第4章 模式识别—
4.1贝叶斯决策

信控学院 蔡利梅

4.1.1贝叶斯决策的基本概念

- **原理**

用概率统计的方法研究随机模式的决策问题。

- **前提条件**

  - 各类别总体的概率分布是已知的；
  - 要决策的类别数是一定的。

- 概率

- 先验概率：预先已知的或者可以估计的模式识别系统位于某种类型的概率
- 类条件概率密度函数：系统位于某种类型条件下模式样本x出现的概率密度分布函数
- 后验概率：系统在某个具体的模式样本X条件下位于某种类型的概率

例：某地癌症发生率为千分之五，设正常为ω_1类，异常为ω_2类，$P(\omega_1)=99.5\%$，$P(\omega_2)=0.5\%$，这是先验概率

正常类中出现某种数据x的占1%，异常类中出现数据x的占95%，$p(x|\omega_1)=1\%$，$p(x|\omega_2)=95\%$，这是类条件概率

出现数据x的是正常类的占67.7%，出现数据x的是异常类的占32.3%，这是后验概率

贝叶斯公式

设试验E的样本空间为S，A为E的事件，$B_1, B_2, \ldots B_n$为S的一个划分，且$P(A)>0, P(B_i)>0$，下列公式为Bayes公式。

$$P(B_i \mid A)=\frac{P(A \mid B_i)\cdot P(B_i)}{\sum_{j=1}^{n} P(A \mid B_j)\cdot P(B_j)}, j=1,2,\ldots,n$$

功能在于将先验概率转化为后验概率

4.1.2最小错误率贝叶斯决策

希望在决策中尽量减少分类错误的概率，因此根据贝叶斯公式建立的**使错误率最小**的分类规则，称之为最小错误率贝叶斯决策。

（1）癌细胞识别实例分析一实例

有要进行识别的细胞，已经经过了预处理，抽取了n个表示细胞的特征，构成n维向量X，判断该细胞为正常或异常细胞。

- 数学表示

设正常细胞属于ω_1类，异常细胞为ω_2类，已知n维特征向量X，求$X \in \omega_1$ or $X \in \omega_2$

- 以往的统计数据
  - 先验概率$P(\omega_1)$和$P(\omega_2)$

根据先验的统计知识做出估计，如某一个地区癌症的发病率为5‰，即：$P(\omega_1)=0.995 \quad P(\omega_2)=0.005$

$P(\omega_1)>P(\omega_2)$只说明是正常细胞的可能性大，不能作为正常或异常的判据。

  - 类条件概率密度$p(X|\omega_1)$和$p(X|\omega_2)$

根据统计资料判断两类中X出现的概率。

假设特征向量 '$X=$阴或$X=$阳'，患者的试验反映为阳性的概率为0.95，正常人试验反映为阳性的概率为0.01，即 $p(X=\text{阳}|\omega_1)=1\%$，$p(X=\text{阳}|\omega_2)=95\%$

设实例中提取出的特征向量‘$X=$阳’，判断$X\in\omega_1$ or $X\in\omega_2$

- **利用Bayes公式求后验概率$P(\omega_i|X)$**

$$P(\omega_1\mid X=\text{阳})=\frac{p(X=\text{阳}\mid\omega_1)\cdot P(\omega_1)}{p(X=\text{阳}\mid\omega_1)\cdot P(\omega_1)+p(X=\text{阳}\mid\omega_2)\cdot P(\omega_2)}$$

$$=\frac{0.01\times0.995}{0.01\times0.995+0.95\times0.005}=0.677$$

$$P(\omega_2\mid X=\text{阳})=1-P(\omega_1\mid X=\text{阳})=0.323$$

- **分析**

根据后验概率，发现细胞不正常的可能性增大了。$P(\omega_1|X)>P(\omega_2|X)$所以判断该细胞为正常的。

实际中仅这个结论不能确诊的，需要更有效的化验。

（2）最小错误率贝叶斯决策规则

$$若 P(\omega_1 \mid X) \underset{<}{\overset{>}{}} P(\omega_2 \mid X)，则 X \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

$$若 P(\omega_1) \cdot p(X \mid \omega_1) \underset{<}{\overset{>}{}} P(\omega_2) \cdot p(X \mid \omega_2)，则 X \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

$$若 l(X) = \frac{p(X \mid \omega_1)}{p(X \mid \omega_2)} \underset{<}{\overset{>}{}} \frac{P(\omega_2)}{P(\omega_1)}，则 X \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

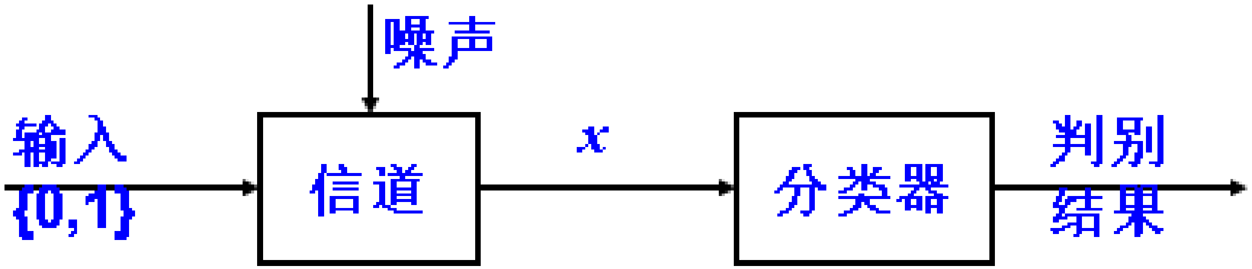
$$若 h(X) = -\ln[l(X)]$$

$$= -\ln p(X \mid \omega_1) + \ln p(X \mid \omega_2) \underset{>}{\overset{<}{}} \ln\left[\frac{P(\omega_1)}{P(\omega_2)}\right]，则 X \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

$$若 P(\omega_i \mid X) = \max P(\omega_j \mid X), j = 1,2,\ldots,c，则 X \in \omega_i$$

（3）实例

例：信号通过一受噪声干扰的信道，输入信号为0或1，噪声为高斯型，均值为0，方差为σ^2，信道输出为x。判断输出x是0还是1。



一般认为x<0.5判为0，x>0.5判为1

解：设送0为ω_1类，送1为ω_2类，送0的先验概率为$P(0)$，送1的先验概率为$P(1)$；

输入信号受正态分布噪声干扰，幅值的概率密度为：

$$p(x|\omega_1)=\frac{1}{\sqrt{2\pi}\sigma}e^{\left(-\frac{x^2}{2\sigma^2}\right)},\ p(x|\omega_2)=\frac{1}{\sqrt{2\pi}\sigma}e^{\left[-\frac{(x-1)^2}{2\sigma^2}\right]}$$

似然比：$l(x)=\dfrac{p(x|\omega_1)}{p(x|\omega_2)}=e^{\left[\frac{1-2x}{2\sigma^2}\right]}$

最小错误率贝叶斯决策：若 $e^{\left[\frac{1-2x}{2\sigma^2}\right]}\gtrless\dfrac{P(1)}{P(0)}$，则 $x\in\begin{cases}\omega_1\\ \omega_2\end{cases}$

假设 $P(0)=P(1)$，则决策变为：

$$若\ \frac{1-2x}{2\sigma^2}\gtrless 0,\ 即 x\lessgtr\frac{1}{2}时,\ x\in\begin{cases}\omega_1\\ \omega_2\end{cases},\ 即 x=\begin{cases}0\\ 1\end{cases}$$

例：地震预报是比较困难的一个课题，可以根据地震与生物异常反应之间的联系来进行研究。

- **根据历史记录的统计，地震前一周内出现生物异常反应的概率为50%，而一周内没有发生地震但也出现了生物异常反应的概率为10%。**
- **假设某一个地区属于地震高发区，发生地震的概率为20%。**
- **问：如果某日观察到明显的生物异常反应现象，是否应当预报一周内将发生地震？**

解：发生地震ω_1类，不发生为ω_2类。

由第二个条件得先验概率：$P(\omega_1)=0.2$和$P(\omega_2)=0.8$

设地震前一周出现生物异常反应为x，由第一个条件得类条件概率：$p(x|\omega_1)=0.5$，$p(x|\omega_2)=0.1$

观测到生物异常，发生地震的概率为：

$$P(\omega_1|x)=\frac{p(x|\omega_1)\cdot P(\omega_1)}{p(x|\omega_1)\cdot P(\omega_1)+p(x|\omega_2)\cdot P(\omega_2)}=\frac{0.5\times0.2}{0.5\times0.2+0.1\times0.8}=\frac{5}{9}$$

不发生地震的概率为： $P(\omega_2|x)=1-P(\omega_1|x)=\frac{4}{9}$

由于$P(\omega_1|x)>P(\omega_2|x)$，判别为要发生地震，预警

（4）验证错分概率

$$P(e)=\int_{-\infty}^{\infty} P(e|X)P(X)dX$$

$$P(e|X)=\begin{cases} P(\omega_1|X), 当P(\omega_2|X)>P(\omega_1|X) \\ P(\omega_2|X), 当P(\omega_1|X)>P(\omega_2|X) \end{cases}$$

对于所有的X值所进行的判断，错误率为最小，从而保证平均错误率P(e)也达到最小。

4.1.3最小风险贝叶斯决策

作出任何决策都有风险，都会带来一定的后果，错误率最小不一定风险也最小，因此，考虑分类错误引起的损失而产生**最小风险**的贝叶斯决策方法。

（1）问题表述

样本X为n维向量：$X=\begin{pmatrix}x_1 & x_2 & \cdots & x_n\end{pmatrix}^T$

状态空间由c个可能状态（类别）组成：$\Omega=\{\omega_1 \quad \omega_2 \quad \cdots \quad \omega_c\}$

对X可能采取的决策：$A=\{\alpha_1 \quad \alpha_2 \quad \cdots \quad \alpha_a\}$

做出某种决策α_i使风险最小

（2）风险定义

- 决策表

	α_1	α_2	…	α_i	…	α_a
ω_1	$\lambda(\omega_1,\alpha_1)$	$\lambda(\omega_1,\alpha_2)$		$\lambda(\omega_1,\alpha_i)$		$\lambda(\omega_1,\alpha_a)$
ω_2	$\lambda(\omega_2,\alpha_1)$	$\lambda(\omega_2,\alpha_2)$		$\lambda(\omega_2,\alpha_i)$		$\lambda(\omega_2,\alpha_a)$
…						
ω_j	$\lambda(\omega_j,\alpha_1)$	$\lambda(\omega_j,\alpha_2)$		$\lambda(\omega_j,\alpha_i)$		$\lambda(\omega_j,\alpha_a)$
…						
ω_c	$\lambda(\omega_c,\alpha_1)$	$\lambda(\omega_c,\alpha_2)$		$\lambda(\omega_c,\alpha_i)$		$\lambda(\omega_c,\alpha_a)$

经过分析研究统计得出。

条件风险

把一个特征向量X作出决策α_i时带来的**条件期望损失**$R(\alpha_i|X)$定义为条件风险

$$R(\alpha_i|X)=E[\lambda(\omega_j,\alpha_i)]$$
$$=\sum_{j=1}^{c}\lambda(\omega_j,\alpha_i)P(\omega_j|x), i=1,2,\dots,a$$

每一个X可以作出决策α_i时的风险大小不同，采用哪种决策将由X的取值而定，因此把决策$\alpha(X)$看成X的函数。

期望风险

对所有的X取值采取相应的决策时，所带来的平均风险

$$R=\int R[\alpha(X)|X]p(X)\mathrm{d}X$$

全概率

（3）决策规则

- **规则**

$$\text{若}R(\alpha_k \mid X) = \min R(\alpha_i \mid X),\ \text{则}\alpha = \alpha_k$$

- **步骤**
  - **求后验概率$P(\omega_j|X)$**
  - **利用决策表及后验概率求条件风险$R(\alpha_i|X)$**
  - **求最小条件风险$R(\alpha_k|X)$，确定作出决策α_k**

（4）例题

例：细胞识别，正常类ω_1，异常类ω_2，观察值为x，$P(\omega_1)$=0.9，$P(\omega_2)$=0.1，$p(x|\omega_1)$=0.2，$p(x|\omega_2)$=0.4，决策表如下，按最小风险贝叶斯决策进行分类。

	α_1	α_2
ω_1	0	1
ω_2	6	0

解：

- 后验概率

$$P(\omega_1|x) = 0.818$$
$$P(\omega_2|x) = 0.182$$

- 损失系数

$$\lambda_{11} = 0 \quad \lambda_{12} = 1$$
$$\lambda_{21} = 6 \quad \lambda_{22} = 0$$

- **条件风险**

$$R(\alpha_1 \mid x)=\sum_{j=1}^{2}\lambda_{j1}P(\omega_j \mid x)$$

$$=\lambda_{21}P(\omega_2 \mid x)=6\times 0.182=1.092$$

$$R(\alpha_2 \mid x)=\sum_{j=1}^{2}\lambda_{j2}P(\omega_j \mid x)=\lambda_{12}P(\omega_1 \mid x)=0.818$$

$\because R(\alpha_1 \mid x)>R(\alpha_2 \mid x)\quad \therefore$ 采取决策α_2，判断细胞异常

结果与按最小错误率贝叶斯决策进行分类截然相反，这是由于λ_{21}和λ_{12}相差太多，损失起了主导因素而造成。

例 2:对于前例中的地震预报问题，

- 假设预报一周内发生地震，可以预先组织抗震救灾，防灾成本会有2500万元，而当地震确实发生时，造成的直接损失会有1000 万元；
- 假设不预报将发生地震而地震又发生了，造成的损失会达到5000万元。
- 请问在观察到明显的生物异常反应后，是否应当预报一周内将发生地震？

解：发生地震ω_1类，不发生为ω_2类，决策α_1为发布地震预报，决策α_2为不发布地震预报，根据题目，得出损失系数λ。

- 发生地震，提前预报，损失系数$\lambda_{11} = 3500$；
- 发生地震，不预报， 损失系数$\lambda_{12} = 5000$；
- 不地震， 提前预报，损失系数$\lambda_{21} = 2500$；
- 不地震， 不预报， 损失系数$\lambda_{22} = 0$

	α_1	α_2
ω_1	3500	5000
ω_2	2500	0

观测到生物异常，发布地震预报的条件风险为：

$$R(\alpha_1 \mid x) = \sum_{j=1}^{2} \lambda_{j1} P(\omega_j \mid x) = 3500 \times \frac{5}{9} + 2500 \times \frac{4}{9} = 3056$$

不发布地震预报的条件风险为：

$$R(\alpha_2 \mid x) = \sum_{j=1}^{2} \lambda_{j2} P(\omega_j \mid x) = 5000 \times \frac{5}{9} = 2778$$

$\because R(\alpha_1|x) > R(\alpha_2|x)$ $\therefore$ 采取决策α_2，不发布地震预报

（5）最小错误率和最小风险两种决策的关系

- 0-1损失函数 $\lambda(\omega_j, \alpha_i) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$
- 条件风险 $R(\alpha_i|X) = \sum_{j=1}^{c} \lambda(\omega_j, \alpha_i) P(\omega_j|X) = \sum_{j=1, j\neq i}^{c} P(\omega_j|X)$

等式右边的求和表示将X归为ω_i的条件错误概率。

如$R(\alpha_k|X) = \min_i R(\alpha_i|X) = min \sum_{j=1,j\neq i}^{c} P(\omega_j|X)$

正好为求最小条件错误概率

最小错误率贝叶斯决策是在0-1损失函数条件下的最小风险贝叶斯决策，即前者是后者的特例。

思考：手写数字识别，采用贝叶斯决策，应该如何设计方案