldsymbol{I}_i - \boldsymbol{W}_i)\|_2^2 \\
&= \sum_{i=1}^{m} (\boldsymbol{Z}(\boldsymbol{I}_i - \boldsymbol{W}_i))^T\boldsymbol{Z}(\boldsymbol{I}_i - \boldsymbol{W}_i) \\
&= \sum_{i=1}^{m} (\boldsymbol{I}_i - \boldsymbol{W}_i)^T\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{I}_i - \boldsymbol{W}_i) \\
&= tr((\boldsymbol{I} - \boldsymbol{W})^T\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{I} - \boldsymbol{W})) \\
&= tr(\boldsymbol{Z}(\boldsymbol{I} - \boldsymbol{W})(\boldsymbol{I} - \boldsymbol{W})^T\boldsymbol{Z}^T) \\
&= tr(\boldsymbol{Z}\boldsymbol{M}\boldsymbol{Z}^T)
\end{aligned}$$

其中，$\boldsymbol{M} = (\boldsymbol{I} - \boldsymbol{W})(\boldsymbol{I} - \boldsymbol{W})^T$。[解析]：约束条件 $\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{I}$ 是为了得到标准化（标准正交空间）的低维数据。

参考文献

[1] Michael Grant. Lagrangian optimization with matrix constrains, 2015. URL: https://math.stackexchange.com/questions/1104376/how-to-set-up-lagrangian-optimization-with-matrix-constrains.

[2] Wikipedia contributors. Frobenius inner product, 2020. URL: https://en.wikipedia.org/wiki/Frobenius_inner_product.

第 11 章　特征选择与稀疏学习

11.1　公式 (11.1)

$$\mathrm{Gain}(A)=\mathrm{Ent}(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\mathrm{Ent}\left(D^v\right)$$

[解析]：此为信息熵的定义式，其中 $p_k,k=1,2,\ldots|\mathcal{Y}|$ 表示 D 中第 i 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（约定 $0\log_2 0=0$）。

11.2　公式 (11.2)

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

[解析]：此为信息熵的定义式，其中 $p_k,k=1,2,\ldots|\mathcal{Y}|$ 表示 D 中第 i 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。此时必有 $p_i=1,p_{\setminus i}=0,i=1,2,\ldots,|\mathcal{Y}|$。

11.3　公式 (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 $w^\top x_i$ 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大小。

11.4　公式 (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为了防止 w 的分量过大而导致过拟合的风险。

11.5　公式 (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 11.6 中的 L_2 正规化项替换成了 L_1 正规化项，也叫 LASSO 回归。关于 L_2 和 L_1 两个正规化项的区别，原书图 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。

11.6　公式 (11.10)

$$\begin{aligned}\hat{f}(\boldsymbol{x})&\simeq f\left(\boldsymbol{x}_k\right)+\left\langle\nabla f\left(\boldsymbol{x}_k\right),\boldsymbol{x}-\boldsymbol{x}_k\right\rangle+\frac{L}{2}\left\|\boldsymbol{x}-\boldsymbol{x}_k\right\|^2\\&=\frac{L}{2}\left\|\boldsymbol{x}-\left(\boldsymbol{x}_k-\frac{1}{L}\nabla f\left(\boldsymbol{x}_k\right)\right)\right\|_2^2+\mathrm{const}\end{aligned}$$

[解析]：首先注意优化目标式和 11.7 LASSO 回归的联系和区别，该式中的 x 对应到式 11.7 的 w，即我们优化的目标。再解释下什么是$L-$Lipschitz 条件，根据维基百科的定义：它是一个比通常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，在 wiki 百科的定义中，式 11.9 应该写成

$$|\nabla f(\boldsymbol{x}') - \nabla f(\boldsymbol{x})| \leqslant L|\boldsymbol{x}' - \boldsymbol{x}| \quad (\forall \boldsymbol{x}, \boldsymbol{x}')$$

移项得

$$\frac{|\nabla f(\boldsymbol{x}') - \nabla f(\boldsymbol{x})|}{|x' - x|} \leqslant L \quad (\forall \boldsymbol{x}, \boldsymbol{x}')$$

由于上式对所有的 x, x' 都成立，由导数的定义，上式可以看成是 $f(x)$ 的二阶导数恒不大于 L。即

$$\nabla^2 f(x) \leqslant L$$

得到这个结论之后，我们来推导式 11.10。由泰勒公式，x_k 附近的 $f(x)$ 通过二阶泰勒展开式可近似为

$$\begin{aligned}
\hat{f}(\boldsymbol{x}) &\simeq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k \rangle + \frac{\nabla^2 f(x_k)}{2} \|\boldsymbol{x} - \boldsymbol{x}_k\|^2 \\
&\leqslant f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k \rangle + \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{x}_k\|^2 \\
&= f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^\top (\boldsymbol{x} - \boldsymbol{x}_k) + \frac{L}{2} (\boldsymbol{x} - \boldsymbol{x}_k)^\top (\boldsymbol{x} - \boldsymbol{x}_k) \\
&= f(x_k) + \frac{L}{2} \left((\boldsymbol{x} - \boldsymbol{x}_k)^\top (\boldsymbol{x} - \boldsymbol{x}_k) + \frac{2}{L} \nabla f(\boldsymbol{x}_k)^\top (\boldsymbol{x} - \boldsymbol{x}_k) \right) \\
&= f(x_k) + \frac{L}{2} \left((\boldsymbol{x} - \boldsymbol{x}_k)^\top (\boldsymbol{x} - \boldsymbol{x}_k) + \frac{2}{L} \nabla f(\boldsymbol{x}_k)^\top (\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{L^2} \nabla f(x_k)^\top \nabla f(x_k) \right) - \frac{1}{2L} \nabla f(x_k)^\top \nabla f(x_k) \\
&= f(x_k) + \frac{L}{2} \left((\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{L} \nabla f(\boldsymbol{x}_k) \right)^\top \left((\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{L} \nabla f(\boldsymbol{x}_k) \right) - \frac{1}{2L} \nabla f(x_k)^\top \nabla f(x_k) \\
&= \frac{L}{2} \left\| \boldsymbol{x} - \left(\boldsymbol{x}_k - \frac{1}{L} \nabla f(\boldsymbol{x}_k) \right) \right\|_2^2 + \text{const}
\end{aligned}$$

其中 $\text{const} = f(x_k) - \frac{1}{2L} \nabla f(x_k)^\top \nabla f(x_k)$

11.7 公式 (11.11)

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{1}{L} \nabla f(\boldsymbol{x}_k)$$

[解析]：这个很容易理解，因为 2 范数的最小值为 0，当 $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{1}{L} \nabla f(\boldsymbol{x}_k)$ 时，$\hat{f}(x_{k+1}) \leqslant \hat{f}(x_k)$ 恒成立，同理 $\hat{f}(x_{k+2}) \leqslant \hat{f}(x_{k+1}), \cdots$，因此反复迭代能够使 $\hat{f}(x)$ 的值不断下降。

11.8 公式 (11.12)

$$\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x}} \frac{L}{2} \left\| \boldsymbol{x} - \left(\boldsymbol{x}_k - \frac{1}{L} \nabla f(\boldsymbol{x}_k) \right) \right\|_2^2 + \lambda \|\boldsymbol{x}\|_1$$

[解析]：式 11.11 是用来优化 $\hat{f}(x)$ 的，而对于式 11.8，优化的函数为 $f(x) + \lambda \|x\|_1$，由泰勒展开公式，优化的目标可近似为 $\hat{f}(x) + \lambda \|x\|_1$，根据式 11.10 可知，$x$ 的更新由式 11.12 决定。

11.9 公式 (11.13)

$$\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x}} \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2 + \lambda\|\boldsymbol{x}\|$$

[解析]：这里将式 11.12 的优化步骤拆分成了两步，首先令 $z = x_k - \frac{1}{L}\nabla f(x_k)$ 以计算 z，然后再求解式 11.13，得到的结果是一致的。

11.10 公式 (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]：令优化函数

$$\begin{aligned} g(\boldsymbol{x}) &= \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2 + \lambda\|\boldsymbol{x}\|_1 \\ &= \frac{L}{2}\sum_{i=1}^{d}\left\|x^i - z^i\right\|_2^2 + \lambda\sum_{i=1}^{d}\left\|x^i\right\|_1 \\ &= \sum_{i=1}^{d}\left(\frac{L}{2}\left(x^i - z^i\right)^2 + \lambda\left|x^i\right|\right) \end{aligned}$$

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

$$\frac{dg\left(x^i\right)}{dx^i} = L\left(x^i - z^i\right) + \lambda sgn\left(x^i\right)$$

其中

$$\text{sign}\left(x^i\right) = \begin{cases} 1, & x^i > 0 \\ -1, & x^i < 0 \end{cases}$$

称为符号函数 [1]，对于 $x_i = 0$ 的特殊情况，由于 $|x_i|$ 在 $x_i = 0$ 点出不光滑，所以其不可导，需单独讨论。令 $\frac{dg\left(x^i\right)}{dx^i} = 0$ 有

$$x^i = z^i - \frac{\lambda}{L}\text{sign}\left(x^i\right)$$

此式的解即为优化目标 $g(x^i)$ 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 $\text{sign}(x^i) = -1$，那么有 $x^i = z^i + \frac{\lambda}{L} > 0$ 与假设矛盾；b. 假设 $x^i > 0$，则 $\text{sign}(x^i) = 1$，那么有 $x^i = z^i - \frac{\lambda}{L} > 0$ 和假设相符和，下面来检验 $x^i = z^i - \frac{\lambda}{L}$ 是否是使函数 $g(x^i)$ 的取得最小值。当 $x^i > 0$ 时，

   $$\frac{dg\left(x^i\right)}{dx^i} = L\left(x^i - z^i\right) + \lambda$$

   在定义域内连续可导，则 $g(x^i)$ 的二阶导数

   $$\frac{d^2g\left(x^i\right)}{dx^{i^2}} = L$$

   由于 L 是 Lipschitz 常数恒大于 0，因为 $x^i = z^i - \frac{\lambda}{L}$ 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 $\text{sign}(x^i) = 1$，那么有 $x^i = z^i - \frac{\lambda}{L} < 0$ 与假设矛盾；b. 假设 $x^i < 0$，则 $\text{sign}(x^i) = -1$，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 $\text{sign}(x^i) = 1$，那么有 $x^i = z^i - \frac{\lambda}{L} \leqslant 0$ 与假设矛盾；b. 假设 $x^i < 0$，则 $\text{sign}(x^i) = -1$，那么有 $x^i = z^i + \frac{\lambda}{L} \geqslant 0$ 与假设矛盾。

4. 最后讨论 $x_i = 0$ 的情况，此时 $g(x^i) = \frac{L}{2}\left(z^i\right)^2$

   - 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

$$\begin{aligned} g(x^i)|_{x^i=0} - g(x^i)|_{x_i=z^i-\frac{\lambda}{L}} &= \frac{L}{2}\left(z^i\right)^2 - \left(\lambda z^i - \frac{\lambda^2}{2L}\right) \\ &= \frac{L}{2}\left(z^i - \frac{\lambda}{L}\right)^2 \\ &> 0 \end{aligned}$$

     因此当 $|z^i| > \frac{\lambda}{L}$ 时，$x_i = 0$ 不会是函数 $g(x_i)$ 的最小值。

   - 当 $-\frac{\lambda}{L} \leqslant z_i \leqslant \frac{\lambda}{L}$ 时，对于任何 $\Delta x \neq 0$ 有

$$\begin{aligned} g(\Delta x) &= \frac{L}{2}\left(\Delta x - z^i\right)^2 + \lambda|\Delta x| \\ &= \frac{L}{2}\left((\Delta x)^2 - 2\Delta x \cdot z^i + \frac{2\lambda}{L}|\Delta x|\right) + \frac{L}{2}\left(z^i\right)^2 \\ &\geq \frac{L}{2}\left((\Delta x)^2 - 2\Delta x \cdot z^i + \frac{2\lambda}{L}\Delta x\right) + \frac{L}{2}\left(z^i\right)^2 \\ &\geq \frac{L}{2}(\Delta x)^2 + \frac{L}{2}\left(z^i\right)^2 \\ &> g(x^i)|_{x^i=0} \end{aligned}$$

     因此 $x^i = 0$ 是 $g(x^i)$ 的最小值点。

综上所述，11.14 成立

11.11 公式 (11.15)

$$\min_{\mathbf{B},\boldsymbol{\alpha}_i} \sum_{i=1}^{m} \|\boldsymbol{x}_i - \mathbf{B}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{i=1}^{m} \|\boldsymbol{\alpha}_i\|_1$$

[解析]：这个式子表达的意思很容易理解，即希望样本 x_i 的稀疏表示 $\boldsymbol{\alpha}_i$ 通过字典 $\mathbf{B}$ 重构后和样本 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 范数项是为了使表示更加稀疏。

11.12 公式 (11.16)

$$\min_{\boldsymbol{\alpha}_i} \|\boldsymbol{x}_i - \mathbf{B}\boldsymbol{\alpha}_i\|_2^2 + \lambda \|\boldsymbol{\alpha}_i\|_1$$

[解析]：为了优化 11.15，我们采用变量交替优化的方式 (有点类似 EM 算法)，首先固定变量 $\mathbf{B}$，则 11.15 求解的是 m 个样本相加的最小值，因为公式里没有样本之间的交互 (即文中所述 $\alpha_i^u \alpha_i^v (u \neq v)$ 这样的形式)，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 11.13，11.14。

11.13 公式 (11.17)

$$\min_{\mathbf{B}} \|\mathbf{X}-\mathbf{B}\mathbf{A}\|_F^2$$

[解析]：这是优化 11.15 的第二步，固定住 $\boldsymbol{\alpha}_i, i=1,2,\dots,m$，此时式 11.15 的第二项为一个常数，优化 11.15 即优化 $\min_{\mathbf{B}} \sum_{i=1}^{m} \|\boldsymbol{x}_i-\mathbf{B}\boldsymbol{\alpha}_i\|_2^2$。其写成矩阵相乘的形式为 $\min_{\mathbf{B}} \|\mathbf{X}-\mathbf{B}\mathbf{A}\|_2^2$，将 2 范数扩展到 F 范数即得优化目标为 $\min_{\mathbf{B}} \|\mathbf{X}-\mathbf{B}\mathbf{A}\|_F^2$。

11.14 公式 (11.18)

$$\begin{aligned}\min_{\mathbf{B}} \|\mathbf{X}-\mathbf{B}\mathbf{A}\|_F^2 &= \min_{\boldsymbol{b}_i} \left\|\mathbf{X}-\sum_{j=1}^{k}\boldsymbol{b}_j\boldsymbol{\alpha}^j\right\|_F^2 \\ &= \min_{\boldsymbol{b}_i} \left\|\left(\mathbf{X}-\sum_{j\neq i}\boldsymbol{b}_j\boldsymbol{\alpha}^j\right)-\boldsymbol{b}_i\boldsymbol{\alpha}^i\right\|_F^2 \\ &= \min_{\boldsymbol{b}_i} \left\|\mathbf{E}_i-\boldsymbol{b}_i\boldsymbol{\alpha}^i\right\|_F^2\end{aligned}$$

[解析]：这个公式难点在于推导 $\mathbf{B}\mathbf{A}=\sum_{j=1}^{k}\boldsymbol{b}_j\boldsymbol{\alpha}^j$。大致的思路是 $\boldsymbol{b}_j\boldsymbol{\alpha}^j$ 会生成和矩阵 $\mathbf{B}\mathbf{A}$ 同样维度的矩阵，这个矩阵对应位置的元素是 $\mathbf{B}\mathbf{A}$ 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：

$$\begin{aligned}\boldsymbol{B}\boldsymbol{A} &= \begin{bmatrix} b_1^1 & b_2^1 & \cdot & \cdot & \cdot & b_k^1 \\ b_1^2 & b_2^2 & \cdot & \cdot & \cdot & b_k^2 \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ b_1^d & b_2^d & \cdot & \cdot & \cdot & b_k^d \end{bmatrix}_{d\times k} \cdot \begin{bmatrix} \alpha_1^1 & \alpha_2^1 & \cdot & \cdot & \cdot & \alpha_m^1 \\ \alpha_1^2 & \alpha_2^2 & \cdot & \cdot & \cdot & \alpha_m^2 \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ \alpha_1^k & \alpha_2^k & \cdot & \cdot & \cdot & \alpha_m^k \end{bmatrix}_{k\times m} \\ &= \begin{bmatrix} \sum_{j=1}^k b_j^1\alpha_1^j & \sum_{j=1}^k b_j^1\alpha_2^j & \cdot & \cdot & \cdot & \sum_{j=1}^k b_j^1\alpha_m^j \\ \sum_{j=1}^k b_j^2\alpha_1^j & \sum_{j=1}^k b_j^2\alpha_2^j & \cdot & \cdot & \cdot & \sum_{j=1}^k b_j^2\alpha_m^j \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ \sum_{j=1}^k b_j^d\alpha_1^j & \sum_{j=1}^k b_j^d\alpha_2^j & \cdot & \cdot & \cdot & \sum_{j=1}^k b_j^d\alpha_m^j \end{bmatrix}_{d\times m}\end{aligned}$$

$$\begin{aligned}\boldsymbol{b_j}\boldsymbol{\alpha^j} &= \begin{bmatrix} b_j^1 \\ b_j^2 \\ \cdot \\ \cdot \\ \cdot \\ b_j^d \end{bmatrix} \cdot \begin{bmatrix} \alpha_1^j & \alpha_2^j & \cdot & \cdot & \cdot & \alpha_m^j \end{bmatrix} \\ &= \begin{bmatrix} b_j^1\alpha_1^j & b_j^1\alpha_2^j & \cdot & \cdot & \cdot & b_j^1\alpha_m^j \\ b_j^2\alpha_1^j & b_j^2\alpha_2^j & \cdot & \cdot & \cdot & b_j^2\alpha_m^j \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ b_j^d\alpha_1^j & b_j^d\alpha_2^j & \cdot & \cdot & \cdot & b_j^d\alpha_m^j \end{bmatrix}_{d\times m}\end{aligned}$$

求和可得：

$$\sum_{j=1}^{k}\boldsymbol{b_j}\boldsymbol{\alpha^j}=\sum_{j=1}^{k}\left(\begin{bmatrix}b_j^1\\b_j^2\\\cdot\\\cdot\\\cdot\\b_j^d\end{bmatrix}\cdot\begin{bmatrix}\alpha_1^j & \alpha_2^j & \cdot & \cdot & \cdot & \alpha_m^j\end{bmatrix}\right)$$

$$=\begin{bmatrix}\sum_{j=1}^{k}b_j^1\alpha_1^j & \sum_{j=1}^{k}b_j^1\alpha_2^j & \cdot & \cdot & \cdot & \sum_{j=1}^{k}b_j^1\alpha_m^j\\\sum_{j=1}^{k}b_j^2\alpha_1^j & \sum_{j=1}^{k}b_j^2\alpha_2^j & \cdot & \cdot & \cdot & \sum_{j=1}^{k}b_j^2\alpha_m^j\\\cdot & \cdot & \cdot & & & \cdot\\\cdot & \cdot & & \cdot & & \cdot\\\cdot & \cdot & & & \cdot & \cdot\\\sum_{j=1}^{k}b_j^d\alpha_1^j & \sum_{j=1}^{k}b_j^d\alpha_2^j & \cdot & \cdot & \cdot & \sum_{j=1}^{k}b_j^d\alpha_m^j\end{bmatrix}_{d\times m}$$

得证。

将矩阵 $\mathbf{B}$ 分解成矩阵列 $\boldsymbol{b}_j, j=1,2,\ldots,k$ 带来一个好处，即和 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 $\min\limits_{\mathbf{B}}\|\ldots\mathbf{B}\ldots\|_F^2$ 转化成了 $\min\limits_{b_i}\|\ldots\boldsymbol{b}_i\ldots\|_F^2$，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。

参考文献

[1] Wikipedia contributors. Sign function, 2020. URL: https://en.wikipedia.org/wiki/Sign_function.

第 12 章 计算学习理论

12.1 公式 (12.1)

$$E(h;\mathcal{D}) = P_{\boldsymbol{x}\sim\mathcal{D}}(h(\boldsymbol{x}) \neq y)$$

[解析]：该式为泛化误差的定义式，所谓泛化误差，是指当样本 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 $\mathcal{D}$，我们拿到的数据集可以看做是从样本分布 $\mathcal{D}$ 中独立同分布采样得到的。在西瓜书中，我们拿到的数据集，称为样例集 D[也叫观测集、样本集，注意与花体 $\mathcal{D}$ 的区别]。

12.2 公式 (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]：该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 $x_i, i = 1, 2, \cdots, m$ 的预测值 $h(\boldsymbol{x}_i)$ 和真实值 y_i 的期望误差。

12.3 公式 (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。

12.4 公式 (12.4)

$$f(\mathbb{E}(x)) \leqslant \mathbb{E}(f(x))$$

[解析]：Jensen 不等式：这个式子可以做很直观的理解，比如说在二维空间上，凸函数可以想象成开口向上的抛物线，假如我们有两个点 x_1, x_2，那么 $f(\mathbb{E}(x))$ 表示的是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。

12.5 公式 (12.5)

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

[解析]：Hoeffding 不等式：对于独立随机变量 $x_1, x_2, \cdots, x_m$ 来说，他们观测值 x_i 的均值 $\frac{1}{m}\sum_{i=1}^{m}x_i$ 总是和他们期望 $\mathbb{E}(x_i)$ 的均值 $\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left(x_i\right)$ 相近，上式从概率的角度对这样一个结论进行了描述：即它们之间差值不小于 ϵ 这样的事件出现的概率不大于 $\exp\left(-2m\epsilon^2\right)$，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。

12.6 公式 (12.7)

$$P\left(f\left(x_1,\ldots,x_m\right)-\mathbb{E}\left(f\left(x_1,\ldots,x_m\right)\right)\geqslant\epsilon\right)\leqslant\exp\left(\frac{-2\epsilon^2}{\sum_i c_i^2}\right)$$

[解析]：McDiarmid 不等式：首先解释下前提条件：

$$\sup_{x_1,\ldots,x_m,x_i'}\left|f\left(x_1,\ldots,x_m\right)-f\left(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_m\right)\right|\leqslant c_i$$

表示当函数 f 某个输入 x_i 变到 x_i' 的时候，其变化的上确 sup 仍满足不大于 c_i。所谓上确界 sup 可以理解成变化的极限最大值，可能取到也可能无穷逼近。当满足这个条件时，McDiarmid 不等式指出：函数值 $f(x_1,\ldots,x_m)$ 和其期望值 $\mathbb{E}\left(f(x_1,\ldots,x_m)\right)$ 也相近，从概率的角度描述是：它们之间差值不小于 ϵ 这样的事件出现的概率不大于 $\exp\left(\frac{-2\epsilon^2}{\sum_i c_i^2}\right)$，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。

12.7 公式 (12.9)

$$P(E(h)\leq\epsilon)\geq 1-\delta$$

[解析]：PAC 辨识的定义：$E(h)$ 表示算法 $\mathcal{L}$ 在用观测集 D 训练后输出的假设函数 h，它的泛化误差 (见公式 12.1)。这个概率定义指出，如果 h 的泛化误差不大于 ϵ 的概率不小于 $1-\delta$，那么我们称学习算法 $\mathcal{L}$ 能从假设空间 $\mathcal{H}$ 中 PAC 辨识概念类 $\mathcal{C}$。

从式 12.10 到式 12.14 的公式是为了回答一个问题：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 $\mathcal{L}$ 以概率 $1-\delta$ 找到目标假设的 ϵ 近似即可。下面就是用数学公式进行抽象

12.8 公式 (12.10)

$$\begin{aligned}P(h(\boldsymbol{x})=y)&=1-P(h(\boldsymbol{x})\neq y)\\&=1-E(h)\\&<1-\epsilon\end{aligned}$$

[解析]：$P(h(\boldsymbol{x})=y)=1-P(h(\boldsymbol{x})\neq y)$ 因为它们是对立事件，$P(h(x)\neq y)=E(h)$ 是泛化误差的定义 (见 12.1)，由于我们假定了泛化误差 $E(h)>\epsilon$，因此有 $1-E(h)<1-\epsilon$。

12.9 公式 (12.11)

$$\begin{aligned}P\left(\left(h\left(\boldsymbol{x}_1\right)=y_1\right)\wedge\ldots\wedge\left(h\left(\boldsymbol{x}_m\right)=y_m\right)\right)&=\left(1-P(h(\boldsymbol{x})\neq y)\right)^m\\&<\left(1-\epsilon\right)^m\end{aligned}$$

[解析]：先解释什么是 h 与 D “表现一致”，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，我们称问题对学习算法是一致的。即 $\left(h\left(\boldsymbol{x}_1\right)=y_1\right)\wedge\ldots\wedge\left(h\left(\boldsymbol{x}_m\right)=y_m\right)$ 为 True。因为每个事件是独立的，所以上式可以写成 $P\left(\left(h\left(\boldsymbol{x}_1\right)=y_1\right)\wedge\ldots\wedge\left(h\left(\boldsymbol{x}_m\right)=y_m\right)\right)=\prod_{i=1}^m P\left(h\left(\boldsymbol{x}_i\right)=y_i\right)$。根据对立事件的定义有：$\prod_{i=1}^m P\left(h\left(\boldsymbol{x}_i\right)=y_i\right)=\prod_{i=1}^m\left(1-P\left(h\left(\boldsymbol{x}_i\right)\neq y_i\right)\right)$，又根据公式 (12.10)，有

$$\prod_{i=1}^m\left(1-P\left(h\left(\boldsymbol{x}_i\right)\neq y_i\right)\right)<\prod_{i=1}^m(1-\epsilon)=(1-\epsilon)^m$$

12.10 公式 (12.12)

$$\begin{aligned} P(h \in \mathcal{H} : E(h) > \epsilon \wedge \widehat{E}(h) = 0) &< |\mathcal{H}|(1-\epsilon)^m \\ &< |\mathcal{H}|e^{-m\epsilon} \end{aligned}$$

[解析]：首先解释为什么” 我们事先并不知道学习算法 $\mathcal{L}$ 会输出 $\mathcal{H}$ 中的哪个假设 “，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，比如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 $P(h \in \mathcal{H} : E(h) > \epsilon \wedge \widehat{E}(h) = 0)$，根据式 12.11，每一个这样的假设 h 都满足 $P(E(h) > \epsilon \wedge \widehat{E}(h) = 0) < (1-\epsilon)^m$，假设一共有 $|\mathcal{H}|$ 这么多个这样的假设 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

$$\begin{aligned} P\left(h \in \mathcal{H} : E(h) > \epsilon \wedge \widehat{E}(h) = 0\right) &= \sum_{i}^{|\mathcal{H}|} P\left(E(h_i) > \epsilon \wedge \widehat{E}(h_i) = 0\right) \\ &< |\mathcal{H}|(1-\epsilon)^m \end{aligned}$$

小于号依据公式 (12.11)。第二个小于号实际上是要证明 $|\mathcal{H}|(1-\epsilon)^m < |\mathcal{H}|e^{-m\epsilon}$，即证明 $(1-\epsilon)^m < e^{-m\epsilon}$，其中 $\epsilon \in (0,1]$，m 是正整数，推导如下：[推导]：当 $\epsilon = 1$ 时，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：令 $f(\epsilon) = \ln(1-\epsilon) + \epsilon$，其中 $\epsilon \in (0,1)$，$f'(\epsilon) = 1 - \frac{1}{1-\epsilon} = 0 \Rightarrow \epsilon = 0$ 取极大值 0，因此 $ln(1-\epsilon) < -\epsilon$ 也即 $|\mathcal{H}|(1-\epsilon)^m < |\mathcal{H}|e^{-m\epsilon}$ 成立。

12.11 公式 (12.13)

$$|\mathcal{H}|e^{-m\epsilon} \leqslant \delta$$

[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 $\mathcal{L}$ 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 $\mathcal{L}$ 生成的假设大于目标假设的 ϵ 近似的概率为 $P\left(h \in \mathcal{H} : E(h) > \epsilon \wedge \widehat{E}(h) = 0\right) < |\mathcal{H}|e^{-m\epsilon}$，因此学习算法 $\mathcal{L}$ 生成的假设落在目标假设的 ϵ 近似的概率为 $1 - P\left(h \in \mathcal{H} : E(h) > \epsilon \wedge \widehat{E}(h) = 0\right) \geq 1 - |\mathcal{H}|e^{-m\epsilon}$，这个概率我们希望至少是 $1-\delta$，因此 $1-\delta \leqslant 1 - |\mathcal{H}|e^{-m\epsilon} \Rightarrow |\mathcal{H}|e^{-m\epsilon} \leqslant \delta$

12.12 公式 (12.14)

$$m \geqslant \frac{1}{\epsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)$$

[推导]：

$$\begin{aligned} |\mathcal{H}|e^{-m\epsilon} &\leqslant \delta \\ e^{-m\epsilon} &\leqslant \frac{\delta}{|\mathcal{H}|} \\ -m\epsilon &\leqslant \ln\delta - \ln|\mathcal{H}| \\ m &\geqslant \frac{1}{\epsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right) \end{aligned}$$

[解析]：这个式子告诉我们，在假设空间 $\mathcal{H}$ 是 PAC 可学习的情况下，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。

12.13 公式 (12.15)

$$P(\widehat{E}(h)-E(h)\geqslant\epsilon)\leqslant\exp\left(-2m\epsilon^2\right)$$

[解析]：参见 12.5

12.14 公式 (12.16)

$$P(E(h)-\widehat{E}(h)\geqslant\epsilon)\leqslant\exp\left(-2m\epsilon^2\right)$$

[解析]：参见 12.5

12.15 公式 (12.17)

$$P(|E(h)-\widehat{E}(h)|\geqslant\epsilon)\leqslant2\exp\left(-2m\epsilon^2\right)$$

[解析]：参见 12.6

12.16 公式 (12.18)

$$\widehat{E}(h)-\sqrt{\frac{\ln(2/\delta)}{2m}}\leqslant E(h)\leqslant\widehat{E}(h)+\sqrt{\frac{\ln(2/\delta)}{2m}}$$

[推导]：令 $\delta=2e^{-2m\epsilon^2}$，则 $\epsilon=\sqrt{\frac{\ln(2/\delta)}{2m}}$，由式 12.17

$$\begin{aligned}
P(|E(h)-\widehat{E}(h)|\geqslant\epsilon)&\leqslant2\exp\left(-2m\epsilon^2\right)\\
P(|E(h)-\widehat{E}(h)|\geqslant\epsilon)&\leqslant\delta\\
P(|E(h)-\widehat{E}(h)|\leqslant\epsilon)&\geqslant1-\delta\\
P(-\epsilon\leqslant E(h)-\widehat{E}(h)\leqslant\epsilon)&\geqslant1-\delta\\
P(\widehat{E}(h)-\epsilon\leqslant E(h)\leqslant\widehat{E}(h)+\epsilon)&\geqslant1-\delta
\end{aligned}$$

带入 $\epsilon=\sqrt{\frac{\ln(2/\delta)}{2m}}$ 得证。这个式子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。

12.17 公式 (12.19)

$$P\left(|E(h)-\widehat{E}(h)|\leqslant\sqrt{\frac{\ln|\mathcal{H}|+\ln(2/\delta)}{2m}}\right)\geqslant1-\delta$$

[推导]：令 $h_1,h_2,\ldots,h_{|\mathcal{H}|}$ 表示假设空间 $\mathcal{H}$ 中的假设，有

$$\begin{aligned}
&P(\exists h\in\mathcal{H}:|E(h)-\widehat{E}(h)|>\epsilon)\\
=&P\left(\left(\left|E_{h_1}-\widehat{E}_{h_1}\right|>\epsilon\right)\vee\ldots\vee\left(|E_{h_{|\mathcal{H}|}}-\widehat{E}_{h_{|\mathcal{H}|}}|>\epsilon\right)\right)\\
\leqslant&\sum_{h\in\mathcal{H}}P(|E(h)-\widehat{E}(h)|>\epsilon)
\end{aligned}$$

这一步是很好理解的，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 $h_i, i\in 1,\dots,|\mathcal{H}|$，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：

$$\begin{aligned}&P(|E(h)-\widehat{E}(h)|\geqslant\epsilon)\leqslant 2\exp\left(-2m\epsilon^2\right)\\&\Rightarrow\sum_{h\in\mathcal{H}}P(|E(h)-\widehat{E}(h)|>\epsilon)\leqslant 2|\mathcal{H}|\exp\left(-2m\epsilon^2\right)\end{aligned}$$

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：

$$\begin{aligned}P(\forall h\in\mathcal{H}:|E(h)-\widehat{E}(h)|\leqslant\epsilon)&=1-P(\exists h\in\mathcal{H}:|E(h)-\widehat{E}(h)|>\epsilon)\\&\geqslant 1-2|\mathcal{H}|\exp\left(-2m\epsilon^2\right)\end{aligned}$$

令 $\delta=2|\mathcal{H}|e^{-2m\epsilon^2}$，则 $\epsilon=\sqrt{\frac{\ln|\mathcal{H}|+\ln(2/\delta)}{2m}}$，带入上式中即可得到

$$P\left(\forall h\in\mathcal{H}:|E(h)-\widehat{E}(h)|\leqslant\sqrt{\frac{\ln|\mathcal{H}|+\ln(2/\delta)}{2m}}\right)\geqslant 1-\delta$$

其中 $\forall h\in\mathcal{H}$ 这个前置条件可以省略。

12.18 公式 (12.20)

$$P\left(E(h)-\min_{h'\in\mathcal{H}}E\left(h'\right)\leqslant\epsilon\right)\geqslant 1-\delta$$

[解析]：这个式子是" 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 $\mathcal{L}$ 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。

12.19 公式 (12.21)

$$\Pi_{\mathcal{H}}(m)=\max_{\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_m\}\subseteq\mathcal{X}}\left|\left\{\left(h\left(\boldsymbol{x}_1\right),\dots,h\left(\boldsymbol{x}_m\right)\right)|h\in\mathcal{H}\right\}\right|$$

[解析]：这个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。比如对于两个样本的二分类问题，一共有 4 中可能的标签组合 $[[0,0],[0,1],[1,0],[1,1]]$，如果假设空间 $\mathcal{H}_1$ 能赋予这两个样本两种标签组合 $[[0,0],[1,1]]$，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。

12.20 公式 (12.22)

$$P(|E(h)-\widehat{E}(h)|>\epsilon)\leqslant 4\Pi_{\mathcal{H}}(2m)\exp\left(-\frac{m\epsilon^2}{8}\right)$$

[解析]：这个式子的前提假设有误，应当写成对假设空间 $\mathcal{H}$，$m\in\mathbb{N}$，$0<\epsilon<1$，存在 $h\in\mathcal{H}$ 详细证明参见原论文 On the uniform convergence of relative frequencies of events to their probabilities [3]

12.21 公式 (12.23)

$$\mathrm{VC}(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\}$$

[解析]：这是 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 2 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。

12.22 公式 (12.24)

$$\Pi_{\mathcal{H}}(m) \leqslant \sum_{i=0}^{d} \begin{pmatrix} m \\ i \end{pmatrix}$$

[解析]：首先解释下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 12.23) $\mathrm{VC}(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\} = 0$ 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 右边为 $\sum_{i=0}^{0} \begin{pmatrix} 1 \\ i \end{pmatrix} = 1$，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 $\Pi_{\mathcal{H}}(1) = 2$，不等式右边为 $\sum_{i=0}^{1} \begin{pmatrix} 1 \\ i \end{pmatrix} = 2$，因此不等式成立。

再介绍归纳过程，这里采样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 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ 和观测集 $D' = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{m-1}\}$，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：

$$\begin{aligned}
\mathcal{H}_{|D} &= \{(h(\boldsymbol{x}_1), h(\boldsymbol{x}_2), \ldots, h(\boldsymbol{x}_m)) \,|h \in \mathcal{H}\} \\
\mathcal{H}_{|D'} &= \{(h(\boldsymbol{x}_1), h(\boldsymbol{x}_2), \ldots, h(\boldsymbol{x}_{m-1})) \,|h \in \mathcal{H}\}
\end{aligned}$$

如果假设 $h \in \mathcal{H}$ 对 x_m 的分类结果为 $+1$，或为 -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 $m = 3$：

$$\begin{aligned}
\mathcal{H}_{|D} &= \{(+,-,-), (+,+,-), (+,+,+), (-,+,-), (-,-,+)\} \\
\mathcal{H}_{|D'} &= \{(+,+), (+,-), (-,+), (-,-)\}
\end{aligned}$$

其中串 $(+,+)$ 在 $\mathcal{H}_{|D}$ 中出现了两次 $(+,+,+), (+,+,-)$，$\mathcal{H}_{|D'}$ 中得其他串 $(+,-), (-,+), (-,-)$ 均只在 $\mathcal{H}_{|D}$ 中出现了一次。这里的原因是每个样本是二分类的，所以多出的样本 x_m 要么取 $+$，要么取 $-$，要么都取到 (至少两个假设 h 对 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，比如在上例中 $\mathcal{H}_{D'|D} = \{(+,+)\}$，有

$$\left|\mathcal{H}_{|D}\right| = \left|\mathcal{H}_{|D'}\right| + \left|\mathcal{H}_{D'|D}\right|$$

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' 的数目为 $m-1$，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：

$$\left|\mathcal{H}_{|D'}\right| \leqslant \Pi_{\mathcal{H}}(m-1) \leqslant \sum_{i=0}^{d} \begin{pmatrix} m-1 \\ i \end{pmatrix}$$

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，因为根据 $\mathcal{H}_{D'|D}$ 的定义，H_D 必对元素 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

$$\left|\mathcal{H}_{D'|D}\right| \leqslant \Pi_{\mathcal{H}}(m-1) \leqslant \sum_{i=0}^{d-1} \begin{pmatrix} m-1 \\ i \end{pmatrix}$$

因此：

$$\begin{aligned}
\left|\mathcal{H}_{|D}\right| &= \left|\mathcal{H}_{|D'}\right| + \left|\mathcal{H}_{D'|D}\right| \\
&\leqslant \sum_{i=0}^{d} \begin{pmatrix} m-1 \\ i \end{pmatrix} + \sum_{i=0}^{d+1} \begin{pmatrix} m-1 \\ i \end{pmatrix} \\
&= \sum_{i=0}^{d} \left(\begin{pmatrix} m-1 \\ i \end{pmatrix} + \begin{pmatrix} m-1 \\ i-1 \end{pmatrix} \right) \\
&= \sum_{i=0}^{d} \begin{pmatrix} m \\ i \end{pmatrix}
\end{aligned}$$

注：最后一步依据组合公式，推导如下：

$$\begin{aligned}
\begin{pmatrix} m-1 \\ i \end{pmatrix} + \begin{pmatrix} m-1 \\ i-1 \end{pmatrix} &= \frac{(m-1)!}{(m-1-i)!i!} + \frac{(m-1)!}{(m-1-i+1)!(i-1)!} \\
&= \frac{(m-1)!(m-i)}{(m-i)(m-1-i)!i!} + \frac{(m-1)!i}{(m-i)!(i-1)!i} \\
&= \frac{(m-1)!(m-i)+(m-1)!i}{(m-i)!i!} \\
&= \frac{(m-1)!(m-i+i)}{(m-i)!i!} = \frac{(m-1)!m}{(m-i)!i!} \\
&= \frac{m!}{(m-i)!i!} = \begin{pmatrix} m \\ i \end{pmatrix}
\end{aligned}$$

12.23 公式 (12.25)

$$\left|\mathcal{H}_{|D}\right| = \left|\mathcal{H}_{|D'}\right| + \left|\mathcal{H}_{D'|D}\right|$$

[解析]：参见 12.24

12.24 公式 (12.26)

$$\left|\mathcal{H}_{|D'}\right| \leqslant \Pi_{\mathcal{H}}(m-1) \leqslant \sum_{i=0}^{d} \begin{pmatrix} m-1 \\ i \end{pmatrix}$$

[解析]：参见 12.24

12.25 公式 (12.27)

$$\left|\mathcal{H}_{D'|D}\right| \leqslant \Pi_{\mathcal{H}}(m-1) \leqslant \sum_{i=0}^{d-1} \begin{pmatrix} m-1 \\ i \end{pmatrix}$$

[解析]：参见 12.24

12.26 公式 (12.28)

$$\Pi_{\mathcal{H}}(m) \leqslant \left(\frac{e \cdot m}{d}\right)^{d}$$

[推导]:

$$\begin{aligned}
\Pi_{\mathcal{H}}(m) &\leqslant \sum_{i=0}^{d} \begin{pmatrix} m \\ i \end{pmatrix} \\
&\leqslant \sum_{i=0}^{d} \begin{pmatrix} m \\ i \end{pmatrix} \left(\frac{m}{d}\right)^{d-i} \\
&= \left(\frac{m}{d}\right)^{d} \sum_{i=0}^{d} \begin{pmatrix} m \\ i \end{pmatrix} \left(\frac{d}{m}\right)^{i} \\
&\leqslant \left(\frac{m}{d}\right)^{d} \sum_{i=0}^{m} \begin{pmatrix} m \\ i \end{pmatrix} \left(\frac{d}{m}\right)^{i} \\
&= \left(\frac{m}{d}\right)^{d} \left(1+\frac{d}{m}\right)^{m} \\
&< \left(\frac{e \cdot m}{d}\right)^{d}
\end{aligned}$$

第一步到第二步和第三步到第四步均因为 $m \geqslant d$，第四步到第五步是由于二项式定理 [4]：$(x+y)^n = \sum_{k=0}^{n} \begin{pmatrix} n \\ k \end{pmatrix} x^{n-k}y^k$，其中令 $k=i, n=m, x=1, y=\frac{d}{m}$ 得 $\left(\frac{m}{d}\right)^d \sum_{i=0}^{m} \begin{pmatrix} m \\ i \end{pmatrix} \left(\frac{d}{m}\right)^i = \left(\frac{m}{d}\right)^d (1+\frac{d}{m})^m$，最后一步的不等式即需证明 $\left(1+\frac{d}{m}\right)^m \leqslant e^d$，因为 $\left(1+\frac{d}{m}\right)^m = \left(1+\frac{d}{m}\right)^{\frac{m}{d}d}$，根据自然对数底数 e 的定义 [5]，$\left(1+\frac{d}{m}\right)^{\frac{m}{d}d} < e^d$，注意原文中用的是 $\leqslant$，但是由于 $e = \lim_{\frac{d}{m}\to 0}\left(1+\frac{d}{m}\right)^{\frac{m}{d}}$ 的定义是一个极限，所以应该是用 $<$。

12.27 公式 (12.29)

$$P\left(E(h) - \widehat{E}(h) \leqslant \sqrt{\frac{8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta}}{m}}\right) \geqslant 1-\delta$$

[推导]：这里应该是作者的笔误，根据式 12.22，$E(h)-\widehat{E}(h)$ 应当被绝对值符号包裹。将式 12.28 带入式 12.22 得

$$P\left(|E(h) - \widehat{E}(h)| > \epsilon\right) \leqslant 4\left(\frac{2em}{d}\right)^{d} \exp\left(-\frac{m\epsilon^2}{8}\right)$$

令 $4\left(\frac{2em}{d}\right)^d \exp\left(-\frac{m\epsilon^2}{8}\right) = \delta$ 可解得

$$\delta = \sqrt{\frac{8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta}}{m}}$$

带入式 12.22，则定理得证。这个式子是用 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 $\sqrt{\frac{\ln m}{m}}$（书上简化为 $\frac{1}{\sqrt{m}}$）。

12.28 公式 (12.30)

$$\widehat{E}(h) = \min_{h' \in \mathcal{H}} \widehat{E}(h')$$

[解析]：这个是经验风险最小化的定义式。即从假设空间中找出能使经验风险最小的假设。

12.29 公式 (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

$$\delta'=\frac{\delta}{2}$$
$$\sqrt{\frac{(\ln 2/\delta')}{2m}}=\frac{\epsilon}{2}$$

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。写成概率的形式即：

$$P\left(|E(g)-\widehat{E}(g)|\leqslant\frac{\epsilon}{2}\right)\geqslant 1-\delta/2$$

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

$$\sqrt{\frac{8d\ln\frac{2em}{d}+8\ln\frac{4}{\delta'}}{m}}=\frac{\epsilon}{2}$$

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 $P\left(E(h)-\widehat{E}(h)\leqslant\frac{\epsilon}{2}\right)\geqslant 1-\delta/2$ 均成立可知则事件 $E(g)-\widehat{E}(g)\geqslant-\frac{\epsilon}{2}$ 和事件 $E(h)-\widehat{E}(h)\leqslant\frac{\epsilon}{2}$ 同时成立的概率为：

$$\begin{aligned}
&P\left(\left(E(g)-\widehat{E}(g)\geqslant-\frac{\epsilon}{2}\right)\wedge\left(E(h)-\widehat{E}(h)\leqslant\frac{\epsilon}{2}\right)\right)\\
=&P\left(E(g)-\widehat{E}(g)\geqslant-\frac{\epsilon}{2}\right)+P\left(E(h)-\widehat{E}(h)\leqslant\frac{\epsilon}{2}\right)-P\left(\left(E(g)-\widehat{E}(g)\geqslant-\frac{\epsilon}{2}\right)\vee\left(E(h)-\widehat{E}(h)\leqslant\frac{\epsilon}{2}\right)\right)\\
\geqslant&1-\delta/2+1-\delta/2-1\\
=&1-\delta
\end{aligned}$$

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

$$P\left(\widehat{E}(g)-E(g)+E(h)-\widehat{E}(h)\leqslant\frac{\epsilon}{2}+\frac{\epsilon}{2}\right)=P\left(E(h)-E(g)\leqslant\widehat{E}(h)-\widehat{E}(g)+\epsilon\right)\geqslant 1-\delta$$

再由 h 和 g 的定义，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 $(1/\epsilon,1/\delta,\text{size}(x),\text{size}(c))$ 的多项式，因此根据定理 12.2，定理 12.5，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。

12.30 公式 (12.32)

$$\sqrt{\frac{(\ln 2/\delta')}{2m}} = \frac{\epsilon}{2}$$

[解析]：参见 12.31

12.31 公式 (12.34)

$$\sqrt{\frac{8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta'}}{m}} = \frac{\epsilon}{2}$$

[解析]：参见 12.31

12.32 公式 (12.36)

$$\begin{aligned}\widehat{E}(h) &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(h\left(\boldsymbol{x}_i\right) \neq y_i\right) \\ &= \frac{1}{m}\sum_{i=1}^{m}\frac{1-y_i h\left(\boldsymbol{x}_i\right)}{2} \\ &= \frac{1}{2} - \frac{1}{2m}\sum_{i=1}^{m} y_i h\left(\boldsymbol{x}_i\right)\end{aligned}$$

[解析]：这里解释从第一步到第二步的推导，因为前提假设是 2 分类问题，$y_k \in \{-1, +1\}$，因此 $\mathbb{I}\left(h(x_i) \neq y_i\right) \equiv \frac{1-y_i h(x_i)}{2}$。这是因为假如 $y_i = +1, h(x_i) = +1$ 或 $y_i = -1, h(x_i) = -1$，有 $\mathbb{I}\left(h(x_i) \neq y_i\right) = 0 = \frac{1-y_i h(x_i)}{2}$；反之，假如 $y_i = -1, h(x_i) = +1$ 或 $y_i = +1, h(x_i) = -1$，有 $\mathbb{I}\left(h(x_i) \neq y_i\right) = 1 = \frac{1-y_i h(x_i)}{2}$。

12.33 公式 (12.37)

$$\arg\max_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} y_i h\left(\boldsymbol{x}_i\right)$$

[解析]：由公式 12.36 可知，经验误差 $\widehat{E}(h)$ 和 $\frac{1}{m}\sum_{i=1}^{m} y_i h\left(\boldsymbol{x}_i\right)$ 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 $\frac{1}{m}\sum_{i=1}^{m} y_i h\left(\boldsymbol{x}_i\right)$ 最大的 h。

12.34 公式 (12.38)

$$\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h\left(\boldsymbol{x}_i\right)$$

[解析]：上确界 sup 这个概念前面已经解释过，见式 12.7 的解析。由于 σ_i 是随机变量，因此这个式子可以理解为求解和随机生成的标签 (即 σ) 最契合的假设 (当 σ_i 和 $h(\boldsymbol{x}_i)$ 完全一致时，他们的内积最大)。

12.35 公式 (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]：这个式子可以用来衡量假设空间 $\mathcal{H}$ 的表达能力，对变量 σ 求期望可以理解为当变量 σ 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 2 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 $D=\{(x_1,y_1),(x_2,y_2),\dots,(x_m,y_m)\}$ 的" 对分 "(12.4 节)，如果一个假设空间的表达能力越强，那么就越有可能对于每一种 σ_i，假设空间中都存在一个 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。

12.36 公式 (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 $\mathcal{F}$ 代替了假设空间 $\mathcal{H}$，函数 f 代替了假设 h，很容易理解，因为假设 h 即可以看做是作用在数据 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 $\mathcal{F}:\mathcal{Z}\to\mathbb{R}$，即映射 f 将样本 z_i 映射到了实数空间，这个时候所有的 σ_i 将是一个标量即 $\sigma_i\in\{+1,-1\}$。

12.37 公式 (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。

12.38 公式 (12.42)

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{z})]&\leqslant\tfrac{1}{m}\textstyle\sum_{i=1}^{m}f\left(\boldsymbol{z}_i\right)+2R_m(\mathcal{F})+\sqrt{\frac{\ln(1/\delta)}{2m}}\\ \mathbb{E}[f(\boldsymbol{z})]&\leqslant\tfrac{1}{m}\textstyle\sum_{i=1}^{m}f\left(\boldsymbol{z}_i\right)+2\widehat{R}_Z(\mathcal{F})+3\sqrt{\frac{\ln(2/\delta)}{2m}}\end{aligned}$$

[解析]：首先令记号

$$\begin{aligned}\widehat{E}_Z(f)&=\frac{1}{m}\sum_{i=1}^{m}f\left(\boldsymbol{z}_i\right)\\ \Phi(Z)&=\sup_{f\in\mathcal{F}}\left(\mathbb{E}[f]-\widehat{E}_Z(f)\right)\end{aligned}$$

即 $\widehat{E}_Z(f)$ 表示函数 f 作为假设下的经验误差，$\Phi(Z)$ 表示泛化误差和经验误差的差的上确界。再令 Z' 为只与 Z 有一个示例 (样本) 不同的训练集，不妨设 $z_m\in Z$ 和 $z'_m\in Z'$ 为不同的示例，那么有

$$\begin{aligned}\Phi\left(Z'\right)-\Phi(Z)&=\sup_{f\in\mathcal{F}}\left(\mathbb{E}[f]-\widehat{E}_{Z'}(f)\right)-\sup_{f\in\mathcal{F}}\left(\mathbb{E}[f]-\widehat{E}_Z(f)\right)\\ &\leqslant\sup_{f\in\mathcal{F}}\left(\widehat{E}_Z(f)-\widehat{E}_{Z'}(f)\right)\\ &=\sup_{f\in\mathcal{F}}\frac{\sum_{i=1}^{m}f(z_i)-\sum_{i=1}^{m}f(z'_i)}{m}\\ &=\sup_{f\in\mathcal{F}}\frac{f\left(z_m\right)-f\left(z'_m\right)}{m}\\ &\leqslant\frac{1}{m}\end{aligned}$$

第一个不等式是因为上确界的差不大于差的上确界 [2]，第四行的等号由于 Z' 与 Z 只有 z_m 不相同，最后一行的不等式是因为前提假设 $\mathcal{F}:\mathcal{Z}\to[0,1]$，即 $f(z_m),f(z'_m)\in[0,1]$。同理

$$\Phi(Z)-\Phi\left(Z'\right)=\sup_{f\in\mathcal{F}}\frac{f\left(z'_m\right)-f\left(z_m\right)}{m}\leqslant\frac{1}{m}$$

综上二式有：

$$|\Phi(Z)-\Phi\left(Z'\right)|\leqslant\frac{1}{m}$$

将 Φ 看做函数 f(注意这里的 f 不是 Φ 定义里的 f)，那么可以套用 McDiarmid 不等式的结论式 12.7

$$P\left(\Phi(Z)-\mathbb{E}_Z[\Phi(Z)]\geqslant\epsilon\right)\leqslant\exp\left(\frac{-2\epsilon^2}{\sum_i c_i^2}\right)$$

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

$$P\left(\Phi(Z)-\mathbb{E}_Z[\Phi(Z)]\leqslant\sqrt{\frac{\ln(1/\delta)}{2m}}\right)\geqslant1-\delta$$

即书中式 12.44 的结论。下面来估计 $\mathbb{E}_Z[\Phi(Z)]$ 的上界：

$$\begin{aligned}
\mathbb{E}_Z[\Phi(Z)]&=\mathbb{E}_Z\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}[f]-\widehat{E}_Z(f)\right)\right]\\
&=\mathbb{E}_Z\left[\sup_{f\in\mathcal{F}}\mathbb{E}_{Z'}\left[\widehat{E}_{Z'}(f)-\widehat{E}_Z(f)\right]\right]\\
&\leqslant\mathbb{E}_{Z,Z'}\left[\sup_{f\in\mathcal{F}}\left(\widehat{E}_{Z'}(f)-\widehat{E}_Z(f)\right)\right]\\
&=\mathbb{E}_{Z,Z'}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\left(f\left(\boldsymbol{z}'_i\right)-f\left(\boldsymbol{z}_i\right)\right)\right]\\
&=\mathbb{E}_{\boldsymbol{\sigma},Z,Z'}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left(f\left(\boldsymbol{z}'_i\right)-f\left(\boldsymbol{z}_i\right)\right)\right]\\
&\leqslant\mathbb{E}_{\boldsymbol{\sigma},Z'}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f\left(\boldsymbol{z}'_i\right)\right]+\mathbb{E}_{\boldsymbol{\sigma},Z}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}-\sigma_i f\left(\boldsymbol{z}_i\right)\right]\\
&=2\mathbb{E}_{\boldsymbol{\sigma},Z}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f\left(\boldsymbol{z}_i\right)\right]\\
&=2R_m(\mathcal{F})
\end{aligned}$$

第二行等式是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 $\mathbb{E}_{Z'\sim\mathcal{D}}[\widehat{E}_Z(f)]=\widehat{E}_Z(f)$。第三行不等式基于上确界函数 sup 是个凸函数，将 $\sup_{f\in\mathcal{F}}$ 看做是凸函数 f，将 $\widehat{E}_{Z'}(f)-\widehat{E}_Z(f)$ 看做变量 x 根据 Jesen 不等式 (式 12.4)，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。第五行引入对 Rademacher 随机变量的期望，由于函数值空间是标量，因为 σ_i 也是标量，即 $\sigma_i\in\{-1,+1\}$，且 σ_i 总以相同概率可以取到这两个值，因此可以引入 $\mathbb{E}_\sigma$ 而不影响最终结果。第六行利用了上确界的和不小于和的上确界 [2]，因为第一项中只含有变量 z'，所以可以将 $\mathbb{E}_Z$ 去掉，因为第二项中只含有变量 z，所以可以将 $\mathbb{E}_{Z'}$ 去掉。第七行利用 σ 是对称的，所以 $-\sigma$ 的分布和 σ 完全一致，所以可以将第二项中的负号去除，又因为 Z 和 Z' 均是从 $\mathcal{D}$ 中 *i.i.d.* 采样得到的数据，因此可以将第一项中的 z'_i 替换成 z，将 Z' 替换成 Z。最后根据定义式 12.41 可得 $\mathbb{E}_Z[\Phi(Z)]=2\mathcal{R}_m(\mathcal{F})$，式 12.42 得证。

12.39 公式 (12.43)

$$\mathbb{E}[f(\boldsymbol{z})] \leqslant \frac{1}{m}\sum_{i=1}^{m} f\left(\boldsymbol{z}_i\right) + 2\widehat{R}_Z(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

[解析]：参见 12.42

12.40 公式 (12.44)

$$\Phi(Z) \leqslant \mathbb{E}_Z[\Phi(Z)] + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

[解析]：参见 12.42

12.41 公式 (12.45)

$$R_m(\mathcal{F}) \leqslant \widehat{R}_Z(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2m}}$$

[解析]：参见 12.42

12.42 公式 (12.46)

$$\Phi(Z) \leqslant 2\widehat{R}_Z(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

[解析]：参见 12.42

12.43 公式 (12.52)

$$R_m(\mathcal{H}) \leqslant \sqrt{\frac{2\ln \Pi_{\mathcal{H}}(m)}{m}}$$

[证明]：比较繁琐，同书上所示，参见 Foundations of Machine Learning[1]

12.44 公式 (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 $\Pi_{\mathcal{H}}(m) \leqslant \left(\frac{e\cdot m}{d}\right)^d$，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 $E(h) \leqslant \widehat{E}(h) + R_m(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2m}}$ 即证。

12.45 公式 (12.57)

$$\begin{aligned}
&\left|\ell\left(\mathfrak{L}_D, \boldsymbol{z}\right) - \ell\left(\mathfrak{L}_{D^i}, \boldsymbol{z}\right)\right| \\
\leqslant &\left|\ell\left(\mathfrak{L}_D, \boldsymbol{z}\right) - \ell\left(\mathfrak{L}_{D^{\setminus i}}, \boldsymbol{z}\right)\right| + \left|\ell\left(\mathfrak{L}_{D^i,\boldsymbol{z}}\right) - \ell\left(\mathfrak{L}_{D^{\setminus i},\boldsymbol{z}}\right)\right| \\
\leqslant &2\beta
\end{aligned}$$

[解析]：根据三角不等式 [6]，有 $|a+b| \leq |a| + |b|$，将 $a = \ell\left(\mathfrak{L}_D, \boldsymbol{z}\right) - \ell\left(\mathfrak{L}_{D^i}\right)$，$b = \ell\left(\mathfrak{L}_{D^i,\boldsymbol{z}}\right) - \ell\left(\mathfrak{L}_{D^{\setminus i},\boldsymbol{z}}\right)$ 带入即可得出第一个不等式，根据 $D^{\setminus i}$ 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, b 的变动均为一个样本，根据式 12.57，$a \leqslant \beta, b \leqslant \beta$，因此 $a + b \leqslant 2\beta$。

12.46 公式 (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]：比较繁琐，同书上所示，参见 Foundations of Machine Learning[1]

12.47 公式 (12.59)

$$\ell(\mathfrak{L},\mathcal{D}) \leqslant \ell_{loo}(\mathfrak{L},D) + \beta + (4m\beta + M)\sqrt{\frac{\ln(1/\delta)}{2m}}$$

[证明]：比较繁琐，同书上所示，参见 Foundations of Machine Learning[1]

12.48 公式 (12.60)

$$\ell(\mathfrak{L},\mathcal{D}) \leqslant \widehat{\ell}(\mathfrak{L},D) + \frac{2}{m} + (4 + M)\sqrt{\frac{\ln(1/\delta)}{2m}}$$

[证明]：将 $\beta = \frac{1}{m}$ 带入至式 12.58 即得证。

12.49 定理 (12.9)

若学习算法 $\mathcal{L}$ 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 $\mathcal{L}$ 满足经验风险最小化 (Empirical Risk Minimization)。由于 $\mathcal{L}$ 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

$$\begin{aligned}\epsilon' &= \tfrac{\epsilon}{2}\\ \tfrac{\delta}{2} &= 2\exp\left(-2m\left(\epsilon'\right)^2\right)\end{aligned}$$

将 $\epsilon' = \frac{\epsilon}{2}$ 带入到 $\frac{\delta}{2} = 2\exp\left(-2m\left(\epsilon'\right)^2\right)$ 可以解得 $m = \frac{2}{\epsilon^2}\ln\frac{4}{\delta}$，由 Hoeffding 不等式 12.6，

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

其中 $\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left(x_i\right) = \ell(g,\mathcal{D})$，$\frac{1}{m}\sum_{i=1}^{m} x_i = \widehat{\ell}(g,\mathcal{D})$，带入可得

$$P(|\ell(g,\mathcal{D}) - \widehat{\ell}(g,D)| \geqslant \frac{\epsilon}{2}) \leqslant \frac{\delta}{2}$$

根据逆事件的概率可得

$$P(|\ell(g,\mathcal{D}) - \widehat{\ell}(g,D)| \leqslant \frac{\epsilon}{2}) \geqslant 1 - \frac{\delta}{2}$$

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

$$\sqrt{m} = \frac{(4+M)\sqrt{\frac{\ln(2/\delta)}{2}} + \sqrt{(4+M)^2\frac{\ln(2/\delta)}{2} - 4\times\frac{\epsilon}{2}\times(-2)}}{2\times\frac{\epsilon}{2}}$$

即 $m = O\left(\frac{1}{\epsilon^2}\ln\frac{1}{\delta}\right)$。

由 $P(|\ell(g,\mathcal{D})-\widehat{\ell}(g,D)|\leqslant\frac{\epsilon}{2})\geqslant 1-\frac{\delta}{2}$ 可以按照同公式 12.31 中介绍的相同的方法推导出

$$P(\ell(\mathfrak{L},\mathcal{D})-\ell(g,\mathcal{D})\leqslant\epsilon)\geqslant 1-\delta$$

又因为 m 为与 $(1/\epsilon,1/\delta,\text{size}(x),\text{size}(c))$ 相关的多项式的值，因此根据定理 12.2，定理 12.5，得到结论 $\mathcal{H}$ 是 (不可知)PAC 可学习的。

参考文献

[1] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. 2018. URL: https://cs.nyu.edu/~mohri/mlbook/.

[2] robjohn. Supremum of the difference of two functions, 2013. URL: https://math.stackexchange.com/questions/246015/supremum-of-the-difference-of-two-functions.

[3] Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015.

[4] Wikipedia contributors. Binomial theorem, 2020. URL: https://zh.wikipedia.org/zh-hans/%E4%BA%8C%E9%A1%B9%E5%BC%8F%E5%AE%9A%E7%90%86.

[5] Wikipedia contributors. E, 2020. URL: https://en.wikipedia.org/wiki/E_(mathematical_constant).

[6] Wikipedia contributors. Triangle inequality, 2020. URL: https://en.wikipedia.org/wiki/Triangle_inequality.

第 13 章　半监督学习

13.1　公式 (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]：高斯混合分布的定义式。

13.2　公式 (13.2)

$$\begin{aligned} f(\boldsymbol{x}) &= \arg\max_{j\in\mathcal{Y}} p(y=j|\boldsymbol{x}) \\ &= \arg\max_{j\in\mathcal{Y}} \sum_{i=1}^{N} p(y=j, \Theta=i|\boldsymbol{x}) \\ &= \arg\max_{j\in\mathcal{Y}} \sum_{i=1}^{N} p(y=j|\Theta=i, \boldsymbol{x}) \cdot p(\Theta=i|\boldsymbol{x}) \end{aligned}$$

[解析]：从公式第 1 行到第 2 行是对概率进行边缘化 (marginalization)；通过引入 Θ 并对其求和 $\sum_{i=1}^{N}$ 以抵消引入的影响。从公式第 2 行到第 3 行推导如下

$$\begin{aligned} p(y=j, \Theta=i|\boldsymbol{x}) &= \frac{p(y=j, \Theta=i, \boldsymbol{x})}{p(\boldsymbol{x})} \\ &= \frac{p(y=j, \Theta=i, \boldsymbol{x})}{p(\Theta=i, \boldsymbol{x})} \cdot \frac{p(\Theta=i, \boldsymbol{x})}{p(\boldsymbol{x})} \\ &= p(y=j|\Theta=i, \boldsymbol{x}) \cdot p(\Theta=i|\boldsymbol{x}) \end{aligned}$$

13.3　公式 (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 13.1

$$p(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i \cdot p\left(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)$$

因此

$$\begin{aligned} p(\Theta=i|\boldsymbol{x}) &= \frac{p(\Theta=i, \boldsymbol{x})}{P(x)} \\ &= \frac{\alpha_i \cdot p\left(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\sum_{i=1}^{N} \alpha_i \cdot p\left(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)} \end{aligned}$$

13.4　公式 (13.4)

$$\begin{aligned} LL\left(D_l \cup D_u\right) = &\sum_{(x_j, y_j)\in D_l} \ln\left(\sum_{i=1}^{N} \alpha_i \cdot p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \cdot p\left(y_j|\Theta=i, \boldsymbol{x}_j\right)\right) \\ &+ \sum_{x_j\in D_u} \ln\left(\sum_{i=1}^{N} \alpha_i \cdot p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)\right) \end{aligned}$$

[解析]：第二项很好解释，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，所有无类别信息的样本 D_u 的似然是所有样本的乘积，因为 ln 函数是单调的，所以也可以将 ln 函数作用于这个乘积消除因为连乘产生的数值计算问题。第一项引入了样本的标签信息，由

$$p(y=j|\Theta=i,\boldsymbol{x})=\begin{cases}1, & i=j\\ 0, & i\neq j\end{cases}$$

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。

13.5 公式 (13.5)

$$\gamma_{ji}=\frac{\alpha_i\cdot p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)}{\sum_{i=1}^{N}\alpha_i\cdot p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)}$$

[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：

$$\begin{aligned}&\alpha_i=\tfrac{l_i}{|D_l|},\ \text{where}\ |D_l|=\textstyle\sum_{i=1}^{N}l_i\\&\boldsymbol{\mu}_i=\tfrac{1}{l_i}\textstyle\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\boldsymbol{x}_j\\&\boldsymbol{\Sigma}_i=\tfrac{1}{l_i}\textstyle\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\top}\end{aligned}$$

13.6 公式 (13.6)

$$\boldsymbol{\mu}_i=\frac{1}{\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}+l_i}\left(\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}\boldsymbol{x}_j+\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\boldsymbol{x}_j\right)$$

[推导]：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：

$$\begin{aligned}LL(D_l)&=\sum_{(\boldsymbol{x_j},y_j\in D_l)}\ln\left(\sum_{s=1}^{N}\alpha_s\cdot p(\boldsymbol{x_j}|\boldsymbol{\mu}_s,\boldsymbol{\Sigma}_s)\cdot p(y_i|\Theta=s,\boldsymbol{x_j})\right)\\&=\sum_{(\boldsymbol{x_j},y_j\in D_l)}\ln\left(\alpha_{y_j}\cdot p(\boldsymbol{x_j}|\boldsymbol{\mu}_{y_j},\boldsymbol{\Sigma}_{y_j})\right)\\LL(D_u)&=\sum_{\boldsymbol{x_j}\in D_u}\ln\left(\sum_{s=1}^{N}\alpha_s\cdot p(\boldsymbol{x_j}|\boldsymbol{\mu}_s,\boldsymbol{\Sigma}_s)\right)\end{aligned}$$

首先，$LL(D_l)$ 对 $\boldsymbol{\mu_i}$ 求偏导，$LL(D_l)$ 求和号中只有 $y_j=i$ 的项能留下来，即

$$\begin{aligned}\frac{\partial LL\left(D_l\right)}{\partial\boldsymbol{\mu}_i}&=\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\frac{\partial\ln\left(\alpha_i\cdot p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)\right)}{\partial\boldsymbol{\mu}_i}\\&=\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\frac{1}{p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)}\cdot\frac{\partial p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)}{\partial\boldsymbol{\mu}_i}\\&=\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\frac{1}{p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)}\cdot p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)\cdot\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\\&=\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\end{aligned}$$

$LL(D_u)$ 对 $\boldsymbol{\mu_i}$ 求导，参考 9.33 的推导：

$$\begin{aligned}\frac{\partial LL\left(D_u\right)}{\partial \boldsymbol{\mu}_i} &= \sum_{\boldsymbol{x}_j \in D_u} \frac{\alpha_i}{\sum_{s=1}^{N} \alpha_s \cdot p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\right)} \cdot p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \cdot \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right) \\ &= \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\end{aligned}$$

综上，

$$\begin{aligned}\frac{\partial LL\left(D_l \cup D_u\right)}{\partial \boldsymbol{\mu}_i} &= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right) + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right) \\ &= \boldsymbol{\Sigma}_i^{-1}\left(\sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right) + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\right) \\ &= \boldsymbol{\Sigma}_i^{-1}\left(\sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{x}_j + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{x}_j - \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{\mu}_i - \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{\mu}_i\right)\end{aligned}$$

令 $\frac{\partial LL(D_l \cup D_u)}{\partial \boldsymbol{\mu}_i} = 0$，两边同时左乘 Σ_i 并移项：

$$\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{\mu}_i + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{\mu}_i = \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{x}_j + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{x}_j$$

上式中，$\boldsymbol{\mu_i}$ 可以作为常量提到求和号外面，而 $\sum_{(x_j, y_j) \in D_l \wedge y_j = i} 1 = l_i$，即第 i 类样本的有标记样本数目，因此

$$\left(\sum_{x_j \in D_u} \gamma_{ji} + \sum_{(x_j, y_j) \in D_l \wedge y_j = i} 1\right) \boldsymbol{\mu}_i = \sum_{x_j \in D_u} \gamma_{ji} \cdot \boldsymbol{x}_j + \sum_{(x_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{x}_j$$

即得式 13.6。

13.7 公式 (13.7)

$$\begin{aligned}\boldsymbol{\Sigma}_i = & \frac{1}{\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i}\left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\top}\right. \\ & \left. + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\top}\right)\end{aligned}$$

[推导]：首先 $LL(D_l)$ 对 $\boldsymbol{\Sigma_i}$ 求偏导，类似于 13.6

$$\begin{aligned}\frac{\partial LL\left(D_l\right)}{\partial \boldsymbol{\Sigma}_i} &= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{\partial \ln\left(\alpha_i \cdot p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)\right)}{\partial \boldsymbol{\Sigma}_i} \\ &= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{1}{p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)} \cdot \frac{\partial p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\partial \boldsymbol{\Sigma}_i} \\ &= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{1}{p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)} \cdot p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \cdot \left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\top} - \boldsymbol{I}\right) \cdot \frac{1}{2} \boldsymbol{\Sigma}_i^{-1} \\ &= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\top} - \boldsymbol{I}\right) \cdot \frac{1}{2} \boldsymbol{\Sigma}_i^{-1}\end{aligned}$$

然后 $LL(D_u)$ 对 $\boldsymbol{\Sigma_i}$ 求偏导，类似于 9.35

$$\frac{\partial LL\left(D_u\right)}{\partial \boldsymbol{\Sigma}_i} = \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right)^{\top} - \boldsymbol{I}\right) \cdot \frac{1}{2} \boldsymbol{\Sigma}_i^{-1}$$

综合可得：

$$\begin{aligned}\frac{\partial LL\left(D_l \cup D_u\right)}{\partial \boldsymbol{\Sigma}_i} &= \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\top}-\boldsymbol{I}\right) \cdot \frac{1}{2}\boldsymbol{\Sigma}_i^{-1} \\ &\quad + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\top}-\boldsymbol{I}\right) \cdot \frac{1}{2}\boldsymbol{\Sigma}_i^{-1} \\ &= \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\top}-\boldsymbol{I}\right) \right. \\ &\quad \left. + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\top}-\boldsymbol{I}\right) \right) \cdot \frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\end{aligned}$$

令 $\frac{\partial LL(D_l \cup D_u)}{\partial \boldsymbol{\Sigma}_i} = 0$，两边同时右乘 $2\Sigma_i$ 并移项：

$$\begin{aligned}\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\top} + \sum_{(\boldsymbol{x}_j, y_j \in D_l \wedge y_j = i} \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\top} \\ = \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{I} + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{I} \\ = \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i \right) \boldsymbol{I}\end{aligned}$$

两边同时左乘以 Σ_i：

$$\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\top} + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i\right)^{\top} = \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i \right) \boldsymbol{\Sigma}_i$$

即得式 13.7。

13.8 公式 (13.8)

$$\alpha_i = \frac{1}{m}\left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i \right)$$

[推导]：类似于式 9.36，写出 $LL(D_l \cup D_u)$ 的拉格朗日形式

$$\begin{aligned}\mathcal{L}\left(D_l \cup D_u, \lambda\right) &= LL\left(D_l \cup D_u\right) + \lambda\left(\sum_{s=1}^{N} \alpha_s - 1\right) \\ &= LL\left(D_l\right) + LL\left(D_u\right) + \lambda\left(\sum_{s=1}^{N} \alpha_s - 1\right)\end{aligned}$$

类似于式 9.37，对 α_i 求偏导。对于 $LL(D_u)$，求导结果与式 9.37 的推导过程一样

$$\frac{\partial LL\left(D_u\right)}{\partial \alpha_i} = \sum_{\boldsymbol{x}_j \in D_u} \frac{1}{\sum_{s=1}^{N} \alpha_s \cdot p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\right)} \cdot p\left(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)$$

对于 $LL(D_l)$，类似于 13.6 和 13.7 的推导过程

$$\begin{aligned}\frac{\partial LL\left(D_l\right)}{\partial \alpha_i} &= \sum_{(x_j,y_j)\in D_l \wedge y_j=i} \frac{\partial \ln\left(\alpha_i \cdot p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)\right)}{\partial \alpha_i} \\ &= \sum_{(\boldsymbol{x}_j,y_j)\in D_l \wedge y_j=i} \frac{1}{\alpha_i \cdot p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)} \cdot \frac{\partial\left(\alpha_i \cdot p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)\right)}{\partial \alpha_i} \\ &= \sum_{(\boldsymbol{x}_j,y_j)\in D_l \wedge y_j=i} \frac{1}{\alpha_i \cdot p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)} \cdot p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right) \\ &= \sum_{(\boldsymbol{x}_j,y_j)\in D_l \wedge y_j=i} \frac{1}{\alpha_i} = \frac{1}{\alpha_i} \cdot \sum_{(\boldsymbol{x}_j,y_j)\in D_l \wedge y_j=i} 1 = \frac{l_i}{\alpha_i}\end{aligned}$$

上式推导过程中，重点注意变量是 α_i ，$p(x_j|\mu_i,\Sigma_i)$ 是常量；最后一行 α_i 相对于求和变量为常量，因此作为公因子提到求和号外面；l_i 为第 i 类样本的有标记样本数目。综合两项结果：

$$\frac{\partial \mathcal{L}\left(D_l \cup D_u, \lambda\right)}{\partial \alpha_i} = \frac{l_i}{\alpha_i} + \sum_{\boldsymbol{x}_j \in D_u} \frac{p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)}{\sum_{s=1}^{N} \alpha_s \cdot p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_s,\boldsymbol{\Sigma}_s\right)} + \lambda$$

令 $\dfrac{\partial LL(D_l \cup D_u)}{\partial \alpha_i} = 0$ 并且两边同乘以 α_i，得

$$\alpha_i \cdot \frac{l_i}{\alpha_i} + \sum_{\boldsymbol{x}_j \in D_u} \frac{\alpha_i \cdot p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)}{\sum_{s=1}^{N} \alpha_s \cdot p\left(\boldsymbol{x}_j|\boldsymbol{\mu}_s,\boldsymbol{\Sigma}_s\right)} + \lambda \cdot \alpha_i = 0$$

结合式 9.30 发现，求和号内即为后验概率 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

$$\sum_{i=1}^{N} l_i + \sum_{i=1}^{N} \sum_{x_i \in D_u} \gamma_{ji} + \sum_{i=1}^{N} \lambda \alpha_i = 0$$

这里 $\sum_{i=1}^{N} \alpha_i = 1$ ，因此 $\sum_{i=1}^{N} \lambda \alpha_i = \lambda \sum_{i=1}^{N} \alpha_i = \lambda$，根据 9.30 中 γ_{ji} 表达式可知

$$\sum_{i=1}^{N} \gamma_{ji} = \sum_{i=1}^{N} \frac{\alpha_i \cdot p(x_j|\mu_i,\Sigma_i)}{\Sigma_{s=1}^{N} \alpha_s \cdot p(x_j|\mu_s,\Sigma_s)} = \frac{\sum_{i=1}^{N} \alpha_i \cdot p(x_j|\mu_i,\Sigma_i)}{\sum_{s=1}^{N} \alpha_s \cdot p(x_j|\mu_s,\Sigma_s)} = 1$$

再结合加法满足交换律，所以

$$\sum_{i=1}^{N} \sum_{x_i \in D_u} \gamma_{ji} = \sum_{x_i \in D_u} \sum_{i=1}^{N} \gamma_{ji} = \sum_{x_i \in D_u} 1 = u$$

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；$\sum_{i=1}^{N} l_i = l$ 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 $l + u + \lambda = 0$ 且 $l + u = m$ 其中 m 为样本总个数，移项即得 $\lambda = -m$ 最后带入整理解得

$$l_i + \sum_{x_j \in D_u} \gamma_{ji} - \lambda \alpha_i = 0$$

整理即得式 13.8。

13.9 公式 (13.9)

$$\min_{\boldsymbol{w},\boldsymbol{b},\boldsymbol{y},\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C_l\sum_{i=1}^{l}\xi_i + C_u\sum_{i=l+1}^{m}\xi_i \qquad \begin{aligned} \text{s.t. } & y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)\geqslant 1-\xi_i, \quad i=1,2,\ldots,l \\ & \hat{y}_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)\geqslant 1-\xi_i, \quad i=l+1,l+2,\ldots,m \\ & \xi_i\geqslant 0, \quad i=1,2,\ldots,m \end{aligned}$$

[解析]：这个公式和公式 6.35 基本一致，除了引入了无标记样本的松弛变量 $\xi_i, i=l+1,\cdots m$ 和对应的权重系数 C_u 和无标记样本的标记指派 $\hat{y}_i$。

13.10 公式 (13.12)

$$\begin{aligned} E(f) &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}\left(f\left(\boldsymbol{x}_i\right)-f\left(\boldsymbol{x}_j\right)\right)^2 \\ &= \frac{1}{2}\left(\sum_{i=1}^{m}d_i f^2\left(\boldsymbol{x}_i\right)+\sum_{j=1}^{m}d_j f^2\left(\boldsymbol{x}_j\right)-2\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f\left(\boldsymbol{x}_i\right)f\left(\boldsymbol{x}_j\right)\right) \\ &= \sum_{i=1}^{m}d_i f^2\left(\boldsymbol{x}_i\right)-\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f\left(\boldsymbol{x}_i\right)f\left(\boldsymbol{x}_j\right) \\ &= \boldsymbol{f}^{\mathrm{T}}(\mathbf{D}-\mathbf{W})\boldsymbol{f} \end{aligned}$$

[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,j，如果它们不相邻，则 $(\mathbf{W})_{ij}=0$，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。下面进行公式的推导，首先由二项展开可得：

$$\begin{aligned} E(f) &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}\left(f\left(\boldsymbol{x}_i\right)-f\left(\boldsymbol{x}_j\right)\right)^2 \\ &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}\left(f^2\left(\boldsymbol{x}_i\right)-2f\left(\boldsymbol{x}_i\right)f\left(\boldsymbol{x}_j\right)+f^2\left(\boldsymbol{x}_j\right)\right) \\ &= \frac{1}{2}\left(\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f^2\left(\boldsymbol{x}_i\right)+\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f^2\left(\boldsymbol{x}_j\right)-2\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f\left(\boldsymbol{x}_i\right)f\left(\boldsymbol{x}_j\right)\right) \end{aligned}$$

由于 $\mathbf{W}$ 是一个对称矩阵，可以通过变量替换得到

$$\begin{aligned} \sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f^2\left(\boldsymbol{x}_j\right) &= \sum_{j=1}^{m}\sum_{i=1}^{m}(\mathbf{W})_{ji}f^2\left(\boldsymbol{x}_i\right) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f^2\left(\boldsymbol{x}_i\right) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f^2\left(\boldsymbol{x}_j\right) \end{aligned}$$

因此 $E(f)$ 可化简为

$$E(f)=\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f^2\left(\boldsymbol{x}_i\right)-\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f\left(\boldsymbol{x}_i\right)f\left(\boldsymbol{x}_j\right)$$

根据定义 $d_i = \sum_{j=1}^{l+u} (\mathbf{W})_{ij}$，且 $m = l + u$ 则

$$\begin{aligned} E(f) &= \sum_{i=1}^{m} d_i f^2(\boldsymbol{x}_i) - \sum_{i=1}^{m}\sum_{j=1}^{m} (\mathbf{W})_{ij} f(\boldsymbol{x}_i) f(\boldsymbol{x}_j) \\ &= \boldsymbol{f}^{\mathrm{T}} \mathbf{D} \boldsymbol{f} - \boldsymbol{f}^{\mathrm{T}} \mathbf{W} \boldsymbol{f} \\ &= \boldsymbol{f}^{\mathrm{T}} (\mathbf{D} - \mathbf{W}) \boldsymbol{f} \end{aligned}$$

13.11 公式 (13.13)

$$\begin{aligned} E(f) &= \left(\boldsymbol{f}_l^{\mathrm{T}} \boldsymbol{f}_u^{\mathrm{T}}\right) \left(\begin{bmatrix} \mathbf{D}_{ll} & \mathbf{0}_{lu} \\ \mathbf{0}_{ul} & \mathbf{D}_{uu} \end{bmatrix} - \begin{bmatrix} \mathbf{W}_{ll} & \mathbf{W}_{lu} \\ \mathbf{W}_{ul} & \mathbf{W}_{uu} \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{f}_l \\ \boldsymbol{f}_u \end{bmatrix} \\ &= \boldsymbol{f}_l^{\mathrm{T}} (\mathbf{D}_{ll} - \mathbf{W}_{ll}) \boldsymbol{f}_l - 2\boldsymbol{f}_u^{\mathrm{T}} \mathbf{W}_{ul} \boldsymbol{f}_l + \boldsymbol{f}_u^{\mathrm{T}} (\mathbf{D}_{uu} - \mathbf{W}_{uu}) \boldsymbol{f}_u \end{aligned}$$

[解析]：这里第一项西瓜书中的符号有歧义，应该表示成 $\begin{bmatrix} \boldsymbol{f}_l^{\mathrm{T}} & \boldsymbol{f}_u^{\mathrm{T}} \end{bmatrix}$ 即一个 $\mathbb{R}^{1\times(l+u)}$ 的行向量。根据矩阵乘法的定义，有：

$$\begin{aligned} E(f) &= \begin{bmatrix} \boldsymbol{f}_l^{\mathrm{T}} & \boldsymbol{f}_u^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{D}_{ll} - \boldsymbol{W}_{ll} & -\boldsymbol{W}_{lu} \\ -\boldsymbol{W}_{ul} & \boldsymbol{D}_{uu} - \boldsymbol{W}_{uu} \end{bmatrix} \begin{bmatrix} \boldsymbol{f}_l \\ \boldsymbol{f}_u \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{f}_l^{\mathrm{T}} (\boldsymbol{D}_{ll} - \boldsymbol{W}_{ll}) - \boldsymbol{f}_u^{\mathrm{T}} \boldsymbol{W}_{ul} & -\boldsymbol{f}_l^{\mathrm{T}} \boldsymbol{W}_{lu} + \boldsymbol{f}_u^{\mathrm{T}} (\boldsymbol{D}_{uu} - \boldsymbol{W}_{uu}) \end{bmatrix} \begin{bmatrix} \boldsymbol{f}_l \\ \boldsymbol{f}_u \end{bmatrix} \\ &= \left(\boldsymbol{f}_l^{\mathrm{T}} (\boldsymbol{D}_{ll} - \boldsymbol{W}_{ll}) - \boldsymbol{f}_u^{\mathrm{T}} \boldsymbol{W}_{ul}\right) \boldsymbol{f}_l + \left(-\boldsymbol{f}_l^{\mathrm{T}} \boldsymbol{W}_{lu} + \boldsymbol{f}_u^{\mathrm{T}} (\boldsymbol{D}_{uu} - \boldsymbol{W}_{uu})\right) \boldsymbol{f}_u \\ &= \boldsymbol{f}_l^{\mathrm{T}} (\boldsymbol{D}_{ll} - \boldsymbol{W}_{ll}) \boldsymbol{f}_l - \boldsymbol{f}_u^{\mathrm{T}} \boldsymbol{W}_{ul} \boldsymbol{f}_l - \boldsymbol{f}_l^{\mathrm{T}} \boldsymbol{W}_{lu} \boldsymbol{f}_u + \boldsymbol{f}_u^{\mathrm{T}} (\boldsymbol{D}_{uu} - \boldsymbol{W}_{uu}) \boldsymbol{f}_u \\ &= \boldsymbol{f}_l^{\mathrm{T}} (\boldsymbol{D}_{ll} - \boldsymbol{W}_{ll}) \boldsymbol{f}_l - 2\boldsymbol{f}_u^{\mathrm{T}} \boldsymbol{W}_{ul} \boldsymbol{f}_l + \boldsymbol{f}_u^{\mathrm{T}} (\boldsymbol{D}_{uu} - \boldsymbol{W}_{uu}) \boldsymbol{f}_u \end{aligned}$$

其中最后一步，$\boldsymbol{f}_l^{\mathrm{T}} \boldsymbol{W}_{lu} \boldsymbol{f}_u = \left(\boldsymbol{f}_l^{\mathrm{T}} \boldsymbol{W}_{lu} \boldsymbol{f}_u\right)^{\mathrm{T}} = f_u^{\mathrm{T}} \boldsymbol{W}_{ul} \boldsymbol{f}_l$，因为这个式子的结果是一个标量。

13.12 公式 (13.14)

$$\begin{aligned} E(f) &= \left(\boldsymbol{f}_l^{\mathrm{T}} \boldsymbol{f}_u^{\mathrm{T}}\right) \left(\begin{bmatrix} \mathbf{D}_{ll} & \mathbf{0}_{lu} \\ \mathbf{0}_{ul} & \mathbf{D}_{uu} \end{bmatrix} - \begin{bmatrix} \mathbf{W}_{ll} & \mathbf{W}_{lu} \\ \mathbf{W}_{ul} & \mathbf{W}_{uu} \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{f}_l \\ \boldsymbol{f}_u \end{bmatrix} \\ &= \boldsymbol{f}_l^{\mathrm{T}} (\mathbf{D}_{ll} - \mathbf{W}_{ll}) \boldsymbol{f}_l - 2\boldsymbol{f}_u^{\mathrm{T}} \mathbf{W}_{ul} \boldsymbol{f}_l + \boldsymbol{f}_u^{\mathrm{T}} (\mathbf{D}_{uu} - \mathbf{W}_{uu}) \boldsymbol{f}_u \end{aligned}$$

[解析]：参考 13.13

13.13 公式 (13.15)

$$\boldsymbol{f}_u = (\mathbf{D}_{uu} - \mathbf{W}_{uu})^{-1} \mathbf{W}_{ul} \boldsymbol{f}_l$$

[解析]：由 13.13，有

$$\begin{aligned} \frac{\partial E(f)}{\partial \boldsymbol{f}_u} &= \frac{\partial \boldsymbol{f}_l^{\mathrm{T}} (\boldsymbol{D}_{ll} - \boldsymbol{W}_{ll}) \boldsymbol{f}_l - 2\boldsymbol{f}_u^{\mathrm{T}} \boldsymbol{W}_{ul} \boldsymbol{f}_l + \boldsymbol{f}_u^{\mathrm{T}} (\boldsymbol{D}_{uu} - \boldsymbol{W}_{uu}) \boldsymbol{f}_u}{\partial \boldsymbol{f}_u} \\ &= -2\boldsymbol{W}_{ul} \boldsymbol{f}_l + 2(\boldsymbol{D}_{uu} - \boldsymbol{W}_{uu}) \boldsymbol{f}_u \end{aligned}$$

令结果等于 0 即得 13.15。

13.14 公式 (13.16)

$$\begin{aligned}\mathbf{P}=\mathbf{D}^{-1}\mathbf{W}&=\begin{bmatrix}\mathbf{D}_{ll}^{-1} & \mathbf{0}_{lu}\\ \mathbf{0}_{ul} & \mathbf{D}_{uu}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{W}_{ll} & \mathbf{W}_{lu}\\ \mathbf{W}_{ul} & \mathbf{W}_{uu}\end{bmatrix}\\&=\begin{bmatrix}\mathbf{D}_{ll}^{-1}\mathbf{W}_{ll} & \mathbf{D}_{ll}^{-1}\mathbf{W}_{lu}\\ \mathbf{D}_{uu}^{-1}\mathbf{W}_{ul} & \mathbf{D}_{uu}^{-1}\mathbf{W}_{uu}\end{bmatrix}\end{aligned}$$

[解析]：根据矩阵乘法的定义计算可得该式，其中需要注意的是，对角矩阵 $\mathbf{D}$ 的拟等于其各个对角元素的倒数。

13.15 公式 (13.17)

$$\begin{aligned}\boldsymbol{f}_u&=\left(\mathbf{D}_{uu}\left(\mathbf{I}-\mathbf{D}_{uu}^{-1}\mathbf{W}_{uu}\right)\right)^{-1}\mathbf{W}_{ul}\boldsymbol{f}_l\\&=\left(\mathbf{I}-\mathbf{D}_{uu}^{-1}\mathbf{W}_{uu}\right)^{-1}\mathbf{D}_{uu}^{-1}\mathbf{W}_{ul}\boldsymbol{f}_l\\&=\left(\mathbf{I}-\mathbf{P}_{uu}\right)^{-1}\mathbf{P}_{ul}\boldsymbol{f}_l\end{aligned}$$

[解析]：第一项到第二项是根据矩阵乘法逆的定义：$(\mathbf{AB})^{-1}=\mathbf{B}^{-1}\mathbf{A}^{-1}$，在这个式子中

$$\begin{aligned}\mathbf{P}_{uu}&=\mathbf{D}_{uu}^{-1}\mathbf{W}_{uu}\\\mathbf{P}_{ul}&=\mathbf{D}_{uu}^{-1}\mathbf{W}_{ul}\end{aligned}$$

均可以根据 $\mathbf{W}_{ij}$ 计算得到，因此可以通过标记 $\mathbf{f}_l$ 计算未标记数据的标签 $\mathbf{f}_u$。

13.16 公式 (13.20)

$$\mathbf{F}^*=\lim_{t\to\infty}\mathbf{F}(t)=(1-\alpha)(\mathbf{I}-\alpha\mathbf{S})^{-1}\mathbf{Y}$$

[解析]：由 13.19

$$\mathbf{F}(t+1)=\alpha\mathbf{S}\mathbf{F}(t)+(1-\alpha)\mathbf{Y}$$

当 t 取不同的值时，有：

$$\begin{aligned}t=0:\mathbf{F}(1)&=\alpha\mathbf{S}\mathbf{F}(0)+(1-\alpha)\mathbf{Y}\\&=\alpha\mathbf{S}\mathbf{Y}+(1-\alpha)\mathbf{Y}\\t=1:\mathbf{F}(2)&=\alpha\mathbf{S}\mathbf{F}(1)+(1-\alpha)\mathbf{Y}=\alpha\mathbf{S}(\alpha\mathbf{S}\mathbf{Y}+(1-\alpha)\mathbf{Y})+(1-\alpha)\mathbf{Y}\\&=(\alpha\mathbf{S})^2\mathbf{Y}+(1-\alpha)\left(\sum_{i=0}^{1}(\alpha\mathbf{S})^i\right)\mathbf{Y}\\t=2:\mathbf{F}(3)&=\alpha\mathbf{S}\mathbf{F}(2)+(1-\alpha)\mathbf{Y}\\&=\alpha\mathbf{S}\left((\alpha\mathbf{S})^2\mathbf{Y}+(1-\alpha)\left(\sum_{i=0}^{1}(\alpha\mathbf{S})^i\right)\mathbf{Y}\right)+(1-\alpha)\mathbf{Y}\\&=(\alpha\mathbf{S})^3\mathbf{Y}+(1-\alpha)\left(\sum_{i=0}^{2}(\alpha\mathbf{S})^i\right)\mathbf{Y}\end{aligned}$$

可以观察到规律

$$\mathbf{F}(t)=(\alpha\mathbf{S})^t\mathbf{Y}+(1-\alpha)\left(\sum_{i=0}^{t-1}(\alpha\mathbf{S})^i\right)\mathbf{Y}$$

则

$$\mathbf{F}^* = \lim_{t\to\infty}\mathbf{F}(t) = \lim_{t\to\infty}(\alpha\mathbf{S})^t\mathbf{Y} + \lim_{t\to\infty}(1-\alpha)\left(\sum_{i=0}^{t-1}(\alpha\mathbf{S})^i\right)\mathbf{Y}$$

其中第一项由于 $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$ 的特征值介于 [-1, 1] 之间 [1]，而 $\alpha \in (0,1)$，所以 $\lim_{t\to\infty}(\alpha\mathbf{S})^t = 0$，第二项由等比数列公式

$$\lim_{t\to\infty}\sum_{i=0}^{t-1}(\alpha\mathbf{S})^i = \frac{\mathbf{I} - \lim_{t\to\infty}(\alpha\mathbf{S})^t}{\mathbf{I}-\alpha\mathbf{S}} = \frac{\mathbf{I}}{\mathbf{I}-\alpha\mathbf{S}} = (\mathbf{I}-\alpha\mathbf{S})^{-1}$$

综合可得式 13.20。

参考文献

[1] Wikipedia contributors. Laplacian matrix, 2020. URL: https://en.wikipedia.org/wiki/Laplacian_matrix.

第 14 章　概率图模型

14.1　公式 (14.1)

$$P\left(x_1, y_1, \ldots, x_n, y_n\right) = P\left(y_1\right) P\left(x_1 | y_1\right) \prod_{i=2}^{n} P\left(y_i | y_{i-1}\right) P\left(x_i | y_i\right)$$

[解析]：所有的相乘关系都表示概率的相互独立。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，输出观测概率和条件转移概率。

14.2　公式 (14.2)

$$P(\mathbf{x}) = \frac{1}{Z} \prod_{Q \in C} \psi_Q\left(\mathbf{x}_Q\right)$$

[解析]：因为各个团之间概率分布相互独立，因此它们连乘可以表示最终的概率。

14.3　公式 (14.3)

$$P(\mathbf{x}) = \frac{1}{Z^*} \prod_{Q \in C*} \psi_Q\left(\mathbf{x}_Q\right)$$

[解析]：意义同式 14.2, 区别在于此处的团为极大团。

14.4　公式 (14.4)

$$P\left(x_A, x_B, x_C\right) = \frac{1}{Z} \psi_{AC}\left(x_A, x_C\right) \psi_{BC}\left(x_B, x_C\right)$$

[解析]：将图 14.3 分解成 x_A, x_C 和 x_B, x_C 两个团。

14.5　公式 (14.5)

$$P\left(x_A, x_B | x_C\right) = \frac{\psi_{AC}\left(x_A, x_C\right)}{\sum_{x'_A} \psi_{AC}\left(x'_A, x_C\right)} \cdot \frac{\psi_{BC}\left(x_B, x_C\right)}{\sum_{x'_B} \psi_{BC}\left(x'_B, x_C\right)}$$

[推导]：参见原书推导。

14.6　公式 (14.6)

$$P\left(x_A | x_C\right) = \frac{\psi_{AC}\left(x_A, x_C\right)}{\sum_{x_A} \psi_{AC}\left(x'_A, x_C\right)}$$

[推导]：参见原书推导。

14.7　公式 (14.7)

$$P\left(x_A, x_B | x_C\right) = P\left(x_A | x_C\right) P\left(x_B | x_C\right)$$

[解析]：可由 14.5、14.6 联立可得。

14.8 公式 (14.8)

$$\psi_Q\left(\mathbf{x}_Q\right)=e^{-H_Q\left(\mathbf{x}_Q\right)}$$

[解析]：此为势函数的定义式，即将势函数写作指数函数的形式。指数函数满足非负性，且便于求导，因此在机器学习中具有广泛应用，例如西瓜书公式 8.5 和 13.11。

14.9 公式 (14.9)

$$H_Q\left(\mathbf{x}_Q\right)=\sum_{u,v\in Q,u\neq v}\alpha_{uv}x_ux_v+\sum_{v\in Q}\beta_vx_v$$

[解析]：此为定义在变量 $\mathbf{x}_Q$ 上的函数 $H_Q\left(\cdot\right)$ 的定义式，第二项考虑单节点，第一项考虑每一对节点之间的关系。

14.10 公式 (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，则该变量独立与其他变量，即该变量只与其邻接变量有关，所以式 14.10 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。

14.11 公式 (14.14)

$$\begin{aligned}P\left(x_5\right)&=\sum_{x_4}\sum_{x_3}\sum_{x_2}\sum_{x_1}P\left(x_1,x_2,x_3,x_4,x_5\right)\\&=\sum_{x_4}\sum_{x_3}\sum_{x_2}\sum_{x_1}P\left(x_1\right)P\left(x_2|x_1\right)P\left(x_3|x_2\right)P\left(x_4|x_3\right)P\left(x_5|x_3\right)\end{aligned}$$

[解析]：在消去变量的过程中，在消去每一个变量时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。

14.12 公式 (14.15)

$$\begin{aligned}P\left(x_5\right)&=\sum_{x_3}P\left(x_5|x_3\right)\sum_{x_4}P\left(x_4|x_3\right)\sum_{x_2}P\left(x_3|x_2\right)\sum_{x_1}P\left(x_1\right)P\left(x_2|x_1\right)\\&=\sum_{x_3}P\left(x_5|x_3\right)\sum_{x_4}P\left(x_4|x_3\right)\sum_{x_2}P\left(x_3|x_2\right)m_{12}\left(x_2\right)\end{aligned}$$

[解析]：变量消去的顺序为从右至左求和号的下标，应当注意 x_4 与 x_5 相互独立，因此可与 x_3 的消去顺序互换，对最终结果无影响。

14.13 公式 (14.16)

$$\begin{aligned} P\left(x_5\right) &= \sum_{x_3} P\left(x_5|x_3\right) \sum_{x_4} P\left(x_4|x_3\right) m_{23}\left(x_3\right) \\ &= \sum_{x_3} P\left(x_5|x_3\right) m_{23}\left(x_3\right) \sum_{x_4} P\left(x_4|x_3\right) \\ &= \sum_{x_3} P\left(x_5|x_3\right) m_{23}\left(x_3\right) \\ &= m_{35}\left(x_5\right) \end{aligned}$$

[解析]：注意到 $\sum_{x_4} P\left(x_4|x_3\right) = 1$。

14.14 公式 (14.17)

$$P\left(x_1, x_2, x_3, x_4, x_5\right) = \frac{1}{Z} \psi_{12}\left(x_1, x_2\right) \psi_{23}\left(x_2, x_3\right) \psi_{34}\left(x_3, x_4\right) \psi_{35}\left(x_3, x_5\right)$$

[解析]：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 为规范化因子确保所有可能性的概率之和为 1。

14.15 公式 (14.18)

$$\begin{aligned} P\left(x_5\right) &= \frac{1}{Z} \sum_{x_3} \psi_{35}\left(x_3, x_5\right) \sum_{x_4} \psi_{34}\left(x_3, x_4\right) \sum_{x_2} \psi_{23}\left(x_2, x_3\right) \sum_{x_1} \psi_{12}\left(x_1, x_2\right) \\ &= \frac{1}{Z} \sum_{x_3} \psi_{35}\left(x_3, x_5\right) \sum_{x_4} \psi_{34}\left(x_3, x_4\right) \sum_{x_2} \psi_{23}\left(x_2, x_3\right) m_{12}\left(x_2\right) \\ &= \cdots \\ &= \frac{1}{Z} m_{35}\left(x_5\right) \end{aligned}$$

[解析]：原理同式 14.15, 区别在于把条件概率替换为势函数。

14.16 公式 (14.19)

$$m_{ij}\left(x_j\right) = \sum_{x_i} \psi\left(x_i, x_j\right) \prod_{k \in n(i) \backslash j} m_{ki}\left(x_i\right)$$

[解析]：该式表示从节点 i 传递到节点 j 的过程，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。应当注意这里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 以外的消息告诉节点 j 即可。

14.17 公式 (14.20)

$$P\left(x_i\right) \propto \prod_{k \in n(i)} m_{ki}\left(x_i\right)$$

[解析]：应当注意这里是正比于而不是等于，因为涉及到概率的规范化。可以这么解释，每个变量可以看作一个有一些邻居的房子，每个邻居根据其自己的见闻告诉你一些事情 (消息)，任何一条消息的可信度应当与所有邻居都有相关性，此处这种相关性用乘积来表达。

14.18 公式 (14.22)

$$\hat{f} = \frac{1}{N}\sum_{i=1}^{N} f\left(x_i\right)$$

[推导]：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

$$\begin{aligned}\hat{f} &= \frac{1}{N}\sum_{j=1}^{M} f\left(x_j\right)\cdot m_j \\ &= \sum_{j=1}^{M} f\left(x_j\right)\cdot \frac{m_j}{N} \\ &\approx \sum_{j=1}^{M} f\left(x_j\right)\cdot p(x_j) \\ &\approx \int f(x)p(x)dx\end{aligned}$$

14.19 公式 (14.26)

$$p\left(\mathbf{x}^t\right)T\left(\mathbf{x}^{t-1}\mid\mathbf{x}^t\right) = p\left(\mathbf{x}^{t-1}\right)T\left(\mathbf{x}^t\mid\mathbf{x}^{t-1}\right)$$

[解析]：假设变量 $\mathbf{x}$ 所在的空间有 n 个状态 $(s_1, s_2, .., s_n)$, 定义在该空间上的一个转移矩阵 $\mathbf{T}\in\mathbb{R}^{n\times n}$ 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

$$\boldsymbol{\pi}\mathbf{T} = \boldsymbol{\pi}$$

其中, $\boldsymbol{\pi}$ 是一个是一个 n 维向量，代表 $s_1, s_2, .., s_n$ 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 $\mathbf{x}^1, \mathbf{x}^2, .., \mathbf{x}^t$, 应当采用哪一个转移矩阵 $\mathbf{T}\in\mathbb{R}^{n\times n}$ 呢?

事实上，转移矩阵只需要满足马尔可夫细致平稳条件

$$\pi_i\mathbf{T}_{ij} = \pi_j\mathbf{T}_{ji}$$

即公式 14.26，这里采用的符号与西瓜书略有区别以便于理解. 证明如下

$$\boldsymbol{\pi}\mathbf{T}_{j\cdot} = \sum_i \pi_i\mathbf{T}_{ij} = \sum_i \pi_j\mathbf{T}_{ji} = \pi_j$$

假设采样得到的序列为 $\mathbf{x}^1, \mathbf{x}^2, .., \mathbf{x}^{t-1}, \mathbf{x}^t$，则可以使用 MH 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。

14.20 公式 (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 Q 和 A 之积，即先验概率和接受概率，便于实际算法的实现。

14.21 公式 (14.28)

$$A(x^*|x^{t-1}) = \min\left(1, \frac{p(x^*)Q(x^{t-1}|x^*)}{p(x^{t-1})Q(x^*|x^{t-1})}\right)$$

[推导]：这个公式其实是拒绝采样的一个 trick，因为基于式 14.27 只需要

$$A(x^*|x^{t-1}) = p(x^*)Q(x^{t-1}|x^*)$$
$$A(x^{t-1}|x^*) = p(x^{t-1})Q(x^*|x^{t-1})$$

即可满足式 14.26，但是实际上等号右边的数值可能比较小，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

$$A(x^*|x^{t-1}) = \frac{p(x^*)Q(x^{t-1}|x^*)}{norm}$$
$$A(x^{t-1}|x^*) = \frac{p(x^{t-1})Q(x^*|x^{t-1})}{norm}$$

其中

$$norm = \max\left(p(x^{t-1})Q(x^*|x^{t-1}), p(x^*)Q(x^{t-1}|x^*)\right)$$

即西瓜书中的 14.28。

14.22 公式 (14.29)

$$p(\mathbf{x}|\Theta) = \prod_{i=1}^{N}\sum_{\mathbf{z}} p\left(x_i, \mathbf{z}|\Theta\right)$$

[解析]：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。

14.23 公式 (14.30)

$$\ln p(\mathbf{x}|\Theta) = \sum_{i=1}^{N} \ln\left\{\sum_{\mathbf{z}} p\left(x_i, \mathbf{z}|\Theta\right)\right\}$$

[解析]：对式 14.29 取对数。

14.24 公式 (14.31)

$$\begin{aligned}\Theta^{t+1} &= \arg\max_{\Theta} \mathcal{Q}\left(\Theta; \Theta^t\right) \\ &= \arg\max_{\Theta} \sum_{\mathbf{z}} p\left(\mathbf{z}|\mathbf{x}, \Theta^t\right) \ln p(\mathbf{x}, \mathbf{z}|\Theta)\end{aligned}$$

[解析]：EM 算法中的 M 步，参见 7.6 节。

14.25 公式 (14.32)

$$\ln p(x) = \mathcal{L}(q) + \mathrm{KL}(q \parallel p)$$

[推导]：根据条件概率公式 $p(x,z)=p(z|x)*p(x)$，可以得到 $p(x)=\frac{p(x,z)}{p(z|x)}$ 然后两边同时作用 ln 函数，可得 $\ln p(x)=\ln\frac{p(x,z)}{p(z|x)}$ 因为 $q(z)$ 是概率密度函数，所以 $1=\int q(z)dz$ 等式两边同时乘以 $\ln p(x)$，因为 $\ln p(x)$ 是不关于变量 z 的函数，所以 $\ln p(x)$ 可以拿进积分里面，得到 $\ln p(x)=\int q(z)\ln p(x)dz$

$$\begin{aligned}\ln p(x)&=\int q(z)\ln p(x)dz\\&=\int q(z)\ln\frac{p(x,z)}{p(z|x)}\\&=\int q(z)\ln\left\{\frac{p(x,z)}{q(z)}\cdot\frac{q(z)}{p(z|x)}\right\}\\&=\int q(z)\left(\ln\frac{p(x,z)}{q(z)}-\ln\frac{p(z|x)}{q(z)}\right)\\&=\int q(z)\ln\left\{\frac{p(x,z)}{q(z)}\right\}-\int q(z)\ln\frac{p(z|x)}{q(z)}\\&=\mathcal{L}(q)+\mathrm{KL}(q\parallel p)\end{aligned}$$

最后一行是根据 $\mathcal{L}$ 和 KL 的定义。

14.26 公式 (14.33)

$$\mathcal{L}(q)=\int q(\mathbf{z})\ln\left\{\frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right\}\mathrm{d}\mathbf{z}$$

[解析]：见 14.32 解析。

14.27 公式 (14.34)

$$\mathrm{KL}(q\|p)=-\int q(\mathrm{z})\ln\frac{p(\mathrm{z}|\mathrm{x})}{q(\mathrm{z})}\mathrm{dz}$$

[解析]：见 14.32 解析。

14.28 公式 (14.35)

$$q(\mathbf{z})=\prod_{i=1}^{M}q_i\left(\mathbf{z}_i\right)$$

[解析]：再一次，条件独立的假设。可以看到，当问题复杂是往往简化问题到最简单最容易计算的局面，实际上往往效果不错。

14.29 公式 (14.36)

$$\begin{aligned}\mathcal{L}(q)&=\int\prod_i q_i\left\{\ln p(\mathbf{x},\mathbf{z})-\sum_i\ln q_i\right\}d\mathbf{z}\\&=\int q_j\left\{\int p(x,z)\prod_{i\neq j}q_i d\mathbf{z_i}\right\}d\mathbf{z_j}-\int q_j\ln q_j d\mathbf{z_j}+\mathrm{const}\\&=\int q_j\ln\tilde{p}(\mathbf{x},\mathbf{z_j})d\mathbf{z_j}-\int q_j\ln q_j d\mathbf{z_j}+\mathrm{const}\end{aligned}$$

[推导]：

$$\mathcal{L}(q)=\int\prod_i q_i\left\{\ln p(\mathbf{x},\mathbf{z})-\sum_i\ln q_i\right\}d\mathbf{z}=\int\prod_i q_i\ln p(\mathbf{x},\mathbf{z})d\mathbf{z}-\int\prod_i q_i\sum_i\ln q_i d\mathbf{z}$$

公式可以看做两个积分相减，我们先来看左边积分 $\int \prod_i q_i \ln p(\mathbf{x},\mathbf{z}) d\mathbf{z}$ 的推导。

$$\begin{aligned}\int \prod_i q_i \ln p(\mathbf{x},\mathbf{z}) d\mathbf{z} &= \int q_j \prod_{i\neq j} q_i \ln p(\mathbf{x},\mathbf{z}) d\mathbf{z} \\ &= \int q_j \left\{ \int \ln p(\mathbf{x},\mathbf{z}) \prod_{i\neq j} q_i d\mathbf{z_i} \right\} d\mathbf{z_j}\end{aligned}$$

即先对 $\mathbf{z_j}$ 求积分，再对 $\mathbf{z_i}$ 求积分，这个就是教材中的 14.36 左边的积分部分。我们现在看下右边积分的推导 $\int \prod_i q_i \sum_i \ln q_i d\mathbf{z}$ 的推导。在此之前我们看下 $\int \prod_i q_i \ln q_k d\mathbf{z}$ 的计算

$$\begin{aligned}\int \prod_i q_i \ln q_k d\mathbf{z} &= \int q_{i'} \prod_{i\neq i'} q_i \ln q_k d\mathbf{z} \\ &= \int q_{i'} \left\{ \int \prod_{i\neq i'} q_i \ln q_k d\mathbf{z_i} \right\} d\mathbf{z_{i'}}\end{aligned}$$

第一个等式是一个展开项，选取一个变量 $q_{i'}, i' \neq k$，由于 $\left\{ \int \prod_{i\neq i'} q_i \ln q_k d\mathbf{z_i} \right\}$ 部分与变量 $q_{i'}$ 无关，所以可以拿到积分外面。又因为 $\int q_{i'} d\mathbf{z_{i'}} = 1$，所以

$$\begin{aligned}\int \prod_i q_i \ln q_k d\mathbf{z} &= \int \prod_{i\neq i'} q_i \ln q_k d\mathbf{z_i} \\ &= \int q_k \ln q_k d\mathbf{z_k}\end{aligned}$$

即所有 k 以外的变量都可以通过上面的方式消除，有了这个结论，我们再来看公式

$$\begin{aligned}\int \prod_i q_i \sum_i \ln q_i d\mathbf{z} &= \int \prod_i q_i \ln q_j d\mathbf{z} + \sum_{k\neq j} \int \prod_i q_i \ln q_k d\mathbf{z} \\ &= \int q_j \ln q_j d\mathbf{z_j} + \sum_{k\neq j} \int q_k \ln q_k d\mathbf{z_k} \\ &= \int q_j \ln q_j d\mathbf{z_j} + \text{const}\end{aligned}$$

其中第二个等式是依据上述规律进行消除，最后将与 q_j 无关的部分写作 const，这个就是 14.36 右边的积分部分。

14.30 公式 (14.37)

$$\ln \tilde{p}\left(\mathbf{x},\mathbf{z}_j\right) = \mathbb{E}_{i\neq j}[\ln p(\mathbf{x},\mathbf{z})] + \text{ const}$$

[解析]：参见 14.36

14.31 公式 (14.38)

$$\mathbb{E}_{i\neq j}[\ln p(\mathbf{x},\mathbf{z})] = \int \ln p(\mathbf{x},\mathbf{z}) \prod_{i\neq j} q_i \mathrm{d}\mathbf{z}_i$$

[解析]：参见 14.36

14.32 公式 (14.39)

$$\ln q_j^*\left(\mathbf{z}_j\right) = \mathbb{E}_{i\neq j}[\ln p(\mathbf{x},\mathbf{z})] + \text{const}$$

[解析]：散度取得极值的条件是两个概率分布相同，见附录 $C.3$。

14.33 公式 (14.40)

$$q_j^*(\mathbf{z}_j) = \frac{\exp\left(\mathbb{E}_{i\neq j}[\ln(p(\mathbf{x},\mathbf{z}))]\right)}{\int \exp\left(\mathbb{E}_{i\neq j}[\ln(p(\mathbf{x},\mathbf{z}))]\right) \mathrm{d}\mathbf{z}_j}$$

[推导]：由 14.39 去对数并积分

$$\begin{aligned}\int q_j^*(\mathbf{z}_j)\mathrm{d}\mathbf{z}_j &= \int \exp\left(\mathbb{E}_{i\neq j}[\ln(p(\mathbf{x},\mathbf{z}))]\right) \cdot \exp(const)\, \mathrm{d}\mathbf{z}_j \\ &= \exp(const) \int \exp\left(\mathbb{E}_{i\neq j}[\ln(p(\mathbf{x},\mathbf{z}))]\right)\, \mathrm{d}\mathbf{z}_j \\ &= 1\end{aligned}$$

所以

$$\exp(const) = \frac{1}{\int \exp\left(\mathbb{E}_{i\neq j}[\ln(p(\mathbf{x},\mathbf{z}))]\right)\, \mathrm{d}\mathbf{z}_j}$$

$$\begin{aligned}q_j^*(\mathbf{z}_j) &= \exp\left(\mathbb{E}_{i\neq j}[\ln(p(\mathbf{x},\mathbf{z}))]\right) \cdot \exp(const) \\ &= \frac{\exp\left(\mathbb{E}_{i\neq j}[\ln(p(\mathbf{x},\mathbf{z}))]\right)}{\int \exp\left(\mathbb{E}_{i\neq j}[\ln(p(\mathbf{x},\mathbf{z}))]\right) \mathrm{d}\mathbf{z}_j}\end{aligned}$$

14.34 公式 (14.41)

$$p(\boldsymbol{W},\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta}|\boldsymbol{\alpha},\boldsymbol{\eta}) = \prod_{t=1}^{T} p(\boldsymbol{\theta}_t|\boldsymbol{\alpha}) \prod_{k=1}^{K} p(\boldsymbol{\beta}_k|\boldsymbol{\eta}) (\prod_{n=1}^{N} P(w_{t,n}|z_{t,n},\boldsymbol{\beta}_k) P(z_{t,n}|\boldsymbol{\theta}_t))$$

[解析]：此式表示 LDA 模型下根据参数 α,η 生成文档 W 的概率。其中 z,β,θ 是生成过程的中间变量。具体的生成步骤可见概率图 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。这里共有三个连乘符号，表示三个相互独立的概率关系。第一个连乘表示 T 个文档每个文档的话题分布都是相互独立的。第二个连乘表示 K 个话题每个话题下单词的分布是相互独立的。最后一个连乘号表示每篇文档中的所有单词的生成是相互独立的。

14.35 公式 (14.42)

$$p\left(\Theta_t|\boldsymbol{\alpha}\right) = \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma\left(\alpha_k\right)} \prod_k \Theta_{t,k}^{\alpha_k - 1}$$

[解析]：参见附录 $C1.6$。

14.36 公式 (14.43)

$$LL(\boldsymbol{\alpha},\boldsymbol{\eta}) = \sum_{t=1}^{T} \ln p\left(\boldsymbol{w}_t|\boldsymbol{\alpha},\boldsymbol{\eta}\right)$$

[解析]：对数似然函数。参见 7.2 极大似然估计。

14.37 公式 (14.44)

$$p(\mathbf{z},\boldsymbol{\beta},\Theta|\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\eta}) = \frac{p(\mathbf{W},\mathbf{z},\boldsymbol{\beta},\boldsymbol{\Theta}|\boldsymbol{\alpha},\boldsymbol{\eta})}{p(\mathbf{W}|\boldsymbol{\alpha},\boldsymbol{\eta})}$$

[解析]：分母为边际分布，需要对变量 $\mathbf{z},\boldsymbol{\beta},\Theta$ 积分或者求和，所以往往难以直接求解。

第 15 章　规则学习

15.1　公式 (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]：似然率统计量 (Likelihood Ratio Statistics) 的定义式。

15.2　公式 (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]：FOIL 增益 (FOIL gain) 的定义式。

15.3　公式 (15.6)

$$(A \vee B) - \{B\} = A$$

[解析]：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。

15.4　公式 (15.7)

$$C = (C_1 - \{L\}) \vee (C_2 - \{\neg L\})$$

[解析]：$C = A \vee B$，把 $A = C_1 - \{L\}$ 和 $L = C_2 - \{\neg L\}$ 带入即得。

15.5　公式 (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。

15.6　公式 (15.10)

$$\frac{p \leftarrow A \wedge B \quad q \leftarrow A}{p \leftarrow q \wedge B \quad q \leftarrow A}$$

[解析]：吸收 (absorption) 操作的定义。

15.7　公式 (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]：辨识 (identification) 操作的定义。

15.8 公式 (15.12)

$$\frac{p \leftarrow A \wedge B \quad p \leftarrow A \wedge q}{q \leftarrow B \quad p \leftarrow A \wedge q \quad q \leftarrow C}$$

[解析]：内构 (intra-construction) 操作的定义。

15.9 公式 (15.13)

$$\frac{p \leftarrow A \wedge B \quad q \leftarrow r \wedge C}{p \leftarrow r \wedge B \quad r \leftarrow A \quad q \leftarrow r \wedge C}$$

[解析]：互构 (inter-construction) 操作的定义。

15.10 公式 (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。

15.11 公式 (15.16)

$$C_2 = (C - (C_1 - \{L_1\})\,\theta_1 \vee \{\neg L_1\theta_1\})\,\theta_2^{-1}$$

[推导]：θ_1 为作者笔误，由 15.9

$$C_2 = (C - (C_1 - \{L_1\})) \vee \{L_2\}$$

因为 $L_2 = (\neg L_1\theta_1)\theta_2^{-1}$，替换得证。

第 16 章 强化学习

16.1 公式 (16.2)

$$Q_n(k) = \frac{1}{n}\left((n-1) \times Q_{n-1}(k) + v_n\right)$$

[推导]:

$$\begin{aligned} Q_n(k) &= \frac{1}{n}\sum_{i=1}^{n} v_i \\ &= \frac{1}{n}\left(\sum_{i=1}^{n-1} v_i + v_n\right) \\ &= \frac{1}{n}\left((n-1) \times Q_{n-1}(k) + v_n\right) \\ &= Q_{n-1}(k) + \frac{1}{n}\left(v_n - Q_{n-1}(k)\right) \end{aligned}$$

16.2 公式 (16.3)

$$\begin{aligned} Q_n(k) &= \frac{1}{n}\left((n-1) \times Q_{n-1}(k) + v_n\right) \\ &= Q_{n-1}(k) + \frac{1}{n}\left(v_n - Q_{n-1}(k)\right) \end{aligned}$$

[推导]: 参见 16.2

16.3 公式 (16.4)

$$P(k) = \frac{e^{\frac{Q(k)}{\tau}}}{\sum_{i=1}^{K} e^{\frac{Q(i)}{\tau}}}$$

[解析]:

$$P(k) = \frac{e^{\frac{Q(k)}{\tau}}}{\sum_{i=1}^{K} e^{\frac{Q(i)}{\tau}}} \propto e^{\frac{Q(k)}{\tau}} \propto \frac{Q(k)}{\tau} \propto \frac{1}{\tau}$$

16.4 公式 (16.7)

$$\begin{aligned} V_T^{\pi}(x) &= \mathbb{E}_{\pi}[\frac{1}{T}\sum_{t=1}^{T} r_t \mid x_0 = x] \\ &= \mathbb{E}_{\pi}[\frac{1}{T}r_1 + \frac{T-1}{T}\frac{1}{T-1}\sum_{t=2}^{T} r_t \mid x_0 = x] \\ &= \sum_{a\in A}\pi(x,a)\sum_{x'\in X} P_{x\to x'}^{a}(\frac{1}{T}R_{x\to x'}^{a} + \frac{T-1}{T}\mathbb{E}_{\pi}[\frac{1}{T-1}\sum_{t=1}^{T-1} r_t \mid x_0 = x']) \\ &= \sum_{a\in A}\pi(x,a)\sum_{x'\in X} P_{x\to x'}^{a}(\frac{1}{T}R_{x\to x'}^{a} + \frac{T-1}{T}V_{T-1}^{\pi}(x')]) \end{aligned}$$

[解析]: 因为

$$\pi(x,a) = P(action = a|state = x)$$

表示在状态 x 下选择动作 a 的概率，又因为动作事件之间两两互斥且和为动作空间，由全概率展开公式

$$P(A)=\sum_{i=1}^{\infty}P(B_i)P(A\mid B_i)$$

可得

$$\begin{aligned}&\mathbb{E}_\pi[\frac{1}{T}r_1+\frac{T-1}{T}\frac{1}{T-1}\sum_{t=2}^{T}r_t\mid x_0=x]\\&=\sum_{a\in A}\pi(x,a)\sum_{x'\in X}P^a_{x\to x'}(\frac{1}{T}R^a_{x\to x'}+\frac{T-1}{T}\mathbb{E}_\pi[\frac{1}{T-1}\sum_{t=1}^{T-1}r_t\mid x_0=x'])\end{aligned}$$

其中

$$r_1=\pi(x,a)P^a_{x\to x'}R^a_{x\to x'}$$

最后一个等式用到了递归形式。

16.5 公式 (16.8)

$$V^\pi_\gamma(x)=\sum_{a\in A}\pi(x,a)\sum_{x'\in X}P^a_{x\to x'}(R^a_{x\to x'}+\gamma V^\pi_\gamma(x'))$$

[推导]:

$$\begin{aligned}V^\pi_\gamma(x)&=\mathbb{E}_\pi[\sum_{t=0}^{\infty}\gamma^t r_{t+1}\mid x_0=x]\\&=\mathbb{E}_\pi[r_1+\sum_{t=1}^{\infty}\gamma^t r_{t+1}\mid x_0=x]\\&=\mathbb{E}_\pi[r_1+\gamma\sum_{t=1}^{\infty}\gamma^{t-1} r_{t+1}\mid x_0=x]\\&=\sum_{a\in A}\pi(x,a)\sum_{x'\in X}P^a_{x\to x'}(R^a_{x\to x'}+\gamma\mathbb{E}_\pi[\sum_{t=0}^{\infty}\gamma^t r_{t+1}\mid x_0=x'])\\&=\sum_{a\in A}\pi(x,a)\sum_{x'\in X}P^a_{x\to x'}(R^a_{x\to x'}+\gamma V^\pi_\gamma(x'))\end{aligned}$$

16.6 公式 (16.10)

$$\begin{cases}Q^\pi_T(x,a)=\sum_{x'\in X}P^a_{x\to x'}\left(\frac{1}{T}R^a_{x\to x'}+\frac{T-1}{T}V^\pi_{T-1}(x')\right)\\Q^\pi_\gamma(x,a)=\sum_{x'\in X}P^a_{x\to x'}\left(R^a_{x\to x'}+\gamma V^\pi_\gamma(x')\right)\end{cases}$$

[推导]: 参见 16.7, 16.8

16.7 公式 (16.14)

$$V^*(x)=\max_{a\in A}Q^{\pi^*}(x,a)$$

[解析]: 为了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 最大的状态 $\max_{a\in A}$。

16.8 公式 (16.16)

$$V^{\pi}(x) \leq V^{\pi'}(x)$$

[推导]:

$$\begin{aligned}
V^{\pi}(x) &\leq Q^{\pi}(x, \pi'(x)) \\
&= \sum_{x' \in X} P_{x \to x'}^{\pi'(x)} (R_{x \to x'}^{\pi'(x)} + \gamma V^{\pi}(x')) \\
&\leq \sum_{x' \in X} P_{x \to x'}^{\pi'(x)} (R_{x \to x'}^{\pi'(x)} + \gamma Q^{\pi}(x', \pi'(x'))) \\
&= \sum_{x' \in X} P_{x \to x'}^{\pi'(x)} (R_{x \to x'}^{\pi'(x)} + \gamma \sum_{x' \in X} P_{x' \to x'}^{\pi'(x')} (R_{x' \to x'}^{\pi'(x')} + \gamma V^{\pi}(x'))) \\
&= \sum_{x' \in X} P_{x \to x'}^{\pi'(x)} (R_{x \to x'}^{\pi'(x)} + \gamma V^{\pi'}(x')) \\
&= V^{\pi'}(x)
\end{aligned}$$

其中，使用了动作改变条件

$$Q^{\pi}(x, \pi'(x)) \geq V^{\pi}(x)$$

以及状态-动作值函数

$$Q^{\pi}(x', \pi'(x')) = \sum_{x' \in X} P_{x' \to x'}^{\pi'(x')} (R_{x' \to x'}^{\pi'(x')} + \gamma V^{\pi}(x'))$$

于是，当前状态的最优值函数为

$$V^{*}(x) = V^{\pi'}(x) \geq V^{\pi}(x)$$

16.9 公式 (16.31)

$$Q_{t+1}^{\pi}(x, a) = Q_{t}^{\pi}(x, a) + \alpha (R_{x \to x'}^{a} + \gamma Q_{t}^{\pi}(x', a') - Q_{t}^{\pi}(x, a))$$

[推导]：对比公式 16.29

$$Q_{t+1}^{\pi}(x, a) = Q_{t}^{\pi}(x, a) + \frac{1}{t+1} (r_{t+1} - Q_{t}^{\pi}(x, a))$$

以及由

$$\frac{1}{t+1} = \alpha$$

可知，若下式成立，则公式 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。