目录

第3章 贝叶斯决策

贝叶斯决策是一种基于概率论的模式分类方法，能够在数学层面为分类问题的求解提供基础理论框架。贝叶斯决策的核心内容包括贝叶斯分类准则与概率模型估计两个方面，前者为样本的分类提供了基于概率的判断规则，后者为概率值提供基于数据的估计方法，两者结合才能实现模式分类任务。作为一种典型的生成类模型，贝叶斯分类器与以线性分类器为代表的判别类模型在数学本质上具有深层次的关联性，这也是本书为什么首先介绍贝叶斯决策论的原因。

3.1 贝叶斯决策论概述

贝叶斯决策论是一套基于概率论的决策方法。在分类问题上，贝叶斯决策论可以根据某种概率的大小判断当前对象的类别。

在人工智能技术发展早期，决策问题完全建立在确定性集合论基础上。简单地说，当我们进行分类任务时，只考虑“如果……就是……，否则……就不是……”的二值逻辑，比如“如果太阳没有落山，就是白天；如果太阳落山了，就是夜晚。”然而在很多真实的分类问题中，观察数据与对象类别的联系存在着不确定性。比如不是蚊虫低飞就一定会下雨，不是瓜蒂蜷曲西瓜就一定会甜，不是发达国家的社会制度就一定先进等等。如何处理分类问题中的不确定性，曾经是困扰学术界的一个难题。基于概率论与数理统计的贝叶斯决策论为这一难题的解决提供了一个合理的数学框架。针对分类任务，贝叶斯决策论主要解决两方面的问题：1）如何描述分类问题中的不确定性，2）如何基于不确定性的描述实现分类任务。

首先，贝叶斯决策论使用概率来描述分类问题中的不确定性，同时使用数理统计方法定量估计概率模型。关于不确定性的理解，学界存在两个不同的学派——“频率学派”与“贝叶斯学派”。频率学派认为某一随机事件对应的概率模型是固定的，而观测值作为概率模型的某种采样是不确定的，当观测的次数趋近于无穷大的时候，就可以通过估计算法无限地接逼近这个概率模型。贝叶斯学派则认为，并不存在一个确定的概率模型，当前观测的结果是无数个可能的概率模型的综合体现，因此不可能也不需要以无限精度逼近某个的概率模型（因为根本不存在所谓唯一的、真实的概率模型），只需要找到所有可能的概率模型中可能性最大的那一个就行。

以上的论述有一点晦涩，我们可以用一个小例子来简单解释一下。小黑向小明展示了 100 次扔硬币的结果，其中 30 次正面，70 次反面。如果小明属于频率学派，他会认为小黑的口袋里只有一枚硬币，这枚硬币被铸造成正面概率为 0.3，背面概率为 0.7 的形态；如果小明属于贝叶斯学派，他会认为小黑口袋里有很多硬币，他们各自形态不同，被小黑选中的概率也不同，总体上表现出来的平均效果是正面概率为 0.3，背面概率为 0.7。之所以要在“频率学派”和“贝叶斯学派”的概念上大费笔墨，是因为他们将分别衍生出“最大似然估计”与“最大后验概率估计”两种不同的概率模型估计方法，介绍这两种方法是本章最重要的任务之一。

在概率估计方法的基础上，如何实现模式分类是贝叶斯决策论要研究的核心问题。在贝叶斯决策论框架下，选用某种概率或基于概率的某种函数，可以为分类问题提供一种公理性准则。分类准则可以是：“在当前观测结果的条件下，目标属于某一个类别的概率最大。”也可以是“在当前观测结果的条件下，目标被错误分类的概率最小。”还可以是“在当前观测结果的条件下，分类结果的平均风险最小。”而上述分类准则又能够为分类器的设计提供指导。

3.2 贝叶斯分类器

贝叶斯分类器是贝叶斯决策论在模式分类问题上的应用形态。贝叶斯分类器的设计与使用主要包含 3 个步骤：1）根据每个类别的训练样本，建立相应的概率模型，2）将需要分类的样本代入每一类的概率模型，并计算某种基于概率值的评价值，3）将基于概率的评价结果带入分类准则，给出最终的分类结果。

3.2.1 贝叶斯分类器基本概念

关于概率模型的描述、估计与应用必然涉及到一些基本概念。本节将使用符号表述与 Iris 数据库示例相结合的方式对这些概念加以解释，如表 3-1 所示。

表 3-1 贝叶斯分类器相关概念

名称	符号	示例（IRIS 数据库）
类别	$\Omega = \{\omega_i, i = 1,2,\dots,M\}$	三个鸢尾花子类：Setosa, Versicolour, Virginica，$M = 3$
样本特征向量	$\boldsymbol{x} = [x_1, x_2, \dots, x_d]^T$	[花萼长度,花萼宽度,花瓣长度,花瓣宽度]^T，特征维度$d = 4$
数据库	$D = \{(\boldsymbol{x}_n, y_n) \mid n = 1, \dots, N\}$	y_n为第n个样本的类别标签，样本总数量$N = 150$
先验概率	$P(\omega_i), i = 1,2,\dots,M$	各类样本占总数的比例，$P(\omega_1) \approx P(\omega_2) \approx P(\omega_3) \approx \frac{1}{3}$
后验概率	$P(\omega_i \mid \boldsymbol{x}), i = 1,2,\dots,M$	特征已经被观测到的某朵花属于某个子类的概率
类条件概率密度函数	$p(\boldsymbol{x} \mid \omega_i), i = 1,2,\dots,M$	某一个子类的特征分布概率密度函数
特征概率密度函数	$p(\boldsymbol{x}) = \sum_{i=1}^{M} p(\boldsymbol{x} \mid \omega_i) P(\omega_i)$	所有子类总的特征分布概率密度函数
贝叶斯分类器	$h(\boldsymbol{x})$	当输入为$\boldsymbol{x}$时，分类器$h(\boldsymbol{x})$的输出为三个子类的某一个

本文中事件发生的后验概率和先验概率均使用大写P，表示离散事件对应的概率质量，例如样本从属于某一个类别的概率值；概率密度函数使用小写p表示，通常对应连续随机变量分布的概率密度等。原则上，概率质量$0 \le P \le 1$，同一个样本从属于不同类别的后验概率和为 1，既$\sum_{i=1}^{M} P(\omega_i \mid \boldsymbol{x}) = 1$；概率密度$p$可以大于 1，但不能小于 0，同一个类别的类条件概率密度函数的积分为 1，既$\int_{\boldsymbol{x}} p(\boldsymbol{x} \mid \omega_i) = 1, i = 1,2,\dots,M$。图 3-1 的柱状图是对 Iris 数据库中三个鸢尾花子类的花萼长度类条件概率密度函数$p(x_1 \mid \omega_i), i = 1,2,3$的某种估计值；图 3-2 则是对 Iris 数据库中全部 150 个样本总体的花萼长度概率密度函数$p(x_1)$的估计。基于上述估计，可以采用贝叶斯分类准则实现具体样本的分类。

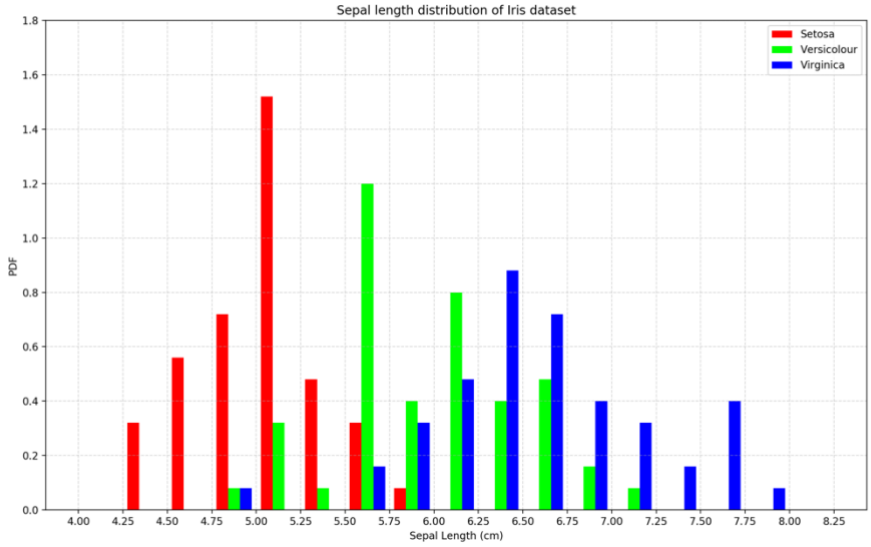


图 3-1 Iris 数据库花萼长度类条件概率密度函数估计结果。

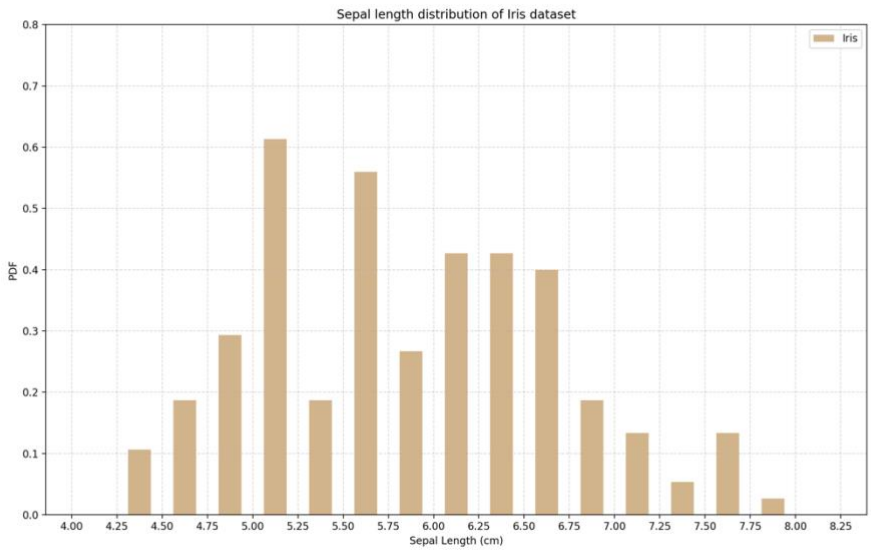


图 3-2 Iris 数据库花萼长度特征概率密度函数估计结果。

3.2.2 最大后验概率分类准则

我们首先给出最基本的贝叶斯分类准则如下：

“在观测到样本特征x的条件下，该样本应被分给使得后验概率$P(\omega_i|x)$最大的类别。”

这句话对应的模式识别任务假设函数$y = h(\boldsymbol{x})$的数学描述如下：

$$h(\boldsymbol{x}) = \omega^* = \underset{\omega_i \in \Omega}{\operatorname{argmax}} P(\omega_i|\boldsymbol{x}) \tag{3.1}$$

其中Ω是所有类别的集合。最大后验概率准则的数学形式看似简单，但在实际应用中要如何根据有限的数据估计后验概率$P(\omega_i|\boldsymbol{x})$是个棘手的问题，这也是机器学习领域研究的核心问题之一。从表 3-1 结合图 3-1、图 3-2 可以看出，虽然直接估计后验概率$P(\omega_i|\boldsymbol{x})$比较困难，但先验概率$P(\omega_i)$、类条件概率密度函数$p(\boldsymbol{x}|\omega_i)$和特征概率密度函数$p(\boldsymbol{x})$的估计相对容易。能否基于这几个容易估计的概率来间接计算后验概率$P(\omega_i|\boldsymbol{x})$的呢？公式(3.2)描述的**“贝叶斯定理”**为该问题提供了解决思路。

$$P(\omega_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{p(\boldsymbol{x})} \tag{3.2}$$

其中，根据概率论中的全概率公式，特征$\boldsymbol{x}$的概率密度函数$p(\boldsymbol{x})$可以写为：

$$p(\boldsymbol{x}) = \sum_{i=1}^{M} p(\boldsymbol{x}|\omega_i)P(\omega_i) \tag{3.3}$$

结合公式(3.2)和(3.3)，可以基于$p(\boldsymbol{x}|\omega_i), P(\omega_i), i = 1, \dots, M$实现后验概率$P(\omega_i|\boldsymbol{x})$的估计，再将其代入公式(3.1)就实现了一个简单的贝叶斯分类器。在实际操作中，我们发现在比较后验概率大小时，公式(3.2)中的分母$p(\boldsymbol{x})$对于每一个类别都相同，不会影响比较结果，可以消去。再进一步，如果不同类别的先验概率都相同，即$P(\omega_i) = P(\omega_j), \forall i \neq j$。则公式(3.1)所描述的分类准则可以简化为：

$$h(\boldsymbol{x}) = \omega^* = \underset{\omega_i \in \Omega}{\operatorname{argmax}} p(\boldsymbol{x}|\omega_i) \tag{3.4}$$

公式(3.4)意味着在各类别先验概率相同的情况，贝叶斯分类器可以通过比较类条件概率密度函数$p(\boldsymbol{x}|\omega_i)$实现分类。步假设在一个二分类问题中，两个类别的特征服从正态分布，例如：仅依靠花萼长度特征对 Setosa 和 Versicolour 两个子类进行二分类的问题，其类条件概率密度曲线如图 3-3 (a)所示。从中不难看出，在正态分布假设下，比较类条件概率密度函数值的大小等价于一个简单的阈值判断：

$$h(x) = \begin{cases} \omega_1 & x < x_0 \\ \omega_2 & x > x_0 \end{cases} \tag{3.5}$$

3.2.3 最小错误概率分类准则

除了上一节提到的“最大后验概率分类准则”，是否还存在其他合理的分类准则呢？答案是肯定的，例如一个好的分类器应该尽可能地减少分类错误。在概率论框架下，分类错误的可能性由错误概率来描述。对于当前样本$\boldsymbol{x}$，如果分给ω_i类，则对应的错误概率定义为：

$$P_e(\omega_i|\boldsymbol{x}) = 1 - P(\omega_i|\boldsymbol{x}) \tag{3.6}$$

公式(3.6)的涵义是“将样本$\boldsymbol{x}$分给ω_i这一行为的错误概率$P_e(\omega_i|\boldsymbol{x})$就是该样本不属于$\omega_i$的概率$1 - P(\omega_i|\boldsymbol{x})$。”因此，最小错误概率分类准则可以描述为：

$$h_e(\boldsymbol{x}) = \underset{\omega_i \in \Omega}{\operatorname{argmin}} P_e(\omega_i|\boldsymbol{x}) \tag{3.7}$$

根据公式(3.1)、(3.6)和(3.7)，可以看出 “最小错误概率准则”和“最大后验概率准则”在本质上是一致的，它们不过是关于同一种原则的两个不同观察角度而已。

- **最小平均错误概率分类器**

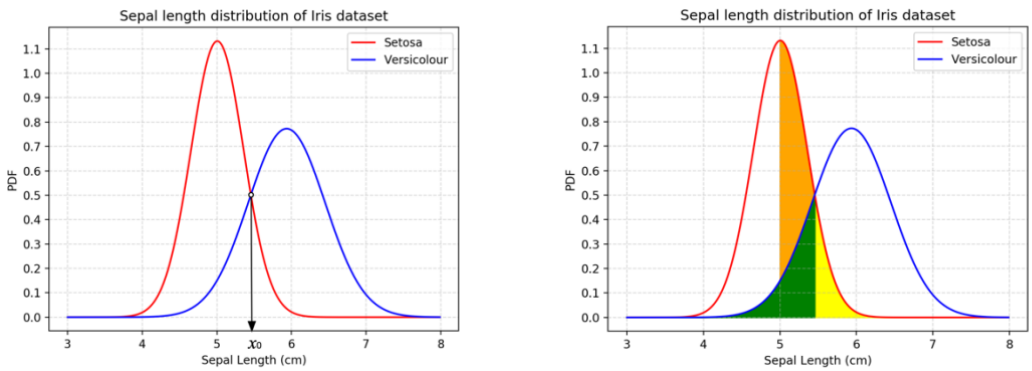
对于所有可能的$\boldsymbol{x}$，$h(\boldsymbol{x})$的平均分类错误概率可以写为：

$$P_E(h) = \mathbb{E}_{\boldsymbol{x}}[P_e(h(\boldsymbol{x})|\boldsymbol{x})] = \int P_e(h(\boldsymbol{x})|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \tag{3.8}$$

显然，以平均分类错误概率最小化为目标的贝叶斯分类器h_E可以定义为：

$$h_E = \underset{h}{\operatorname{argmin}} P_E(h) = \underset{h}{\operatorname{argmin}} \left[\int P_e(h(\boldsymbol{x})|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}\right] \tag{3.9}$$

对比公式(3.9)和(3.7)，会发现“最小错误概率分类准则” $h_e(x)$与“最小平均错误概率分类器” $h_E(x)$是等价的，因为使得每一个$\boldsymbol{x}$的错误概率最小就能使得整个分类器的平均错误概率最小。为了便于理解公式(3.9)，我们尝试以图 3-3 给出的二分类问题为例加以说明。假设 Setosa 为ω_1，Versicolour 为ω_2，按照最小错误概率分类准则给出的结果，将公式(3.8)分为两个积分区域。



(a) (b)

图 3-3 正态分布假设下的 Iris 数据库部分样本花萼长度类条件概率密度函数。(a)最大后验概率分类准则对应的阈值，(b)最小平均错误概率分类器的解释。

$$P_E(h_e) = \int_{h_e(\boldsymbol{x})=\omega_1} P_e(\omega_1|\boldsymbol{x})p(x)d\boldsymbol{x} + \int_{h_e(\boldsymbol{x})=\omega_2} P_e(\omega_2|\boldsymbol{x})p(x)d\boldsymbol{x} \tag{3.10}$$

由于$P_e(\omega_i|\boldsymbol{x}) = 1 - P(\omega_i|\boldsymbol{x})$，且$\sum_{i=1}^{2} P(\omega_i|\boldsymbol{x}) = 1$，可以推导出：

$$P_e(\omega_1|\boldsymbol{x}) = P(\omega_2|\boldsymbol{x}),\ P_e(\omega_2|\boldsymbol{x}) = P(\omega_1|\boldsymbol{x}) \tag{3.11}$$

将公式(3.11)代入公式(3.10)，可以得到：

$$P_E(h_e) = P_{21}(h_e) + P_{12}(h_e) \tag{3.12}$$

其中：

$$P_{21}(h_e)=\int_{h_e(\boldsymbol{x})=\omega_1}P(\omega_2|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}=\int_{h_e(\boldsymbol{x})=\omega_1}p(\boldsymbol{x}|\omega_2)P(\omega_2)d\boldsymbol{x}$$
$$P_{12}(h_e)=\int_{h_e(\boldsymbol{x})=\omega_2}P(\omega_1|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}=\int_{h_e(\boldsymbol{x})=\omega_2}p(\boldsymbol{x}|\omega_1)P(\omega_1)d\boldsymbol{x} \tag{3.13}$$

对照公式(3.13)与图 3-3 (a)[1]，很容易看出对于“最小错误概率准则”$h_e(x)$，两个积分区域的分界点为类条件概率密度函数$p(\boldsymbol{x}|\omega_1)$和$p(\boldsymbol{x}|\omega_2)$的曲线交点横坐标$x_0$。$P_{21}$等于绿色阴影面积，表示$\omega_2$类被误判为$\omega_1$类的平均错误概率；$P_{12}$对应于黄色阴影面积，表示$\omega_1$类被误判为$\omega_2$类的平均错误概率；两者相加的总面积则为平均错误概率P_E。如果不采用最小错误概率分类准则，例如将分界点向左移动，放在$x=5$的位置（如图 3-3 (b)所示），则在区间段$[5,x_0]$，分类器$h(\boldsymbol{x})=\omega_2$，相比于最小错误概率分类准则$h_e(\boldsymbol{x})$，将会额外增加错误概率$\Delta P_E$，其大小等于橙色区域面积，如公式(3.14)所示：

$$\Delta P_E=\int_5^{x_0}\big(p(x|\omega_1)P(\omega_1)-p(x|\omega_2)P(\omega_2)\big)dx \tag{3.14}$$

同理，将分界点向右移动也会产生类似的效果。这是因为在橙色区域对应的区间段内，分类器$h(\boldsymbol{x})$没有给出使错误概率最小的分类结果，从而增加了平均分类错误概率。从上面的案例分析可以得到如下结论：

“最小错误概率分类准则h_e在分类结果上等价于最小平均错误概率分类器h_E”

3.2.4 最小风险分类准则

在真实世界的某些决策任务中，不仅需要考虑分类结果的对错，还需要考虑分类行为对应的风险。以肿瘤诊断为例，将恶性肿瘤误诊为良性肿瘤会贻误治疗时机，造成难以挽回的后果，其风险通常大于将良性肿瘤误诊为恶性肿瘤。针对此类问题，可以采用最小风险分类准则执行模式分类任务。仍以二分类问题为例，定义**风险矩阵**如下：

$$\Lambda=\begin{bmatrix}\lambda_{11} & \lambda_{12}\\ \lambda_{21} & \lambda_{22}\end{bmatrix} \tag{3.15}$$

其中λ_{ij}表示将ω_i类样本分给ω_j类的“风险”。在肿瘤诊断任务中，假设良性为ω_1类，恶性为ω_2类，则一般有$\lambda_{21}>\lambda_{12}$。此外，需要注意的是，在某些分类任务中，即便是正确的分类也可能带来风险，既λ_{11}和λ_{22}也可能不为零。仍以肿瘤诊断为例，一个模式识别系统将良性肿瘤正确的识别为良性，患者得知结果后可能会因此疏忽自己的健康状况，未采取合理的治疗方案，从而导致几个月后良性肿瘤发展成恶性肿瘤；反之一个恶性肿瘤被正确诊断为恶性，患者得知后失去希望，自暴自弃，导致肿瘤快速恶化。

设将样本$\boldsymbol{x}$分给ω_1类的风险为ℓ_1，分给ω_2类的风险为ℓ_2，根据风险矩阵的定义，则有：

[1] 在 Iris 数据库中，Setosa 和 Versicolour 两类的先验概率相同，因此可以忽略其影响，只观察类条件概率密度函数$p(\boldsymbol{x}|\omega_i)$

$$\ell_1(\boldsymbol{x}) = \lambda_{11}P(\omega_1|\boldsymbol{x}) + \lambda_{21}P(\omega_2|\boldsymbol{x}),\ \ell_2(\boldsymbol{x}) = \lambda_{12}P(\omega_1|\boldsymbol{x}) + \lambda_{22}P(\omega_2|\boldsymbol{x}) \tag{3.16}$$

则最小风险分类准则可以描述为

$$h_r(\boldsymbol{x}) = \begin{cases} \omega_1 & \ell_1(\boldsymbol{x}) < \ell_2(\boldsymbol{x}) \\ \omega_2 & \ell_1(\boldsymbol{x}) > \ell_2(\boldsymbol{x}) \end{cases} \tag{3.17}$$

显然，分类器$h_r(\boldsymbol{x})$也对应了一个分界点x_r，在该点有：

$$\ell_1(\boldsymbol{x}) = \lambda_{11}P(\omega_1|\boldsymbol{x}) + \lambda_{21}P(\omega_2|\boldsymbol{x}) = \lambda_{12}P(\omega_1|\boldsymbol{x}) + \lambda_{22}P(\omega_2|\boldsymbol{x}) = \ell_2(\boldsymbol{x}) \tag{3.18}$$

根据贝叶斯公式可以推出：

$$r_1P(\omega_1)p(\boldsymbol{x}|\omega_1) = r_2P(\omega_2)p(\boldsymbol{x}|\omega_2) \tag{3.19}$$

其中

$$r_1 = \lambda_{12} - \lambda_{11}, r_2 = \lambda_{21} - \lambda_{22} \tag{3.20}$$

可以将r_1和r_2分别视为类别ω_1和ω_2的某种风险系数，r_i的数值越小，代表把样本分给ω_i类的风险越大。

♦ 例题-基于不同准则的 Iris 数据库分类

某花卉工厂从养殖基地采购了两类鸢尾花共 1000 支，其中 Setosa 类（ω_1）600 支，Versicolour 类（ω_2）400 支。根据统计，两类鸢尾花的花萼长度基本服从正态分布：

$$P(x|\omega_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right], i = 1,2$$

其中 $\mu_1 = 5$ cm, $\sigma_1 = 1$ cm；$\mu_2 = 7$ cm, $\sigma_2 = 2$ cm。花卉工厂将鸢尾花分类卖给某花店，并针对分类错误签订了罚款协议：ω_1 类被错分为 ω_2 类，每支罚款 3 元；如 ω_2 类被错分为 ω_1 类，每朵罚款 1 元。请根据最小平均错误概率准则和最小平均风险准则分别设计两个分类器，并为花卉工厂的分拣工人提供最优的分类方案。

解答：

(a) 最小错误概率分类准则

因为最小错误概率分类准则与最大后验概率分类准则在分类结果上完全等价，在正态分布假设下，表现为一个阈值分类器，如公式(3.5)所示。在阈值x_e处，根据贝叶斯公式，应有：

$$p(\omega_1|x_e) = p(\omega_2|x_e)$$
$$p(x_e|\omega_1)P(\omega_1) = p(x_e|\omega_2)P(\omega_2)$$

根据两类鸢尾花的数量比例有$P(\omega_1) = 0.6, P(\omega_2) = 0.4$，带入正态分布概率公式有：

$$\frac{0.6}{\sqrt{2\pi}}\exp\left[-\frac{(x_e-5)^2}{2\times 1}\right] = \frac{0.4}{2\sqrt{2\pi}}\exp\left[-\frac{(x_e-7)^2}{2\times 4}\right]$$

解得$x_e \approx 6.5(\text{cm})$。

(b) 最小风险分类准则

根据题意设$\lambda_{12}=3,\lambda_{21}=1,\lambda_{11}=\lambda_{22}=0$。根据公式(3.19)和(3.20)，在阈值$x_{\mathrm{r}}$处应有：

$$(\lambda_{12}-\lambda_{11})p(x_{\mathrm{r}}|\omega_1)P(\omega_1)=(\lambda_{21}-\lambda_{22})p(x_{\mathrm{r}}|\omega_2)P(\omega_2)$$

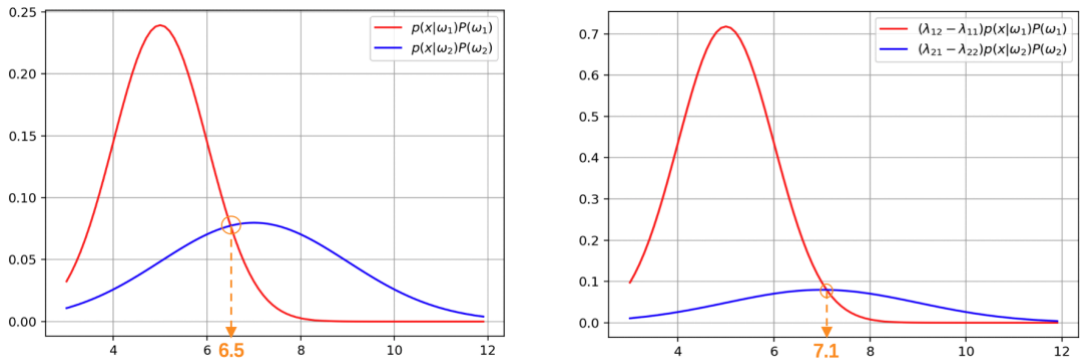
代入正态分布概率公式有：

$$\frac{3\times 0.6}{\sqrt{2\pi}}\exp\left[-\frac{(x_{\mathrm{r}}-5)^2}{2\times 1}\right]=\frac{1\times 0.4}{2\sqrt{2\pi}}\exp\left[-\frac{(x_{\mathrm{r}}-7)^2}{2\times 4}\right]$$

解得$x_{\mathrm{r}}\approx 7.1(\mathrm{cm})$。

(c) 结果分析与方案总结

如图 3-4 所示，最小风险分类准则给出的分类阈值x_{r}大于最小错误法律分类准则的阈值x_{e}，这意味着前者更倾向于把样本分配给ω_1类。这是因为ω_1类错分给ω_2类的罚款更多。这个结果与我们的直觉经验一致。



(a) (b)

图 3-4 基于最小风险的贝叶斯分类器的阈值变化情况。(a)最小错误概率分类准则的分类阈值；(b)最小风险分类准则的分类阈值。

将最小风险分类准则从二分类问题扩展到多分类问题，则将样本$\boldsymbol{x}$分给ω_j类的风险为：

$$\ell_j(\boldsymbol{x})=\sum_{i=1}^{M}\lambda_{ij}\,P(\omega_i|\boldsymbol{x}) \tag{3.21}$$

则最小风险分类准则可以描述为

$$h_r(\boldsymbol{x})=\underset{\omega_j\in\Omega}{\operatorname{argmin}}\,\ell_j(\boldsymbol{x}) \tag{3.22}$$

- **最小平均风险分类器**

在最小风险分类准则基础上，可以构造**“最小平均风险分类器”**。在公式(3.13)的基础上，进一步定义概率$P_{ij}(h)$为分类器$h(\boldsymbol{x})$把ω_i类样本分为ω_j类的概率，如公式(3.23)所示。

$$P_{ij}(h)=\int_{h(\boldsymbol{x})=\omega_j} p(\omega_i|\boldsymbol{x})P(\boldsymbol{x})d\boldsymbol{x}=\int_{h(\boldsymbol{x})=\omega_j} p(\boldsymbol{x}|\omega_i)P(\omega_i)d\boldsymbol{x},\qquad i,j=1,2,\dots,M \tag{3.23}$$

其中$i\neq j$对应于某种错误分类结果，$i=j$则对应正确的分类结果。则分类器$h(\boldsymbol{x})$的平均分类风险可以写为：

$$R(h)=\sum_{j=1}^{M}\sum_{i=1}^{M}\lambda_{ij}P_{ij}(h) \tag{3.24}$$

最小平均风险分类器h_R定义如下：

$$h_R=\underset{h}{\operatorname{argmin}}\,R(h) \tag{3.25}$$

参考“最小平均错误概率分类器”与“最小错误概率分类准则”之间的等价性，“最小平均风险分类器”也等价于“最小风险分类准则”。要证明这一关系，可以对平均风险$R(h)$进行另一种形式的分解。对于某个分类器$h(\boldsymbol{x})$，根据其分类结果，可以将样本空间X分为M个区域$A_j=\{\boldsymbol{x}|h(\boldsymbol{x})=\omega_j\},j=1,\dots,M$。设区域$A_j$的平均风险$\mathcal{L}_j$可以写为：

$$\mathcal{L}_j=\int_{A_j}\left(\sum_{i=1}^{M}\lambda_{ij}P(\omega_i|\boldsymbol{x})\right)p(\boldsymbol{x})d\boldsymbol{x}=\int_{A_j}\ell_j(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \tag{3.26}$$

其中，$\ell_j(\boldsymbol{x})$表示样本$\boldsymbol{x}$被分给第j类的风险，定义如公式(3.21)所示。则整个分类器$h(\boldsymbol{x})$的平均风险$R(h)$可以重新写为：

$$R(h)=\sum_{j=1}^{M}\mathcal{L}_j=\sum_{j=1}^{M}\left[\int_{A_j}\ell_j(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}\right] \tag{3.27}$$

对比公式(3.22)给出的最小风险分类准则$h_r(\boldsymbol{x})$，可以发现：对于每一个具体的样本$\boldsymbol{x}$，如果都能选择使得分类风险$\ell_j(\boldsymbol{x})$最小的类别ω_j，则可以使得平均分类风险$R(h)$最小化。因此我们可以得到类似的结论：

“最小风险分类准则h_r在分类结果上等价于最小平均风险分类器h_R”

这里需要注意的是，在公式(3.7)和(3.22)，最小错误概率分类准则和最小风险分类准则对应的分类器$h_e(\boldsymbol{x})$和$h_r(\boldsymbol{x})$都是以输出的类别标签ω来定义的，既分类准则并没有给出分类器的具体数学形式，只给出了分类结果。而公式(3.9)和(3.27)分别定义的最小平均错误概率分类器h_E和最小平均风险分类器h_R则对应某种最优的函数形式h。这是因为它们是基于所有样本的某种平均指标来设计的，而非单独针对某一个样本$\boldsymbol{x}$，因此一般需要定义出某种需要优化的函数形式以及相应的目标函数。由于以上的差别，我们一般认为$h_e(\boldsymbol{x})$和$h_r(\boldsymbol{x})$是“分类准则”，而h_E和h_R是“分类器”。在后续的研究中，我们会发现大多数基于机器学习的分类算法都采用了“分类器”而非“分类准则”的设计思路，这种思路的核心步骤是如何构造待优化的函数以及如何设计一个合理的目标函数。

3.3 概率模型估计

在贝叶斯决策论的框架下，无论采用最大后验概率准则、最小错误概率准则还是最小风险准则，核心问题都在于如何获取后验概率$P(\omega_i|\boldsymbol{x})$的估计值。在贝叶斯公式的帮助下，这一问题被转化为类条件概率密度函数$p(\boldsymbol{x}|\omega_i)$和先验概率$P(\omega_i)$的估计问题。先验概率可以通过统计当前类别样本在数据集中的比例进行估计，因此最核心也是最具挑战性的问题是如何实现类条件概率密度$p(\boldsymbol{x}|\omega_i)$的估计。这就是概率模型估计问题。

3.3.1 概率模型概述

估计类条件概率密度$p(\boldsymbol{x}|\omega_i)$的首要任务是建立一个概率模型，也就是表示概率的数学形式，常见的概率模型设计思路有两种。一种使用带有参数的函数形式来描述$p(\boldsymbol{x}|\omega_i)$，例如以均值矢量$\boldsymbol{\mu}_i$和协方差矩阵$\Sigma_i$为参数的多元高斯函数。用参数向量$\boldsymbol{\theta}_i$表示$\omega_i$类概率模型的所有参数，则类条件概率密度$p(\boldsymbol{x}|\omega_i)$可以表示为有参数的概率密度函数$p(\boldsymbol{x};\boldsymbol{\theta}_i)$，或记为$p(\boldsymbol{x}|\boldsymbol{\theta}_i)$；前者是从参数化函数的角度表示，后者是从类条件概率的角度表示。因此，只要能够估计出参数向量$\boldsymbol{\theta}_i$，再将样本$\boldsymbol{x}$代入概率模型参数就可以得到类条件概率密度函数值$p(\boldsymbol{x}|\omega_i) = p(\boldsymbol{x};\boldsymbol{\theta}_i)$。此类方法统称为有参数概率模型估计。

另一种思路则放弃了参数化的函数形式，采用表格、数列、散点等离散的数据形式表征概率密度$p(\boldsymbol{x}|\omega_i)$，例如图 3-1 中使用的直方图，其数据形式是一个有限行数的表格，每一行对应于样本特征空间中的某一段区域B_j，并记录了该区域对应的概率密度函数p_j。对于一个需要查询的样本$\boldsymbol{x}$，如果满足$\boldsymbol{x} \in B_j$，则概率密度估计值为$p(\boldsymbol{x}|\omega_i) = p_j$。除直方图外，还有一些其他的无需有参函数的方法也能实现概率估计，此类方法统称为无参数概率模型估计。

3.3.2 有参数概率估计方法

对于有参数概率模型而言，有三种经典的参数估计方法，分别称为最大似然估计（Maximum Likelihood Estimation，简称 MLE），最大后验概率估计（Maximum A-posteriori Probability Estimation，简称 MAP）和贝叶斯估计（Bayesian Estimation，或称贝叶斯推论 Bayesian Inference）。这三者在数学方法上一脉相承，但在理念上各有不同。

- **最大似然估计**

最大似然估计是从频率学派观点衍生出的一种概率模型估计方法，认为每个随机自然事件对应一个概率分布，如果在独立重复实验中事件发生的概率的极限趋近于某组特定的参数对应的概率密度函数，则认为这些参数就是描述该事件发生的概率模型参数。

要理解最大似然估计，首先要理解似然性（Likelihood）的概念，及其与概率（Probability）在概念定义和形式上的异同。

1）当模型参数向量$\boldsymbol{\theta}$固定，观测值$\boldsymbol{x}$为变量时，$p(\boldsymbol{x}|\boldsymbol{\theta})$称为类条件概率密度函数或简称类条件概率，用于定量描述参数$\boldsymbol{\theta}$确定的$\boldsymbol{\omega}$类条件下，观测值$\boldsymbol{x}$的不确定性分布；

2）当观测值$\boldsymbol{x}$固定，模型参数向量$\boldsymbol{\theta}$为变量时，$\ell(\boldsymbol{x};\boldsymbol{\theta})$称为似然性函数或简称似然性，用于定量描述概率模型参数$\boldsymbol{\theta}$的不确定性分布。

从以上定义可以看出，似然性$\ell(\boldsymbol{x};\boldsymbol{\theta})$与类条件概率$p(\boldsymbol{x}|\boldsymbol{\theta})$在函数形式上完全一致，但前者以$\boldsymbol{\theta}$为自变量，$\boldsymbol{x}$是参数；后者以$\boldsymbol{x}$为自变量，$\boldsymbol{\theta}$是参数。两者表示完全不同的概念。

在似然性概念的基础上，最大似然估计是指基于当前观测数据$\boldsymbol{x}$，能够使似然性函数$\ell(\boldsymbol{x};\boldsymbol{\theta})$达到最大的$\boldsymbol{\theta}_{\mathrm{MLE}}$就是对模型参数$\boldsymbol{\theta}$的最优估计，其在数学形式上等价于使得当前观测数据集的类条件概率最大的参数$\boldsymbol{\theta}$。设当前观测数据为$X=\{\boldsymbol{x}_i|i=1,\ldots,N\}$，且每次观测事件均相互独立，根据概率论的独立事件联合概率分布公式，则参数$\boldsymbol{\theta}$的似然性可以写为：

$$L(X;\boldsymbol{\theta})=P(X|\boldsymbol{\theta})=\prod_{i=1}^{N}p(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{3.28}$$

则最大似然估计的结果$\boldsymbol{\theta}_{\mathrm{MLE}}$即使得似然性$L(X;\boldsymbol{\theta})$最大的$\boldsymbol{\theta}$值，如公式(3.29)所示。

$$\boldsymbol{\theta}_{\mathrm{MLE}}=\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\,L(X;\boldsymbol{\theta}) \tag{3.29}$$

由于似然性函数的任意单调递增函数$f\big(L(X;\boldsymbol{\theta})\big)$与原似然函数$L(X;\boldsymbol{\theta})$具有相同的最优解，从简化计算的角度出发，设计对数似然性（Log-Likelihood）函数$LL(X;\boldsymbol{\theta})$，则最大似然估计可以写为公式(3.30)：

$$\boldsymbol{\theta}_{\mathrm{MLE}}=\underset{\boldsymbol{\theta}}{\operatorname{argmax}}[LL(X;\boldsymbol{\theta})]=\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\big[\ln\big(L(X;\boldsymbol{\theta})\big)\big]=\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[\sum_{i=1}^{N}\ln\big(p(\boldsymbol{x}_i|\boldsymbol{\theta})\big)\right] \tag{3.30}$$

- **最大后验概率估计**

与来自于频率学派的最大似然估计不同，最大后验概率估计是从贝叶斯学派观点衍生出的概率模型估计方法。这种方法认为不存在一个客观的概率模型，也就是说自然世界中并不存在一个由固定参数$\boldsymbol{\theta}$描述的概率模型$p(\boldsymbol{x}|\boldsymbol{\theta})$用于产生当前的观测数据$X$。观测数据集$X$所表现出来的随机性很大程度上来源于观测者知识的不完备性。举例说明，一个人扔骰子的点数表面看起来似乎是随机的，但这种随机性不仅仅来源于概率密度函数$p(\boldsymbol{x}|\boldsymbol{\theta})$本身的随机性，也包含了骰子被扔出来的一瞬间的速度、加速度、筛子本身的形状、弹性、密度分布、空气流动模式、桌面硬度和弹性等知识所对应的参数$\boldsymbol{\theta}$的不确定性。因此概率模型参数的估计既要考虑类条件概率密度函数$p(X|\boldsymbol{\theta})$，也要考虑参数$\boldsymbol{\theta}$的先验概率$p(\boldsymbol{\theta})$，因此最大后验概率估计要寻找的是使$p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})$最大化的参数$\boldsymbol{\theta}_{\mathrm{MAP}}$。考虑到$p(X)$与参数$\boldsymbol{\theta}$无关，因此最大后验概率估计的目标函数可以由$p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})$转化为后验概率$p(\boldsymbol{\theta}|X)$，如公式(3.31)所示。

$$\boldsymbol{\theta}_{\mathrm{MAP}}=\underset{\boldsymbol{\theta}}{\operatorname{argmax}}[p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})]=\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[\frac{p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(X)}\right]=\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\,p(\boldsymbol{\theta}|X) \tag{3.31}$$

显然公式(3.31)的解就是使得后验概率$p(\boldsymbol{\theta}|X)$最大化的解，这也是为什么这种方法被称为最大后验概率估计的原因之一。需要提到的是，后验概率$p(\boldsymbol{\theta}|X)$的最大化与$p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})$的最大化在数学求解上虽然是一致的，但其实从内涵角度看有细微的差别。最大后验概率估计是从贝叶斯学派的观点出发，希望找到一个与观测者知识和经验的吻合度最大的解，其原始目的并非是要寻找使得后验概率最大的参数$\boldsymbol{\theta}$。

● 贝叶斯估计

在介绍了最大后验概率估计方法后，善于思考的同学应该会想到一个问题：既然贝叶斯学派认为类别ω对应的参数向量$\boldsymbol{\theta}$是一个随机向量而非固定值，那为什么一定要求出一个固定的估计值$\boldsymbol{\theta}_{\mathrm{MAP}}$呢？这个问题提得很有道理。让我们不忘初心，回头看看当初提出概率模型估计问题的初衷是什么。

假设所有的观测数据X均来源于同一个类别ω，当给定一个新样本$\boldsymbol{x}$的时候，贝叶斯决策论要求首先能够计算出类条件概率密度函数$p(\boldsymbol{x}|\omega)$的值，之后才能实现分类。根据贝叶斯学派的观点，类别ω并不是对应一个固定的参数$\boldsymbol{\theta}$，而需要用当前观测数据X下参数向量$\boldsymbol{\theta}$的随机分布$p(\boldsymbol{\theta}|X)$来描述。而整个概率模型估计的最终目的是以观测数据 X 为条件，对新样本$\boldsymbol{x}$的类条件概率密度$p(\boldsymbol{x}|\omega)$进行估计的问题，因此可以写为公式(3.32)：

$$p(\boldsymbol{x}|\omega) \approx p(x|X) = \int_{\boldsymbol{\theta}\in\Theta} p(x|\boldsymbol{\theta})p(\boldsymbol{\theta}|X)d\boldsymbol{\theta} \tag{3.32}$$

其中Θ表示参数向量$\boldsymbol{\theta}$的空间，即所有可能的参数向量$\boldsymbol{\theta}$的集合。

基于当前观测数据X可以推测出参数向量$\boldsymbol{\theta}$的分布概率$p(\boldsymbol{\theta}|X)$，针对每一个服从改分布的参数向量$\boldsymbol{\theta}_m \in \Theta$，估计新样本$\boldsymbol{x}$的类条件概率密度函数$p(\boldsymbol{x}|\boldsymbol{\theta}_m)$。所以每一个可能的参数向量$\boldsymbol{\theta}_m$对应于一条从观测数据 X 到新样本 $\boldsymbol{x}$ 的推论路径，综合所有路径得到的平均情况（基于参数向量$\boldsymbol{\theta}$的分布求所有推论路径结果的数学期望，如图 3-5 所示。）就是贝叶斯估计的结果，因此贝叶斯估计也称为贝叶斯推论（Bayesian Inference，简称 BI）。

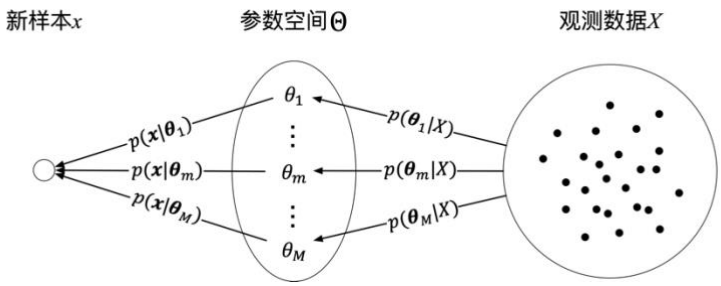


图 3-5 贝叶斯估计原理示意图

根据贝叶斯定理，公式(3.32)中的后验概率$p(\boldsymbol{\theta}|X)$可以写为：

$$p(\boldsymbol{\theta}|X) = \frac{p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(X)} = \frac{p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\Theta} p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{3.33}$$

将公式(3.33)代入公式(3.32)可以求出贝叶斯估计的最终结果$p(\boldsymbol{x}|X)$，从而实现贝叶斯分类。虽然贝叶斯估计似乎比最大后验概率估计更合理，但因数学形式复杂，通常只在一些比较特殊的概率分布假设下才能得到解析的函数表达式；当$p(\boldsymbol{x}|\boldsymbol{\theta})$和$p(\boldsymbol{\theta})$的数学形式比较复杂时，缺少解析式的$p(x|X)$概率估计将会是非常棘手的问题。

♦ 例题-正态分布假设下的最大似然估计

假设类别ω服从d维多元正态分布$\mathcal{N}(\boldsymbol{\mu},\Sigma),\boldsymbol{\mu}\in\Re^d,\Sigma\in\Re^{d\times d}$，对类别$\omega$的随机观测样本集合为$X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_N\}$，$\boldsymbol{x}_i\in\Re^d,i=1,2,\dots,N$。类别$\omega$的类条件概率密度函数可以写为：

$$p(\boldsymbol{x}|\omega)=p(\boldsymbol{x}|\boldsymbol{\theta})=\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{3.34}$$

其中$|\Sigma|$表示矩阵Σ的行列式，请利用最大似然估计方法给出两组参数的最优估计。

解答：

根据公式(3.34)给出的对数似然性函数可以写为：

$$\ln\big(L(X;\boldsymbol{\theta})\big)=\sum_{i=1}^{N}\ln\big(p(\boldsymbol{x}_i|\boldsymbol{\theta})\big)=\sum_{i=1}^{N}\ln\left(\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right)\right) \tag{3.35}$$

根据最大似然估计方法，令$\ln\big(L(X;\boldsymbol{\theta})\big)$对模型参数向量$\boldsymbol{\theta}=[\boldsymbol{\mu},\Sigma]^T$中的每一个分量求偏导等于0，则该方程的解即为最大似然估计的解$\boldsymbol{\theta}_{MLE}$。首先求解均值向量$\boldsymbol{\mu}$的偏导方程。

$$\begin{aligned}
\frac{\partial\ln\big(L(X;\boldsymbol{\theta})\big)}{\partial\boldsymbol{\mu}}&=0\\
\frac{\sum_{i=1}^{N}\partial\ln\left(\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right)\right)}{\partial\boldsymbol{\mu}}&=0\\
\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})&=0\\
\boldsymbol{\mu}_{\mathrm{MLE}}&=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i
\end{aligned} \tag{3.36}$$

其次，求解协方差矩阵Σ的偏导方程：

$$\begin{aligned}
\frac{\partial\ln\big(L(X;\boldsymbol{\theta})\big)}{\partial\Sigma}&=0\\
\frac{\sum_{i=1}^{N}\partial\ln\left(\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right)\right)}{\partial\Sigma}&=0\\
\sum_{i=1}^{N}\left(\frac{1}{2}\Sigma^{-1}-\frac{1}{2}\Sigma^{-2}(\boldsymbol{x}_i-\boldsymbol{\mu})^T(\boldsymbol{x}_i-\boldsymbol{\mu})\right)&=0\\
\Sigma_{\mathrm{MLE}}&=\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})^T(\boldsymbol{x}_i-\boldsymbol{\mu})
\end{aligned} \tag{3.37}$$

公式(3.36)、(3.37)的推导结果与我们在数理统计课程中学习到的算数均值和方差的定义一致，它们可以视为正态分布假设下的均值与方差的最大似然估计结果。

3.3.3 高斯混合模型

在采用最大似然法进行概率估计是首先要给出关于类条件概率密度函数$p(\boldsymbol{x}|\omega)$的参数化函数形式$p(\boldsymbol{x};\boldsymbol{\theta})$，例如上一节例题中就假设某一类别服从正态分布，$p(\boldsymbol{x};\boldsymbol{\theta})$就是以均值向量$\boldsymbol{\mu}$和协方差矩阵Σ为参数的高斯函数。然而类别的定义千差万别，数据的来源多种多样，很难事先确定某一类别的概率密度函数服从某一种具体的函数形式。因此，我们需要建立一种描述概率密度函数的通用数学形式。这就是本节要讨论的内容——高斯混合模型（Gaussian Mixture Model，简称 GMM）。

高斯混合模型的基本想法是：“任意分布的概率密度函数可由多个高斯函数的线性组合近似拟合”。这一理念对于我们来说并不陌生，在高等数学中的泰勒展开和傅里叶展开都采用了相似的思想。在此类方法中，参与线性组合的基本函数被称为基函数，例如泰勒展开中的幂函数，傅里叶展开中的三角函数，高斯混合模型中的高斯函数。由于中心极限定理预言了多个独立的随机变量的联合分布会趋近于正态分布（也就是其概率密度函数趋近于高斯函数），而自然界中可以观测到的事件在物理层面上通常是由多种更基本的对象和事件混合而成，因此它们的概率密度函数通常接近于某种高斯分布。因此，选择高斯混合模型作为概率密度函数的一般性描述具有较好的合理性。

基于上述分析，任意类别ω的类条件概率密度函数可以近似地写为：

$$p(\boldsymbol{x}|\omega) = p(\boldsymbol{x};\Theta) \approx \sum_{j=1}^{M} p(\boldsymbol{x}|j,\Theta)P_j \tag{3.38}$$

公式(3.38)表示ω类的分布可以视为M个不同的高斯函数的线性组合。其中，Θ表示 GMM 模型的所有参数，记为$\Theta = \{\boldsymbol{\theta}_j, P_j | j = 1,2,\dots,M\}$，$\boldsymbol{\theta}_j$是第$j$个分布的参数向量，$P_j$是第$j$个分布的先验概率（或理解为比例权重），满足$\sum_{j=1}^{M} P_j = 1$；$p(\boldsymbol{x}|j,\Theta)$表示第$j$个分布的类条件概率密度函数，定义为一个高斯函数$G(\boldsymbol{x};\boldsymbol{\mu}_j,\Sigma_j)$：

$$p(\boldsymbol{x}|j,\Theta) = p(\boldsymbol{x}|\boldsymbol{\theta}_j) = G(\boldsymbol{x};\boldsymbol{\mu}_j,\Sigma_j) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_j|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right] \tag{3.39}$$

其中，$\boldsymbol{\theta}_j = \{\boldsymbol{\mu}_j \in \Re^d, \Sigma_j \in \Re^{d\times d}\}$是第$j$个高斯分布的参数向量。

理论上讲，令公式(3.38)中的概率密度函数$p(\boldsymbol{x};\Theta)$对各个参数分量求导等于 0，求解非线性联立方程组可以得到高斯混合模型的参数Θ的最大似然估计值。然而在实际计算中，由于 GMM 模型本身数学形式的复杂性，相应的偏导方程组无法直接推导出解析解，需要利用数值计算方法进行求解。虽然使用常规的梯度下降最优化数值计算方法也能够求解 GMM 模型的最大似然估计问题，但其效率难以保证，在真正的应用中，我们通常会采用一种巧妙的迭代优化算法进行求解，这就是下一节要介绍的 EM 算法。

3.3.4 EM 算法

EM 算法既 Expectation Maximization 算法，表示期望最大化。该算法是 1977 年由 Dempster 等人提出的一种迭代优化算法，主要用于解决含有隐变量的概率模型参数最大似然估计问题。

隐变量（Hidden Variable）有时又称为潜变量（Latent Variable），是指对数据观测结果有影响但却无法直接观测到的变量。隐变量这一概念是 EM 算法之所以可以用于解决 GMM 问题的关键。

让我们用一个简单的例子来解释 GMM 模型中的隐变量。假设在学校体检过程中，医生从班级中随机选择同学测量身高。如选择的是男同学，则测量结果服从均值$\mu_m = 175\text{cm}$，标准差$\sigma_m = 8\text{cm}$的正态分布G_m；如选择的是女同学，则测量结果服从均值$\mu_f = 165\text{cm}$，标准差$\sigma_f = 6\text{cm}$的正态分布G_f；班级中男女比例为 7:3。假设医生在记录数据时只记录了身高数据，没有记录性别信息，并要求仅根据身高数据对当前班级同学身高分布的概率密度函数进行参数估计。这就是一个典型 GMM 问题，其中没有被记录下来的性别信息就是一个隐变量。如果性别这个隐变量已知，则基于 GMM 模型的概率估计问题就可以转化为相对简单的两个相互独立的正态分布——男生的身高分布和女生的身高分布——的参数估计问题。遗憾的是性别这个隐变量目前无法直接观测，所以我们需要引入专门用来处理隐变量的 EM 算法。

为了解释为什么 EM 算法可以解决 GMM 模型的最大似然估计问题，需要首先定义一些变量和相关概念。假设X表示可以直接观测到的随机变量数据，Z表示隐变量数据，模型参数仍用向量Θ表示，则合并后的数据$(X;Z)$称为完全数据(Complete-data)，观测数据X称为不完全数据(Incomplete-data)。在上面给出的身高测量的例子中，身高数据就是可观测数据X，性别数据就是隐变量数据 Z。而相应的参数$\mu_m, \mu_f, \sigma_m, \sigma_f$和男女类别的采样比例$P_m, P_f$共同组成了参数向量$\Theta$。将隐变量$Z$视为随机变量，则不完全数据$X$关于参数$\Theta$的对数似然函数可以写作:

$$LL(X;\Theta) = \ln P(X|\Theta) = \ln \sum_z P(X,Z|\Theta) = \ln \left(\sum_z P(X|Z,\Theta) P(Z|\Theta) \right) \tag{3.40}$$

由于隐变量 Z 未知，且包含有连加式（如Z为连续随机变量，则为积分式）的对数形式，公式(3.40)的极大值很难通过求解目标函数偏导等于 0 的方程得到。针对这一相对复杂的最优化问题，EM 算法给出了一种逐步迭代优化的求解思路，核心步骤如下：

1） 基于第t次迭代的参数估计值Θ^t与观测数据X，估计隐变量Z的条件概率分布$P(Z|X,\Theta^t)$；

2） 基于条件概率分布$P(Z|X,\Theta^t)$的估计值，计算完全数据(X,Z)的对数似然函数$\ln P(X,Z|\Theta)$的数学期望，记为$Q(\Theta,\Theta^t)$；

3） 求解使$Q(\Theta,\Theta^t)$极大化的解作为第$t+1$次迭代的参数估计值Θ^{t+1}；

4） 重复步骤 1)~3)，直至参数Θ收敛。

关于上述算法能够实现对数似然函数$LL(X;\Theta)$极大化的严格数学证明相对复杂，本书只给出面向模式分类任务的标准 EM 算法步骤与公式。

- **EM 算法**

输入:	观测样本数据$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ 隐变量定义$Z = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$ 完全数据联合条件分布概率模型$P(X,Z|\Theta)$ 隐变量条件分布概率模型$P(Z|X,\Theta)$
步骤:	

1	初始化参数Θ^0，开始迭代：
2	Repeat：
3	E 步骤：计算数学期望$Q(\Theta,\Theta^t)=\mathbb{E}_{Z\sim P(Z\|X,\Theta^t)}[\ln P(X,Z\|\Theta)]$
4	M 步骤：计算$t+1$轮的参数估计值$\Theta^{t+1}=\mathrm{argmax}_{\boldsymbol{\theta}}[Q(\Theta,\Theta^t)]$
5	until：任意一条终止条件满足
6	条件 1): $\|\Theta^{t+1}-\Theta^t\|<\epsilon_1$
7	条件 2): $\|Q(\Theta^{t+1},\Theta^t)-Q(\Theta^t,\Theta^t)\|<\epsilon_2$
输出:	参数向量Θ^t

图 3-6 EM 算法步骤

● GMM+EM 算法推导

EM 算法理论上可以解决带有隐变量的、任意具有可导函数形式的概率分布的估计问题。但在实际应用中，通常将 EM 算法与 GMM 模型相结合对数学形式未知的概率分布进行估计。具体方法主要分为以下三个步骤。

1）变量命名与参数初始化

- $\boldsymbol{x}_i$：第i个训练样本，$i=1,\dots,N$；
- j：第j个高斯分布的序号，$j=1,\dots,M$；
- j_i：第i个样本对应的高斯分布的序号，$j_i\in\{1,\dots,M\}$；
- t：迭代次数的序号；
- $\boldsymbol{\mu}_j^t$：t时刻第j个高斯分布的均值向量估计值；
- Σ_j^t：t时刻第j个高斯分布的协方差矩阵估计值；
- P_j^t：t时刻第j个高斯分布的先验概率估计值

第t轮参数估计值Θ^t，$t=0$时需要随机初始化

从上述定义可以看出，不完全数据$X=\{x_i\}_{i=1}^N$，隐变量$Z=J=\{j_i\}_{i=1}^N$是关于每个具体样本$\boldsymbol{x}_i$来自于哪个高斯分布的信息。

2）E 步骤

根据 EM 算法定义，E 步骤既推导对数似然函数的数学期望表达式的过程。首先给出完全数据集的对数似然函数在 GMM 假设下的表达式：

$$LL(X,J|\Theta)=\sum_{i=1}^{N}\ln\left[p(\boldsymbol{x}_i|j_i,\Theta)P_j\right] \tag{3.41}$$

由于并未观测到j_i的真实值，因此需要在所有的隐变量j_i服从分布$P(J|X,\Theta^t)$的条件下，基于当前的参数估计值Θ^t，推导参数Θ的数学期望$Q(\Theta;\Theta^t)$：

$$Q(\Theta;\Theta^t)=\mathbb{E}_{J\sim P(J|X,\Theta^t)}\left[\sum_{i=1}^{N}\ln\left(p(\boldsymbol{x}_i|j_i,\Theta)P_j\right)\right]=\sum_{i=1}^{N}\sum_{j=1}^{M}\ln\left[p(\boldsymbol{x}_i|j,\Theta)P_j\right]P(j|x_i;\Theta^t) \tag{3.42}$$

公式(3.42)表示，应在当前参数估计值Θ^t的条件下评估每一个训练样本$\boldsymbol{x}_i$属于每一个高斯分布的后验概率$P(j|x_i;\Theta^t)$，并以其为权重来估计样本$\boldsymbol{x}_i$的对数似然函数$\ln\left[p(\boldsymbol{x}_i|j,\Theta)P_j\right]$在所有高斯分布上的加权平均值。根据贝叶斯公式，后验概率$P(j|x_i;\Theta^t)$可以写为：

$$P(j|x_i;\Theta^t)=\frac{p(\boldsymbol{x}_i|j,\Theta^t)P_j^t}{\sum_{j=1}^{K}p(\boldsymbol{x}_i|j,\Theta^t)P_j^t} \tag{3.43}$$

将后验概率$P(j|x_i;\Theta^t)$简记为γ_{ji}^t，将公式(3.39)带入公式(3.42)，可以得到$Q(\Theta;\Theta^t)$的具体数学形式如公式(3.44)所示，其中需要被优化的参数为$\Theta=\{\boldsymbol{\mu}_j,\Sigma_j,P_j\}_{j=1}^{M}$。

$$\begin{aligned}Q(\Theta;\Theta^t)&=\sum_{j=1}^{K}\left[\ln P_j\sum_{i=1}^{N}\gamma_{ji}^t+\sum_{i=1}^{N}\left(\gamma_{ji}^t\ln p(\boldsymbol{x}_i|j,\Theta)\right)\right]\\&=\sum_{j=1}^{K}\left[\ln P_j\sum_{i=1}^{N}\gamma_{ji}^t-\frac{1}{2}\sum_{i=1}^{N}\left[\gamma_{ji}^t\left(\ln 2\pi+\ln|\Sigma_j|+(\boldsymbol{x}_i-\boldsymbol{\mu}_j)^T\Sigma_j^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_j)\right)\right]\right]\end{aligned} \tag{3.44}$$

3）M 步骤

根据 EM 算法，M 步骤是寻找使得目标函数$Q(\Theta;\Theta^t)$最大化的参数Θ 的过程，可以通过求解$Q(\Theta;\Theta^t)$对$\boldsymbol{\mu}_j,\Sigma_j,P_j$的偏导方程实现。

首先另$Q(\Theta;\Theta^t)$对均值向量$\boldsymbol{\mu}_j$求导等于 0，带入公式(3.44)，推导得到：

$$\frac{\partial Q(\Theta;\Theta^t)}{\partial\boldsymbol{\mu}_j}=0\rightarrow-\Sigma_j^{-1}\sum_{i=1}^{N}\left[\gamma_{ji}^t(\boldsymbol{x}_i-\boldsymbol{\mu}_j)\right]=0\rightarrow\frac{\sum_{i=1}^{N}\gamma_{ji}^t\boldsymbol{x}_i}{\sum_{i=1}^{N}\gamma_{ji}^t}=\hat{\boldsymbol{\mu}}_j,j=1,\dots,M \tag{3.45}$$

同理，另$Q(\Theta;\Theta^t)$分别对Σ_j求导等于 0，推导得到：

$$\begin{aligned}\frac{\partial Q(\Theta;\Theta^t)}{\partial\boldsymbol{\mu}_j}&=0\\-\frac{1}{2}\Sigma_j^{-1}\sum_{i=1}^{N}\gamma_{ji}^t-\frac{1}{2}\Sigma_j^{-1}\left(\sum_{i=1}^{N}\left[\gamma_{ji}^t(\boldsymbol{x}_i-\boldsymbol{\mu}_j)(\boldsymbol{x}_i-\boldsymbol{\mu}_j)^T\right]\right)\Sigma_j^{-1}&=0\\\frac{\sum_{i=1}^{N}\gamma_{ji}^t(\boldsymbol{x}_i-\boldsymbol{\mu}_j)(\boldsymbol{x}_i-\boldsymbol{\mu}_j)^T}{\sum_{i=1}^{N}\gamma_{ji}^t}&=\hat{\Sigma}_j\end{aligned} \tag{3.46}$$

参数P_j的求解与前两个参数略有不同，主要是因为 GMM 模型对参数P_j给出了$\sum_{j=1}^{M}P_j=1$的约束条件，需要引入拉格朗日法求解该有约束最优化问题。首先，忽略公式(3.44)与参数P_j无关的加法项；其次等式约束条件$\sum_{j=1}^{M}P_j=1$，构造拉格朗日乘子式$Q(\Theta;\Theta^t,\lambda)$:

$$Q(\Theta;\Theta^t,\lambda)=\sum_{j=1}^{M}\left[\ln P_j\sum_{i=1}^{N}\gamma_{ji}^t\right]+\lambda\left[\left(\sum_{j=1}^{M}P_j\right)-1\right] \tag{3.47}$$

通过求解$Q(\Theta;\Theta^t,\lambda)$对$P_j$的偏导方程，得到：

$$\begin{aligned}\frac{\partial Q(\Theta;\Theta^t,\lambda)}{\partial P_j}&=0\\-\frac{\sum_{i=1}^{N}\gamma_{ji}^t}{\lambda}&=\hat{P}_j\end{aligned} \tag{3.48}$$

将公式(3.48)的结果带入约束条件$\sum_{j=1}^{M} P_j = 1$，可以解出：

$$\lambda = -\sum_{j=1}^{M}\sum_{i=1}^{N}\gamma_{ji}^{t} = -\sum_{i=1}^{N}\sum_{j=1}^{M}P(j|x_i;\Theta^t) = -N \tag{3.49}$$

将公式(3.49)带入公式(3.48)，推导得到：

$$\hat{P}_j = \frac{\sum_{i=1}^{N}\gamma_{ji}^{t}}{N} \tag{3.50}$$

至此，我们得到了第t轮的参数最优估计值$\hat{\boldsymbol{\mu}}_j, \hat{\Sigma}_j$和$\hat{P}_j$，将其分别用于更新第 t+1 轮的参数估计值$\boldsymbol{\mu}_j^{t+1}, \Sigma_j^{t+1}, P_j^{t+1}$。

4）EM 算法步骤总结

基于上述推导过程和结果，可以将 GMM+EM 算法步骤总结如图 3-7 所示，其中虚线代表了公式带入的关系与顺序。根据 EM 算法，不断重复上述过程直到满足终止条件，就可以实现 GMM 模型的参数估计。

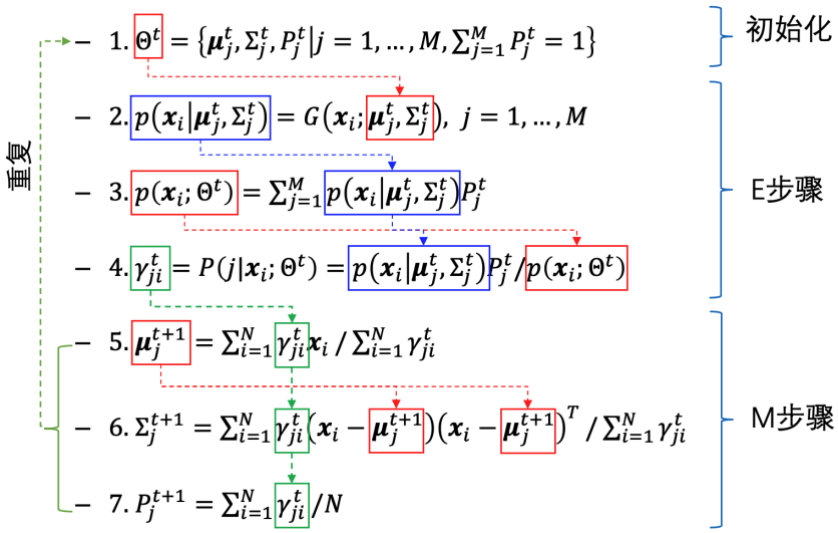


图 3-7 GMM+EM 算法流程示意图

● 小结——有参数概率模型估计方法

至此，我们对基于有参数函数的概率模型估计方法作一个小结。在给定概率模型的参数化函数形式的条件下，我们可以采用最大似然估计（MLE）、最大后验概率估计（MAP）和贝叶斯估计（BE）三种方法中的任意一种对概率函数中的参数进行估计。但现实问题中，观测变量的分布具体服从哪一种函数形式通常是无法事先确定的。因此我们引入了高斯混合模型（GMM）。该模型作为一种通用的函数表达形式，理论上能够以足够的精度对任意概率分布进行拟合。然而高斯混合模型的最大似然估计求解过程相对复杂，且无法直接求得解析解。因此我们又介绍了期望最大化算法（EM），该算法能够通过一种迭代优化的方式对 GMM 的

最大似然解进行相对快速和稳定地计算。

虽然 EM 算法的计算效率相对较高，但在实际应用中仍然存在两个问题：1）超参数M——预定义的高斯分布的个数——难以确定；2）当数据库规模较大、维度较高或超参数M的数值较大时，EM 算法的收敛速度和可靠性难以满足实际应用需求。因此，单纯依靠“GMM+EM”的有参概率模型估计方法，并不能保证在所有的贝叶斯决策问题中都取得令人满意的结果，需要探索其他新的概率估计方法。

3.3.5 无参数概率模型估计方法

有参概率估计方法在实际应用时可能遭遇两类问题：1）对需要估计的随机变量缺少足够的先验知识，因此无法为概率模型找到合适的参数化函数，也无法预估概率密度函数的泛化能力；2）概率模型的参数优化问题过于复杂，具体表现为计算量过大，优化过程收敛速度慢，容易陷入局部最优解等。本章要介绍的无参数概率模型估计方法在某些方面可以解决上述问题。顾名思义，无参数概率模型估计是指无需为类条件概率密度建立参数化函数形式的概率估计方法，此类方法一般利用表格或者样本作为概率模型的基本表达形式。本节将要介绍的直方图、Parzen 窗与 k-近邻估计方法，均属于此类方法。

- **直方图法**

直方图法（Histogram）将数据空间分割为一系列等大的格子（通常称为 bin），并统计每个格子中观测样本的数量作为概率估计的依据。以一维数据为例，设观测样本总数量为N，每个 bin 的中心位置为$\hat{x}_j$，宽度均为h，则这个格子可以记为$B_j = [\hat{x}_j - h/2, \hat{x}_j + h/2)$。需要注意的是这里使用了左闭右开的区间来定义$B_j$。如果选用两侧闭区间，可能导致同一个样本在两个相邻的格子上被重复统计；如果选用两侧开区间，会遗漏正好处在两个相邻区间边界上的样本。如果将论域分为M个格子，要求论域$X = \cup_{j=1}^{M} B_j$，$|B_j| = |B_k|, B_j \cap B_k = \emptyset, \forall j \neq k$。假设落入$B_j$的样本数量记为$N_i$，则基于直方图的概率密度估计值可以写为：

$$\hat{p}(x|x \in B_j) = \hat{p}(\hat{x}_j) \approx \frac{1}{h}\frac{N_j}{N} \tag{3.51}$$

公式(3.51)说明某个具体样本x的概率密度可以用它所在的格子中心位置$\hat{x}_j$的概率密度估计值近似，之所以要除以格子的宽度h，是为了保证概率密度$p(x)$在论域X上的积分为 1，既满足$\int_{x \in X} p(x)dx = 1$。对于维度为$d$的数据而言，一个格子对应于一个以$\hat{\boldsymbol{x}}_j \in \Re^d$为中心，边长为$h$的$d$维超立方体，则相应的概率密度函数估计公式可以写为：

$$\hat{p}(\boldsymbol{x}|\boldsymbol{x} \in B_j) = p(\hat{\boldsymbol{x}}_j) \approx \frac{1}{h^d}\frac{N_j}{N} \tag{3.52}$$

直方图是一种相当便捷且直观的概率统计方法，其估计结果可以用一个简单的表格来表示，对于具体的待估计样本只需要在表格中找到相应的格子就能实现概率估计值的快速查询。对于二维以下的观测数据而言，直方图可以直接转化为对应的柱状图，从而获得直观的可视化效果。直方图方法的最大挑战是“维度诅咒”问题。假设d维观测数据在一个维度上需要分为M个格子，则整个观测空间中的 bin 的总数量为M^d，这意味着直方图对应的表格的存储和查询计算量都随着维度d呈指数上升。以 Iris 数据集为例，假设$M = 10$，则需要建立10^4个格

子。此时格子的数量已经远远大于样本数量——150，大量的 bin 中都不会落入样本，这意味着直方图的概率统计结果将不再可靠。因此直方图方法通常只应用于不超过 2 维且样本数量比较丰富的观测数据。此外，格子的大小h的选择也是个相当棘手的问题。如果h取值太大，则概率估计相对稳定（误差方差较小），但精度较差（误差均值较大）；如果h取值太小，虽则估计精度较高，但估计的稳定性会变差，甚至很多格子中根本不会落入样本。

- **Parzen 窗法**

在分析直方图法的缺陷时，我们发现即便h的取值比较适中，当样本$\boldsymbol{x}$靠近格子边界时，由于距离格子中心$\hat{\boldsymbol{x}}_j$较远，与那些靠近格子中心的样本相比，其概率密度估计结果的准确性和可靠性仍会比较差。能够找到一种新的估计方法，使得不同位置的样本在估计精度上有更好的一致性呢？Parzen 窗算法对该问题提出了一种新的解决思路。对于一个待估计的样本$\boldsymbol{x} \in \Re^d$，基于 Parzen 窗法的概率密度估计可由公式(3.53)计算得到。

$$\hat{p}(\boldsymbol{x}) = \frac{1}{h^d}\left(\frac{1}{N}\sum_{i=1}^{N}\mathcal{K}(\boldsymbol{x},\boldsymbol{x}_i)\right) \tag{3.53}$$

其中：

$$\mathcal{K}(\boldsymbol{x},\boldsymbol{x}_i) = \begin{cases} 1 & \boldsymbol{x}_i \in \Omega(\boldsymbol{x}) \\ 0 & \text{otherwise} \end{cases} \tag{3.54}$$

公式(3.53)和(3.54)的含义是以当前待观测样本$\boldsymbol{x}$为中心构建一个边长为h的d维超立方体，记为$\Omega(\boldsymbol{x})$（也就是我们所说的 Parzen 窗口），并统计观测数据集中落入$\Omega(\boldsymbol{x})$的样本的数量用于估计样本$\boldsymbol{x}$的概率密度函数。Parzen 窗法中的“窗”就相当于直方图法中的“格子”，但直方图法的格子的位置是固定的，而 Parzen 窗法中的窗的位置是可移动的。这样每次在估计一个具体的样本$\boldsymbol{x}$时，Parzen 窗的窗口中心都正好处在样本$\boldsymbol{x}$上，这样就可以有效解决直方图法中当样本$\boldsymbol{x}$距离 bin 中心位置较远时的估计精度不足的问题。

由于 Parzen 窗法并不需要实现建立M^d个格子，而是在给出查询样本$\boldsymbol{x}$后才临时构建一个 Parzen 窗进行概率统计，因此可以有效解决“维度诅咒”问题，这是该方法相对于直方图法的巨大优势；但在另一方面，Parzen 窗法在每一次概率查询任务中，都需要重新计算所有样本是否落入当前的窗口内，而不像直方图法那样直接查表即可，因此单词查询计算量较大，且需要存储全部的 N 个训练样本。

此外，与直方图类似，Parzen 窗方法也面临着窗口大小h的两难选择问题。尤其在一些数据分布相对稀疏的区域，传统的 Parzen 窗估计得到的概率密度值经常为 0。为解决这一问题，研究者提出了相对灵活的改进 Parzen 窗方法，可以通过调整函数$\mathcal{K}(\boldsymbol{x},\boldsymbol{x}_i)$，使传统的方形 Parzen 窗变为一个高斯窗口，如公式(3.55)所示。

$$\mathcal{K}_g(\boldsymbol{x},\boldsymbol{x}_i) = \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}}\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{x}_i\|^2\right) \tag{3.55}$$

由于高斯函数在有限范围内不会取值为 0，因此即便在相对稀疏的区域，高斯 Parzen 窗仍然能够统计出有效的概率密度函数估计值。由于$\mathcal{K}_g(\boldsymbol{x},\boldsymbol{x}_i)$的形式类似于核函数，因此高斯

Parzen 窗法一般被称为**核密度估计法**（Kernel Density Estimation, KDE）。然而即便采用了核密度估计法，仍然存在着一个参数σ用于控制高斯窗口的大小，显然参数σ的选择与参数h类似，仍是一个两难问题。

- **K-近邻法**

为了解决窗口大小的选择问题，研究者提出了一种 K-近邻法（K-Nearest Neighbor, KNN）用于概率估计。与直方图或 Parzen 窗采用固定大小的窗口不同，K-近邻使用了一种大小可自动调节的采样策略。对于一个待估计样本$\boldsymbol{x}=[x_1,x_2,\dots,x_d]$，在观测数据集中选择 K 个距离该样本最近的样本（**K-近邻样本**），记为$\boldsymbol{x}^j=\left[x_1^j,x_2^j,\dots,x_d^j\right],j=1,2,\dots,K$。找到一个以$\boldsymbol{x}$为中心的最小的 d 维超立方体窗口$\Omega_K(\boldsymbol{x})$使得$\boldsymbol{x}^j\in\Omega_k(\boldsymbol{x}),\forall j=1,\dots,K$，则该立方体的边长$h_K$可以由公式(3.56)计算得到。

$$h_K=\max_{j,l}\{|x_l^j-x_l|\},\qquad j=1,\dots,K;l=1,\dots,d \tag{3.56}$$

则新样本$\boldsymbol{x}$的概率密度估计值为：

$$\hat{p}(\boldsymbol{x})=\frac{1}{h_K^d}\frac{K}{N} \tag{3.57}$$

公式(3.57)的含义是在保证采样窗口内观测样本数量为K的条件下，改变采样窗口的大小来实现概率密度估计。K-近邻方法可以很好的解决窗口大小的选择的问题，从而产生更加稳定和精准的估计结果。但在计算过程中，不但需要对每一个观测样本到当前待估计样本的距离进行计算，还需要根据距离进行排序以确定 K-近邻样本，因此计算速度比 Parzen 窗方法更慢。

- **无参概率模型估计方法对比分析**

基于上述介绍，我们发现直方图、Parzen 窗（含核密度估计法）和 K-近邻三种方法本质上都采用了窗口采样的方式进行概率密度估计，但在窗口的位置、大小的选择上采用了不同的策略，如图 3-8 所示。直方图法的采样窗口位置和大小是固定的，查询样本$\boldsymbol{x}$采用自身所在的窗口B_j的中心位置$\hat{\boldsymbol{x}}_j$的概率密度估计值作为估计结果；Parzen 窗方法以查询样本$\boldsymbol{x}$为采样窗口的中心，以固定的大小创建一个采样窗口进行概率密度估计；K-近邻方法同样以样本$\boldsymbol{x}$为采样窗口的中心，但窗口的大小可变，以保证窗口内的观测样本个数为K时能够取到的最小窗口为准。从算法思路上看，概率估计问题上：

1）直方图方法误差较大但计算量小；

2）Parzen 窗相对准确但计算量存在明显的增加；

3）K-近邻方法最为准确和稳定但计算量也最大。

相比于有参数的概率估计方法，以直方图、Parzen 窗和 K-近邻估计为代表的无参数概率估计方法最大的优势在于无需对类条件概率密度$p(\boldsymbol{x}|\omega)$的函数形式进行假设，也没有复杂的优化或学习过程。直方图法计算速度最快，但精度低，可靠性差，存在“维度诅咒”风险，很难应用于高维数据；Parzen 窗和 K-近邻法在每次概率查询时需要对所有已知样本进行统计，因此需要存储所有数据且计算速度较慢。因此相比于有参数概率模型估计方法，上述三种无

参概率模型估计算法可谓各有优劣，需要根据具体的问题条件酌情选用。

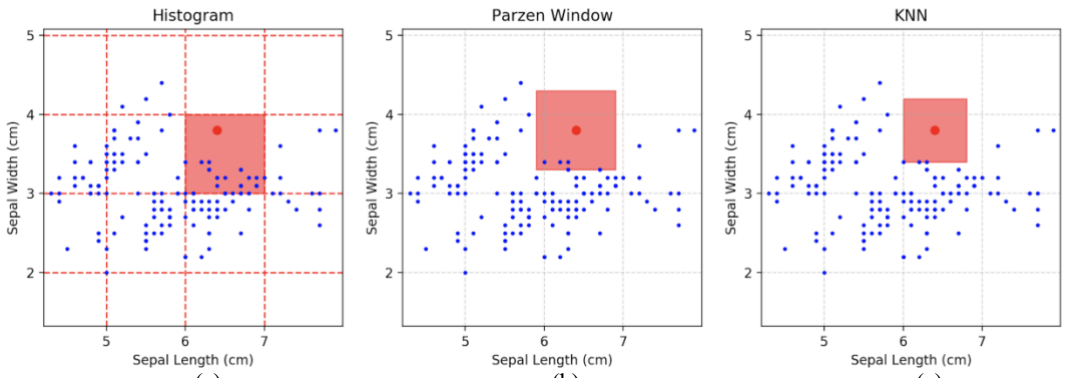


图 3-8 无参数概率密度估计方法示意图,(a)直方图法，(b)Parzen 窗法，(c)K 近邻法。

3.3.6 朴素贝叶斯分类器

无论是以最大似然估计和高斯混合模型为代表的有参数概率估计方法，还是以直方图、Parzen 窗、K-近邻估计为代表的无参数概率估计方法，在处理高维数据的概率估计问题时都存在一定的困难。朴素贝叶斯分类器（Naïve Bayesian Classifier，NBC）为这一问题提供了一个解决方案。针对本节提到的核心问题——类条件概率$p(\boldsymbol{x}|\omega)$的估计问题，朴素贝叶斯分类器给出的方案是，对于观测随机变量$\boldsymbol{x}=[x_1,x_2,\dots,x_d]^T$:

$$p(\boldsymbol{x}|\omega)=\prod_{i=1}^{d}p(x_i|\omega) \tag{3.58}$$

根据概率论的基础知识，公式(3.58)的成立是有条件的，即要求随机变量$\boldsymbol{x}$的每一个分量$x_i,i=1,2,\dots,d$都相互独立。由于样本特征向量$\boldsymbol{x}$的每一个分量x_i在模式识别任务中都可以视为识别对象的一个属性，这个条件被称为“属性条件独立性假设”（Attribute Conditional Independence Assumption, AICA）。我们都知道万事万物是普遍联系的。因此假设同一个对象的两个属性相互独立显然是一种过于理想化的假设。Naïve Bayesian Classifier 中的“Naïve”一词原意为天真的、幼稚的，正是对这种过于理想化的假设的一种说明。但为了方便交流，本书仍然使用“朴素贝叶斯分类器”这一学术圈普遍接受的翻译。

朴素贝叶斯分类器使用最大后验概率分类准则：

$$\omega^*=\underset{\omega}{\operatorname{argmax}}\,P(\omega|\boldsymbol{x})=\underset{\omega}{\operatorname{argmax}}\,p(\boldsymbol{x}|\omega)P(\omega) \tag{3.59}$$

利用公式(3.58)，将$p(\boldsymbol{x}|\omega)$转化为$p(x_i|\omega)$，在通过构造对数似然函数，则公式(3.59)的最优化问题转化为：

$$\omega^*=\underset{\omega}{\operatorname{argmax}}\left(\ln P(\omega)+\sum_{i=1}^{d}\ln(p(x_i|\omega))\right) \tag{3.60}$$

根据公式(3.60)，只需要能够估计样本的每个属性x_i的类条件概率密度$p(x_i|\omega)$，就可以基于朴素贝叶斯分类器实现高维数据的分类。原则上，本章学习的各类概率估计方法都可以应用于该任务。考虑到属性x_i为一维的标量，不涉及到“维度诅咒”问题，因此从算法复杂度、存储量、计算量等各方面看，直方图法都具有比较明显的优势。但从公式(3.58)和(3.60)分析，如果单个属性x_i的类条件概率密度$p(x_i|\omega)=0$，则所有属性的联合条件概率$p(\boldsymbol{x}|\omega)=0$，这回导致朴素贝叶斯分类器的分类结果不可靠。而当样本数量有限时，落入直方图的某个格子B_j的样本数量$N_j=0$的情况很有可能出现，根据公式(3.51)就会导致概率估计结果为零。为了避免上述情况，在与朴素贝叶斯分类器配合时，直方图的某个格子对应的概率密度估计值一般采用以下计算方式：

$$p(x_i|x_i\in B_{i,j},\omega)\approx\frac{N_{i,j}+\lambda}{N+M_i\lambda}\frac{1}{|B_{i,j}|} \tag{3.61}$$

其中，$B_{i,j}$表示属性x_i的论域上的第j个格子，$|B_{i,j}|$表示格子的大小；$N_{i,j}$表示训练样本集中第i个属性x_i落入格子$B_{i,j}$的样本数量；M_i是x_i对应的直方图格子的数量；$\lambda>0$，称为**拉普拉斯平滑项**。根据公式(3.61)，$p(x_i|\omega)$在该属性的论域上的积分仍为 1，且当$N_{i,j}=0$时，$p(x_i|\omega)$的估计值也不为零，因此很好地解决了直方图与朴素贝叶斯分类器结合过程中出现的问题。

3.3.7 贝叶斯网络

虽然朴素贝叶斯分类器极大地简化了高维数据的概率估计问题，但属性条件独立性假设过于严格，在很多实际应用中无法满足。对于观测样本不同属性之间普遍存在的非独立性（或称为依赖关系），学术界提出了“贝叶斯网络”加以描述和解决。

首先忽略类别条件ω，根据概率论中的链式法则，样本$\boldsymbol{x}$的概率可以写为：

$$p(\boldsymbol{x})=p(x_1,\dots,x_d)=p(x_d|x_{d-1},\dots,x_1)p(x_{d-1}|x_{d-2},\dots,x_1)\dots p(x_2|x_1)p(x_1) \tag{3.62}$$

公式(3.62)将d个随机变量$x_i,i=1,\dots,d$的联合分布转化为$d-1$个条件概率$p(x_i|x_{i-1},\dots,x_1),i=2,\dots,d$和一个边缘概率$p(x_1)$的乘积，且该法则与上述$d$个随机变量的具体排序无关。在很多具体问题中，当前属性x_i并非与$x_1,\dots,x_{i-1}$这$i-1$个属性都存在依赖关系，因此可以将与属性x_i有依赖关系的其他属性集合定义为A_i，则有：

$$p(x_i|x_{i-1},\dots,x_1)=p(x_i|A_i),A_i\subseteq\{x_{i-1},\dots,x_1\} \tag{3.63}$$

则公式(3.62)可以进一步写为：

$$p(x_1,\dots,x_d)=p(x_1)\prod_{i=2}^{d}p(x_i|A_i) \tag{3.64}$$

显然，只要能确定每一个属性x_i对应的依赖属性集合A_i，并估计出相应的条件概率$p(x_i|A_i)$，则原始的多属性联合概率估计问题就可以转化为相对简单的多个概率分布乘积问题。为了简洁地表达不同属性之间的依赖关系，贝叶斯网络模型引入了图（Graph）的概念。

图主要由节点和边组成的，一般表示为集合$G=(V,E)$，V为节点集合，E为连接节点的边的集合。将每个属性x_i看作一个节点v_i，如果属性x_i依赖于属性x_j，则构造一条从节点v_j指向结

点v_i的单向边。根据$A_i, i = 2, \ldots, d$的具体分布情况，可以画出一个“图模型”。对 Iris 数据库做一个合理的扩展，可以举例说明如何构造一个图模型。在 Iris 数据库有的四个属性基础上，通过观测再增加花冠高度和平均日照强度这两个属性，从而重新构造一个数据集，记为 Iris-plus 数据库，其中x_1为花冠高度，x_2为平均日照强度，x_3为花萼宽度，x_4为花萼长度，x_5为花瓣宽度，x_6为花瓣长度。基于数据分析结合植物学知识，做出以下依赖性假设：

1）花瓣长度x_6依赖于花瓣宽度x_5和平均日照强度x_2，则有$A_6 = \{x_5, x_2\}$；

2）花瓣宽度x_5依赖于平均日照强度x_2，则有$A_5 = \{x_2\}$；

3）花萼长度x_4依赖于花萼宽度x_3和花冠高度x_1，则有$A_4 = \{x_3, x_1\}$；

4）花萼宽度x_3依赖于花冠高度x_1，则有$A_3 = \{x_1\}$；

5）平均日照强度x_2依赖于花冠高度x_1，则有$A_2 = \{x_1\}$；

则对应于 Iris-plus 数据库的贝叶斯网络可以表示为以下图模型。

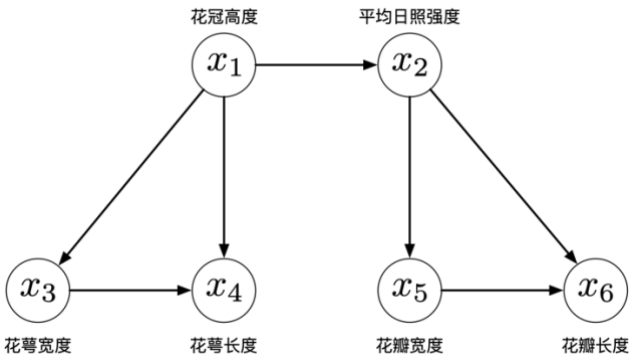


图 3-9 面向 Iris-plus 数据库的贝叶斯网络图模型

根据公式(3.64)，则 Iris-plus 数据库六个属性的联合概率可以写为：

$$p(x_1, x_2, x_3, x_4, x_5, x_6) = p(x_6|x_5, x_2)p(x_5|x_2)p(x_4|x_3, x_1)p(x_3|x_1)p(x_2|x_1)p(x_1) \tag{3.65}$$

相比于原始问题中的 6 维数据概率估计问题，经过贝叶斯网络转化后的问题中每个条件概率最多只包含 3 个属性，可以使用类似于直方图的条件概率表进行统计，这样做可以大量节省存储与计算成本，降低“维度诅咒”风险。假设上述问题中每个属性x_i对应的直方图格子数量均为M。则原始问题的格子总数量为M^6，而根据公式(3.65)，贝叶斯网络需要统计的格子总数为$2M^3 + 3M^2 + M$，朴素贝叶斯分类器需要统计的数量则为$6M$。当$M = 10$时，三种方法统计的格子总数量比例约为$16667:39:1$。显然，朴素贝叶斯分类器计算量最小；贝叶斯网络次之；而原始问题的计算量要远大于两者。

总的来说，贝叶斯网络为多变量（高维）数据的概率估计提供了一种折中的方法。相比于朴素贝叶斯分类器，贝叶斯网络由于不需要属性条件独立性假设，因此应用范围更广或者说估计精度更高；相比于基于 EM 算法的 GMM 模型，由于不需要在高维数据空间进行参数向量的迭代优化，因此不容易陷入局部最小，算法速度也更快；相比于直方图、Parzen 窗或 K-近邻估计，贝叶斯网络不存在明显的维度诅咒、样本数量不足等问题，也不存在每次估计都需要对所有样本进行统计的计算速度问题。但如何确定贝叶斯网络的结构是一个巨大的挑

战，因此对于超高维数据——如图像、语音、文本等非结构化数据——的模式识别任务，贝叶斯网络并不是一个很好的选择，但其研究思想在机器学习领域仍然具有深远的影响，

■ 知识拓展-贝叶斯网络结构学习

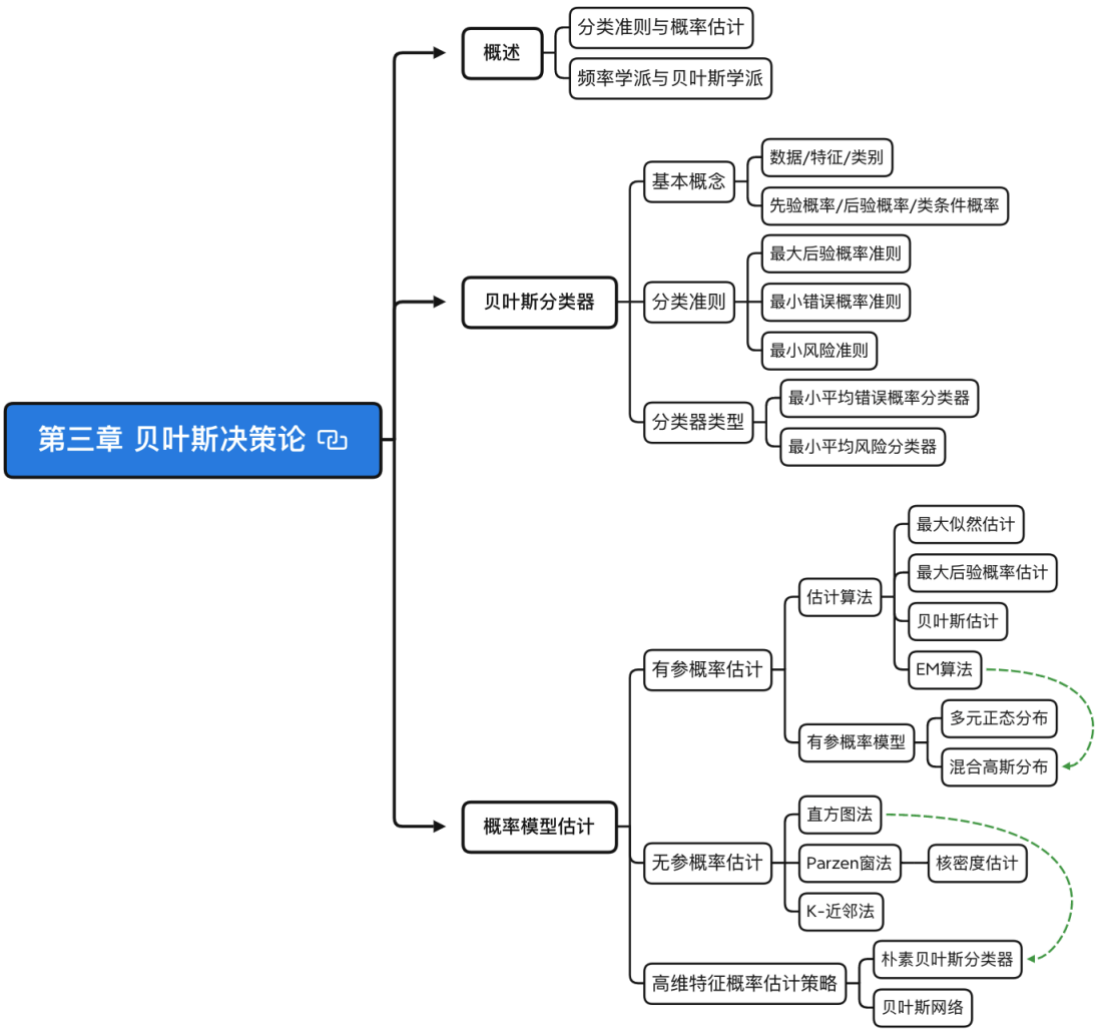
尽管贝叶斯网络在理论上具有相当好的性能，但在实际应用中需要首先解决一个问题——如何确定贝叶斯网络的图模型结构。解决这个问题的具体方法相对复杂，不属于本书要详细介绍的范畴。但为了避免读者在这一问题上出现知识体系的逻辑漏洞，这里将对如何确定贝叶斯网络的结构这一问题做简单的介绍。由于本书关注的是基于机器学习的模式识别方法，因此暂不考虑单纯依靠人的经验、知识或物理因果关系确定网络结构的方法，只讨论基于数据统计、分析与优化的结构学习方法。

第一种方法从全局入手，将整个贝叶斯网络看作一个变量，将基于贝叶斯网络的概率模型估计效果的评价指标看作是一个目标函数，则不同的网络结构就对应于不同的解，网络结构的学习过程则是目标函数的优化过程，能够使得评价指标最优的结构就是贝叶斯网络结构学习的目标。作为目标函数的评价指标被称为评分函数，通常分为贝叶斯评分函数与基于信息论的评分函数两类。该方法将贝叶斯网络的学习问题转化为一个离散优化问题，由于贝叶斯网络结构的候选空间通常极大，难以穷举，因此离散优化过程通常还需要相应的启发式随机搜索策略（贪婪算法、随机抽样、遗传算法、蚁群算法等）。因此，此类方法又称为基于评分搜索的结构学习方法。

第二种方法从局部入手，主要通过不同变量（属性）之间的条件独立性测试对网络结构提出约束条件，最后构造并优化一个有向无环图来尽可能多的涵盖已经统计出来的变量关系约束条件。条件独立性测试主要采用基于信息论的互信息或条件互信息作为度量，结构的优化可以采用从完全图开始逐步删减边的策略，也可以采用从稀疏图开始逐渐增添边的策略，也可以采用两者结构的多阶优化策略。代表性工作有 SGS 算法，PC 算法和 TPDA 算法。由于此类方法重点关注网络结构的约束条件，因此又称为基于约束条件的结构学习方法。

第三种方法，则是对前两种思路的结合，相当于有约束条件下的评分优化问题。通常的策略是首先基于约束条件对网络结构进行初始化，之后采用评分搜索策略对网络结构进行局部优化。这一过程也可以反复多次迭代，从而形成多阶段算法。

本章思维导图


第三章 贝叶斯决策论
概述
分类准则与概率估计
频率学派与贝叶斯学派
贝叶斯分类器
基本概念
数据/特征/类别
先验概率/后验概率/类条件概率
分类准则
最大后验概率准则
最小错误概率准则
最小风险准则
分类器类型
最小平均错误概率分类器
最小平均风险分类器
概率模型估计
有参概率估计
估计算法
最大似然估计
最大后验概率估计
贝叶斯估计
EM算法
有参概率模型
多元正态分布
混合高斯分布
无参概率估计
直方图法
Parzen窗法
核密度估计
K-近邻法
高维特征概率估计策略
朴素贝叶斯分类器
贝叶斯网络

本章习题

一、填空题

1. 利用肿瘤大小预测其为良性还是恶性，则恶性肿瘤中直径大小为 8 厘米的案例的比例可以作为对（　　）概率的估计，所有肿瘤中恶性肿瘤的比例可以作为对（　　）概率的估计。
2. 某支鸢尾花从属于 Setosa 类的概率为 0.8，则将其分给 Setosa 类的错误概率是（　）。
3. 用多元高斯函数去描述 Iris 数据集中的 Virginica 类的类条件概率密度，一共需要估计（　　）个参数；用混合高斯模型去描述整个 Iris 数据库的概率密度分布，一共需要估计（　　）个参数（注：参数个数以标量个数为准）

二、判断题

4. 在分类问题中，类条件概率是概率质量，后验概率是概率密度。（　）
5. 最小平均错误概率分类器和最小错误概率分类准则给出的分类结果是一致的。（　）
6. 最大似然函数的自变量是概率模型参数，样本$\boldsymbol{x}_i, i = 1, \ldots, N$是参数。（　）

三、选择题

7. 某工厂拟根据有无噪声来诊断机器故障，共统计了 100 台机器，发现无噪声的机器 40 台，已知故障机器有噪声的类条件概率为 0.9，无故障机器无噪声的类条件概率为 0.7，则有噪声机器为故障机器的后验概率为（　）

A. 0.6　　B. 0.75　　C. 0.8　　D. 无法确定

8. 一维小样本库共有 10 个样本{1,1,2,3,7,8,5,5,7,4}。使用直方图法，设 bin 宽度为2，左开右闭，5 个 bin 中心分别为 0,2,4,6,8，则x=3 处的概率密度估计值为（　）

A. 0.1　　B. 0.2　　C. 0.3　　D. 0.4

9. 已知 5 个二维样本：[0,0],[0.5,1],[1,1],[1,0.5],[0,1]。使用 KNN 算法，设$K = 3$，则点[0.5,0.6]处的概率密度函数估计值为：（　）

A. 0.6　　B. 3　　C. 0.9375　　D. 2.4

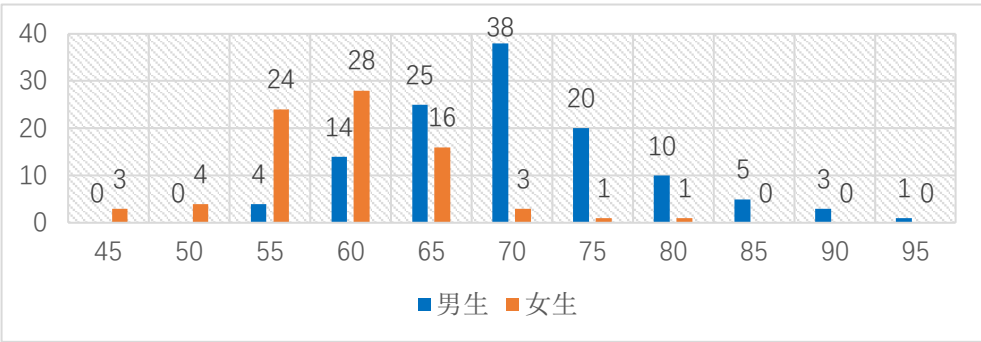
四、简答题

10. 请在现实生活中举一个不适合使用最小错误概率分类准则，而适合使用最小风险分类准则的模式分类任务案例，并分析原因。
11. 如果对于某个数据集的概率分布模型，最大似然估计和最大后验概率估计给出结果完全相同，试解释可能造成这一结果的原因（可画图说明）。
12. 请自行查阅资料，列举一种适合使用贝叶斯估计的概率模型假设，并推导相应的贝叶斯估计解析表达式。

五、计算题

13. 假设一个学校里有 60%的男生和 40%的女生；女生点外卖的人数和吃食堂的人数相等，男生中吃外卖的比例约为 80%。假设你在校门口的取餐点捡到一份无人领取的外卖，你更倾向于送到男生宿舍还是女生宿舍？请给出计算依据。

14. 某专业共有学生 200 人，其中男生 120 人，女生 80 人，他们的体重分布如下表所示：



（1）请计算男生的先验概率、女生体重处于 55 公斤区间段的类条件概率、男学生体重为 55 公斤的类条件概率密度函数估计值、任意学生体重处于 55 公斤区间段的概率、某个体重为 55 公斤的学生属于女生类别的后验概率，并写出计算过程。

（2）请根据最大后验概率预测某个体重为 55 公斤的学生的性别，并写出计算过程。

（3）设男生为ω_1类，女生为ω_2类，给出预测学生性别的风险矩阵为：

$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix}$$

请根据最小风险分类准则，给出预测结果，并列出计算过程。

15. 已知两个类别ω_1和ω_2，其先验概率相等，两类的类条件概率服从正态分布，其中$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$，$\mu_2 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$，$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$，试对样本$\boldsymbol{x} = \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix}$进行分类，并给出计算过程。

16. 两个学生玩猜硬币的游戏；共猜了 30 次，其中正面 17 次，反面 13 次。假设一枚硬币扔出正面的概率为θ，单次扔硬币的结果服从伯努利分布：

$$P(x;\theta) = \theta^x(1-\theta)^{1-x}, x = 0 \; or \; 1$$

其中，$x = 1$表示扔出正面，$x = 0$表示扔出反面；

（1）根据题意，写出似然性函数的数学表达式。

（2）写出该问题的最大似然估计目标函数的具体数学形式。

（3）采用最大似然估计求取参数θ的值？

（4）假设θ的分布服从$\mathcal{N}(0.4, 0.2^2)$，请利用最大后验概率估计方法估计θ的取值。

17. 将 Iris 数据库按照 7:3 的比例分为训练集与测试集，利用直方图+朴素贝叶斯分类器的分类方案构建分类器，并给出测试结果。