第3章 贝叶斯决策

CHAPTER 3

决策是机器学习中一个相对独立的部分。当机器学习的模型已经确定，对于新的输入可计算模型输出，模型的输出代表什么，即对模型输出做出最后判断，这是决策过程要做的事情。针对不同的模型，决策过程的作用是不一样的。对于有的模型，模型输出直接表示了明确的结果，不需要一个附加的决策过程；而对于其他模型，尤其是概率类模型，往往需要对模型输出做出一个最终的决策，这是决策过程的作用。在机器学习中，决策往往是一个独立且相对简单的单元，本章讨论决策问题，集中在贝叶斯决策。

3.1 机器学习中的决策

ML05-统计基础-2

一般来讲，机器学习是通过训练过程得到描述问题的模型，可将模型表示为一种数学关系。当给出新的输入数据时，可按照模型需要的格式将输入数据转换为模型可接受的输入特征向量，计算模型的输出。所谓决策，就是对于模型的输出给出一个判决结果。

决策就是要做出最后的结论，对于分类，要给出类型的结果；对于回归，要给出输出值。对于一个模型，从其输出是否确定的角度，可将模型分为概率和非概率模型。对于非概率模型，模型是一个确定性的判别函数，该模型通过训练过程直接得到确定的函数关系 $\hat{y}=f(\boldsymbol{x})$，其中 $\boldsymbol{x}$ 为输入特征向量。当通过训练得到模型后，给出一个新的 $\boldsymbol{x}$，函数产生结果 $\hat{y}$。对于分类问题，$\hat{y}$ 取离散值并表示类型；对于回归问题，$\hat{y}$ 得到连续的输出值。对于这类确定性模型，决策是直接的，一般不需要进一步决策。

对于概率模型，训练过程中给出的模型是输出 y 的一种概率表示。有两类基本的概率模型：一类是生成模型，给出的是联合概率 $p(\boldsymbol{x},y)$；另一类是判别模型(注意与确定性判别函数是有区别的)，给出的是后验概率 $p(y|\boldsymbol{x})$。目前的概率模型中，判别模型应用更广泛。以判别模型为例，假设通过训练过程得到了后验概率表示式 $p(y|\boldsymbol{x})$，我们将首先针对二分类问题说明决策过程。设分别用 C_1 和 C_2 表示类型，则对于新的 $\boldsymbol{x}$，可计算 $p(y=C_1|\boldsymbol{x})$(简记为 $p(C_1|\boldsymbol{x})$)和 $p(y=C_2|\boldsymbol{x})$(简记为 $p(C_2|\boldsymbol{x})$)，由这些概率怎样确定输入 $\boldsymbol{x}$ 对应哪一类呢？这需要通过决策理论做出最后的判决。例如，$p(C_1|\boldsymbol{x})=0.6$，$p(C_2|\boldsymbol{x})=0.4$，是否一定会判决为类型 C_1 呢？

对于概率模型，怎样做出最后的决策呢？为了得出最后的结论，需要给出问题的评价函数，一般可以用风险函数作为评价函数，通过最小化风险函数的后验概率期望(即贝叶斯风险函数)获得判决准则，然后利用判决准则对模型输出的结果做出结论。由于主要使用后验

概率做出决策，并采用贝叶斯风险函数作为评价函数，故将所讨论的决策问题称为贝叶斯决策。分类和回归的决策方法和评价函数不同，将单独予以处理。

前文提到的生成模型主要是针对监督学习情况。实际上，生成模型是机器学习中一类重要的模型。更一般地，给出训练样本集$\{\boldsymbol{x}_n\}_{n=1}^N$，这里$\boldsymbol{x}_n$表示一般化的样本向量，它是通过对一个概率分布$p(\boldsymbol{x})$采样所获得的，但对我们来讲$p(\boldsymbol{x})$是未知的，通过样本集学习出$p(\boldsymbol{x})$，这是生成模型的一般含义。若$p(\boldsymbol{x})$可由高斯分布表示，则生成模型是简单的，但在机器学习中常遇到各种非常复杂的数据，其背后的$p(\boldsymbol{x})$同样极其复杂，这种情况下生成模型的学习具有很高的挑战性。稍后，3.3节针对高斯情况，第5章针对简单的离散分布（朴素贝叶斯），以分类为例给出简单的生成模型算法，第11章将介绍一种很有效的针对无监督情况的生成模型学习，即生成对抗网络。

3.2 分类的决策

假设学习阶段通过训练已得到模型的联合概率$p(\boldsymbol{x},y)$（对于生成模型）或后验概率$p(y|\boldsymbol{x})$（对于判别模型），需要对类型输出做出最终判决，即决策。

首先讨论二分类问题。以下使用联合概率导出结论，但实际上对于分类决策只需要后验概率。

在讨论的开始，首先假设特征输入$\boldsymbol{x}$和类型C的联合概率$p(\boldsymbol{x},C)$已知，由于是二分类问题，C只有C_1和C_2两个取值，故可以分别写出两种类型的联合概率值$p(\boldsymbol{x},C_1)$和$p(\boldsymbol{x},C_2)$。对于分类问题，一个最直接的评价函数是错误分类率，错误分类率等于两部分之和：$\boldsymbol{x}$属于C_1类却被分类为C_2和$\boldsymbol{x}$属于C_2却被分类为C_1。决策理论的目标是找到一个判决准则，使错误分类率最小，即最小错误分类率（Minimum Misclassification Rate，MMR）准则。

设输入特征向量$\boldsymbol{x}$是D维向量，其输入空间是D维向量空间的一个区域R，通过决策理论，可将区域R划分为两个不重叠区域R_1和R_2。当$\boldsymbol{x}\in R_1$时，判断类型输出为C_1；当$\boldsymbol{x}\in R_2$时，判断类型输出为C_2。划分区域的准则就是MMR。

为了便于理解，图3.2.1所示为$\boldsymbol{x}$为标量情况下的概率密度函数$p(x,C_1)$和$p(x,C_2)$。假如已经做出了区域划分R_1和R_2，那么当$\boldsymbol{x}\in R_1$但其真实是属于C_2，则对应一个错误的分类，其错误概率可表示为

$$p(\boldsymbol{x}\in R_1,C_2)=\int_{R_1}p(\boldsymbol{x},C_2)\,\mathrm{d}\boldsymbol{x}$$

反之，当$\boldsymbol{x}\in R_2$但真实是属于C_1类时，则对应一个错误分类，其错误概率为

$$p(\boldsymbol{x}\in R_2,C_1)=\int_{R_2}p(\boldsymbol{x},C_1)\,\mathrm{d}\boldsymbol{x}$$

将两者合并一起，总的错误分类率p_e为

$$\begin{aligned}p_\mathrm{e}&=p(\boldsymbol{x}\in R_1,C_2)+p(\boldsymbol{x}\in R_2,C_1)\\&=\int_{R_1}p(\boldsymbol{x},C_2)\,\mathrm{d}\boldsymbol{x}+\int_{R_2}p(\boldsymbol{x},C_1)\,\mathrm{d}\boldsymbol{x}\end{aligned}\tag{3.2.1}$$

以上假设已划分出R_1和R_2，从而写出了错误分类率公式。现在我们反过来，通过错误分类率公式，选择R_1和R_2使p_e最小。通过观察图3.2.1和式(3.2.1)发现，若想p_e最

小，只需这样选择 R_1 和 R_2：将满足 $p(\boldsymbol{x},C_1)>p(\boldsymbol{x},C_2)$ 的取值集合取为 R_1，反之取为 R_2，一般将 $p(\boldsymbol{x},C_1)=p(\boldsymbol{x},C_2)$ 的点任意分配给 R_1 或 R_2。

由此可得到判决准则，当给出一个新的 $\boldsymbol{x}$，若

$$p(\boldsymbol{x},C_1)>p(\boldsymbol{x},C_2) \tag{3.2.2}$$

则分类为 C_1，反之分类为 C_2。由概率公式 $p(\boldsymbol{x},C_i)=p(C_i|\boldsymbol{x})p(\boldsymbol{x})$，将式(3.2.2)表示为后验概率形式为，即若

$$p(C_1|\boldsymbol{x})>p(C_2|\boldsymbol{x}) \tag{3.2.3}$$

则分类为 C_1，否则分类为 C_2。应用 MMR 准则的决策公式为式(3.2.2)或式(3.2.3)。目前分类算法中，判别模型应用更多，故式(3.2.3)更常用。由于式(3.2.3)也表示了式(3.2.2)的含义，因此若非特殊需要，总是以式(3.2.3)表示决策公式。

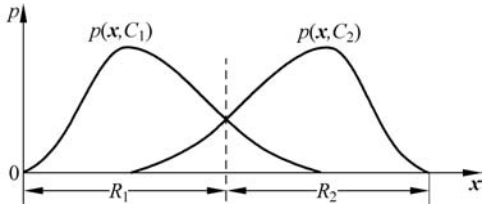


图 3.2.1　联合概率和判决

以上结果可直接推广到多类情况，设有 $\{C_1,C_2,\cdots,C_K\}$ 种类型，最后分类结果为 C_{i^*}，则

$$C_{i^*}=\arg\max_{C_i}\{p(C_i|\boldsymbol{x})\} \tag{3.2.4}$$

以上给出了在最小错误分类率准则下的判决准则，结果非常直观，即将后验概率最大的类作为分类输出。回到本章的开始，若一个机器学习模型是概率模型，对于新的 $\boldsymbol{x}$ 可分别计算分类为 C_i 的后验概率，则决策准则将后验概率最大的类作为最终类输出。

以上的基本决策原理的前提条件是假设所有错误的代价是平等的，这在很多实际应用中不符合现实，下面讨论两种更实际的判决方式。

3.2.1　加权错误率准则

在实际应用中，一些错误比另一些错误代价更大。例如，一辆无人驾驶汽车的刹车系统，为了方便说明，一个简化的模型输出只有两类：刹车或不刹车，这可看作分类问题。应刹车时判决为不刹车，比不应刹车时判决为刹车往往代价更大，所以要对刹车判决的不同错误定义不同的代价，如图 3.2.2 所示。

在图 3.2.2 中，刹车被错判为不刹车的代价是不刹车被错判为刹车的代价的 10 倍，这是一个主观的加权。对于实际问题，如刹车问题，可通过预先得到的大量交通事故数据按所关心的指标给出加权矩阵的统计值。对于更一般的多类型情况，将加权矩阵表示为 $\boldsymbol{L}$，矩阵的各元素表示为 $L_{kj}=L(C_j|C_k)$，即将 C_k 分类为 C_j 的代价加权值。考虑所有的 C_k 和 C_j 的组合，得到总期望损失为

	刹车	不刹车
刹车	0	10
不刹车	1	0

图 3.2.2　刹车决策的错误代价加权矩阵

$$E[\boldsymbol{L}]=\sum_k\sum_j L_{kj}\int_{R_j}p(\boldsymbol{x},C_k)\mathrm{d}\boldsymbol{x} \tag{3.2.5}$$

将式(3.2.5)重组为

$$E[\boldsymbol{L}]=\sum_j\int_{R_j}\sum_k L_{kj}p(\boldsymbol{x},C_k)\mathrm{d}\boldsymbol{x}=\sum_j\int_{R_j}\left[\sum_k L_{kj}p(C_k\mid\boldsymbol{x})\right]p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{3.2.6}$$

分类为 C_j 的风险定义为

$$R(C_j\mid\boldsymbol{x})=\sum_k L_{kj}p(C_k\mid\boldsymbol{x}) \tag{3.2.7}$$

可见为了使式(3.2.6)的结果最小,划分 R_j 的准则是将 $R(C_j\mid\boldsymbol{x})$最小的区间划分为 R_j。由于 C_j 表示所有可能的类,故分类为 C_{i^*} 的决策公式为

$$C_{i^*}=\arg\min_{C_j}\left\{R(C_j\mid\boldsymbol{x})=\sum_k L_{kj}p(C_k\mid\boldsymbol{x})\right\} \tag{3.2.8}$$

由于每个 $p(C_k\mid\boldsymbol{x})$在学习过程都已经训练得到,L_{kj} 是预先确定的,式(3.2.8)的决策是简单的加权求和与比较运算。

例 3.2.1 讨论式(3.2.8)在二分类情况下的特殊形式。只有两类时,式(3.2.7)的风险值只有两个,即

$$\begin{cases}R(C_1\mid\boldsymbol{x})=L_{11}p(C_1\mid\boldsymbol{x})+L_{21}p(C_2\mid\boldsymbol{x})\\R(C_2\mid\boldsymbol{x})=L_{12}p(C_1\mid\boldsymbol{x})+L_{22}p(C_2\mid\boldsymbol{x})\end{cases} \tag{3.2.9}$$

由式(3.2.8),若要分类结果为 C_1,则只需 $R(C_1\mid\boldsymbol{x})<R(C_2\mid\boldsymbol{x})$,将(3.2.9)各式代入并整理得

$$(L_{12}-L_{11})p(C_1\mid\boldsymbol{x})>(L_{21}-L_{22})p(C_2\mid\boldsymbol{x}) \tag{3.2.10}$$

(1) 情况 1。取 $L_{12}=L_{21}=1$,$L_{22}=L_{11}=0$,则式(3.2.10)简化为 $p(C_1\mid\boldsymbol{x})>p(C_2\mid\boldsymbol{x})$,即在各种错误等代价的二分类问题时,式(3.2.8)与式(3.2.3)等价。

(2) 情况 2。若取 $L_{12}=10$,$L_{21}=1$,$L_{22}=L_{11}=0$,则式(3.2.10)简化为 $p(C_1\mid\boldsymbol{x})>0.1p(C_2\mid\boldsymbol{x})$,即可判断为 C_1,这里的加权用的是图 3.2.2 的有关刹车的加权矩阵,可见在该损失加权的条件下,$p(C_1\mid\boldsymbol{x})=0.1$ 就可以决策为刹车。

由贝叶斯公式,可将式(3.2.10)写为

$$(L_{12}-L_{11})p(\boldsymbol{x}\mid C_1)p(C_1)>(L_{21}-L_{22})p(\boldsymbol{x}\mid C_2)p(C_2) \tag{3.2.11}$$

整理得到分类为 C_1 的条件为

$$\frac{p(\boldsymbol{x}\mid C_1)}{p(\boldsymbol{x}\mid C_2)}>\frac{(L_{21}-L_{22})}{(L_{12}-L_{11})}\frac{p(C_2)}{p(C_1)} \tag{3.2.12}$$

式(3.2.12)中利用了类条件概率(密度)$p(\boldsymbol{x}\mid C_i)$和类先验概率 $p(C_i)$,称为似然比准则。

3.2.2 拒绝判决

在各种误分类代价相等的情况下,在二分类时只要满足式(3.2.3)即可分为类型 C_1,如 $p(C_1\mid\boldsymbol{x})=0.51$ 即可分类为 C_1。当两类的后验概率很接近时,分类结果可信度不高,错误分类率也较大,在一些需要高可靠分类的应用中,这种分类结果显然无法接受,故在很多情况下,可能对一定的后验概率范围拒绝做出判决。如图 3.2.3 所示,在 $p(\boldsymbol{x}\mid C_i)$均小于一个预定的阈值 θ(如 $\theta=0.9$)时拒绝做出判决。对于多分类问题,只有至少有一个 $p(\boldsymbol{x}\mid C_i)$

$\geqslant\theta$ 时，才利用式(3.2.4)做判决，否则拒绝判决。

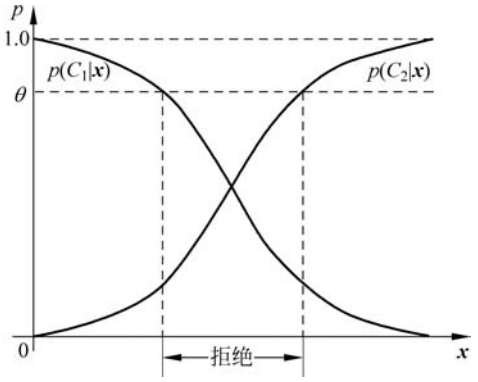


图 3.2.3　拒绝判决

拒绝判决是一个有意义但需要谨慎使用的原则，其使用与所面对的问题的代价分析有关。例如，在一个邮件自动分拣的邮政编码识别系统中，假设一封信的第 1 位邮政编码数字被自动分类为 1 和 7 的概率最大但很接近，该信可以由自动分拣系统拒绝判决，转为人工服务，显然，人工服务的成本比自动分拣高，但远低于一封信被错误投递的代价。

拒绝判决可降低错误分类率，极端的例子是，拒绝做任何判决则错误分类率为 0，但这样的系统毫无意义。故选择拒绝判决以及拒绝判决的阈值与应用是密切相关的，需要在实际系统设计中谨慎选择。

3.3　回归的决策

对于回归问题，本书介绍的回归模型较多的是直接得到回归函数 $\hat{y}=g(\boldsymbol{x})$，也有一些方法是首先通过学习过程得到联合概率 $p(\boldsymbol{x},y)$或后验概率 $p(y|\boldsymbol{x})$，对这种模型首先需要选择一种评价性能的函数，通过决策给出回归的连续输出值 $\hat{y}$。在回归情况下最常用的评价函数之一是均方误差(Mean Square Error，MSE)。回归输出 $\hat{y}$ 与真实 y 的均方误差定义为

$$\mathrm{mse}(\hat{y})=\iint(y-\hat{y})^2p(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y \tag{3.3.1}$$

若要求一个 $\hat{y}$ 使均方误差最小，可令将(3.3.1)两侧对 $\hat{y}$ 求导且令导数为 0，将得到一个解为 $\hat{y}=g(\boldsymbol{x})$。

利用贝叶斯公式，有

$$\mathrm{mse}(\hat{y})=\iint(y-\hat{y})^2p(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y=\int\left[\int(y-\hat{y})^2p(y\mid\boldsymbol{x})\mathrm{d}y\right]p_x(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{3.3.2}$$

将上述等式对 $\hat{y}$ 求导，并交换积分和求导顺序，得

$$\frac{\partial\mathrm{mse}(\hat{y})}{\partial\hat{y}}=\int\left[\frac{\partial}{\partial\hat{y}}\int(\hat{y}-y)^2p(y\mid\boldsymbol{x})\mathrm{d}y\right]p_x(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{3.3.3}$$

为求最小均方估计 $\hat{y}$，只需令式(3.3.3)为 0，因为对所有 $\boldsymbol{x}$，$p_x(\boldsymbol{x})\geqslant0$，故欲使$\dfrac{\partial\mathrm{mse}(\hat{y})}{\partial\hat{y}}=0$，只需令

$$\frac{\partial}{\partial \hat{y}}\int (y-\hat{y})^2 p(y \mid \boldsymbol{x})\mathrm{d}y = 0 \tag{3.3.4}$$

将式(3.3.4)中求导和积分次序交换,得

$$\frac{\partial}{\partial \hat{y}}\int (y-\hat{y})^2 p(y \mid \boldsymbol{x})\mathrm{d}y = -2\int (y-\hat{y}) p(y \mid \boldsymbol{x})\mathrm{d}y = 0$$

得到

$$\hat{y} = \int y p(y \mid \boldsymbol{x})\mathrm{d}y = E_{y|x}(y \mid \boldsymbol{x}) \tag{3.3.5}$$

这是最小均方误差(Minimum Mean Square Error,MMSE)意义下回归的最优输出值,称为后验期望输出。在参数估计问题中,若用参数 θ 代替回归输出 y,则同样的结论称为 MMSE 贝叶斯参数估计器。

对于一个回归学习系统,通过学习过程得到了后验概率 $p(y|\boldsymbol{x})$,则给出一个新的特征向量输入 $\boldsymbol{x}$ 后,回归的输出是 y 的后验条件期望值 $\hat{y} = E_{y|\boldsymbol{x}}(y|\boldsymbol{x})$。将回归输出 $\hat{y}$ 代入式(3.3.1),得到最小均方误差为

$$\mathrm{mmse}(\hat{\theta}) = \iint (y - E(y|\boldsymbol{x}))^2 p(\boldsymbol{x}, y)\mathrm{d}\boldsymbol{x}\mathrm{d}y$$

例 3.3.1 对于回归问题,仍以高斯分布为例。若有一个回归问题,则通过学习过程得到的后验概率为

$$p(y \mid \boldsymbol{x}) = N(y \mid \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}, \sigma^2_{y|\boldsymbol{x}})$$

其中,$\boldsymbol{w}$ 为通过训练得到的权系数向量。由高斯分布的特点和式(3.3.5)可得,使 MSE 最小的回归输出为

$$\hat{y} = E_{y|\boldsymbol{x}}(y \mid \boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$$

方差 $\sigma^2_{y|\boldsymbol{x}}$ 刻画了回归输出的不确定性大小。

本节注释 1 介绍决策理论后,可对机器学习的框架再做一个概要讨论。通过机器学习解决一个实际问题,大致分为 3 个步骤:①针对要解决的问题收集数据,预处理数据(数据清洗、标注等),确定解决问题的算法模型,如选择采用监督学习、选择神经网络模型、支持向量机模型或其他模型;②训练过程,用样本集对模型进行训练,选择模型规模和参数,对确定性模型得到 $\hat{y} = f(\boldsymbol{x})$ 的判别函数,对概率模型得到联合概率 $p(\boldsymbol{x}, y)$ 或后验概率 $p(y|\boldsymbol{x})$;③推断或预测过程,给出新的特征输入 $\boldsymbol{x}$,对确定性模型直接得到结果,对概率模型计算得到后验概率 $p(y|\boldsymbol{x})$,通过后验概率和风险函数获得判决准则做出决策。对于复杂问题,以上 3 个步骤也可能要反复,直至得到需要的结果。决策理论是机器学习过程中最后一步的组成部分,总的来讲是比较容易的一部分,本章给出了决策理论的一个概要介绍,后续章节直接应用这些结果,一般来讲,若不与具体应用环境结合,就采用最简单的决策公式(见式(3.2.4))。

本节注释 2 在许多实际应用中,对于决策的目标函数设计,有一些特定的针对性指标,这些内容与应用密切关联。比较广泛的应用包括自然语言处理、计算机视觉、推荐系统、信息检索等,也有很多更专门的应用,如医学诊断、DNA 分析、雷达目标分类、通信信道建模、震动检测等。本书作为机器学习的基本教程,不讨论与这些专门应用对应的一些特殊指标,有兴趣的读者可参考这些领域的专门文献。

3.4　高斯情况下的分类决策

由于对机器学习需求的广泛性，所面对的数据类型也是非常广泛的，既有物理传感器采集的物理信号，也有社会调查所获得的不同人群的各类数据，或是电商平台记录的用户购物数据。面对如此广泛的数据类型，没有一个学习模型是通用的，这也是“没有免费的午餐”定理所阐述的原则。因此，本书会介绍多种不同的学习模型，以适应不同的应用需求。大多数的学习模型表现出不同的复杂性。但在数据的概率分布服从高斯分布时，问题往往变得简单且具有有效的闭式解。

本节以高斯情况为例，进一步理解决策理论在分类的应用，并导出最基本的学习模型。假设所要分类的数据集中包含 K 种类型 $\{C_i\}_{i=1}^{K}$，类 C_i 的出现概率为 $p(C_i)$。高斯情况是指在类型确定为 C_i 时，类条件概率 $p_{\boldsymbol{x}}(\boldsymbol{x}\mid y=C_i)$（简写为 $p_{\boldsymbol{x}}(\boldsymbol{x}\mid C_i)$）服从高斯分布，即

$$p(\boldsymbol{x}\mid C_i)=\frac{1}{(2\pi)^{M/2}\mid\boldsymbol{\Sigma}_i\mid^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right) \tag{3.4.1}$$

其中，$\{\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\}_{i=1}^{K}$ 表示各类条件概率的参数。若假设式(3.4.1)中的各参数均已知，给出一个新的特征向量 $\boldsymbol{x}$，利用 3.2 节的决策准则，给出 $\boldsymbol{x}$ 所对应的类型。

决策准则式（见式(3.2.4)）需要用后验概率进行判决，由于 $p(\boldsymbol{x})$ 不影响结果，则

$$p(C_i\mid\boldsymbol{x})\propto p(\boldsymbol{x}\mid C_i)p(C_i) \tag{3.4.2}$$

对于该问题，为了导出更加直接的决策准则，首先定义对应后验概率式(3.4.2)的判别函数 $g_i(\boldsymbol{x})$ 为

$$\begin{aligned}g_i(\boldsymbol{x})&=\ln[p(\boldsymbol{x}\mid C_i)p(C_i)]\\&=-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)-\frac{1}{2}\ln\mid\boldsymbol{\Sigma}_i\mid-\frac{M}{2}\ln 2\pi+\ln(p(C_i))\end{aligned} \tag{3.4.3}$$

由式(3.4.3)可导出更直接的分类判决准则。下面首先讨论二分类问题，分别讨论 $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\boldsymbol{\Sigma}$ 和 $\boldsymbol{\Sigma}_1\neq\boldsymbol{\Sigma}_2$ 两种情况，然后推广到多分类情况。

3.4.1　相同协方差矩阵情况的二分类

在二分类问题中，只有 C_1 和 C_2 两种类型，假设 $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\boldsymbol{\Sigma}$，由 $\boldsymbol{\mu}_1$ 和 $\boldsymbol{\mu}_2$ 区分两类。注意到当利用式(3.2.3)进行判决时，若 $p(C_1\mid\boldsymbol{x})>p(C_2\mid\boldsymbol{x})$，则判决为类型 C_1，这等价于 $g_1(\boldsymbol{x})>g_2(\boldsymbol{x})$。由于只需比较 $g_i(\boldsymbol{x})$ 的大小，故将式(3.4.3)中各 $g_i(\boldsymbol{x})$ 的相同项丢弃，重写简化的 $g_i(\boldsymbol{x})$ 为

$$g_i(\boldsymbol{x})=-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)+\ln(p(C_i)),\quad i=1,2 \tag{3.4.4}$$

注意到，式(3.4.4)等号右侧第一项展开后，二次项 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$ 与类型无关，也可删去，这样 $g_i(\boldsymbol{x})$ 进一步简化为

$$g_i(\boldsymbol{x})=\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x}+w_{i0} \tag{3.4.5}$$

其中，系数和偏置分别为

$$\begin{cases}\boldsymbol{w}_i=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\\ w_{i0}=-\dfrac{1}{2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i+\ln[p(C_i)],\quad i=1,2\end{cases} \tag{3.4.6}$$

若分类输出判决为 C_1，则需要

$$g_1(\boldsymbol{x}) > g_2(\boldsymbol{x}) \tag{3.4.7}$$

将式(3.4.6)代入式(3.4.7)加以整理，得判决为 C_1 的条件为

$$g(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0 > 0 \tag{3.4.8}$$

式(3.4.8)的系数为

$$\begin{cases} \boldsymbol{w} = \boldsymbol{w}_1 - \boldsymbol{w}_2 = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ w_0 = w_{10} - w_{20} = -\dfrac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \dfrac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\left[\dfrac{p(C_1)}{p(C_2)}\right] \end{cases} \tag{3.4.9}$$

对于高斯分布，若已知 $\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$ 和 $\boldsymbol{\Sigma}$，对于一个新的特征输入 $\boldsymbol{x}$，进行式(3.4.8)的判决，若成立，则输出类型 C_1，否则输出类型 C_2。

在机器学习的应用中，一般并不知道 $\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$，$p(C_1)$，$p(C_2)$ 和 $\boldsymbol{\Sigma}$ 这些参数，而是存在一组训练样本

$$\boldsymbol{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N} \tag{3.4.10}$$

设训练样本是 IID 的，下面通过训练样本估计这些参数，这是例 2.3.4 的一个推广。在例 2.3.4 中没有类型标注 y_n。为了估计上述参数，需要表示联合分布 $p(\boldsymbol{x}_n, y_n)$，然后通过 MLE 估计参数。这里，y_n 作为标注，$y_n = 1$ 表示类型 C_1，$y_n = 0$ 表示类型 C_2，为表示简洁，类型概率表示为

$$p(C_1) = p(y_n = 1) = \pi, \quad p(C_2) = p(y_n = 0) = 1 - \pi \tag{3.4.11}$$

则

$$\begin{aligned} p(\boldsymbol{x}_n, y_n = 1) &= p(\boldsymbol{x}_n, C_1) = p(\boldsymbol{x}_n \mid C_1)\, p(C_1) \\ &= \pi N(\boldsymbol{x}_n \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \end{aligned} \tag{3.4.12}$$

$$\begin{aligned} p(\boldsymbol{x}_n, y_n = 0) &= p(\boldsymbol{x}_n, C_2) = p(\boldsymbol{x}_n \mid C_2)\, p(C_2) \\ &= (1 - \pi) N(\boldsymbol{x}_n \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \end{aligned} \tag{3.4.13}$$

y_n 只有两个取值，相当于伯努利分布，故联合分布 $p(\boldsymbol{x}_n, y_n)$ 为

$$p(\boldsymbol{x}_n, y_n \mid \pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = [\pi N(\boldsymbol{x}_n \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{y_n} [(1 - \pi) N(\boldsymbol{x}_n \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1 - y_n} \tag{3.4.14}$$

考虑样本是 I. I. D 的，则对数似然函数为

$$\begin{aligned} &\ln[p(\boldsymbol{X}, y \mid \pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})] \\ &= \ln \prod_{n=1}^{N} [\pi N(\boldsymbol{x}_n \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{y_n} [(1 - \pi) N(\boldsymbol{x}_n \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1 - y_n} \\ &= \sum_{n=1}^{N} \{ y_n \ln[\pi N(\boldsymbol{x}_n \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma})] + (1 - y_n) \ln[(1 - \pi) N(\boldsymbol{x}_n \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma})] \} \end{aligned} \tag{3.4.15}$$

式(3.4.15)对各参数求偏导数并令其为 0，分别得到各参数的估计值为

$$\hat{\pi} = \frac{1}{N}\sum_{n=1}^{N} y_n = \frac{N_1}{N} \tag{3.4.16}$$

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{N_1}\sum_{n=1}^{N} y_n \boldsymbol{x}_n, \quad \hat{\boldsymbol{\mu}}_2 = \frac{1}{N_2}\sum_{n=1}^{N} (1 - y_n)\boldsymbol{x}_n \tag{3.4.17}$$

$$\hat{\boldsymbol{\Sigma}}=\frac{1}{N}\Big[\sum_{n=1}^{N}y_n(\boldsymbol{x}_n-\hat{\boldsymbol{\mu}}_1)(\boldsymbol{x}_n-\hat{\boldsymbol{\mu}}_1)^{\mathrm{T}}+\sum_{n=1}^{N}(1-y_n)(\boldsymbol{x}_n-\hat{\boldsymbol{\mu}}_2)(\boldsymbol{x}_n-\hat{\boldsymbol{\mu}}_2)^{\mathrm{T}}\Big] \tag{3.4.18}$$

其中，N_1 为样本集中属于 C_1 的样本数目；N_2 为样本集中属于 C_2 的样本数目。

将估计的参数代入式(3.4.9)计算 $\boldsymbol{w}$ 和 w_0，则式(3.4.8)的判决方程就确定了，给出新的特征输入，就可以做出分类决策。注意到在等协方差矩阵的高斯情况下，判决方程是一个线性函数，令

$$g(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+w_0=0 \tag{3.4.19}$$

得到一个 $\boldsymbol{x}$ 空间的超平面，该平面将空间分划成两个区域，$g(\boldsymbol{x})>0$ 的区域属于类型 C_1，$g(\boldsymbol{x})<0$ 的区域属于 C_2，位于超平面上的点可任意判决为 C_1 或 C_2。在这种情况下，依据式(3.2.3)得到的判决方程式(3.4.8)已经退化成一个确定性的线性判决函数，由其可进行分类判决，判决函数使用起来更简单直接，但已失去后验概率所具有的丰富内涵。而后验概率自身有着丰富的内涵，可以进行拒绝判决，可以结合加权损失，也可以通过概率原理集成多个分类器。

在高斯情况下，由以上已得结果可以导出后验概率 $p(C_i|\boldsymbol{x})$。先看 $p(C_1|\boldsymbol{x})$，由贝叶斯公式

$$p(C_1\mid\boldsymbol{x})=\frac{p(\boldsymbol{x},C_1)}{p(\boldsymbol{x})}=\frac{p(\boldsymbol{x}|C_1)p(C_1)}{p(\boldsymbol{x}|C_1)p(C_1)+p(\boldsymbol{x}|C_2)p(C_2)} \tag{3.4.20}$$

用一点数学技巧，得

$$p(C_1\mid\boldsymbol{x})=\frac{1}{1+\dfrac{p(\boldsymbol{x}|C_2)p(C_2)}{p(\boldsymbol{x}|C_1)p(C_1)}}=\frac{1}{1+\mathrm{e}^{-a(\boldsymbol{x})}}=\sigma[a(\boldsymbol{x})] \tag{3.4.21}$$

这里使用了

$$a(\boldsymbol{x})=\ln\frac{p(\boldsymbol{x}|C_1)p(C_1)}{p(\boldsymbol{x}|C_2)p(C_2)} \tag{3.4.22}$$

并且使用了一个函数定义，即

$$\sigma(a)=\frac{1}{1+\mathrm{e}^{-a}} \tag{3.4.23}$$

该函数称为 Sigmoid 函数，其详细的讨论和性质将在第 5 章给出，它是机器学习中广泛使用的函数，这里暂不做深入讨论，只需要注意到 $\sigma(0)=0.5$，$a>0$ 时，$\sigma(a)>0.5$。$p(C_2|\boldsymbol{x})$ 可以表示为

$$p(C_2\mid\boldsymbol{x})=1-p(C_1\mid\boldsymbol{x}) \tag{3.4.24}$$

式(3.4.21)给出了后验概率的表达式，其中参数 $a(\boldsymbol{x})$ 由式(3.4.22)计算，不难验证，将 $p(\boldsymbol{x}|C_i)$，$p(C_i)$ 代入式(3.4.22)整理得(推导细节留作习题)

$$a(\boldsymbol{x})=g(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+w_0 \tag{3.4.25}$$

式(3.4.25)的系数 $\boldsymbol{w}^{\mathrm{T}}$ 和 w_0 在式(3.4.9)已求得。注意到，由式(3.4.21)可见，当 $a(\boldsymbol{x})>0$ 时，$p(C_1|\boldsymbol{x})>0.5$，分类判决为 C_1，这与式(3.4.8)结果一致。在更多应用中，后验概率比式(3.4.8)的判别方程内涵更丰富。

3.4.2 不同协方差矩阵情况的二分类

在$\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$的条件下，除数学表达上略复杂一些，过程与相同协方差矩阵情况相似。略去与比较大小无关的项，$g_i(\boldsymbol{x})$表示为

$$g_i(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}_i\boldsymbol{x} + \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x} + w_{i0}, \quad i = 1,2 \tag{3.4.26}$$

其中，系数和偏置分别为

$$\begin{cases} \boldsymbol{W}_i = -\dfrac{1}{2}\boldsymbol{\Sigma}_i^{-1} \\ \boldsymbol{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i \\ w_{i0} = -\dfrac{1}{2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \dfrac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln[p(C_i)] \end{cases} \tag{3.4.27}$$

对于输入$\boldsymbol{x}$，若分类输出判决为C_1，则需要

$$g(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{x} + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0 > 0 \tag{3.4.28}$$

其中，权系数为

$$\begin{cases} \boldsymbol{W} = -\dfrac{1}{2}(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}) \\ \boldsymbol{w} = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 \\ w_0 = -\dfrac{1}{2}\boldsymbol{\mu}_1{}^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \dfrac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \dfrac{1}{2}\ln\dfrac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} + \ln\dfrac{p(C_1)}{p(C_2)} \end{cases} \tag{3.4.29}$$

在$\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$的条件下，对参数$\boldsymbol{\mu}_i$和$\boldsymbol{\Sigma}_i$的估计更简单，按照$y_n$取1或0将样本集分为$\boldsymbol{D}_1$和$\boldsymbol{D}_2$，直接利用例2.3.4的结果，用各子样本集$\boldsymbol{D}_i$直接估计$\boldsymbol{\mu}_i$和$\boldsymbol{\Sigma}_i$，$\pi$的估计仍用式(3.4.16)。容易验证，只需要用式(3.4.28)的$g(\boldsymbol{x})$替代式(3.4.8)的$g(\boldsymbol{x})$，则后验概率$p(C_1|\boldsymbol{x})$的表达式不变，仍如式(3.4.21)和式(3.4.25)所示。

3.4.3 多分类情况

可将二分类的结果直接推广到有K个类型的多分类问题，对于判别函数$g_i(\boldsymbol{x})$，只需将式(3.4.26)中的$i=1,2$扩展为$i=1,2,\cdots,K$，并取$g_i(\boldsymbol{x})$最大的类为输出类型。对于后验概率$p(C_i|\boldsymbol{x})$，可得到结果(推导细节留作习题)

$$p(C_i \mid \boldsymbol{x}) = \frac{\exp[g_i(\boldsymbol{x})]}{\sum_{k=1}^{K}\exp[g_k(\boldsymbol{x})]} \tag{3.4.30}$$

以上已对高斯情况的分类问题做了较详细的讨论，利用了决策理论的结果。可以看到在满足式(3.4.1)假设的情况下，对于给出如式(3.4.10)所示的样本集，可估计式(3.4.1)的所有参数，也可估计$p(C_i)$的概率。如果以基本的最小错误分类率准则进行分类，在高斯情况下，可得到简单的判决函数$g_i(\boldsymbol{x})$；若需要更丰富信息的后验概率$p(C_i|\boldsymbol{x})$，也得到了后验概率的解析公式。

对于高斯情况，在给出式(3.4.10)的训练样本集后，可以估计出式(3.4.14)的联合概率密度函数$p(\boldsymbol{x},y)$和各种情况下的类后验概率$p(C_i|\boldsymbol{x})$，获得所谓的生成模型和判决模型都不困难。但对于其他复杂概率分布，一般生成模型的学习是很困难的。

若高斯假设与实际数据相符合，则对于分类问题，本节给出的结果是性能良好的，对于复杂的实际情况，高斯假设只有一定的符合度或符合度较差，这种情况下，通过本节的方法很难得到满意的较低的错误分类率，需要探索更多类型的分类算法。本书第 5 章介绍了一些基本的分类算法，后续章节给出了各种更专门的分类算法，如支持向量机、神经网络、决策树和集成学习算法。

3.5　KNN 方法

第 1 章中以 KNN 分类器为例对分类器功能进行了说明，利用 3.2 节的决策理论可进一步解释 KNN 分类器的原理。

设训练样本集 $\boldsymbol{D}=\{(\boldsymbol{x}_n,y_n)\}_{n=1}^{N}$ 对应 $C_1,C_2,\cdots,C_J$ 共 J 种类型，总样本数为 N，各类型对应的样本数分别为 $N_1,N_2,\cdots,N_J$。对于一个给定的 $\boldsymbol{x}$，以其为中心的超球体内包括 K 个样本，体积为 V。设近邻的 K 个样本中，标注为各类型的样本数分别为 $K_1,K_2\cdots,K_J$，利用这些数据进行概率估计，显然有

$$\hat{p}(\boldsymbol{x})=\frac{K}{NV},\quad \hat{p}(\boldsymbol{x}\mid C_j)=\frac{K_j}{N_jV},\quad \hat{p}(C_j)=\frac{N_j}{N} \tag{3.5.1}$$

故后验概率为

$$\hat{p}(C_j\mid\boldsymbol{x})=\frac{\hat{p}(\boldsymbol{x}\mid C_j)\hat{p}(C_j)}{\hat{p}(\boldsymbol{x})}=\frac{K_j}{K} \tag{3.5.2}$$

按照 3.2 节分类错误率最小的决策准则(见式(3.2.4))，如果一个类型 C_{j^*} 的后验概率最大，则输出为 C_{j^*} 类。在 KNN 算法中，由式(3.5.2)可见，$\hat{p}(C_{j^*}\mid\boldsymbol{x})$最大对应近邻样本数 K_{j^*} 最多，则分类为 C_{j^*}。可见 KNN 分类器是一种建立在后验概率最大化基础上的分类器，只是后验概率的估计采用了 KNN 概率估计。

当 $K=1$ 时，称为最近邻分类器，即将待分类的输入特征向量分类为距离最近的训练样本的类型。可以证明，用简单的最近邻分类器，当 $N\to\infty$时，分类误差不大于最优分类误差的 2 倍；若取较大的 K，分类误差进一步降低。尽管简单，但在一些应用中 KNN 分类器可以获得可接受的分类效果。

KNN 方法同样可以用于回归估计，若样本集 $\boldsymbol{D}=\{(\boldsymbol{x}_n,y_n)\}_{n=1}^{N}$ 的标注是实数，对于一个给定的 $\boldsymbol{x}$，得到以其为中心的超球体，其内包括 K 个样本，记 $\boldsymbol{x}$ 的 K 个近邻样本集合为 $\boldsymbol{D}_K(\boldsymbol{x})$，则 KNN 回归输出为

$$\hat{y}(\boldsymbol{x})=\frac{1}{K}\sum_{(\boldsymbol{x}_i,y_i)\in\boldsymbol{D}_K(\boldsymbol{x})}y_i \tag{3.5.3}$$

显然，这个输出近似等于 $E(y\mid\boldsymbol{x})$，是利用 K 个近邻样本的标注值对 $E(y\mid\boldsymbol{x})$的估计，$E(y\mid\boldsymbol{x})$为式(3.3.2)给出的回归问题的最优决策值。

*3.6　概率图模型概述

机器学习中常使用概率模型描述数据，前面也看到可通过后验概率进行决策，总之，概率模型是机器学习中常用的工具。在机器学习中，常遇到高维随机向量 $\boldsymbol{x}$，当维数很高时，$\boldsymbol{x}$

的概率函数非常复杂，学习和推断过程也变得非常复杂，在这种情况下，若能将概率函数分解为多因子乘积的形式，将可能降低问题的复杂度。本节讨论用图的方式表示概率函数，称为概率图模型。概率图模型已经发展为一个相对独立且深入广博的子领域，限于本书的篇幅限制，本节仅给出其非常简略的介绍，对于希望更加深入了解概率图模型的读者，可参考在本章小结中列出的参考书。

本节简要介绍贝叶斯网络（或称为有向图模型）、无向图模型，及其学习和推断的原理。

3.6.1 贝叶斯网络

图是由节点和边组成的，在有向图中边是带有方向的。贝叶斯网络是一个有向无环图（Directed Acyclic Graphs，DAG），其中每个节点代表一个随机变量（更一般地，一个节点可代表一组随机变量，目前为了叙述简单，首先考虑一个节点仅代表一个随机变量的情况），节点之间的连线表示两个随机变量之间有直接关系。

观察一个最简单的图，如图 3.6.1 所示，只有两个节点 a 和 b，分别表示两个同名的随机变量 a 和 b，箭头从节点 a 指向节点 b，表示节点 a 是节点 b 的父节点。节点 a 没有父节点，可以用 $p(a)$ 的因子表示，而节点 b 有父节点 a，可由条件概率 $p(b|a)$ 表示该关系，则该图所表示的联合概率可表示为这两个因子的积，即

图 3.6.1 两个节点的贝叶斯网络

$$p(a,b)=p(a)p(b\mid a) \tag{3.6.1}$$

式(3.6.1)正是联合概率的积公式。图 3.6.1 和式(3.6.1)都过于简单，难以说明更多问题。

2.1 节给出了随机变量集 $\{x_1,\cdots,x_{D-1},x_D\}$ 的链式分解公式，重写如下。

$$p(x_1,\cdots,x_{D-1},x_D)=p(x_D\mid x_{D-1},\cdots,x_1)\cdots p(x_2\mid x_1)p(x_1) \tag{3.6.2}$$

为了表述简单和清楚，给出 4 个随机变量的例子，即

$$p(x_1,x_2,x_3,x_4)=p(x_4\mid x_3,x_2,x_1)p(x_3\mid x_2,x_1)p(x_2\mid x_1)p(x_1) \tag{3.6.3}$$

可用图 3.6.2(a)的贝叶斯网络表示式(3.6.3)的因子分解形式。其中，节点 x_4 有 3 个父节点 x_3,x_2,x_1，故表示为条件概率 $p(x_4|x_3,x_2,x_1)$；节点 x_3 有两个父节点 x_2,x_1，表示为条件概率 $p(x_3|x_2,x_1)$；节点 x_2 只有一个父节点 x_1，表示为条件概率 $p(x_2|x_1)$；x_1 节点没有父节点，故只表示为 $p(x_1)$，将所有这些因子相乘得到式(3.6.3)的链式分解。图 3.6.2(a)是一个有全连接的图，即每两个节点之间都有连线（仅从小序号节点指向大序号节点）。对于更多的节点，这种全连接图对应式(3.6.2)的链式分解，即全连接图没有给出关于概率结构的特殊知识。

如果一种节点间连接的贝叶斯网络如图 3.6.2(b)所示，即图中一些节点之间没有连线，说明这两个节点之间没有直接关系。x_1 和 x_2 节点均没有父节点，故各自的因子为 $p(x_1)$和 $p(x_2)$，x_3 节点有两个父节点 x_2,x_1，对应因子 $p(x_3|x_2,x_1)$，x_4 节点只有一个父节点 x_3，对应因子 $p(x_4|x_3)$，故图 3.6.2(b)所表示的贝叶斯网络的联合概率函数为

$$p(x_1,x_2,x_3,x_4)=p(x_4\mid x_3)p(x_3\mid x_2,x_1)p(x_2)p(x_1) \tag{3.6.4}$$

比较式(3.6.4)和式(3.6.3)，相当于图 3.6.2(b)表示的概率结构将 $p(x_4|x_3,x_2,x_1)$ 简化为 $p(x_4|x_3)$，将 $p(x_2|x_1)$简化为 $p(x_2)$。

为了更清楚地理解图 3.6.2(a)和图 3.6.2(b)所表示的概率结构所带来的不同复杂度，

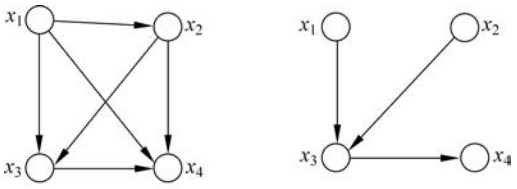


图 3.6.2　4 个节点的贝叶斯网络实例

以 $p(x_1,x_2,x_3,x_4)$ 为例进行说明。设 x_i 是离散的且只取两个值 1 或 0，为了表示 $\{x_1,x_2,x_3,x_4\}$ 的联合概率，需要 2^4-1 即 15 个参数(由于概率和为 1，故减少 1 个参数)。若用图 3.6.2(a)的有向图表示，对应式(3.6.3)的因子分解，在因子 $p(x_4|x_3,x_2,x_1)$ 中，由于 x_3,x_2,x_1 有 8 种组合，故描述 $p(x_4=1|x_3,x_2,x_1)$ 需要 8 个参数(不需要用新参数描述 $p(x_4=0|x_3,x_2,x_1)$)，其他因子分别需要 4、2、1 个参数，共需要 15 个参数，全连接图模型与直接的联合概率表示对比没有减少参数。但对于图 3.6.2(b)的模型，其对应的分解式(3.4.6)右侧中各因子分别需要 2、4、1、1 个参数，故描述其联合概率只需 8 个参数，需要的参数量明显下降。这里只是以 $D=4$ 作为维数做了简单说明，在机器学习中，D 表示的维数可能很大，如大于 1000 甚至更高，则利用概率图描述的结构，可能大大减少需要估计的模型参数。

对于一般的 D 个随机变量，记为 $\boldsymbol{x}=[x_1,\cdots,x_{D-1},x_D]^{\mathrm{T}}$，其可以用具有 D 个节点的有向图表示，相应地，通过图可以将联合概率分解为

$$p(\boldsymbol{x})=\prod_{k=1}^{D}p(x_k|\boldsymbol{x}_{fk}) \tag{3.6.5}$$

其中，对于节点 x_k，设其父节点集合为 $\boldsymbol{x}_{fk}$，则每个因子写为条件概率 $p(x_k|\boldsymbol{x}_{fk})$。

例如，图 3.6.3 是一个具有 D 个节点的有向图模型，图中只有序号相邻的随机变量之间有连线，显然，该图表示的联合概率为

$$p(\boldsymbol{x})=p(x_1)\prod_{k=2}^{D}p(x_k|x_{k-1}) \tag{3.6.6}$$

仍假设 x_k 是仅取两个值的离散随机变量，描述一般的 $p(\boldsymbol{x})$ 需要 2^D-1 个参数，若 $D=1000$ 则参数多到无法存储，但式(3.6.6)或图 3.6.3 所表示的概率结构仅需要 $2D-1$ 个参数，而这种只有近邻之间有直接相关性的情况，可代表许多数据类型。

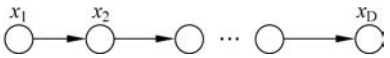
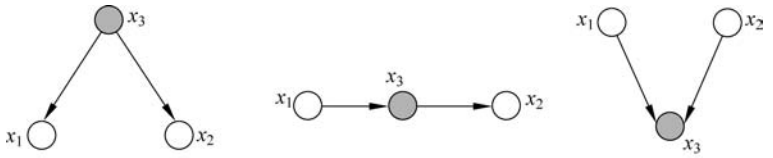


图 3.6.3　一个只有相邻节点有连线的图结构

除了贝叶斯网络刻画的概率结构可能大大降低表示联合概率的复杂性外，一个有趣的应用是，可从图中直接判断两组随机变量子集之间是否在一个条件子集下是独立的。为此，我们首先考查 3 种基本结构。如图 3.6.4 所示，有 3 个随机变量，以及 3 种不同的连接方式。我们判断在给出 x_3 时，x_1 和 x_2 是否条件独立，即

$$p(x_1,x_2\mid x_3)\triangleq p(x_1\mid x_3)\,p(x_2\mid x_3) \tag{3.6.7}$$



(a) “尾-尾”连接　　(b) “尾-头”连接　　(c) “头-头”连接

图 3.6.4　3 种基本结构

分别讨论 3 种连接方式。

(1) “尾-尾”连接。将图 3.6.4(a)的连接方式称为“尾-尾”连接，即从 x_1 到 x_2 的通道中，作为条件的节点 x_3 连接的都是箭头的尾部。对于图 3.6.4(a)，联合概率可写为

$$p(x_1,x_2,x_3)=p(x_3)\,p(x_1\mid x_3)\,p(x_2\mid x_3) \tag{3.6.8}$$

由贝叶斯公式

$$p(x_1,x_2\mid x_3)=\frac{p(x_1,x_2,x_3)}{p(x_3)}=\frac{p(x_3)\,p(x_1\mid x_3)\,p(x_2\mid x_3)}{p(x_3)}$$

$$=p(x_1\mid x_3)\,p(x_2\mid x_3) \tag{3.6.9}$$

可知在 x_3 已知的条件下 x_1 和 x_2 是独立的

$$p(x_1,x_2\mid x_3)=p(x_1\mid x_3)\,p(x_2\mid x_3) \tag{3.6.10}$$

注意，这种独立性是以 x_3 作为条件。若没有 x_3 作为条件，一般 $p(x_1,x_2)\neq p(x_1)p(x_2)$，可以说是 x_3 的被观察到并作为条件“阻断”了 x_1 和 x_2 的联系，使 x_3 作为条件时 x_1 和 x_2 条件独立。

(2) “尾-头”连接。图 3.6.4(b)的连接方式称为“尾-头”连接(反方向的“头-尾”连接，结果相同)。类似地，可以验证：在“尾-头”连接的情况下，若在 x_3 已知的条件下 x_1 和 x_2 是独立的，即式(3.6.10)成立。同样地，x_3 的被观察到并作为条件“阻断”了 x_1 和 x_2 的通道。

(3) “头-头”连接。图 3.6.4(c)的“头-头”连接方式与以上两种情况正好相反。可验证：在不考虑条件的情况下，可得到

$$p(x_1,x_2)=p(x_1)\,p(x_2) \tag{3.6.11}$$

但在 x_3 已知的条件下 x_1 和 x_2 是不独立的，即

$$p(x_1,x_2\mid x_3)\neq p(x_1\mid x_3)\,p(x_2\mid x_3) \tag{3.6.12}$$

与前两种情况相反，x_3 不作为条件时，它阻断从了 x_1 和 x_2 的通道，使之相互独立；但当 x_3 作为条件时，它不再阻断 x_1 和 x_2 的通道，使 x_1 和 x_2 不条件独立。

以上对 3 种基本连接方式给出了其独立性判断，对于“尾-头”和“头-头”连接的验证留作习题。

从以上 3 种基本连接方式出发，可给出更一般的条件独立的判断方法，称为“D-分离”(D-separation)条件。对于一个给定的贝叶斯网络，设有 3 个节点集合 A，B，C，各集合不相交(两两之间交集为空)，将集合 C 作为条件，讨论集合 A 和 B 的条件独立性。

若通过一些中间节点(不考虑连线方向)连接了集合 A 中的一个节点和集合 B 中的一个节点，则称 A 和 B 之间有一条通道。如果满足以下条件之一，则称集合 C 阻断了一条通道。

(1) 遇到“头-尾”或“尾-尾”节点，该节点属于集合 C。

(2) 遇到“头-头”节点，该节点或其子孙节点均不属于集合 C。

若 A 和 B 之间的全部通道都被集合 C 阻断了，则称 C 条件下集合 A 和 B 独立，即

$$p(A,B\,|\,C)=p(A\,|\,C)\,p(B\,|\,C)$$

D-分离条件给出了多个节点组成的随机变量子集之间的条件独立关系，可在更复杂的有向图中进行条件独立性判断。下面给出两个实例，说明通过有向图模型建模独立性关系，这两个实例对应的学习算法均为机器学习中有影响的经典算法。

1. 朴素贝叶斯模型

讨论的第 1 个实例是朴素贝叶斯模型。例如，设计一个垃圾邮件检测器，相当于一个分类问题，输入特征向量为 $\boldsymbol{x}=[x_1,\cdots,x_{D-1},x_D]^{\mathrm{T}}$，输出为 y，$y=1$ 表示垃圾邮件，$y=0$ 表示正常邮件。预设一个关键词汇表，共有 D 个词，当词汇表中第 i 个词出现在邮件中，则对应 $x_i=1$，否则 $x_i=0$，即 x_i 是离散随机变量，仅有两个取值。描述联合概率 $p(\boldsymbol{x},y)$需要 $2^{D+1}-1$ 个参数，由于词汇表单词数目 D 比较大(如数千量级)，描述联合概率非常复杂。因此，我们采用简化的结构，给出如图 3.6.5 所示的关系，即当检测器输出 y 作为条件时，由 y 表示的是否为垃圾邮件分别对各输入分量 x_i 的概率有影响。例如，x_{102} 代表词汇“化妆品”，当 $y=1$ 时，$x_{102}=1$ 的概率更高，而各词汇的概率互相没有影响。图 3.6.5 给出了这种假设的贝叶斯网络表示，由 D-分离原则，y 作为条件时各 x_i 的条件概率是互相独立的(“尾-尾”连接)，即联合概率为

$$p(\boldsymbol{x},y)=p(y)p(\boldsymbol{x}\mid y)=p(y)\prod_{k=1}^{D}p(x_k\mid y) \tag{3.6.13}$$

为了完整描述式(3.6.13)的联合概率，只需要以下参数集

$$\begin{cases}\mu_{i|1}\triangleq p(x_i=1\,|\,y=1)\\ \mu_{i|0}\triangleq p(x_i=1\,|\,y=0) \quad i=1,2,\cdots,D\\ p(y=1)=\pi\end{cases} \tag{3.6.14}$$

即只需要 $2D+1$ 个参数。若通过样本集学习得到式(3.6.14)的参数，则对于给出的一个新的邮件，抽取其特征向量 $\boldsymbol{x}$，由后验概率 $p(y\,|\,\boldsymbol{x})$可判断该邮件是否为垃圾邮件。我们将在 5.4 节给出朴素贝叶斯学习的详细讨论，其基础就是图 3.6.5 的概率结构假设。

本例为了直观说明问题，只假设了 x_i 是仅取两个值的离散随机变量，实际上更一般的朴素贝叶斯假设只限制于图 3.6.5 的结构，至于 x_i 是离散变量还是连续变量，都有同样的独立条件。

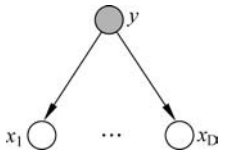


图 3.6.5　朴素贝叶斯模型

2. 隐马尔可夫模型

用贝叶斯网络建模的第 2 个实例是隐马尔可夫模型(Hidden Markov Models，HMM)。HMM 用于建模序列数据，即按照时间顺序排列的数据向量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N$，具有时序关系，这里每个 $\boldsymbol{x}_i$ 自身是一个向量，即 $\boldsymbol{x}_{i+1}$ 是在 $\boldsymbol{x}_i$ 之后出现的，相互具有时间相关性。为了描述 $\boldsymbol{x}_{i+1}$，需要条件概率

$$p(\boldsymbol{x}_{i+1}\,|\,\boldsymbol{x}_i,\boldsymbol{x}_{i-1},\cdots,\boldsymbol{x}_1) \tag{3.6.15}$$

即 $\boldsymbol{x}_{i+1}$ 的取值概率与从 $\boldsymbol{x}_i$ 起直到 $\boldsymbol{x}_1$ 的所有以前向量均有关系。建模这样的序列关系，目

前最常用的技术是循环神经网络(RNN),在 RNN 广泛应用之前,HMM 是序列建模的最常用方法。

为了描述序列的时间依赖性,又使问题处理简单,一个有效的方法是引入一系列隐变量。对于序列建模,可引入离散隐变量 $\boldsymbol{z}_i$。所谓隐变量,是指无法观测的随机变量,但其与序列变量 $\boldsymbol{x}_i$ 直接相关。可通过一系列隐变量 $\boldsymbol{z}_i$ 产生一系列观测向量 $\boldsymbol{x}_i$,将序列向量的产生用图 3.6.6 的贝叶斯网络表示。注意,与前面的例子相比,每个节点表示一个随机向量,而不是单一的随机变量。对具有任意时间依赖关系的序列进行建模,建立序列一般的时间依赖关系,可以利用隐变量带来的条件独立性以便于表示。例如,由图 3.6.6 中连接可见,从 $\boldsymbol{z}_{n-1}$ 到 $\boldsymbol{z}_{n+1}$ 经过 $\boldsymbol{z}_n$,这是"尾-头"连接,故满足条件独立性

$$p(\boldsymbol{z}_{n+1},\boldsymbol{z}_{n-1}\mid\boldsymbol{z}_n)=p(\boldsymbol{z}_{n+1}\mid\boldsymbol{z}_n)\,p(\boldsymbol{z}_{n-1}\mid\boldsymbol{z}_n) \tag{3.6.16}$$

类似地,如下条件独立性成立。

$$p(\boldsymbol{x}_n,\boldsymbol{x}_{n-1}\mid\boldsymbol{z}_n)=p(\boldsymbol{x}_n\mid\boldsymbol{z}_n)\,p(\boldsymbol{x}_{n-1}\mid\boldsymbol{z}_n) \tag{3.6.17}$$

图 3.6.6 的图模型表示了 HMM,由 HMM 图模型的节点因子关系,可得到联合概率为

$$p(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N,\boldsymbol{z}_1,\cdots,\boldsymbol{z}_N)=p(\boldsymbol{z}_1)\prod_{n=2}^{N}p(\boldsymbol{z}_n\mid\boldsymbol{z}_{n-1})\prod_{n=1}^{N}p(\boldsymbol{x}_n\mid\boldsymbol{z}_n) \tag{3.6.18}$$

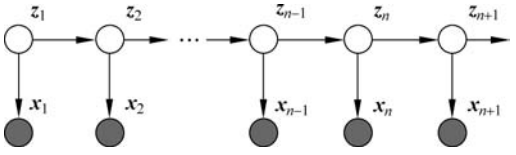


图 3.6.6 HMM 结构

在实际中,由于只能观测到多组序列样本 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N$ 集合,可由样本集估计 HMM 的参数,由于引入隐变量带来的式(3.6.18)的因子分解,可导出有效的参数学习算法。对隐变量的参数学习的一种有效方法是采用期望最大算法(EM),本书在第 12 章详细讨论 EM 算法,用 EM 算法估计 HMM 参数的学习算法可参考本章小结中推荐的参考书。

3.6.2 无向图模型

另外一种概率图模型是无向图,也称为马尔可夫随机场或马尔可夫网络,其节点表示一个或一组随机变量,两个节点之间通过无箭头的连线连接,表示两个节点之间有直接联系。图 3.6.7 所示为一个具有 5 个节点的无向图的示例。

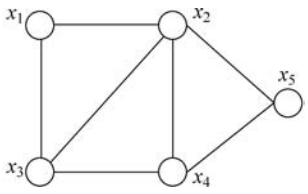


图 3.6.7 无向图示例

无向图同样可以描述节点之间的条件独立性和概率函数的因子分解特性。无向图对于节点子集之间的条件独立性的判断更加简单。例如,图 3.6.7 中,设 3 个随机变量子集分别为 $A=\{x_1\}$,$B=\{x_2,x_3\}$ 和 $C=\{x_4,x_5\}$,以 B 作为条件,考查 A 和 C 的条件独立性。若从子集 A 的节点到子集 C 的节点的通道,都需要通过子集 B 的节点,则说明 B 阻断了 A 和 C,则 B 作为条件时 A 和 C 独立;否则,只要有一条通道不经过子集 B 的节点,B 就没有阻断 A 和 C 的通路,则 B 作为条件时 A 和 C 不独立。图 3.6.7 的例子中,B 阻断了 A 和 C 的通路,故有 $p(A,B\mid C)=$

$p(A|C)p(B|C)$。在无向图的条件独立性判断中，不需要区分“头-头”或“尾-尾”之类的区别，将更加简单。

为了通过无向图表示概率的因子分解，引入“团”(Clique)的概念，一个团是指一个节点子集，该子集中的任何一对节点间有连线。进一步，可引入“最大团”(Maximal Clique)的概念，对于一个给定的无向图，一个最大团是指一个节点子集构成的团，图中不再有一个其他节点能够与该子集构成更大的团。有这些定义，可见图 3.6.7 的无向图有如下子集构成了团：$\{x_1,x_2\}$，$\{x_2,x_3\}$，$\{x_1,x_3\}$，$\{x_2,x_4\}$，$\{x_3,x_4\}$，$\{x_2,x_5\}$，$\{x_4,x_5\}$，$\{x_1,x_2,x_3\}$，$\{x_2,x_3,x_4\}$，$\{x_2,x_4,x_5\}$；而最大团有 3 个：$\{x_1,x_2,x_3\}$，$\{x_2,x_3,x_4\}$，$\{x_2,x_4,x_5\}$。用 $\boldsymbol{x}_c$ 表示一个最大团，例如，本例中 $\boldsymbol{x}_1=\{x_1,x_2,x_3\}$，$\boldsymbol{x}_2=\{x_2,x_3,x_4\}$，$\boldsymbol{x}_3=\{x_2,x_4,x_5\}$，则可由最大团的势函数 $\psi_c(\boldsymbol{x}_c)$ 作为因子，将无向图所表示的概率函数表示为各最大团势函数的积，即

$$p(\boldsymbol{x})=\frac{1}{Z}\prod_c\psi_c(\boldsymbol{x}_c) \tag{3.6.19}$$

其中，Z 为归一化常数。若 $\boldsymbol{x}$ 是离散的，则

$$Z=\sum_{\boldsymbol{x}}\prod_c\psi_c(\boldsymbol{x}_c) \tag{3.6.20}$$

若 $\boldsymbol{x}$ 是连续的，只需要将求和改为积分即可。为了使 $p(\boldsymbol{x})$ 为合格的概率函数，要求势函数 $\psi_c(\boldsymbol{x}_c)\geqslant 0$。故一种常用的势函数由指数函数来定义，即

$$\psi_c(\boldsymbol{x}_c)=\exp[-E(\boldsymbol{x}_c)] \tag{3.6.21}$$

其中，$E(\boldsymbol{x}_c)$ 称为团的能量函数。用指数函数表示势函数的概率函数称为玻尔兹曼分布(Boltzmann Distribution)。将式(3.6.21)代入式(3.6.19)得到指数势函数的概率表示为

$$p(\boldsymbol{x})=\frac{1}{Z}\exp\Big[-\sum_c E(\boldsymbol{x}_c)\Big]=\frac{1}{Z}\exp[-E(\boldsymbol{x})] \tag{3.6.22}$$

其中，$E(\boldsymbol{x})$ 为总能量函数。

$$E(\boldsymbol{x})=\sum_c E(\boldsymbol{x}_c) \tag{3.6.23}$$

对于图 3.6.7 的无向图，用指数函数可以将概率函数写为

$$\begin{aligned}p(\boldsymbol{x})&=\frac{1}{Z}\exp[-E(x_1,x_2,x_3)]\exp[-E(x_2,x_3,x_4)]\exp[-E(x_2,x_4,x_5)]\\&=\frac{1}{Z}\exp[-E(x_1,x_2,x_3)-E(x_2,x_3,x_4)-E(x_2,x_4,x_5)]\end{aligned}$$

在实际中，根据不同的应用需求，可选择不同的能量函数(及其参数)，无向图模型有非常灵活的构成方式，并已获得许多应用成果。

下面通过一个例子说明无向图的建模过程，主要说明怎样根据应用选择能量函数和节点连接关系，从而由团因子构成概率函数。注意到，选择能量函数后，式(3.6.19)的因子相乘变化为式(3.6.23)的能量和的形式，以下例子中，直接使用能量和形式。

例 3.6.1　条件随机场模型。

以信息检索的一个应用场景为例进行说明，不关注信息检索的细节，只简单说明一下变

量的背景，主要看怎样构成一个概率函数。这里讨论的条件随机场是一种对序列数据进行学习的概率模型，其定义了给定观测序列后输出序列的条件分布。在信息检索的模型学习中，对于给出的一个查询 q，样本集给出一组文档，用向量 $\boldsymbol{x}_i$ 表示一个文档的特征向量（输入向量），y_i 表示文档对于该查询的得分（高得分说明该文档与查询匹配度高），将 y_i 看作输出（学习过程中相当于标注值），对于一个查询给出的一组文档，用 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$ 表示一个查询对应的文档集合的输入向量，用 $\boldsymbol{y}=\{y_1,y_2,\cdots,y_n\}$ 表示各文档对应的得分输出，在训练集中每个文档得分输出由专家打分作为标注值。用无向图的一个节点分别表示 $\boldsymbol{x}_i$ 和 y_i，假设 $\boldsymbol{x}_i$ 和 y_i 是有连接的，一些 y_i 之间可能有连接（但只假设两两之间有部分连接）。这种连接关系使 $\{\boldsymbol{x}_i,y_i\}$ 为团，另外一些有连接的 $\{y_i,y_j\}(i\neq j)$ 构成团。无向图结构如图 3.6.8 所示。

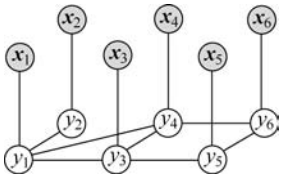


图 3.6.8 作为无向图模型的条件随机场实例

所谓条件随机场，是直接由图 3.6.8 的无向图表示 $\boldsymbol{X}$ 作为条件的 $\boldsymbol{y}$ 的概率函数，由团的组成可得

$$p(\boldsymbol{y}\mid\boldsymbol{X})=\frac{1}{Z(\boldsymbol{X})}\exp\left[-\sum_i\sum_{k=1}^{K}\alpha_k f_k(y_i,\boldsymbol{x}_i)-\sum_{i,j}\beta g(y_i,y_j)\right] \tag{3.6.24}$$

实际上，对于团 $\{\boldsymbol{x}_i,y_i\}$ 定义了一组能量函数的和 $\sum_{k=1}^{K}\alpha_k f_k(y_i,\boldsymbol{x}_i)$，对于团 $\{y_i,y_j\}(i\neq j)$，定义了能量函数 $\beta g(y_i,y_j)$，α_k 和 β 是需要学习的参数。由于是建模条件概率，故归一化因子 $Z(\boldsymbol{X})$ 写为 $\boldsymbol{X}$ 的函数。由于得分 y 是连续值，故 $Z(\boldsymbol{X})$ 的计算式为

$$Z(\boldsymbol{X})=\int_y\exp\left[-\sum_i\sum_{k=1}^{K}\alpha_k f_k(y_i,\boldsymbol{x}_i)-\sum_{i,j}\beta g(y_i,y_j)\right]\mathrm{d}y \tag{3.6.25}$$

在实际应用中，$f_k(y_i,\boldsymbol{x}_i)$ 和 $g(y_i,y_j)$ 函数可适当选择，以获得良好的效果，一类信息检索的应用情况下，效果良好的一种函数选择为

$$f_k(y_i,\boldsymbol{x}_i)=(y_i-x_{i,k})^2 \tag{3.6.26}$$

$$g(y_i,y_j)=\frac{1}{2}S_{i,j}(y_i-y_j)^2 \tag{3.6.27}$$

其中，$x_{i,k}$ 为 $\boldsymbol{x}_i$ 第 k 分量的值；K 为 $\boldsymbol{x}_i$ 的维数；$S_{i,j}$ 为预定的权系数，评价两个文档的相似性。在学习阶段，由样本集 $\{\boldsymbol{X}_q,\boldsymbol{y}_q\}_{q=1}^{N}$ 可学习模型的参数，当实际对新的输入 $\boldsymbol{X}$ 做检索时，可推断输出 $\boldsymbol{y}$ 为

$$\begin{aligned}\boldsymbol{y}&=\arg\max_{\boldsymbol{y}}[p(\boldsymbol{y}\mid\boldsymbol{X})]\\&=\arg\max_{\boldsymbol{y}}\left[-\sum_i\sum_{k=1}^{K}\alpha_k f_k(y_i,\boldsymbol{x}_i)-\sum_{i,j}\beta g(y_i,y_j)\right]\end{aligned} \tag{3.6.28}$$

本例给出的是无向图应用的一个完整工作的一部分，对于其学习或推断的细节，感兴趣的读者可阅读相关文献[①]。

① 见本书参考文献[101]。

3.6.3　图模型的学习与推断

对概率图模型的参数可进行学习，利用已确定了参数的概率图模型可进行推断，本节对此问题仅进行简要介绍。

1. 学习

构成图模型的概率表达式中包含待确定参数。例如，在朴素贝叶斯实例中，式(3.6.14)所示的一组参数；以及在图 3.6.8 所示的无向图实例中，式(3.6.24)中的参数 α_k 和 β。

首先讨论有向图的参数学习。为了表示图模型中包含的参数，可将有向图的概率表示(式(3.6.5))重新表示为以下参数化形式。

$$p(\boldsymbol{x};\boldsymbol{\theta})=\prod_{k=1}^{D}p(x_k\mid\boldsymbol{x}_{fk};\boldsymbol{\theta}_k) \tag{3.6.29}$$

在机器学习应用中，需通过 IID 的样本集 $\boldsymbol{X}=\{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$ 学习图模型中的参数，这里为了避免与图中变量的下标混淆，用上标(n)表示样本序号。当样本集确定后，关于参数的对数似然函数表示为

$$\log p(\boldsymbol{X};\boldsymbol{\theta})=\sum_{n=1}^{N}\log p(\boldsymbol{x}^{(n)};\boldsymbol{\theta})=\sum_{n=1}^{N}\sum_{k=1}^{D}\log p(x_k^{(n)}\mid\boldsymbol{x}_{fk}^{(n)};\boldsymbol{\theta}_k) \tag{3.6.30}$$

由于因子的参数$\boldsymbol{\theta}_k$是各自独立的，故最大似然的参数求解为

$$\hat{\boldsymbol{\theta}}_k=\arg\max_{\boldsymbol{\theta}_k}\sum_{n=1}^{N}\log p(x_k^{(n)}\mid\boldsymbol{x}_{fk}^{(n)};\boldsymbol{\theta}_k),\quad k=1,2,\cdots,D \tag{3.6.31}$$

对于无向图，若采用指数函数表示的势函数，则式(3.6.22)式重写为

$$p(\boldsymbol{x};\boldsymbol{\theta})=\frac{1}{Z(\boldsymbol{\theta})}\exp\Big[-\sum_{c}E(\boldsymbol{x}_c;\boldsymbol{\theta}_c)\Big] \tag{3.6.32}$$

注意到，由归一化因子的定义可知，$Z(\boldsymbol{\theta})$是随$\boldsymbol{\theta}$变化的，其对数似然函数为

$$\log p(\boldsymbol{X};\boldsymbol{\theta})=\sum_{n=1}^{N}\log p(\boldsymbol{x}^{(n)};\boldsymbol{\theta})=-\sum_{n=1}^{N}\sum_{c}E(\boldsymbol{x}_c^{(n)};\boldsymbol{\theta}_c)-NZ(\boldsymbol{\theta}) \tag{3.6.33}$$

5.4 节将对朴素贝叶斯假设给出其参数学习的详细过程，对于图 3.6.8 的无向图实例，其学习过程可参考相关文献[①]。

当在图模型中引入隐变量时，如 HMM，则用 EM 算法学习模型参数是有效的，EM 算法的详细介绍可参考 12.2 节。

2. 推断

当通过学习过程确定了图模型的参数后，可通过图模型进行推断。所谓推断，是指给出图模型表示的随机变量集合 $\boldsymbol{x}$ 的部分观测值后，去推断另一些感兴趣的变量的条件概率。这里可将 $\boldsymbol{x}$ 分为互不重叠的 3 个子集 $\boldsymbol{x}=\{\boldsymbol{x}_{\mathrm{o}},\boldsymbol{x}_{\mathrm{q}},\boldsymbol{x}_{\mathrm{u}}\}$，其中 $\boldsymbol{x}_{\mathrm{o}}$ 表示本次推断观测到取值的变量集合；$\boldsymbol{x}_{\mathrm{q}}$ 表示本次推断感兴趣的变量集合；$\boldsymbol{x}_{\mathrm{u}}$ 表示本次推断时既没有观测到取值也不感兴趣的变量集合。所谓一次推断的任务是从 $p(\boldsymbol{x}_{\mathrm{q}},\boldsymbol{x}_{\mathrm{o}},\boldsymbol{x}_{\mathrm{u}})$的表示出发，计算条件概

① 见本书参考文献[101]。

率 $p(\boldsymbol{x}_{\mathrm{q}}\mid\boldsymbol{x}_{\mathrm{o}})$，由贝叶斯公式可得

$$p(\boldsymbol{x}_{\mathrm{q}}\mid\boldsymbol{x}_{\mathrm{o}})=\frac{p(\boldsymbol{x}_{\mathrm{q}},\boldsymbol{x}_{\mathrm{o}})}{p(\boldsymbol{x}_{\mathrm{o}})}=\frac{\sum_{\boldsymbol{x}_{\mathrm{u}}}p(\boldsymbol{x}_{\mathrm{q}},\boldsymbol{x}_{\mathrm{o}},\boldsymbol{x}_{\mathrm{u}})}{\sum_{\boldsymbol{x}_{\mathrm{q}},\boldsymbol{x}_{\mathrm{u}}}p(\boldsymbol{x}_{\mathrm{q}},\boldsymbol{x}_{\mathrm{o}},\boldsymbol{x}_{\mathrm{u}})} \tag{3.6.34}$$

式(3.6.34)预设了随机变量是离散的，若需要边际化的随机变量是连续的，用积分替代以上求和。

以上推断过程中，若每个随机变量可取 K 个值，有 V 个变量，直接运算需要的计算复杂度为 $O(K^V)$，当 V 很大时，运算开销非常大。若利用图模型的消息传递，可有效控制计算复杂度。若概率图是如图 3.6.3 所示的链图，计算复杂度可限制在 $O(KV)$。若概率图是一种宽度较小的树形结构，则仍可大量节省运算量。在图模型上有效计算边际概率的算法之一是和积算法(Sum-Product)，其详细介绍可参考相关文献[①]。

若在一个简单的二分类问题中使用图模型，可将 $\boldsymbol{x}$ 分为两个子集 $\boldsymbol{x}=\{\boldsymbol{x}_i,y\}$，这里用 $\boldsymbol{x}_i$ 表示分类问题的输入特征向量，y 表示分类输出，仅取 0 和 1 两个值，则给出输入 $\boldsymbol{x}_i$，y 的条件概率为

$$p(y\mid\boldsymbol{x}_i)=\frac{p(y,\boldsymbol{x}_i)}{p(y=0,\boldsymbol{x}_i)+p(y=1,\boldsymbol{x}_i)} \tag{3.6.35}$$

若已由图模型的参数学习确定了 $p(y,\boldsymbol{x}_i)$，则分类输出 y 的后验概率 $p(y\mid\boldsymbol{x}_i)$ 的计算变得较为简单。有了 $p(y\mid\boldsymbol{x}_i)$，由 3.2 节的决策原理确定输出的类型。5.4 节的朴素贝叶斯方法将描述一个简单图模型的建模、学习和推断的完整实例。

概率图模型涉及的内容广泛而深刻，除了对于一个给定模型的学习和推断外，还可能从数据中学习出有效的图结构，本书仅给出了图模型的一个极其简略的介绍，有兴趣的读者可进一步阅读本章小结中推荐的参考书。

3.7 本章小结

决策是机器学习中一个相对独立的部分，通过训练过程确定了一个机器学习模型，若该模型是概率模型，当有一个新的输入特征向量时，可在推断过程计算出分类或回归的后验概率，则决策过程确定最终的输出。对于分类问题，若各类错误是同等重要的，则简单的最大后验即可确定分类输出；对于回归问题，则计算后验期望并作为回归输出。通过基本的决策理论讨论了高斯分布和 KNN 方法的决策实例。

概率结构对许多学习算法和推断算法有重要影响，概率图模型以相对直观的方法描述了概率结构，概率图模型已发展出非常完整的理论和算法，并不断与其他机器学习方法相结合扩展出新的技术，本节只给出了非常简略的介绍。

多个领域的著作均对决策理论给出了深入讨论，统计学的著作中有对决策理论的细致讨论，如 Casella 等的 *Statistical Inference*；侧重统计模式识别的著作常有对决策理论的详细讨论，如 Duda 等的 *Pattern Classification*；信号处理类的著作中，也有很深入的决策理

① 见本书参考文献[13,93]。

论的讨论，如 Poor 的 *An Introduction to Signal Detection and Estimation*。对于概率图模型，Bishop 的 *Pattern Recognition and Machine Learning* 和 Murphy 的 *Machine Learning* 这两本经典机器学习著作都给出了中等深度的介绍，也都给出了通过 EM 算法和概率图模型估计 HMM 参数的算法；而 Koller 等的著作 *Probabilistic Graphical Models*：*Principles and Techniques* 则给出了概率图模型更加全面、系统和深入的介绍。

习题

1. 对于特征向量 $\boldsymbol{x}$ 是一个一维标量的情况，设样本集中只有两种类型，类条件概率分别为

$$p(x\mid C_1)=\frac{1}{2\times(2\pi)^{1/2}}\exp\left[-\frac{1}{8}(x+1)^2\right]$$

$$p(x\mid C_2)=\frac{1}{2\times(2\pi)^{1/2}}\exp\left[-\frac{1}{8}(x-2)^2\right]$$

且 $p(C_1)=p(C_2)=0.5$。

(1) 在错误分类率最小意义下，给出将新的输入 x 判决为 C_1 或 C_2 的判决边界。

(2) 在问题(1)求出的判决边界的条件下，求总的错误分类率。

2. 试推导：由式(3.4.22)的定义可得到式(3.4.25)。

3. 对于高斯分布的多分类问题，证明其后验概率 $p(C_i\mid\boldsymbol{x})$ 可表示为

$$p(C_i\mid\boldsymbol{x})=\frac{p(\boldsymbol{x}\mid C_i)p(C_i)}{\sum_{k=1}^{K}p(\boldsymbol{x}\mid C_k)p(C_k)}=\frac{\exp(a_i)}{\sum_{k=1}^{K}\exp(a_k)}$$

其中，$a_k=\ln p(\boldsymbol{x}\mid C_k)p(C_k)$，并证明

$$a_i=g_i(\boldsymbol{x})=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}_i\boldsymbol{x}+\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x}+w_{i0}\quad i=1,\cdots,K$$

其中，系数如式(3.4.27)所示。

4. 一个二分类问题，特征向量 $\boldsymbol{x}$ 是二维的，有样本集

$$\begin{aligned}\boldsymbol{D}&=\{(\boldsymbol{x}_n,y_n)\}_{n=1}^{N}\\&=\{([0.5,1.5]^{\mathrm{T}},1),([1.5,1.5]^{\mathrm{T}},1),([0.5,0.5]^{\mathrm{T}},1),([1.5,0.5]^{\mathrm{T}},1),\\&\quad([1.5,0.5]^{\mathrm{T}},0),([2.5,0.5]^{\mathrm{T}},0),([1.5,-0.5]^{\mathrm{T}},0),([2.5,-0.5]^{\mathrm{T}},0)\}\end{aligned}$$

类条件概率服从高斯分布，且 $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\boldsymbol{\Sigma}$。

(1) 求通过样本得到的判别函数 $g(\boldsymbol{x})$，在 $\boldsymbol{x}$ 平面上画出分类为 C_1 和 C_2 的区间。

(2) 求后验概率 $p(C_1\mid\boldsymbol{x})$ 表达式。

(3) 对于新的特征输入 $\boldsymbol{x}=[2.0,-0.3]^{\mathrm{T}}$，可分为哪一类？计算后验概率 $p(C_1\mid\boldsymbol{x})$。

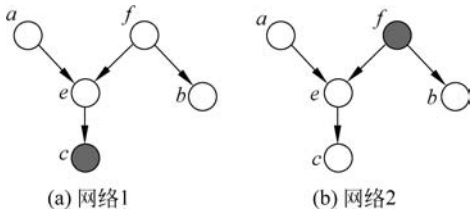
5. 图 3.6.4 的 3 种基本连接中，对于图 3.6.4(b)的“尾-头”连接方式，证明：在 x_3 已知的条件下 x_1 和 x_2 是独立的；对于图 3.6.4(c)的“头-头”连接方式，证明：在不考虑条件的情况下，可得到

$$p(x_1,x_2)=p(x_1)p(x_2)$$

但在 x_3 已知的条件下 x_1 和 x_2 是不独立的，即

$$p(x_1, x_2 \mid x_3) \neq p(x_1 \mid x_3)\, p(x_2 \mid x_3)$$

6. 如图所示，利用 D-分离原则，证明：对于图(a)，$p(a, b \mid c) \neq p(a \mid c)\, p(b \mid c)$；对于图(b)，$p(a, b \mid f) = p(a \mid f)\, p(b \mid f)$。



(a) 网络1　　(b) 网络2

第 6 题图