实验 2：贝叶斯分类器的合成设计

1 简介

在感知器算法的学习中，我们试图寻找能将两类样本分开的超平面，这种算法对于两个线性可分的类别有很好的区分效果，但并不能给出结果的可靠程度，如果类别的数量变多且不是线性可分时，这种算法的结果就很难让人信服。在这种情况下，我们想知道样本属于每种类别的概率是多少，通过概率的大小来判断样本属于哪种类别，这个概率可以是根据之前判断结果得到先验概率，也可以是类别的条件概率密度函数，而贝叶斯分类器是通过结合这两者得到的后验概率，将样本分类到最可能的类别中。

2 贝叶斯决策原理

假设样本属于 m 个类别$\omega_1, \omega_2, \dots, \omega_m$，且每个类别的先验概率$P(\omega_i), i = 1,2, \dots, m$已知。先验概率是根据先前经验得到的每个类别发生的概率，和样本的特征无关，不能用于判断样本的类别。在每个类别中，样本特征服从某一概率密度函数$P(X|\omega_i)$，它表示的是在某一类别中某一随机特征向量出现的概率，因此也不能直接用来判别样本的类别。我们想要知道的是在已知某一样本特征X的条件下，这一样本属于类别i的概率$P(\omega_i|X)$，我们称之为后验概率。利用贝叶斯公式(1)我们可以得到后验概率的算法：

$$P(\omega_i|X) = \frac{P(X|\omega_i)P(\omega_i)}{P(X)} = \frac{P(X|\omega_i)P(\omega_i)}{\sum_{i=1}^{m} P(X|\omega_i)P(\omega_i)} \tag{1}$$

我们从所有类别的后验概率$P(\omega_i|X)$中选出最大的一项，那么这一类别就是这个样本最可能属于的类别。对于两种类别ω_1, ω_2的情况，分类规律可描述为：

如果$P(\omega_1|X) > P(\omega_2|X)$，则X属于类别 1；

如果$P(\omega_1|X) < P(\omega_2|X)$，则X属于类别 2；

对于多类情况，我们用a_k表示判断X属于类别ω_k，即：

$$a_k = \arg\max_i P(\omega_i|X) \tag{2}$$

注意到在所有的类别中$\mathrm{P(X)}$都是相同的，所以实际上我们只需要比较$\mathrm{P(X|\omega_i)P(\omega_i)}$的大小即可。

在这种分类规律下，判断错误的概率为：$\mathrm{P_e} = \sum_{k=1}^{m} P(\mathrm{a_k}, \mathrm{X} \in \omega_\mathrm{i}, i \neq k)$，可以证明，上述分类方法的判断错误概率为最小的，因此我们称这种分类方法为最小分类错误法。

在实际情况当中，每一个判断结果都会有一定的代价，且不同的结果往往有不同的代价，我们用λ_{ik}表示判别结果为k，但实际上样本X属于ω_i的代价，因此当我们做出判断$\mathrm{a_k}$时，产生的代价为$R(a_k|X) = \sum_{i=1}^{m} \lambda_{\mathrm{ik}} \mathrm{P}(\omega_\mathrm{i}|\mathrm{X})$，如果我们希望代价越小越好，那么我们可以得到样本的分类规律为：

$$\mathrm{a_k^*} = \arg\min_j R\left(a_j|X\right), \mathrm{j} = 1,2,\ldots,\mathrm{m} \tag{3}$$

在设计分类器的过程中，我们首先要知道每个类别里特征向量的概率密度函数$\mathrm{P(X|\omega_i)}$，很多情况下这个函数没有办法准确知道，但实践中大多数的概率密度函数都近似为正态密度函数，在样本数量足够多的情况下，这种假设的可信度较高，因此我们在设计贝叶斯分类器时也假定样本服从近似正态分布，概率密度函数与样本均值μ和方差矩阵Σ 有关。

在给定了每个类别的一组 p 维样本$\{X_1, X_2, \ldots, X_n\}$之后，我们可以得到$\mu$和$\Sigma$的极大似然估计为：

$$\hat{\mu} = E(X_i); \qquad \Sigma = E\left((X_i - \mu)^T (X_i - \mu)\right) \tag{4}$$

在得到了这两个参数之后，我们就能估计这一类别的概率密度函数：

$$\mathrm{P(X|\omega_i)} = \frac{1}{(2\pi)^{\frac{p}{2}} |\sum|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X-\mu)^T \sum^{-1} (X-\mu)\right) \tag{5}$$

对于非正态分布的样本也可以通过参数估计的方法找到其概率密度函数，但分类过程没有太大的区别，因此本文分析的样本特征均假设服从正态分布。

3 实验过程

我们首先对一维特征、两个类别的样本进行分类，进而推广得到多维特征、多个类别的分类方法。

3.1 一维特征、两个类别的样本分类

假设我们有两类细胞，其中正常细胞属于ω_1，异常细胞属于ω_2。在某一数据集中，我们得到了两种类别的样本特征，其中正常细胞的样本特征为：$\Omega_1 = \{-3.9847, -3.5549, -1.2401, -0.9780, -0.7932, -2.8531, -2.7605, -3.7287, -3.5414, -2.2692, -3.4549, -3.0752, -3.9934, -0.9780, -1.5799, -1.4885, -0.7431, -0.4221, -1.1186, -2.3462, -1.0826, -3.4196, -1.3193, -0.8367, -0.6579, -2.9683\}$，异常细胞的样本特征为$\Omega_2 = \{2.8792, 0.7932, 1.1882, 3.0682, 4.2532, 0.3271, 0.9846, 2.7648, 2.6588\}$，并且类别的先验概率已知$P(\omega_1) = 0.9$，$P(\omega_2) = 0.1$。当决策类别与实际类别不符时产生的代价如表 1 所示。

表 1　不同决策类别的代价

实际类别		ω_1	ω_2
决策类别	a_1	0	1
	a_2	6	0

首先对样本的均值和方差进行参数估计，得到$\widehat{\mu_1} = -2.1226$，$\widehat{\mu_2} = 2.1019$，$\widehat{\sigma_1} = 1.1912$，$\widehat{\sigma_2} = 1.2410$，再根据正态分布函数，我们得到如图 1 所示的概率密度曲线，其中红色实线是正常细胞的概率密度曲线，蓝色实线是异常细胞的概率密度曲线。

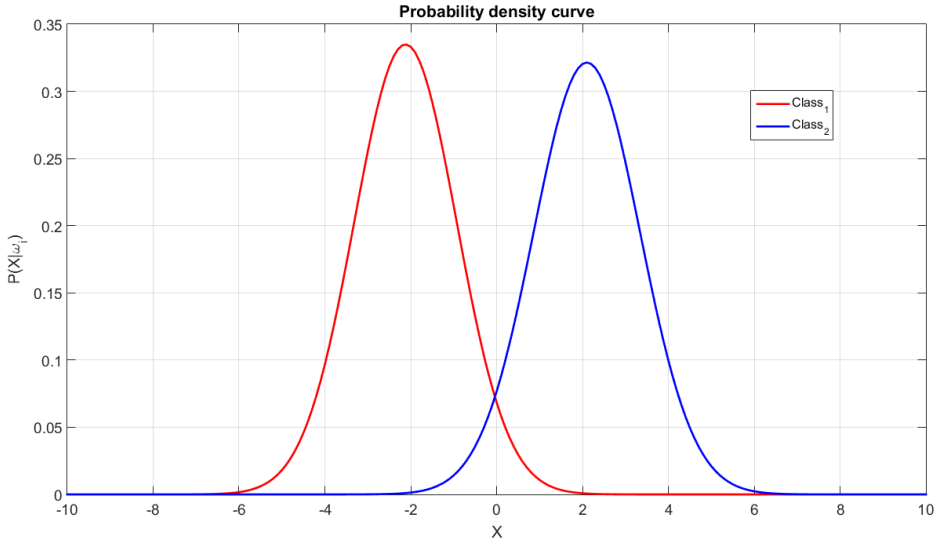

图 1　正常细胞和异常细胞的样本特征概率分布曲线

利用贝叶斯公式(1)，可得到给定样本特征X的条件下，判断样本类别的后验概率$P(\omega_i|X)$，令$G(X)$为分类函数，在最小分类错误法的定义下，分类函数：

$$G_{LE}(X) = P(\omega_1|X) - P(\omega_2|X) \tag{6}$$

当$G_{LE}(X) > 0$时，判断X属于类别 1 正常细胞；

当$G_{LE}(X) < 0$时，判断X属于类别 2 异常细胞；

如图 2 所示，给定细胞特征X的情况下，判别为正常细胞的后验概率曲线为红色实线，判别为异常细胞的后验概率曲线为蓝色实线，黑色虚线为分类函数曲线，当$X = 1.319$时，$G_{LE}(X) = 0$，因此称$X = 1.319$为最小错误法的决策面。

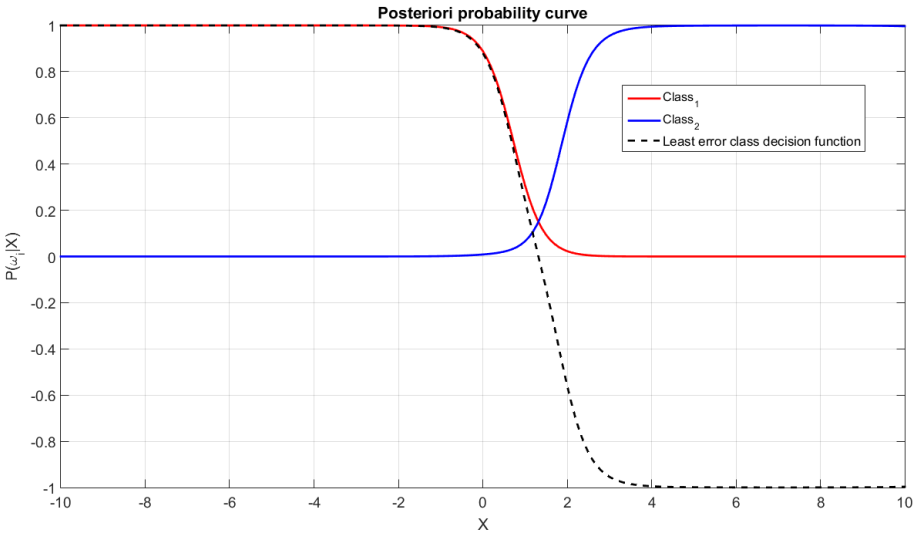

图 2　样本特征的后验概率分布曲线及最小错误法的判别曲线

以最小错误法的分类函数来判断给定数据集的样本，得到表 2 所示的结果。

表 2　最小错误法贝叶斯分类器判别结果

类别	正常细胞ω_1	异常细胞ω_2
样本总数	26	9
判别错误个数	0	4

当考虑每种决策的代价时，我们以最小代价作为判别依据，分类函数：

$$G_{LR}(X) = R(a_1|X) - R(a_2|X) = (\lambda_{11} - \lambda_{12})P(\omega_1|X) + (\lambda_{21} - \lambda_{22})P(\omega_2|X) \quad (7)$$

当$G_{LR}(X) < 0$时，判断X属于类别 1 正常细胞；

当$G_{LR}(X) > 0$时，判断X属于类别 2 异常细胞。

如图 3 所示，给定细胞特征X的情况下，红色实线为判断是正常细胞的代价函数曲线，蓝色实线为判断是异常细胞的代价函数曲线，黑色虚线为分类函数曲线，当$X = 1.68$时，$G_{LR}(X) = 0$，因此称$X = 1.68$为最小代价法的决策面。

以最小代价法的决策面来判断给定数据集的样本，得到表 3 所示的结果。

表 3　最小代价法贝叶斯分类器判别结果

类别	正常细胞ω_1	异常细胞ω_2
样本总数	26	9
判别错误个数	0	4
平均代价	0.1143	

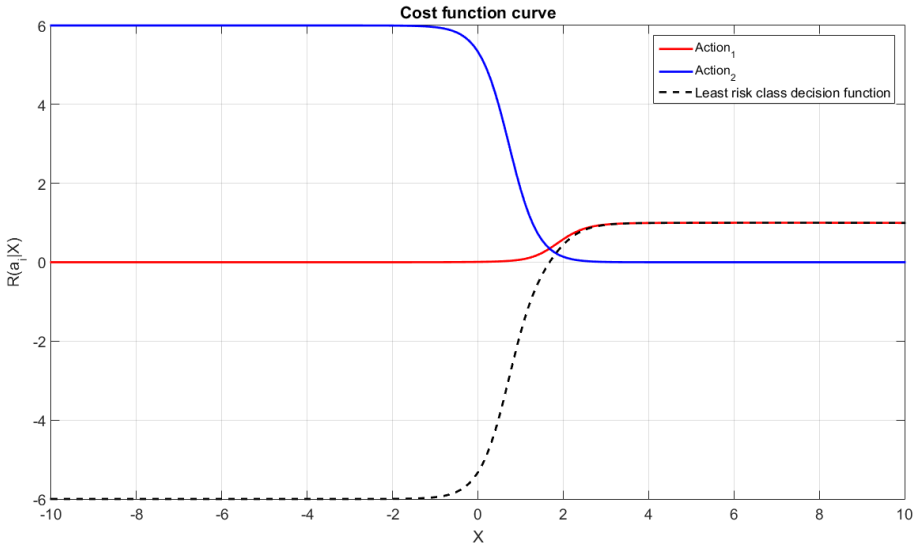

图 3　决策代价函数曲线及最小代价法的分类函数曲线

3.2 一维特征、多个类别的样本分类

以三个类别的样本为例进行分类，更多类别的样本分类方法与三类的区别较小，因此可以用类似方法处理。

假设人群样本特征为一维特征向量 X，其中正常人群属于类别 1 ω_1，心脏病患者属于类别 2 ω_2，癫痫患者属于类别 3 ω_3，并且$\{X|\omega_1\}$服从正态分布$\mathrm{N}(0,4)$；$\{X|\omega_2\}$服从正态分布$\mathrm{N}(-3,1)$；$\{X|\omega_3\}$服从正态分布$\mathrm{N}(3,1)$。根据经验，我们假设先验概率为$\mathrm{P}(\omega_1)=0.5;\mathrm{P}(\omega_2)=0.25;\mathrm{P}(\omega_3)=0.25$。这三类人群特征的概率密度函数如图 4 所示。

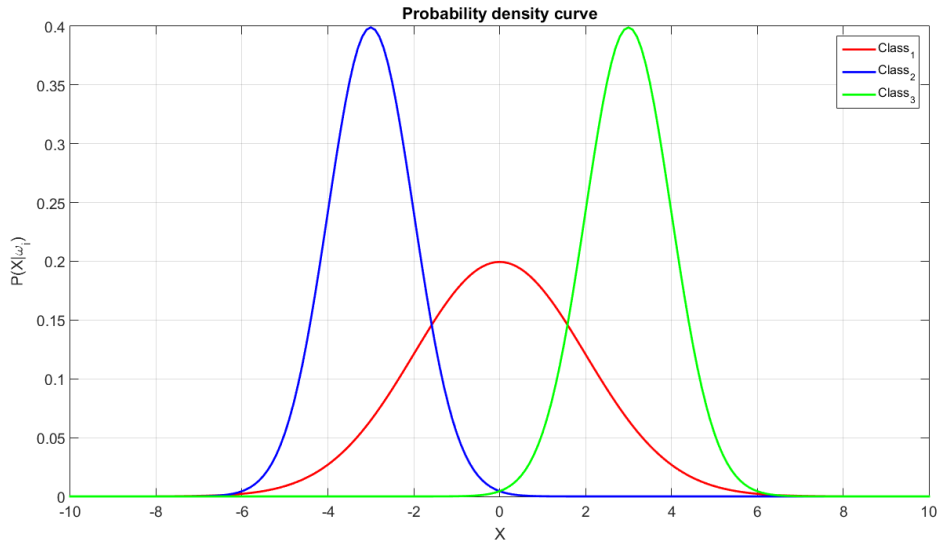

图 4　三类人群特征的概率分布曲线

现在需要根据已知人群特征X来判断人群类型，诊断结果会产生如表 2 所示的代价。

表 4　不同决策类别的代价

实际类别		ω_1	ω_2	ω_3
决策类别	a_1	0	6	4
	a_2	1	0	3
	a_3	1	4	0

与两个类别的判断方法不同的是，分类函数(6)和(7)只能将相邻的两种类别区分开来，但这种区分方式并不能直接得到最小错误法或最小代价法的决策面，这时应采用多类决策函数。基于最小错误法的决策函数为：$a_k = \arg\max_i P(\omega_i|X), i = 1,2,3$，$a_k$是在给定样本特征 X 的条件下，后验概率最大的决策类别；基于最小错误法的决策函数为：$a_k^* = \arg\min_j R(a_j|X), j = 1,2,3$，$a_k^*$是在给定样本特征 X 的条件下，代价最小的决策类别。

在给定样本特征 X 的条件下，三类人群的后验概率曲线和最小错误法的判别曲线如图 5 所示，图中黑色虚线表示的是最大后验概率曲线$\max_i P(\omega_i|X)$，从图中可以得到决策范围：$A_1 = [-10, -4.29) \cup (-2.85, 2.85) \cup (4.29,\ 10]$；$A_2 = (-4.29, -2.85)$；$A_3 = (2.85,\ 4.29)$，决策面分别是$X = -4.29, -2.85, 2.85, 4.29$。

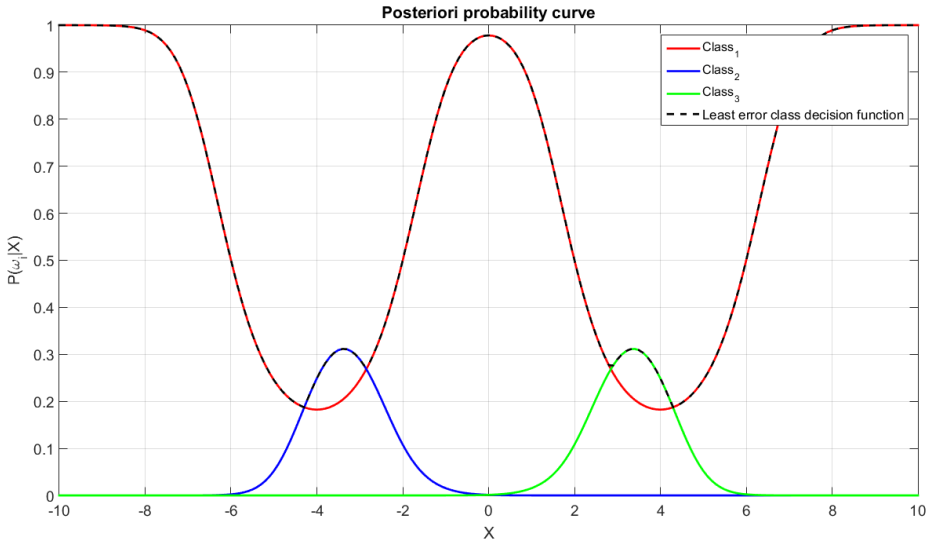

图 5　在样本特征 X 的条件下不同人群类别的后验概率曲线及最小错误法的决策函数曲线

在给定样本特征 X 的条件下，三种决策类别的代价函数曲线和最小代价法的判别曲线如图 6 所示，图中黑色虚线表示的是最小代价函数曲线$\min_j R(a_j|X)$，从图中可以得到决策范围：$A_1 = [-10, -5.08) \cup (-1.89, 2.06) \cup (4.95,\ 10]$；

$A_2 = (-5.08, -1.89)$; $A_3 = (2.06,\ 4.95)$，决策面分别是$X = -5.08,\ -1.89,\ 2.06,\ 4.95$。

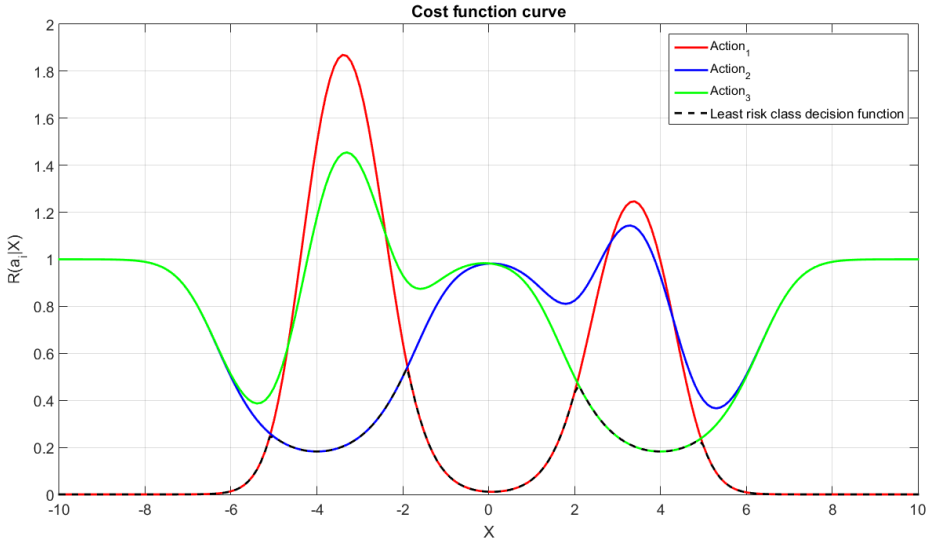

图 6　在样本特征 X 的条件下不同决策类别的代价函数曲线及最小代价法的决策函数曲线

随机生成三个类别人群的样本，每个类别各有 50 个样本，每个样本的特征在附件中给出，分别用最小错误法和最小代价法的决策面进行判断，得到表 5 所示结果。

表 5　贝叶斯分类器判别结果

样本类别	正常人群ω_1	心脏病患者ω_2	癫痫患者ω_3
样本总数	50	50	50
最小错误法判断错误个数	5	29	22
最小代价法判断错误个数	13	9	6
平均代价	0.6067		

3.3 多维特征、多个类别的样本分类

当样本特征从一维变到更高维时，分类方法没有明显区别，只是决策面从点变成了曲线和曲面，因为大于二维的特征很难直观地表示出来，所以这里只对二维特征进行分类，更高维的特征处理方法与二维类似。

还是以三类人群样本为例，这时我们增加一维独立的特征，假设正常人群ω_1的特征服从正态分布$N\left([0,0],\begin{bmatrix}4 & 0\\ 0 & 1\end{bmatrix}\right)$；心脏病患者$\omega_2$的特征服从正态分布$N\left([-3,-2],\begin{bmatrix}1 & 0\\ 0 & 2\end{bmatrix}\right)$；癫痫患者$\omega_3$的特征服从正态分布$N\left([3,2],\begin{bmatrix}1 & 0\\ 0 & 2\end{bmatrix}\right)$。先验概率仍为$P(\omega_1) = 0.5; P(\omega_2) = 0.25; P(\omega_3) = 0.25$。这三类人群特征的概率密度函数如图 7 所示，不同诊断结果产生的代价仍与表 4 中的数值相同。

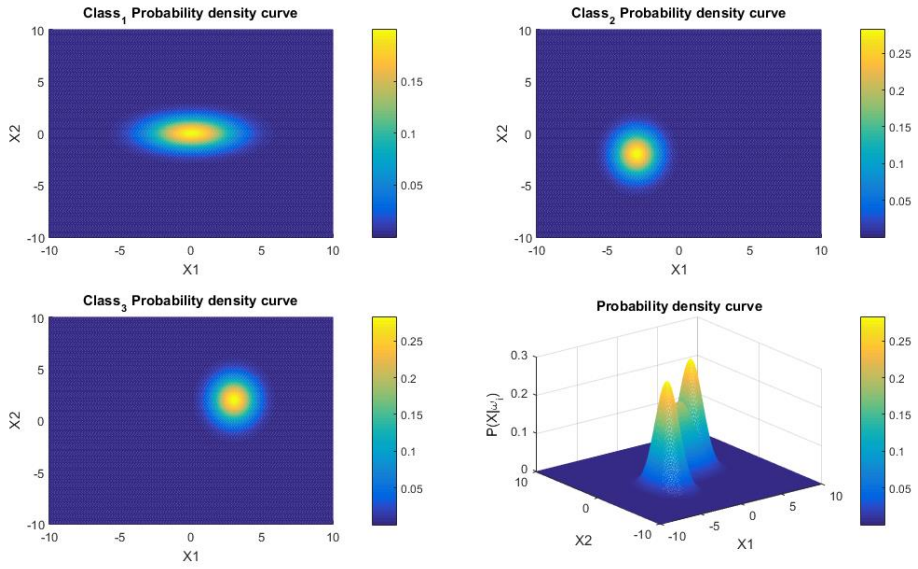

图 7　三类人群样本的二维特征概率密度函数曲面

分类方法与用在一维特征多类样本的方法相同，基于最小错误率法的分类器应首先计算在给定多维样本特征X的条件下每种类别的后验概率，这时的决策函数为：$a_k = \arg\max_i P(\omega_i|X), i = 1,2,3$。三类人群的后验概率和决策面如图 8 所示，在三类后验概率曲面中，颜色越亮的位置后验概率越高，在右下角的图中，样本特征在蓝色区域中时正常人群ω_1的后验概率最大，绿色区域中心脏病患者ω_2的后验概率最大，黄色区域中癫痫患者ω_3的后验概率最大。由于决策面曲线的方程较难精确获得，因此只能利用决策函数来判断样本的类别。

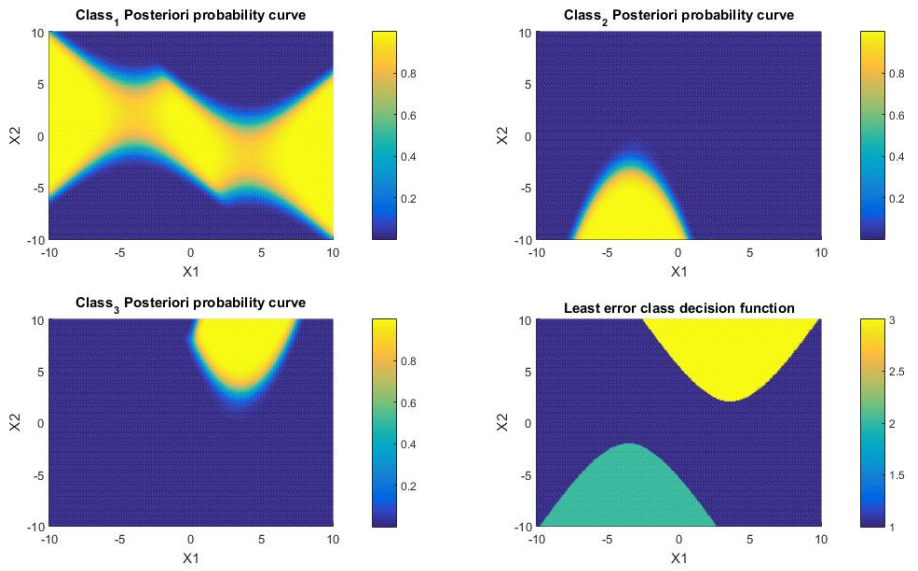

图 8　在样本特征给定时三类人群的后验概率曲面及最小错误法贝叶斯分类器决策面

基于最小代价法的分类器应首先计算在给定多维样本特征X的条件下每种决策的代价，这时的决策函数为：$\mathrm{a_k^*} = \arg\min_j R(a_j|X), \mathrm{j} = 1,2,3$，三类人群的后验概率和决策面如图 9 所示，在三类决策代价函数曲面中，颜色越暗的位置决策代价越低。

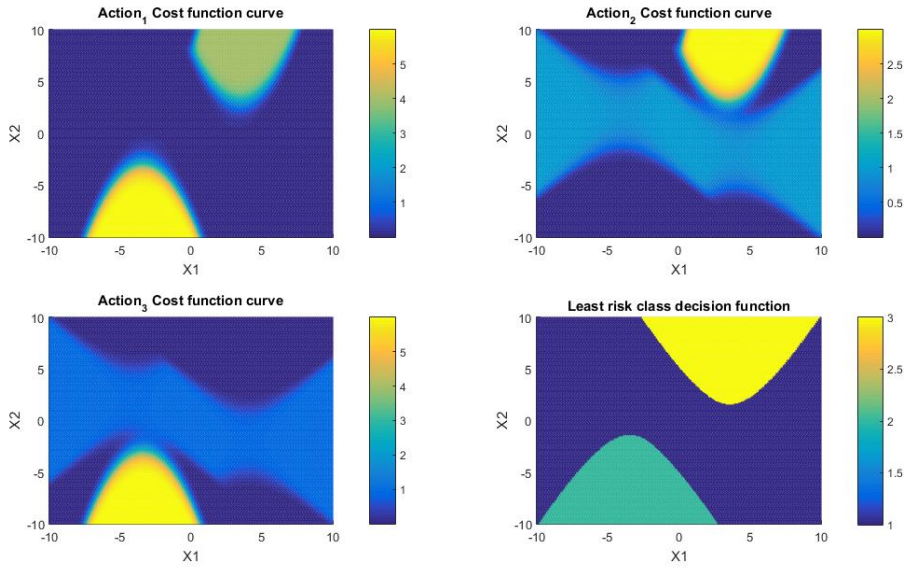

图 9　在样本特征给定时三种决策类型的代价函数曲面及最小代价法贝叶斯分类器决策面

随机生成三个类别人群的样本，每个类别各有 50 个样本，每个样本的特征在附件中给出，分别用最小错误法和最小代价法的决策函数进行判断，得到表 6 所示结果。

表 6 贝叶斯分类器判别结果

样本类别	正常人群ω_1	心脏病患者ω_2	癫痫患者ω_3
样本总数	50	50	50
最小错误法判断错误个数	0	23	15
最小代价法判断错误个数	2	13	12
平均代价	0.8533		

4　结论

本文通过实验分析了近似服从高斯分布的一维特征两类样本、一维特征三类样本、以及两维特征三类样本的后验概率和代价函数，并用基于最小错误率法和最小代价法的贝叶斯分类器对样本进行分类。在比较了这几种不同的数据情况之后，我们得到以下结论：

1. 对于多类别样本的分类器，通用贝叶斯决策函数为$a_k = \arg\max_i P(\omega_i|X)$（最小错误率）或$a_k^* = \arg\min_j R(a_j|X)$（最小代价），当样本仅有两种类别时，决策函数可以简化为$G_{LE}(X) = P(\omega_1|X) - P(\omega_i|X)$和$G_{LR}(X) = R(a_1|X) - R(a_2|X)$；
2. 对于多维特征的两类样本，当$G_{LE}(X) = 0$时可以得到决策面，当特征仅有一维时，决策面是一个特征点，可以很容易的判断样本的类别，当特征是多维时，则很难得到决策面的直接表示方式；
3. 从实验结果可以看出，当两种类型的样本特征差别不大时，贝叶斯分类器的仍然会出现很多判别错误，当样本特征维数增加时，错误率会有所降低，因此要想获得更高的正确率，还需要提取更多差别较大的特征。