

第3章 贝叶斯分类器

模式识别的分类问题可以看作是一种决策,即根据识别对象的观测特征做出其应属于哪一类的决策。贝叶斯决策理论是处理模式识别分类问题的基本理论之一,对分类器的设计具有重要的指导意义。贝叶斯决策理论的基本思想是,给定具有特征向量的待识别样本,计算其属于某一类的概率,并将它属于某一类的概率值作为后续分类决策的依据。在具体实际应用过程中,往往需要首先根据训练样本来估计各类的概率密度函数,然后再进行分类决策。使用贝叶斯分类器时,要求满足以下两个前提条件:

- (1) 每类样本的概率密度函数是已知的;
- (2) 样本的类别数是已知的。

本章将详细介绍在满足以上两个条件的前提下,贝叶斯分类器的设计方法。

3.1 基本概念

介绍基本概念之前,先来看一个简单的例子。

假如路边有一棵树,需要判断树是什么种类,这就是一个分类问题。如果知道这棵树只可能是苹果树或柿子树,这就是一个二分类问题。我们讨论以下两种情况。

(1) 假如现在只知道它是一棵树,没有任何其他信息,也许会猜测这是柿子树,依据是这条路上柿子树的数量远比苹果树多。在这个过程中,你基于对过往经验的分析,认为这是柿子树的概率比苹果树的概率更大,因此做出了概率较大的决策。这些概率是事先根据已发生的事实统计的,所以称为先验概率。令苹果树和柿子树分别为 ω_1 类和 ω_2 类,那么它们的先验概率可以用 $P(\omega_1)$ 和 $P(\omega_2)$ 来表示。

(2) 假如现在可以观察树上果实的颜色,以此来判断树的种类。我们把果实的颜色记为 x ,需要探讨的是在知道果实颜色为 x 的情况下,这棵树属于 ω_1 类和 ω_2 类的概率。将这两个概率分别记为 $P(\omega_1 | x)$ 和 $P(\omega_2 | x)$,这就是后验概率。根据概率论中的贝叶斯公式有

$$P(\omega_i | x) = \frac{p(x | \omega_i)P(\omega_i)}{p(x)}, \quad i = 1, 2 \quad (3-1)$$

其中, $p(x)$ 是所有树果实颜色的概率密度,称为总体密度; $p(x | \omega_i)$ 是第 i 类树果实颜色的概率密度,称为类条件概率密度; $P(\omega_i)$ 是第 i 类树的先验概率。接下来将用学术语言解释这几个概念。



3.1.1 先验概率

先验概率是根据以往经验和分析得到的概率,往往通过大量抽样实验估计得到。对于 c 类的分类问题,用 ω_i 表示第 i 个类别,则 ω_i 类的先验概率用 $P(\omega_i)$ 表示,且满足

$$\sum_{i=1}^c P(\omega_i) = 1, \quad i = 1, 2, \dots, c \quad (3-2)$$

在实际应用中,先验概率一般不作为分类决策的唯一依据。但倘若用于统计先验概率的样本数量足够大时,可将其作为分类决策的主要因素。

3.1.2 类条件概率密度

类条件概率密度是指该类样本的特征在特征空间的分布属性,即 ω_i 类样本的类条件概率密度就是在特征空间中, ω_i 类样本的特征 $\mathbf{x}$ 出现的概率密度。常用 $p(\mathbf{x} | \omega_i)$ 来表示 ω_i 类的类条件概率密度函数。 ω_i 类的类条件概率密度只与 ω_i 类样本的分布有关系,与其他类的样本分布无关。例如:对于细胞状态的识别,假设描述细胞状态的特征用 $\mathbf{x}$ 来表示,细胞有正常 ω_1 和异常 ω_2 两个类别。那么, ω_1 类的类条件概率密度 $p(\mathbf{x} | \omega_1)$ 表示 ω_1 类样本的分布属性,与 ω_2 类的样本无关,反之亦然。 $p(\mathbf{x} | \omega_1)$ 和 $p(\mathbf{x} | \omega_2)$ 之间没有任何关系,一般情况下, $p(\mathbf{x} | \omega_1) + p(\mathbf{x} | \omega_2) \neq 1$ 。

在实际应用中,如果统计数据满足正态分布,可采用正态密度函数作为类条件概率密度的函数形式。在正态密度函数中,只有期望 μ 和方差 σ^2 是未知的参数。可以通过对大量实验样本的统计估计出这两个参数,进而确定类条件概率密度函数。

3.1.3 后验概率

后验概率是指在事件发生后,由某个因素引起这一事件发生的概率,即在某一属性 $\mathbf{x}$ 被观测到的条件下, ω_i 类别发生的概率,常用 $P(\omega_i | \mathbf{x})$ 表示 ω_i 的后验概率。例如,对于上面细胞状态识别的例子, $p(\omega_1 | \mathbf{x})$ 表示特征是 $\mathbf{x}$ 的样本属于 ω_1 类的概率, $p(\omega_2 | \mathbf{x})$ 表示特征是 $\mathbf{x}$ 的样本属于 ω_2 类的概率。由于细胞必属于 ω_1 类和 ω_2 类中的一类,因此 $p(\omega_1 | \mathbf{x}) + p(\omega_2 | \mathbf{x}) = 1$ 。

3.1.4 贝叶斯公式

贝叶斯公式由英国数学家贝叶斯于 1963 年提出,其可以将先验概率 $P(\omega_i)$ 、类条件概率密度 $p(\mathbf{x} | \omega_i)$ 和后验概率 $P(\omega_i | \mathbf{x})$ 联合起来。假设有 c 类样本,利用贝叶斯公式可以得到利用先验概率 $P(\omega_i)$ 和类条件概率密度 $p(\mathbf{x} | \omega_i)$ 计算后验概率 $P(\omega_i | \mathbf{x})$ 的方法,即

$$P(\omega_i | \mathbf{x}) = \frac{p(\mathbf{x} | \omega_i)P(\omega_i)}{p(\mathbf{x})}, \quad (i = 1, 2, \dots, c) \quad (3-3)$$

其中, $p(\mathbf{x})$ 是所有类别的概率密度,可由式(3-4)计算:

$$p(\mathbf{x}) = \sum_{i=1}^c p(\mathbf{x} | \omega_i)P(\omega_i) \quad (3-4)$$



3.2 贝叶斯决策

3.1 节介绍了先验概率、类条件概率密度、后验概率的基本概念以及贝叶斯公式。依据贝叶斯公式,可以利用已知或估计出的先验概率和类条件概率密度计算各类样本的后验概率,并使用某种准则完成分类识别,这就是贝叶斯决策的基本思想。贝叶斯决策理论也称作统计决策理论,设计者按照具体问题具体分析,会采用不同的决策准则,从而产生不同的判别结果。其中,最小错误率贝叶斯决策和最小风险贝叶斯决策是最基本的两种方法。下面将分别介绍。

3.2.1 最小错误率贝叶斯决策

当按照某一种分类准则对目标样本进行分类时,一般都存在决策判断错误的概率。最小错误率贝叶斯决策方法的出发点就是使分类决策的错误率最小。例如,对于一个二类的分类问题,待识别样本记作 $\mathbf{x}$, 两个类别分别记作 ω_1 和 ω_2 , 两类的先验概率用 $P(\omega_1)$ 和 $P(\omega_2)$ 来表示。如果直接使用先验概率进行决策,则决策准则可以表示为

$$\begin{cases} \text{若 } P(\omega_1) > P(\omega_2), \text{ 则 } \mathbf{x} \in \omega_1 \\ \text{若 } P(\omega_2) > P(\omega_1), \text{ 则 } \mathbf{x} \in \omega_2 \end{cases} \quad (3-5)$$

对于两分类问题,两类的先验概率满足 $P(\omega_1) + P(\omega_2) = 1$ 。如果决策 $\mathbf{x} \in \omega_1$, 则分类错误率为 $P(\text{error}) = 1 - P(\omega_1) = P(\omega_2)$; 如果决策 $\mathbf{x} \in \omega_2$, 则分类错误率为 $P(\text{error}) = 1 - P(\omega_2) = P(\omega_1)$ 。当 $P(\omega_1) > P(\omega_2)$ 时,把 $\mathbf{x}$ 划分到 ω_2 类带来的错误率更大,所以应把 $\mathbf{x}$ 划分到错误率更低的 ω_1 类。但是,如果此时按照式(3-5)所示的决策规则,会把所有的样本都归于 ω_1 类,而根本没有达到要把两类样本分开的目的。这是由于先验概率提供的有效信息太少。因此,我们还需要利用更多的观测信息来进行分类决策。

如果两类样本的类条件概率密度分别为 $p(\mathbf{x} | \omega_1)$ 和 $p(\mathbf{x} | \omega_2)$, 则利用两类样本的类条件概率密度分别为 $p(\mathbf{x} | \omega_1)$ 和 $p(\mathbf{x} | \omega_2)$, 利用贝叶斯公式,可由先验概率和类条件概率密度计算出后验概率 $P(\omega_1 | \mathbf{x})$ 和 $P(\omega_2 | \mathbf{x})$ 。利用后验概率,可以定义两类别分类问题的决策规则如下:

$$\begin{cases} \text{若 } P(\omega_1 | \mathbf{x}) > P(\omega_2 | \mathbf{x}), \text{ 则 } \mathbf{x} \in \omega_1 \\ \text{若 } P(\omega_2 | \mathbf{x}) > P(\omega_1 | \mathbf{x}), \text{ 则 } \mathbf{x} \in \omega_2 \end{cases} \quad (3-6)$$

式(3-6)就是二分类情况下最小错误率贝叶斯决策规则,即:在先验概率和类条件概率密度已知的情况下,利用贝叶斯公式计算后验概率,并通过比较样本属于两类的后验概率,将待识别样本决策为后验概率大的一类。该决策规则可以使分类器的总体错误率最小。

对于 c 类别的分类问题,最小错误率贝叶斯决策规则可表示为

$$\text{若 } P(\omega_i | \mathbf{x}) = \max_{j=1,2,\dots,c} \{P(\omega_j | \mathbf{x})\}, \text{ 则 } \mathbf{x} \in \omega_i \quad (3-7)$$

利用贝叶斯公式 $P(\omega_i | \mathbf{x}) = \frac{p(\mathbf{x} | \omega_i)P(\omega_i)}{\sum_{j=1}^c p(\mathbf{x} | \omega_j)P(\omega_j)}$, 可以得到以下最小错误率贝叶斯决策规则的等价形式。



(1) 由于贝叶斯公式的分母与分类结果无关,所以决策时只需要比较贝叶斯公式分子的大小即可。决策规则可写为

$$\text{若 } p(\mathbf{x} | \omega_i)P(\omega_i) = \max_{j=1,2,\dots,c} \{p(\mathbf{x} | \omega_j)P(\omega_j)\}, \text{ 则 } \mathbf{x} \in \omega_i \quad (3-8)$$

(2) 对于二分类问题,式(3-6)可以表示为如下等价形式:

$$\text{若 } l(\mathbf{x}) = \frac{p(\mathbf{x} | \omega_1)}{p(\mathbf{x} | \omega_2)} > \frac{P(\omega_2)}{P(\omega_1)} = \lambda, \text{ 则 } \mathbf{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases} \quad (3-9)$$

其中, $l(\mathbf{x})$ 称为似然比, λ 称为似然比阈值。对待识别样本 $\mathbf{x}$ 进行分类时,首先计算其对应的似然比 $l(\mathbf{x})$,然后将似然比与阈值 λ 进行比较。若似然比大于阈值,则决策 $\mathbf{x} \in \omega_1$; 若似然比小于阈值,则决策 $\mathbf{x} \in \omega_2$ 。

(3) 在某些特定情况下,用对数形式进行计算会更加方便。对式(3-9)两边取自然对数,有

$$\text{若 } \ln l(\mathbf{x}) = \ln p(\mathbf{x} | \omega_1) - \ln p(\mathbf{x} | \omega_2) > \frac{\ln P(\omega_2)}{\ln P(\omega_1)}, \text{ 则 } \mathbf{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases} \quad (3-10)$$

下面探讨分类决策的错误率问题,并证明最小错误率贝叶斯决策确实能使分类错误率最小。这里,仅给出一维情况下的分析和证明,其结果不难推广到多维。

首先应指出,这里分类决策的错误率是指平均错误率,用 $P(e)$ 来表示,其定义为

$$P(e) = \int_{-\infty}^{\infty} P(e | \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (3-11)$$

其中, $P(e | \mathbf{x})$ 表示观测值为 $\mathbf{x}$ 时的条件错误概率, $p(\mathbf{x})$ 为观测值 $\mathbf{x}$ 出现的概率密度函数, $\int_{-\infty}^{\infty} P(e | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$ 表示在整个 d 维特征空间上的积分。

对于两类的分类问题,由最小错误率贝叶斯决策规则可知:当 $P(\omega_1 | \mathbf{x}) > P(\omega_2 | \mathbf{x})$ 时,决策 $\mathbf{x} \in \omega_1$,此时样本 $\mathbf{x}$ 的条件错误概率为 $P(\omega_2 | \mathbf{x})$;反之,当 $P(\omega_2 | \mathbf{x}) > P(\omega_1 | \mathbf{x})$ 时,决策 $\mathbf{x} \in \omega_2$,此时样本 $\mathbf{x}$ 的条件错误概率为 $P(\omega_1 | \mathbf{x})$ 。因此,条件错误概率 $P(e | \mathbf{x})$ 可以表示为

$$P(e | \mathbf{x}) = \begin{cases} P(\omega_1 | \mathbf{x}), & \text{当 } P(\omega_2 | \mathbf{x}) > P(\omega_1 | \mathbf{x}) \\ P(\omega_2 | \mathbf{x}), & \text{当 } P(\omega_1 | \mathbf{x}) > P(\omega_2 | \mathbf{x}) \end{cases} \quad (3-12)$$

令两类样本的分界面为 t ,其可由方程 $P(\omega_1 | \mathbf{x}) = P(\omega_2 | \mathbf{x})$ 来确定。当样本特征向量 $\mathbf{x}$ 的维数为一维时, t 为 x 轴上的一个点,且 t 将 x 轴分为两个判决区域: R_1 和 R_2 。其中, R_1 为 ω_1 类的判决区域 $(-\infty, t)$, R_2 为 ω_2 类的判决区域 (t, ∞) 。该两类问题的错误率可由式(3-11)和式(3-12)推得

$$\begin{aligned} P(e) &= \int_{R_1} P(\omega_2 | \mathbf{x}) p(\mathbf{x}) d\mathbf{x} + \int_{R_2} P(\omega_1 | \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\ &= \int_{R_1} p(\mathbf{x} | \omega_2) P(\omega_2) d\mathbf{x} + \int_{R_2} p(\mathbf{x} | \omega_1) P(\omega_1) d\mathbf{x} \\ &= P(\omega_2) P_2(e) + P(\omega_1) P_1(e) \end{aligned} \quad (3-13)$$

其中, $P_1(e) = \int_{R_2} p(\mathbf{x} | \omega_1) d\mathbf{x}$, $P_2(e) = \int_{R_1} p(\mathbf{x} | \omega_2) d\mathbf{x}$ 。

对于区域 R_1 内任意的 $\mathbf{x}$ 值,都有 $P(\omega_1 | \mathbf{x}) > P(\omega_2 | \mathbf{x})$, $P_2(e)$ 在每个 $\mathbf{x}$ 值处都取最



小者。同样,对于区域 R_2 内的任意 x 值,都有 $P(\omega_2 | x) > P(\omega_1 | x)$, $P_1(e)$ 在每个 x 值处都取最小者。也就是说,对于样本空间中每个任意的 x 值,错误率都取最小者。因此,分类决策的平均错误率 $P(e)$ 也必然为最小。如图 3-1 所示, t 为一维特征空间的分界点。显然, t 的位置不同,错误率也不同。在图 3-1 中,左半边的阴影部分面积代表 $P(\omega_2) P_2(e)$,右半边阴影部分面积代表 $P(\omega_1) P_1(e)$,两者之和为总错误率 $P(e)$ 。

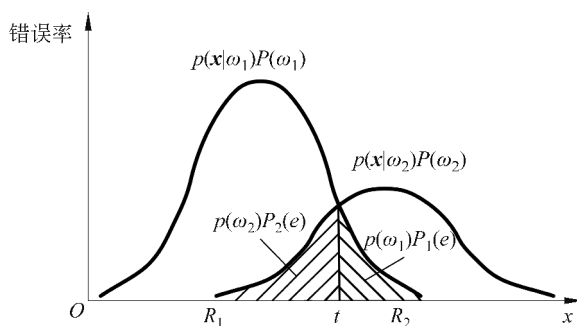


图 3-1 最小错误率贝叶斯决策错误率

例 3.1 假设存在一种机器设备,其能正常工作的概率为 70%,发生异常的概率为 30%,且发现机器在运作过程中有时存在噪声。根据历史统计得知,正常工作情况下出现噪声的概率为 10%,异常状态下出现噪声的概率为 80%。问:若发现一台机器设备出现噪声,利用最小错误率贝叶斯决策规则,应该把它判别为正常工作状态还是异常状态?

解: 设正常工作状态为 ω_1 , 异常状态为 ω_2 , 则两类别的先验概率为

$$P(\omega_1) = 0.7, P(\omega_2) = 0.3$$

将“出现噪声”这一事件设为 x 。当 $x = 1$ 时,表示出现了噪声;当 $x = 0$ 时,表示没有出现噪声。根据题意可知,“出现噪声”的类条件概率密度为

$$p(x = 1 | \omega_1) = 0.1, p(x = 1 | \omega_2) = 0.8$$

利用贝叶斯公式,分别计算出“出现噪声”时的后验概率:

$$P(\omega_1 | x = 1) = \frac{p(x = 1 | \omega_1)P(\omega_1)}{\sum_{j=1}^2 p(x = 1 | \omega_j)P(\omega_j)} = \frac{0.7 \times 0.1}{0.7 \times 0.1 + 0.3 \times 0.8} \approx 0.226$$

$$P(\omega_2 | x = 1) = \frac{p(x = 1 | \omega_2)P(\omega_2)}{\sum_{j=1}^2 p(x = 1 | \omega_j)P(\omega_j)} = \frac{0.3 \times 0.8}{0.7 \times 0.1 + 0.3 \times 0.8} \approx 0.774$$

根据最小错误率贝叶斯决策规则: 因为 $P(\omega_2 | x) > P(\omega_1 | x)$, 所以 $x \in \omega_2$, 即当出现噪声时,将机器判别为异常状态。

3.2.2 最小风险贝叶斯决策

分析例 3.1 可以发现,最小错误率贝叶斯决策的分类结果取决于观测值对各类后验概率的大小,并没有考虑错误判断所带来的“风险”。在实际应用中,有时关心的不仅仅是错误率,还包括错误决策所带来的损失。显然,将机器的异常状态误判成正常工作状态所带来的风险与将机器的正常工作状态误判成异常状态所带来的风险是不同的,前者带来的损失是



更大的。因此,不同的决策所带来的损失是不同的,做决策时,要考虑错误决策所带来的损失。最小风险贝叶斯决策正是考虑各种分类错误引起的损失而提出的一种决策规则。

假定有 c 类样本,用 $\omega_j (j = 1, 2, \dots, c)$ 表示类别,用 $\alpha_i (i = 1, 2, \dots, k)$ 表示可以做出的决策。在实际应用中,有些样本可能不能决策其属于任何一类,有时也可以在决策时把几类样本合并成同一个大类,所以 k 不一定等于 c 。对于给定样本 $\mathbf{x}$, 令损失函数 $\lambda(\alpha_i, \omega_j)$ 表示对于 ω_j 类的样本 $\mathbf{x}$, 采取决策 α_i 所带来的损失。每个决策都会带来一定的损失,它是由样本的真实状态 ω_j 和决策 α_i 来决定的。 $\lambda(\alpha_i, \omega_j)$ 可以用表格的形式给出,即决策表,如表 3-1 所示。

表 3-1 损失函数 $\lambda(\alpha_i, \omega_j)$ 的决策表

决 策	类 型			
	ω_1	ω_2	$\dots$	ω_c
α_1	$\lambda(\alpha_1, \omega_1)$	$\lambda(\alpha_1, \omega_2)$	$\dots$	$\lambda(\alpha_1, \omega_c)$
α_2	$\lambda(\alpha_2, \omega_1)$	$\lambda(\alpha_2, \omega_2)$	$\dots$	$\lambda(\alpha_2, \omega_c)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
α_k	$\lambda(\alpha_k, \omega_1)$	$\lambda(\alpha_k, \omega_2)$	$\dots$	$\lambda(\alpha_k, \omega_c)$

给定样本 $\mathbf{x}$, 假设它的各个状态的后验概率 $P(\omega_j | \mathbf{x}) (j = 1, 2, \dots, c)$ 已经确定。在利用最小风险贝叶斯决策方法进行判决时,由于引入了风险的概念,就不能只根据后验概率的大小进行决策,而必须考虑所采取的决策是否使损失最小。对于样本 $\mathbf{x}$, 如果采取决策 $\alpha_i (i = 1, 2, \dots, k)$, 从表 3-1 所示的决策表可以看出其对应 c 个 $\lambda(\alpha_i, \omega_j)$ 。因此,采取决策 α_i 的期望损失(条件风险)可定义为

$$R(\alpha_i | \mathbf{x}) = E[\lambda(\alpha_i, \omega_j) | \mathbf{x}] = \sum_{j=1}^c \lambda(\alpha_i, \omega_j) P(\omega_j | \mathbf{x}) \quad (3-14)$$

由式(3-14)可以看出,对应 $\mathbf{x}$ 的不同取值,采取决策 α_i 时的条件风险值是不同的。所以,最终采取哪种决策是随 $\mathbf{x}$ 的取值而定的。可将决策 α 看成随 $\mathbf{x}$ 变化的函数,记为 $\alpha(\mathbf{x})$ 。因此,对特征空间中所有可能的样本 $\mathbf{x}$ 采取决策所造成的期望损失(期望风险)可由式(3-15)计算:

$$R(\alpha) = \int R(\alpha(\mathbf{x}) | \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (3-15)$$

其中,积分运算是在整个特征空间进行的。期望风险 $R(\alpha)$ 表示的是对整个特征空间上所有 $\mathbf{x}$ 的取值采取决策 $\alpha(\mathbf{x})$ 所带来的损失;而条件风险 $R(\alpha_i | \mathbf{x})$ 表示对特定的某一样本 $\mathbf{x}$ 采取决策 α_i 所带来的损失。显然,进行决策时,希望采取一系列的决策 $\alpha(\mathbf{x})$ 的期望风险是最小的。如果采取每个决策时都使其条件风险最小,则对所有 $\mathbf{x}$ 做出决策时,其期望风险也必然是最小的。因此,最小风险贝叶斯决策就是最小化期望风险,即

$$\min_{\alpha} R(\alpha) \quad (3-16)$$

要使期望风险 $R(\alpha)$ 最小,就是要对所有的 $\mathbf{x}$ 使得 $R(\alpha(\mathbf{x}) | \mathbf{x})$ 最小。所以,最小风险贝叶斯决策规则可以表述为



$$\text{若 } R(\alpha_i | \mathbf{x}) = \min_{j=1,2,\dots,k} \{R(\alpha_j | \mathbf{x})\}, \text{ 则 } \alpha = \alpha_i \quad (3-17)$$

在已知先验概率和类条件概率密度的条件下,对样本 $\mathbf{x}$,最小风险贝叶斯决策的步骤可总结如下。

- (1) 利用贝叶斯公式计算后验概率 $P(\omega_j | \mathbf{x}), j = 1, 2, \dots, c$;
- (2) 由决策表,利用式(3-14)计算条件风险 $R(\alpha_i | \mathbf{x}), i = 1, 2, \dots, k$;
- (3) 根据决策规则(3-17),选择风险最小的决策。

例 3.2 在例 3.1 给出条件的基础上,利用表 3-2 所示的决策表,按照最小风险贝叶斯决策进行分类。

表 3-2 决策表

决 策	类 型	
	ω_1	ω_2
α_1	0	300
α_2	50	10

解: 已知条件为

$$\begin{aligned} P(\omega_1) &= 0.7, \quad P(\omega_2) = 0.3 \\ p(x=1 | \omega_1) &= 0.1, \quad p(x=1 | \omega_2) = 0.8 \\ \lambda_{11} &= 0, \quad \lambda_{12} = 300 \\ \lambda_{21} &= 50, \quad \lambda_{22} = 10 \end{aligned}$$

根据例 3.1 的计算结果,可知后验概率为

$$P(\omega_1 | \mathbf{x}) = 0.226, \quad P(\omega_2 | \mathbf{x}) = 0.774$$

根据式(3-14)计算条件风险,可得

$$\begin{aligned} R(\alpha_1 | \mathbf{x}) &= \sum_{j=1}^2 \lambda(\alpha_1, \omega_j) P(\omega_j | \mathbf{x}) = 0 \times 0.226 + 300 \times 0.774 = 232.2 \\ R(\alpha_2 | \mathbf{x}) &= \sum_{j=1}^2 \lambda(\alpha_2, \omega_j) P(\omega_j | \mathbf{x}) = 50 \times 0.226 + 10 \times 0.774 = 19.04 \end{aligned}$$

由于 $R(\alpha_1 | \mathbf{x}) > R(\alpha_2 | \mathbf{x})$, 所以决策为 ω_1 带来的风险大于决策为 ω_2 带来的风险。即当出现噪声时,将机器归为异常状态类别是风险最小的策略。

这里需要说明的是,在实际应用中,决策表通常是系统设计者和领域专家共同协商设定的。

下面讨论最小错误率贝叶斯决策和最小风险贝叶斯决策之间的关系。若损失函数 $\lambda(\alpha_i, \omega_j)$ 是一个 0-1 函数,且决策数目与类别数目一致,即

$$\lambda(\alpha_i, \omega_j) = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases} \quad (3-18)$$

式(3-18)表明,做出正确决策时损失为 0,而做出错误决策时损失为 1。这时条件风险变为

$$R(\alpha_i | \mathbf{x}) = \sum_{j=1}^c \lambda(\alpha_i, \omega_j) P(\omega_j | \mathbf{x}) = \sum_{j=1, j \neq i}^c P(\omega_j | \mathbf{x}) = 1 - P(\omega_i | \mathbf{x}) \quad (3-19)$$



对于样本 x ，基于最小风险贝叶斯决策方法的决策规则要求选择使条件风险最小的决策 α_i 。由式(3-19)可知，要使条件风险 $R(\alpha_i | x)$ 最小，就是要使后验概率 $P(\omega_i | x)$ 最大。而基于最小错误率贝叶斯决策方法的决策规则就是将样本归为具有最大后验概率的一类。因此，在 0-1 损失函数的情况下，最小风险贝叶斯决策方法的结果与最小错误率贝叶斯决策方法的结果相同，即两种方法此时是等价的。

3.3 基于正态分布的最小错误率贝叶斯分类器

本节介绍正态分布概率模型下的最小错误率贝叶斯分类器。正态分布也称作高斯分布，之所以引入正态分布来解决问题，是因为对大量实际数据来说，正态分布假设是一种合理的近似。另外，正态分布在数学上具有很多好的性质，便于计算分析。

对于最小错误率贝叶斯决策，判别函数可以定义为

$$g_i(x) = p(x | \omega_i)P(\omega_i), \quad i = 1, 2, \dots, c \quad (3-20)$$

其决策规则为

$$\text{若 } g_i(x) > g_j(x), \quad i = 1, 2, \dots, c, \quad j \neq i, \text{ 则 } x \in \omega_i \quad (3-21)$$

决策面方程为

$$g_i(x) > g_j(x) \quad (3-22)$$

设 x 为维数为 n 的特征向量，且 $p(x | \omega_i)$ 服从正态分布，即

$$p(x | \omega_i) \sim N(\mu_i, \Sigma_i) \quad (3-23)$$

其中， μ_i 是 ω_i 类样本的均值向量， Σ_i 是 ω_i 类样本的协方差矩阵，则 ω_i 类的判别函数可表示为

$$g_i(x) = \frac{1}{(2\pi)^{n/2} |\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)\right] P(\omega_i) \quad (3-24)$$

其中， Σ_i^{-1} 是协方差矩阵 Σ_i 的逆矩阵， $|\Sigma_i|$ 是协方差矩阵 Σ_i 对应行列式的值。为了便于计算，两边取对数，可得判别函数的等价形式：

$$g_i(x) = -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) - \frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i) \quad (3-25)$$

其中， $\frac{n}{2} \ln 2\pi$ 与类别无关，不影响分类决策，可以去掉。因此，式(3-25)可以进一步简化为

$$g_i(x) = -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i) \quad (3-26)$$

将式(3-26)代入式(3-22)，可得决策面方程为

$$-\frac{1}{2} [(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) - (x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j)] - \frac{1}{2} \ln \frac{|\Sigma_i|}{|\Sigma_j|} + \ln \frac{P(\omega_i)}{P(\omega_j)} = 0 \quad (3-27)$$

为了进一步理解多元正态分布下的判别函数和决策面，下面分几种情况讨论。

1. $\Sigma_i = \sigma^2 I$

这种情况意味着每一类的协方差矩阵都是相等的，且类内各特征维度间相互独立，具有



相同的方差。从几何形状上看,相当于各个样本都集中在以该类的均值 μ_i 点为中心的同等大小和形状的超球体内。此时,协方差矩阵 Σ_i 为

$$\Sigma_i = \sigma^2 \mathbf{I} = \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix} \quad (3-28)$$

进一步简化式(3-26),去掉与类别无关的项 $-\frac{1}{2} \ln |\Sigma_i|$, 判别函数可表示为

$$g_i(\mathbf{x}) = -\frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) + \ln P(\omega_i) \quad (3-29)$$

将式(3-28)代入式(3-29),可得

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2} (\mathbf{x} - \mu_i)^T (\mathbf{x} - \mu_i) + \ln P(\omega_i) \quad (3-30)$$

其中, $(\mathbf{x} - \mu_i)^T (\mathbf{x} - \mu_i)$ 为 $\mathbf{x}$ 到类 ω_i 的均值向量 μ_i 的欧氏距离的平方,即

$$(\mathbf{x} - \mu_i)^T (\mathbf{x} - \mu_i) = \|\mathbf{x} - \mu_i\|^2 \quad (3-31)$$

对于待识别样本 $\mathbf{x}$, 决策规则为

$$\text{若 } g_k(\mathbf{x}) = \max_i g_i(\mathbf{x}), \quad \text{则 } \mathbf{x} \in \omega_k \quad (3-32)$$

其中, $i = 1, 2, \dots, c$ 。相应的决策面方程为

$$g_i(\mathbf{x}) = g_j(\mathbf{x}) \quad (3-33)$$

下面再分两种情况讨论。

(1) c 类的先验概率不等,即 $P(\omega_i) \neq P(\omega_j)$ 。

当 c 类的先验概率 $P(\omega_i), i = 1, 2, \dots, c$ 相互之间不相等时,进一步化简如式(3-30)所示的判别函数,并忽略与类别无关的项,可得

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2} (-2\mu_i^T \mathbf{x} + \mu_i^T \mu_i) + \ln P(\omega_i) = \mathbf{w}_i^T \mathbf{x} + w_{i0} \quad (3-34)$$

其中, $\mathbf{w}_i = \frac{1}{\sigma^2} \mu_i, w_{i0} = -\frac{1}{2\sigma^2} \mu_i^T \mu_i + \ln P(\omega_i)$ 。由式(3-32)可以看出,判别函数 $g_i(\mathbf{x})$ 是关于 $\mathbf{x}$ 的线性函数,所以此时最小错误率贝叶斯分类器为线性分类器。

将式(3-34)代入式(3-33),决策面方程可表示为

$$\mathbf{w}^T (\mathbf{x} - \mathbf{x}_0) = 0 \quad (3-35)$$

其中, $\mathbf{w} = \mu_i - \mu_j, \mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2(\mu_i - \mu_j)}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)}$ 。由式(3-35)可知, $\mathbf{w}$ 是由点 μ_j 到点 μ_i 的向量,决策面通过 $\mathbf{x}_0$ 点,且与向量 $\mathbf{w}$ 正交。由于 $\Sigma_i = \sigma^2 \mathbf{I}$, 特征向量之间的协方差都为 0,所以此时等概率面皆为超球体。当不同类的先验概率不同时,其决策界面将会远离先验概率较大的均值点。如图 3-2 所示,对于两分类问题, $P(\omega_2) > P(\omega_1)$, 决策面将会远离 ω_2 类的均值点 μ_2 。

(2) c 类的先验概率相等,即 $P(\omega_i) = P(\omega_j)$ 。

当 c 类的先验概率 $P(\omega_i), i = 1, 2, \dots, c$ 都相等时,则可以去除判别函数式(3-30)中的 $\ln P(\omega_i)$ 项,进一步化简为

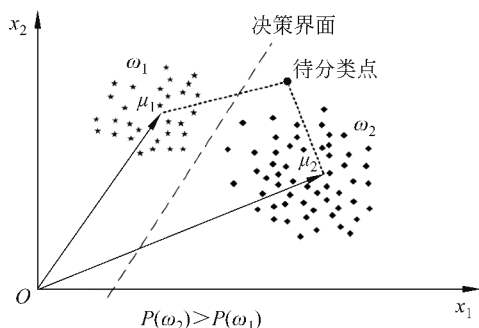


图 3-2 先验概率不同时的决策面

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2} (\mathbf{x} - \boldsymbol{\mu}_i)^T (\mathbf{x} - \boldsymbol{\mu}_i) \quad (3-36)$$

进一步化简,可得

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} \quad (3-37)$$

由式(3-37)可以看出,此时分类决策的结果只与每类的均值 $\boldsymbol{\mu}_i$ 有关。在对样本 $\mathbf{x}$ 进行分类时,只要计算样本 $\mathbf{x}$ 到各类均值 $\boldsymbol{\mu}_i$ 的欧氏距离的平方 $\|\mathbf{x} - \boldsymbol{\mu}_i\|^2$, 并把样本 $\mathbf{x}$ 归到具有 $\min_{i=1,2,\dots,c} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$ 的类。因此,这种分类器又叫作“最小距离分类器”。

此时,决策面方程为

$$\mathbf{w}^T (\mathbf{x} - \mathbf{x}_0) = 0 \quad (3-38)$$

其中, $\mathbf{w} = \mathbf{u}_i - \mathbf{u}_j$, $\mathbf{x}_0 = 1/2(\mathbf{u}_i + \mathbf{u}_j)$, 由式(3-38)可以看出,决策面通过点 $\boldsymbol{\mu}_i$ 和点 $\boldsymbol{\mu}_j$ 连线的中点 $\mathbf{x}_0$, 并与点 $\boldsymbol{\mu}_i$ 和点 $\boldsymbol{\mu}_j$ 的连线正交。如图 3-3 所示,对于两分类问题,当 $P(\omega_1) = P(\omega_2)$ 时,决策面为 $\boldsymbol{\mu}_1$ 和 $\boldsymbol{\mu}_2$ 连线的垂直平分线。

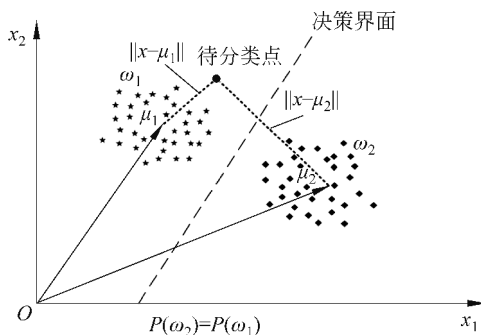


图 3-3 先验概率相同时的决策面

2. $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$

在这种情况下,每类的协方差矩阵均相等。从几何形状上看,相当于各个样本都集中在以该类的均值 $\boldsymbol{\mu}_i$ 点为中心的同等大小和形状的超椭圆体内。

由于 $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \dots = \boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$, 所以 $\boldsymbol{\Sigma}_i$ 与类别无关,判别函数式(3-26)可化简为

$$g(\mathbf{x}) = -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) + \ln P(\omega_i) \quad (3-39)$$



同样,下面分两种情况进行讨论。

(1) c 类的先验概率不等,即 $P(\omega_i) \neq P(\omega_j)$ 。

当 c 类的先验概率 $P(\omega_i), i=1,2,\dots,c$ 相互之间不相等时,决策面方程式(3-27)可化简为

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0 \quad (3-40)$$

$$\text{其中, } \mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j), \mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln \frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}。$$

由式(3-40)可知,决策面通过 $\mathbf{x}_0$ 点,且与向量 $\mathbf{w}$ 正交。当不同类的先验概率不同时, $\mathbf{x}_0$ 点不是点 $\boldsymbol{\mu}_i$ 和点 $\boldsymbol{\mu}_j$ 连线的中点。由于 $\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ 通常不在 $(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ 方向,所以决策面通常不与 $(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ 正交。如图 3-4 所示,对于两分类问题, $P(\omega_2) > P(\omega_1)$, 决策面不通过 $\boldsymbol{\mu}_1$ 和 $\boldsymbol{\mu}_2$ 连线的中点,而是远离先验概率较大的均值点 $\boldsymbol{\mu}_2$ 。

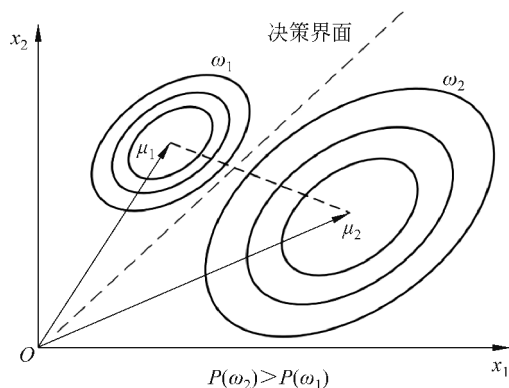


图 3-4 先验概率不同时的决策面

(2) c 类的先验概率相等,即 $P(\omega_i) = P(\omega_j)$ 。

当 c 类的先验概率 $P(\omega_i), i=1,2,\dots,c$ 都相等时,则可以去除判别函数式(3-39)中的 $\ln P(\omega_i)$ 项,进一步化简为

$$g_i(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_i)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) = \gamma^2 \quad (3-41)$$

其中, γ 表示马氏距离(Mahalanobis Distance)。在对样本 $\mathbf{x}$ 进行分类时,只需要计算出样本 $\mathbf{x}$ 到每一类均值点 $\boldsymbol{\mu}_i$ 的马氏距离的平方 γ^2 ,并将样本 $\mathbf{x}$ 归类为 γ^2 最小的类别。此时,决策面方程为

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0 \quad (3-42)$$

其中, $\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j), \mathbf{x}_0 = 1/2(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)$ 。由式(3-42)可以看出,决策面通过点 $\boldsymbol{\mu}_i$ 和点 $\boldsymbol{\mu}_j$ 连线的中点,并与向量 $\mathbf{w}$ 正交。如图 3-5 所示,对于两分类问题,当 $P(\omega_1) = P(\omega_2)$ 时,决策面过 $\boldsymbol{\mu}_1$ 和 $\boldsymbol{\mu}_2$ 的连线中点,但不与 $\boldsymbol{\mu}_1$ 和 $\boldsymbol{\mu}_2$ 的连线正交。

3. 各类协方差矩阵不相等

这种情况是多元正态分布的一般情况,即 $\mathbf{\Sigma}_i$ 任意,各类的协方差矩阵不相等。此时,判别函数为二次型函数,可表示为

$$g_i(\mathbf{x}) = \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0} \quad (3-43)$$

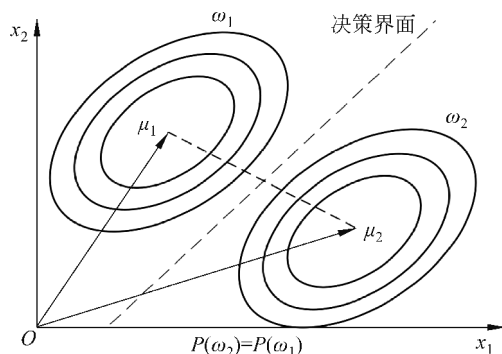


图 3-5 先验概率相同时的决策面

其中, $\mathbf{W}_i = -\frac{1}{2} \boldsymbol{\Sigma}_i^{-1}$, $\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i$, $w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$ 。

决策面方程为

$$\mathbf{x}^T (\mathbf{W}_i - \mathbf{W}_j) \mathbf{x} + (\mathbf{w}_i - \mathbf{w}_j)^T \mathbf{x} + w_{i0} - w_{j0} = 0 \quad (3-44)$$

由式(3-44)决定的决策面为超曲面,随着 $\boldsymbol{\Sigma}_i$ 、 $\boldsymbol{\mu}_i$ 、 $P(\omega_i)$ 的不同而呈现出不同的超二次曲面,如超球面、超椭球面、超抛物面、超双曲面或超平面。图 3-6 给出了二元正态分布下两类决策面的 5 种形式。在这 5 种形式中,假定变量 x_1 和变量 x_2 是类条件独立的,即协方差矩阵为对角阵;两类的先验概率是相等的,即不同的决策面仅仅是由于方差的不同引起的。

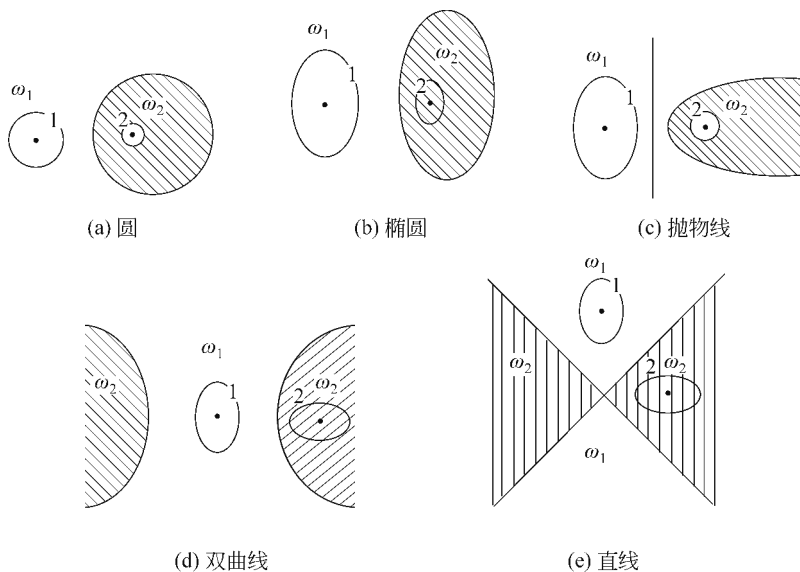


图 3-6 正态分布下几种决策面的形式

例 3.3 设有两类正态分布的样本,每类有 4 个样本:

$$\omega_1: [0, 0, 0]^T, [1, 1, 0]^T, [1, 0, 1]^T, [1, 0, 0]^T$$

$$\omega_2: [0, 0, 1]^T, [0, 1, 1]^T, [0, 1, 0]^T, [1, 1, 1]^T$$

若两类样本的先验概率相等,试求利用最小错误率贝叶斯决策进行分类的决策面方程。

解: 由已知条件可得



$$x_{11} = [0, 0, 0]^T, x_{12} = [1, 1, 0]^T$$

$$x_{13} = [1, 0, 1]^T, x_{14} = [1, 0, 0]^T$$

$$x_{21} = [0, 0, 1]^T, x_{22} = [0, 1, 1]^T$$

$$x_{23} = [0, 1, 0]^T, x_{24} = [1, 1, 1]^T$$

两类样本的均值向量为

$$\mu_1 = \frac{1}{4} \sum_{i=1}^4 x_{1i} = \frac{1}{4} [3, 1, 1]^T$$

$$\mu_2 = \frac{1}{4} \sum_{i=1}^4 x_{2i} = \frac{1}{4} [1, 3, 3]^T$$

由公式 $\Sigma_i = E[(x - \mu_i)(x - \mu_i)^T]$ 计算两类样本的协方差矩阵为

$$\Sigma_1 = \Sigma_2 = \frac{1}{16} \begin{bmatrix} 3 & 1 & 1 \\ 1 & 3 & -1 \\ 1 & -1 & 3 \end{bmatrix}$$

由已知条件和上述计算结果可知,两类样本的先验概率相同,且协方差矩阵相等,所以决策面方程为

$$w^T(x - x_0) = 0$$

其中, $w = \Sigma^{-1}(\mu_i - \mu_j)$, $x_0 = \frac{1}{2}(\mu_i + \mu_j)$ 。即

$$g_1(x) - g_2(x) = (\mu_1 - \mu_2)^T \Sigma_1^{-1} x - \frac{1}{2}(\mu_1^T \Sigma_1^{-1} \mu_1 - \mu_2^T \Sigma_1^{-1} \mu_2) = 0$$

经计算,可得决策面方程为

$$2x_1 - 2x_2 - 2x_3 + 1 = 0$$

3.4 朴素贝叶斯分类器

在实际应用中,通常需要利用样本的多个属性进行分类决策,而属性之间往往具有一定的关联性。因此在利用贝叶斯公式估计后验概率时,类条件概率密度是所有属性的联合概率,很难从有限的训练样本中直接估计得到。为了简化这个问题的求解,可以考虑采用属性条件独立性假设,即对所有已知类别的样本,假设所有的属性之间相互独立,即每个属性独立地对分类结果产生影响。在此属性条件独立性假设的基础上,利用贝叶斯决策理论进行分类,称为朴素贝叶斯分类器。尽管属性条件独立性假设相当苛刻,大多数实际分类系统中属性之间往往是互相联系、存在依赖的,并不满足此假设。但是,研究发现:在大多数情况下,尤其是应用于大型数据库时,朴素贝叶斯分类器表现出良好的性能。

假设有独立同分布的 c 类样本,每个样本 x 具有 d 维属性特征,即 $x = [x_1, x_2, \dots, x_d]$ 。样本的总个数为 n , 每类样本的个数为 $n_i (i = 1, 2, \dots, c)$, 则第 i 类样本的先验概率为

$$P(\omega_i) = \frac{n_i}{n} \quad (3-45)$$



对于离散属性的取值 x_j 来说,令 n_i^j 为第 i 类样本中属性取值为 x_j 的样本个数,则类条件概率可由式(3-46)估计出:

$$P(x_j | \omega_i) = \frac{n_i^j}{n_i} \quad (3-46)$$

对于连续属性,可考虑概率密度函数的取值。假定对于第 i 类样本,属性 x_j 服从正态分布,即 $\sigma_{i,j}^2 p(x_j | \omega_i) \cdot N(u_{i,j}, \sigma_{i,j}^2)$,则有属性 x_j 对应的类条件概率密度函数为

$$p(x_j | \omega_i) = \frac{1}{\sqrt{2\pi}\sigma_{i,j}} \exp\left(-\frac{(x_j - \mu_{i,j})^2}{2\sigma_{i,j}^2}\right) \quad (3-47)$$

其中, $\mu_{i,j}$ 和 $\sigma_{i,j}^2$ 分别为第 i 类样本在属性 x_j 上取值的均值和方差。

基于属性条件独立性假设,贝叶斯公式可改写为

$$P(\omega_i | x) = \frac{P(\omega_i) \prod_{j=1}^d p(x_j | \omega_i)}{p(x)}, \quad i = 1, 2, \dots, c \quad (3-48)$$

其中, d 为属性的数目, x_j 为 x 在第 j 个属性上的取值。

对于所有类别来说, $p(x)$ 是相同的,其与最后的分类结果无关。因此,朴素贝叶斯分类器的判别函数可定义为

$$g_i(x) = P(\omega_i) \prod_{j=1}^d P(x_j | \omega_i) \quad (3-49)$$

相应的决策规则可定义为

$$\text{若 } g_k(x) = \max_{i=1,2,\dots,c} g_i(x), \text{ 则 } x \in \omega_k \quad (3-50)$$

朴素贝叶斯分类器的步骤可总结如下。

- (1) 对每个类别,计算先验概率 $P(\omega_i), i = 1, 2, \dots, c$;
- (2) 对每个特征属性,计算其对应的类条件概率密度 $p(x_j | \omega_i), i = 1, 2, \dots, c$;
- (3) 利用式(3-49)计算每类样本对应的判别函数 $g_i(x)$;
- (4) 根据决策规则(3-50),对待识别样本进行分类。

例 3.4 现给定一个果树数据集,共有 2 个类别,其中规定苹果树为正例(+)、柿子树为反例(-)。数据集共有 8 个已知类别标号的样本(序号为 1~8),每个样本有 3 个属性特征,即:树干、果实和高度,且每种树的属性“高度”服从正态分布,样本集如表 3-3 所示。试用朴素贝叶斯分类器对序号为 9 的样本进行分类。

表 3-3 果树数据集

编号	树干	果实	高度/m	树种	编号	树干	果实	高度/m	树种
1	灰褐	青色	6.3	+	6	浅褐	青色	9.1	+
2	黑褐	青色	10.3	-	7	浅褐	橘色	12.5	-
3	黑褐	红色	4.8	+	8	灰褐	橘色	11.4	-
4	灰褐	红色	8.2	+	9	浅褐	青色	7.4	?
5	黑褐	橘色	8.6	-					



解：由表 3-3 可计算两类样本的先验概率分别为

$$P(\text{树种} = +) = \frac{4}{8} = 0.5$$

$$P(\text{树种} = -) = \frac{4}{8} = 0.5$$

由于每类样本都服从正态分布，可分别计算出两类样本属性“高度”所对应的类条件概率密度的均值和标准差。令 n_1 表示“树种 = +”的样本个数， n_2 表示“树种 = -”的样本个数。分析表 3-3 可知， $n_1 = 4, n_2 = 4$ 。令 x_i^+ 表示“树种 = +”的样本的“高度”属性值， x_i^- 表示“树种 = -”的样本的“高度”属性值，则当“树种 = +”时，其对应的均值和标准差分别为

$$\begin{aligned}\mu_2 &= \frac{1}{n_2} \sum_{i=1}^4 x_i^+ = \frac{1}{4} (10.3 + 8.6 + 12.5 + 11.4) = 10.7 \\ \sigma_2 &= \sqrt{\frac{1}{n_2 - 1} \sum_{i=1}^4 (x_i^- - \mu_2)^2} \\ &= \sqrt{\frac{1}{3} [(10.3 - 10.7)^2 + (8.6 - 10.7)^2 + (12.5 - 10.7)^2 + (11.4 - 10.7)^2]} \\ &\approx 1.663\end{aligned}$$

接着，计算每个属性的类条件概率密度：

$$P_{\text{浅褐}|+} = P(\text{树干} = \text{浅褐} | \text{树种} = +) = \frac{1}{4} = 0.25$$

$$P_{\text{浅褐}|-} = P(\text{树干} = \text{浅褐} | \text{树种} = -) = \frac{1}{4} = 0.25$$

$$P_{\text{青色}|+} = P(\text{果实} = \text{青色} | \text{树种} = +) = \frac{2}{4} = 0.5$$

$$P_{\text{青色}|-} = P(\text{果实} = \text{青色} | \text{树种} = -) = \frac{1}{4} = 0.25$$

$$\begin{aligned}p_{\text{高度}:7.4|+} &= P(\text{高度} = 7.4 | \text{树种} = +) \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(7.4 - \mu_1)^2}{2\sigma_1^2}\right) \\ &= \frac{1}{\sqrt{2\pi} \times 1.927} \exp\left(-\frac{(7.4 - 7.1)^2}{2 \times 1.927^2}\right) \approx 0.202\end{aligned}$$

$$\begin{aligned}p_{\text{高度}:7.4|-} &= P(\text{高度} = 7.4 | \text{树种} = -) \\ &= \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(7.4 - \mu_2)^2}{2\sigma_2^2}\right) \\ &= \frac{1}{\sqrt{2\pi} \times 1.663} \exp\left(-\frac{(7.4 - 10.7)^2}{2 \times 1.663^2}\right) \approx 0.009\end{aligned}$$

可以求得样本 x 属于每一类的后验概率为

$$P(+|x) = P(\text{树种} = +) \cdot P_{\text{浅褐}|+} \cdot P_{\text{青色}|+} \cdot p_{\text{高度}:7.4|+} \approx 0.013$$

$$P(-|x) = P(\text{树种} = -) \cdot P_{\text{浅褐}|-} \cdot P_{\text{青色}|-} \cdot p_{\text{高度}:7.4|-} \approx 2.81 \times 10^{-4}$$

由于 $P(+|x) > P(-|x)$ ，因此，应将序号为 9 的样本判别为苹果树(+)。



在例 3.4 中,测试样本中属性的取值应均在训练集的某一类样本中出现过。倘若出现了训练样本中从未出现的属性值,使用朴素贝叶斯分类器进行决策将出现问题。例如,如果测试样本中含有“果实=红色”的属性时,计算对应的条件概率:

$$P_{\text{红色}|-} = P(\text{果实} = \text{红色} | \text{树种} = -) = \frac{0}{4} = 0$$

在这种情况下,根据式(3-46)进行连乘的结果均为 0,该样本的其余属性无论取什么值,分类的结果都将是“树种=+”。这也是朴素贝叶斯分类器的局限所在。为了避免这一情况发生,在估计概率值时可以进行“平滑”处理。具体地,式(3-45)和式(3-46)可以分别修正为

$$\hat{P}(\omega_i) = \frac{n_i + 1}{n + c} \quad (3-51)$$

$$\hat{P}(x_j | \omega_i) = \frac{n_i^j + 1}{n_i + c_i} \quad (3-52)$$

其中, c 为样本的类别数, c_i 为第 i 个属性可能的取值数。

对于例 3.4,类先验概率可以估计为

$$\hat{P}(\text{树种} = +) = \frac{4 + 1}{8 + 2} = 0.5$$

$$\hat{P}(\text{树种} = -) = \frac{4 + 1}{8 + 2} = 0.5$$

由于果实的颜色有“青色”“红色”“橘色”三种取值,因此对于“果实=红色”属性,对应的条件概率可以估计为

$$\hat{P}_{\text{红色}|-} = \hat{P}(\text{果实} = \text{红色} | \text{树种} = -) = \frac{0 + 1}{4 + 3} \approx 0.143$$

由以上分析可知,通过引入先验知识,“平滑”处理可以避免某些条件概率为 0 的问题。尤其是当训练样本集增大时,“平滑”处理过程中引入的先验知识的影响会逐渐削弱,使得估计值逐步逼近真实值。

3.5 Python 实现

3.5.1 最小错误率贝叶斯决策和最小风险贝叶斯决策

假定某个局部区域细胞识别中正常 ω_1 和异常 ω_2 的两类先验概率分别为

$$P(\omega_1) = 0.9$$

$$P(\omega_2) = 0.1$$

现有一系列待识别细胞,其观察值为 x ,如表 3-4 所示。已知类条件概率密度函数 $P(X | \omega_1)$ 和 $P(X | \omega_2)$ 分别服从正态分布 $N(-2, 0.25)$ 、 $N(2, 4)$ 。条件风险决策表如表 3-5 所示。试对上述观察值 x 分别按照最小错误率贝叶斯决策和最小风险贝叶斯决策进行分类。

表 3-4 待识别细胞的观察值 x

-3.9847	-3.5549	-1.2401	-0.9780	-0.7932	-2.8531
---------	---------	---------	---------	---------	---------



续表

-2.7605	-3.7287	-3.5414	-2.2692	-3.4549	-3.0752
-3.9934	2.8792	0.9780	0.7932	1.1882	3.0682
-1.5799	-1.4885	-0.7431	-0.4221	-1.1186	4.2532

表 3-5 条件风险决策表

	ω_1	ω_2
α_1	0	6
α_2	1	0

程序代码及实验结果。

(1) 程序代码。

最小错误率贝叶斯程序如下。

```
import numpy as np
from scipy.stats import norm
import matplotlib.pyplot as plt
#输入数据
Data=[-3.9847, -3.5549, -1.2401, -0.9780, -0.7932, -2.8531,
-2.760, -3.7287, -3.5414, -2.2692, -3.4549, -3.0752,
-3.9934, 2.8792, -0.9780, 0.7932, 1.1882, 3.0682,
-1.5799, -1.4885, -0.7431, -0.4221, -1.1186, 4.2532]

#两个类型的先验概率
Pw1, Pw2=0.9, 0.1

#两个类型的类概率密度函数服从(u, sig^2)的正态分布,均值 u,标准差 sig
u1, sig1=-2, 0.5
u2, sig2=2, 2

#决策风险表
r11, r12=0, 6
r21, r22=1, 0

#为类型 1 和类型 2 开辟内存,以便后续使用
P1_save=[]
P2_save=[]

for x in Data:
#计算后验概率
#P(w1/x)=(p(x/w1)*p(w1))/p(x)=(p(x/w1)*p(w1))/((p(x/w1)*p(w1))+p(x/
#w2)*p(w2))
#P(w2/x)=(p(x/w2)*p(w2))/p(x)=(p(x/w2)*p(w2))/((p(x/w1)*p(w1))+p(x/
#w2)*p(w2))
P1=(norm.pdf(x, u1, sig1)*Pw1)/(norm.pdf(x, u1, sig1)*Pw1+norm.pdf(x,
u2, sig2)*Pw2)
```



```

P2=(norm.pdf(x, u2, sig2) * Pw2) / (norm.pdf(x, u1, sig1) * Pw1 + norm.pdf(x,
u2, sig2) * Pw2)

if P1 > P2:
    P1_save.append(x)
    print("观测数据:", x, "被分为类型 1 的后验概率为:", P1, "被分为类型 2 的后验概
率为:", P2)
    print("所以观测数据", x, "应该为类型 1")
else:
    P2_save.append(x)
    print("观测数据:", x, "被分为类型 1 的后验概率为:", P1, "被分为类型 2 的后验概
率为:", P2)
    print("所以观测数据", x, "应该为类型 2")

#在-5 到 5 之间以 0.01 步长采样进行画图
data_sample=np.arange(-5, 5, 0.01)
P1_sample=[[]for i in range(len(data_sample))]
P2_sample=[[]for i in range(len(data_sample))]
for i in range(0, len(data_sample)):

    P1=(norm.pdf(data_sample[i], u1, sig1) * Pw1) / (norm.pdf(data_sample[i],
u1, sig1) * Pw1 + norm.pdf(data_sample[i], u2, sig2) * Pw2)
    P2=(norm.pdf(data_sample[i], u2, sig2) * Pw2) / (norm.pdf(data_sample[i],
u1, sig1) * Pw1 + norm.pdf(data_sample[i], u2, sig2) * Pw2)
    P1_sample[i]=P1
    P2_sample[i]=P2

#使用 matplotlib 库函数进行画图
plt.plot(data_sample, P1_sample, "r", label="R1")
plt.plot(data_sample, P2_sample, "b", label="R2")
plt.plot(P1_save, len(P1_save) * [0], "r.", label="w1")
plt.plot(P2_save, len(P2_save) * [0], "b^", label="w2")
#设置曲线图的属性,不能使用中文
plt.xlabel("x")      #设置 x 轴的文字
plt.ylabel("R")      #设置 y 轴的文字
plt.legend()
plt.show()

```

最小风险贝叶斯程序如下。

```

import numpy as np
from scipy.stats import norm
import matplotlib.pyplot as plt

#输入数据
Data=[-3.9847, -3.5549, -1.2401, -0.9780, -0.7932, -2.8531,
      -2.760,  -3.7287, -3.5414, -2.2692, -3.4549, -3.0752,
      -3.9934,  2.8792, -0.9780,  0.7932,  1.1882,  3.0682,
      -1.5799, -1.4885, -0.7431, -0.4221, -1.1186,  4.2532]

#两个类型的先验概率

```



```
Pw1, Pw2 = 0.9, 0.1

# 两个类型的类概率密度函数服从(u, sig^2)的正态分布, 均值 u, 标准差 sig
u1, sig1 = -2, 0.5
u2, sig2 = 2, 2

# 决策风险表
r11, r12 = 0, 6
r21, r22 = 1, 0

# 为类型 1 和类型 2 开辟内存, 以便后续使用
R1_save = []
R2_save = []

for x in Data:
    # 计算后验概率
    #  $P(w1/x) = (p(x/w1) * p(w1)) / p(x) = (p(x/w1) * p(w1)) / ((p(x/w1) * p(w1)) + (p(x/w2) * p(w2)))$ 
    #  $P(w2/x) = (p(x/w2) * p(w2)) / p(x) = (p(x/w2) * p(w2)) / ((p(x/w1) * p(w1)) + (p(x/w2) * p(w2)))$ 
    P1 = (norm.pdf(x, u1, sig1) * Pw1) / (norm.pdf(x, u1, sig1) * Pw1 + norm.pdf(x, u2, sig2) * Pw2)
    P2 = (norm.pdf(x, u2, sig2) * Pw2) / (norm.pdf(x, u1, sig1) * Pw1 + norm.pdf(x, u2, sig2) * Pw2)
    # 计算决策风险
    #  $R(w1/x) = P(w1/x) * r11 + P(w2/x) * r12$ 
    #  $R(w2/x) = P(w1/x) * r21 + P(w2/x) * r22$ 
    R1 = P1 * r11 + P2 * r12
    R2 = P1 * r21 + P2 * r22
    if R1 > R2:
        R2_save.append(x)
        print("观测数据:", x, "被分为类型 1 的风险为:", R1, "被分为类型 2 的风险为:", R2)
    else:
        R1_save.append(x)
        print("观测数据:", x, "被分为类型 1 的风险为:", R1, "被分为类型 2 的风险为:", R2)

# 在-5 到 5 之间以 0.01 步长采样进行画图
data_sample = np.arange(-5, 5, 0.01)
R1_sample = [[] for i in range(len(data_sample))]
R2_sample = [[] for i in range(len(data_sample))]
for i in range(0, len(data_sample)):
    P1 = (norm.pdf(data_sample[i], u1, sig1) * Pw1) / (norm.pdf(data_sample[i], u1, sig1) * Pw1 + norm.pdf(data_sample[i], u2, sig2) * Pw2)
    P2 = (norm.pdf(data_sample[i], u2, sig2) * Pw2) / (norm.pdf(data_sample[i], u1, sig1) * Pw1 + norm.pdf(data_sample[i], u2, sig2) * Pw2)
    R1_sample[i] = P1 * r11 + P2 * r12
    R2_sample[i] = P1 * r21 + P2 * r22
```



```
#使用 matplotlib 库函数进行画图
plt.plot(data_sample, R1_sample, "r", label="R1")
plt.plot(data_sample, R2_sample, "b", label="R2")
plt.plot(R1_save, len(R1_save) * [0], "r.", label="w1")
plt.plot(R2_save, len(R2_save) * [0], "b^", label="w2")
#设置曲线图的属性,不能使用中文
plt.xlabel("x")      #设置 x 轴的文字
plt.ylabel("R")       #设置 y 轴的文字
plt.legend()
plt.show()
```

(2) 实验结果。

程序的运行结果如下。

① 最小错误率贝叶斯程序运行结果。

由于篇幅原因,这里只给出前五个样本的分类结果。

观测数据: -3.9847, 被分为类型 1 的后验概率为: 0.5455062299841593, 被分为类型 2 的后验概率为: 0.4544937700158406;

所以观测数据 -3.9847 应该为类型 1。

观测数据: -3.5549, 被分为类型 1 的后验概率为: 0.931195825156259, 被分为类型 2 的后验概率为: 0.068804174843741;

所以观测数据 -3.5549 应该为类型 1。

观测数据: -1.2401, 被分为类型 1 的后验概率为: 0.9768175399712223, 被分为类型 2 的后验概率为: 0.02318246002877768;

所以观测数据 -1.2401 应该为类型 1。

观测数据: -0.978, 被分为类型 1 的后验概率为: 0.9310615582911299, 被分为类型 2 的后验概率为: 0.06893844170887009;

所以观测数据 -0.978 应该为类型 1。

观测数据: -0.7932, 被分为类型 1 的后验概率为: 0.8383540747852113, 被分为类型 2 的后验概率为: 0.16164592521478868;

所以观测数据 -0.7932 应该为类型 1。

分类结果示意图如图 3-7 所示, 其中圆点表示该数据分类至 ω_1 类, 叉号表示该数据分类至 ω_2 类。R1 代表各点分为 ω_1 类的后验概率, R2 代表各点分为 ω_2 类的后验概率。

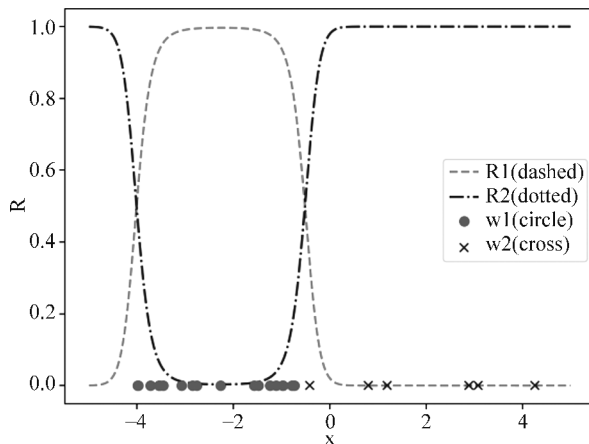


图 3-7 最小错误率分类结果