00 年试题

一、答：五种常用决策是

1. 基于最小错误率的贝叶斯决策。利用概率论中的贝叶斯公式，得出使错误率为最小的分类规则。
2. 基于最小风险的贝叶斯决策。引入了损失函数，得出使决策风险最小的分类。当在 0-1 损失函数条件下，基于最小风险的贝叶斯决策变成基于最小错误率的贝叶斯决策。
3. 在限定一类错误率条件下使另一类错误率为最小的两类别决策。、
4. 最小最大决策。使最大可能的风险为最小的决策。
5. 序贯分类方法

二、答：

①

解：利用贝叶斯公式，分别计算出 ω_1 及 ω_2 的后验概率。

$$P(\omega_1|\boldsymbol{x})=\frac{p(\boldsymbol{x}|\omega_1)P(\omega_1)}{\sum_{j=1}^{2}p(\boldsymbol{x}|\omega_j)P(\omega_j)}=\frac{0.2\times0.9}{0.2\times0.9+0.4\times0.1}=0.818$$

$$P(\omega_2|\boldsymbol{x})=1-p(\omega_1|\boldsymbol{x})=0.182$$

根据贝叶斯决策规则式(2-2)，有

$$P(\omega_1|\boldsymbol{x})=0.818>P(\omega_2|\boldsymbol{x})=0.182$$

所以合理的决策是把 $\boldsymbol{x}$ 归类于正常状态。

②

解：已知条件为

$$P(\omega_1)=0.9,\qquad P(\omega_2)=0.1$$

$$p(\boldsymbol{x}|\omega_1)=0.2,\qquad p(\boldsymbol{x}|\omega_2)=0.4$$

$$\lambda_{11}=0,\qquad \lambda_{12}=6$$

$$\lambda_{21}=1,\qquad \lambda_{22}=0$$

根据例 2.1 的计算结果可知后验概率为

$$P(\omega_1|\boldsymbol{x})=0.818,\qquad P(\omega_2|\boldsymbol{x})=0.182$$

再按式(2-15)计算出条件风险

$$R(\alpha_1|\boldsymbol{x})=\sum_{j=1}^{2}\lambda_{1j}P(\omega_j|\boldsymbol{x})=\lambda_{12}P(\omega_2|\boldsymbol{x})=1.092$$

$$R(\alpha_2|\boldsymbol{x})=\lambda_{21}P(\omega_1|\boldsymbol{x})=0.818$$

由于 $R(\alpha_1|\boldsymbol{x})>R(\alpha_2|\boldsymbol{x})$

即决策为 ω_2 的条件风险小于决策为 ω_1 的条件风险，因此我们采取决策行动 α_2，即判断待识别的细胞 $\boldsymbol{x}$ 为 ω_2 类——异常细胞。

将本例与例 2.1 相对比，其分类结果正好相反，这是因为这里影响决策结果的因素又多了一个，即"损失"。而且两类错误决策所造成的损失相差很悬殊，因此"损失"就起了主导作用。

三、答:总体分布形式为:

$$p(x|\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}$$

最大似然方程:

$$\nabla_\mu H(\mu) = \sum_{k=1}^{N} \nabla_\mu \ln p(x|\mu) = 0$$

即:

$$\sum_{k=1}^{N} (x_k - \hat{\mu}) = 0$$

所以:

$$\hat{\mu} = \frac{1}{N} \sum_{k=1}^{N} x_k$$

四、答:

在实际中有很多模式识别问题并不是线性可分的,这时就需要采用非线性分类器。比如当两类样本分布具有多峰性质并互相交错时,简单的线性判别函数往往会带来较大的分类错误。这时,树分类器作为一种分段线性分类器,常常能有效地应用于这种情况。本章重点

五、答:

$$\mu_1 = \frac{1}{4}\begin{bmatrix}3\\1\\1\end{bmatrix}, \qquad \mu_2 = \frac{1}{4}\begin{bmatrix}1\\3\\3\end{bmatrix}, \qquad \mu = \frac{1}{2}\begin{bmatrix}1\\1\\1\end{bmatrix}$$

$$S_b = \frac{1}{2}\sum_{i=1}^{2}(\mu_i - \mu)(\mu_i - \mu)^T = \frac{1}{4}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T = \frac{1}{16}\begin{bmatrix}1 & -1 & 0\\-1 & 1 & 1\\-1 & 1 & 1\end{bmatrix}$$

$$S_w = \frac{1}{2}(\Sigma_1 + \Sigma_2) = \frac{1}{16}\begin{bmatrix}1 & -1 & 0\\-1 & 1 & 1\\-1 & 1 & 1\end{bmatrix}, \qquad S_w^{-1} = \begin{bmatrix}8 & -4 & -4\\-4 & 8 & 4\\-4 & 4 & 8\end{bmatrix}$$

因为$S_w^{-1}S_b$的秩为 1,所以$S_w^{-1}S_b$只有一个非零本征值,W 是 D×1 矩阵,即 W=w。为求$S_w^{-1}S_b$的本征值应解方程:

$$S_w^{-1}S_b w = \lambda_1 w$$

或

$$\frac{1}{4}S_w^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w = \lambda_1 w$$

$\frac{1}{4}(\mu_1 - \mu_2)^T w$是标量，所以

$$w = S_w^{-1}(\mu_1 - \mu_2) = \begin{bmatrix}8\\-8\\-8\end{bmatrix}$$

01 年

一、答：

1. 基于最小错误率的贝叶斯决策。利用概率论中的贝叶斯公式，得出使错误率为最小的分类规则。
2. 最小最大决策。使最大可能的风险为最小的决策。
3. 现再将问题转化为参数估计问题：设有一个样本集 $\mathscr{X}$（而不是一个 $\boldsymbol{x}$），要求我们找出估计量 $\hat{\theta}$（而不是选出最佳决策 α_k），用来估计 $\mathscr{X}$ 所属总体分布的某个真实参数 θ（而不是真实状态 ω_k）使带来的贝叶斯风险最小。这就是贝叶斯估计的概念。可见贝叶斯决策和贝叶斯估计两者都立足于使贝叶斯风险最小，只是要解决的问题不同而已。前者是要决策 $\boldsymbol{x}$ 的真实状态，而后者则是估计 $\mathscr{X}$ 所属总体分布的参数。这样，它们的各变量间就存在下表中一一对应的关系：
4. 我们可以考虑把 d 维空间的样本投影到一条直线上，形成一维空间，即把维数压缩到一维。这在数学上总是容易办到的。然而，即使样本在 d 维空间里形成若干紧凑的互相分得开的集群，若把它们投影到一条任意的直线上，也可能使几类样本混在一起而变得无法识别。但在一般情况下，总可以找到某个方向，使在这个方向的直线上，样本的投影能分开得最好。问题是如何根据实际情况找到这条最好的、最易于分类的投影线。这就是 Fisher 法所要解决的基本问题（见图 4.3）。
5. $$J_s(\boldsymbol{a}) = \|\boldsymbol{e}\|^2 = \|Y\boldsymbol{a} - \boldsymbol{b}\|^2 = \sum_{n=1}^{N}(\boldsymbol{a}^T\boldsymbol{y}_n - b_n)^2$$
6. **特征选择**　从一组特征中挑选出一些最有效的特征以达到降低特征空间维数的目的，这个过程叫特征选择。

二、答：

二元正态分布协方差矩阵 Σ 及其逆矩阵 Σ^{-1} 为

$$\Sigma=\begin{bmatrix}\sigma_{11}^2 & \sigma_{12}^2\\ \sigma_{12}^2 & \sigma_{22}^2\end{bmatrix} \tag{2-67}$$

$$\Sigma^{-1}=\frac{1}{|\Sigma|}\begin{bmatrix}\sigma_{22}^2 & -\sigma_{12}^2\\ -\sigma_{12}^2 & \sigma_{11}^2\end{bmatrix} \tag{2-68}$$

根据边缘分布定义

$$\begin{aligned}
p(x_1)&=\int_{-\infty}^{\infty}p(x_1,x_2)\mathrm{d}x_2\\
&=\frac{1}{2\pi|\Sigma|^{\frac{1}{2}}}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2|\Sigma|}[\sigma_{22}^2(x_1-\mu_1)^2+\sigma_{11}^2(x_2-\mu_2)^2\right.\\
&\qquad\left.-2\sigma_{12}^2(x_1-\mu_1)(x_2-\mu_2)]\right\}\mathrm{d}x_2\\
&=\frac{1}{2\pi|\Sigma|^{\frac{1}{2}}}\int_{-\infty}^{\infty}\exp\left\{-\frac{\sigma_{11}^2}{2|\Sigma|}\left[(x_2-\mu_2)^2-2\frac{\sigma_{12}^2}{\sigma_{11}^2}(x_1-\mu_1)(x_2-\mu_2)\right.\right.\\
&\qquad\left.\left.+\frac{\sigma_{22}^2}{\sigma_{11}^2}(x_1-\mu_1)^2\right]\right\}\mathrm{d}x_2\\
&=\frac{1}{(2\pi)^{\frac{1}{2}}\sigma_{11}}\exp\left\{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_{11}}\right)^2\right\}\\
&\quad\cdot\frac{\sigma_{11}}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}}\int_{-\infty}^{\infty}\exp\left\{\frac{-\sigma_{11}^2}{2|\Sigma|}\left[(x_2-\mu_2)-\frac{\sigma_{12}^2}{\sigma_{11}^2}(x_1-\mu_1)\right]^2\right\}\mathrm{d}x_2\\
&=\frac{1}{(2\pi)^{\frac{1}{2}}\sigma_{11}}\exp\left\{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_{11}}\right)^2\right\}
\end{aligned} \tag{2-69}$$

其中

$$|\Sigma|=\sigma_{11}^2\sigma_{22}^2-\sigma_{12}^4$$

所以 x_i 的边缘分布为

$$p(x_1)\sim N(\mu_1,\sigma_{11}^2) \tag{2-70}$$

也就是说边缘分布 $p(x_1)$ 服从以均值为 μ_1，方差为 σ_{11}^2 的正态分布。同理可以推出 x_2 的边缘分布为 $p(x_2)\sim N(\mu_2,\sigma_{22}^2)$。

关于给定 x_1 的条件下 x_2 的分布，有定义

$$p(x_2|x_1)=p(x_1,x_2)/p(x_1) \tag{2-71}$$

根据边缘分布的推导过程可以写出

$$\begin{aligned}
p(x_2|x_1)&=\frac{\sigma_{11}}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{\sigma_{11}^2}{2|\Sigma|}\left[(x_2-\mu_2)-\frac{\sigma_{12}^2}{\sigma_{11}^2}(x_1-\mu_1)\right]^2\right\}\\
&=K\cdot\exp\left\{-\frac{\sigma_{11}^2}{2|\Sigma|}\left[x_2-\left(\mu_2+\frac{\sigma_{12}^2}{\sigma_{11}^2}(x_1-\mu_1)\right)\right]^2\right\}
\end{aligned} \tag{2-72}$$

同理可以写出给定 x_2 条件下 x_1 的分布

$$p(x_1|x_2)=\frac{\sigma_{22}}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{\sigma_{22}^2}{2|\Sigma|}\left[(x_1-\mu_1)-\frac{\sigma_{12}^2}{\sigma_{22}^2}(x_2-\mu_2)\right]^2\right\} \tag{2-73}$$

它们也服从正态分布。

三、答 ：

假定样本是按概率

$$p(\boldsymbol{x}) = p(\boldsymbol{x}|\omega_1)P(\omega_1) + p(\boldsymbol{x}|\omega_2)P(\omega_2)$$

独立抽取的，并定义用 MSE 法得到的线性判别函数 $\boldsymbol{g}(\boldsymbol{x})=\boldsymbol{a}^T\boldsymbol{y}$ 与贝叶斯判别函数 $g_0(\boldsymbol{x})$ 之间的均方误差为

$$e^2 = \int[\boldsymbol{a}^T\boldsymbol{y} - g_0(\boldsymbol{x})]^2 p(\boldsymbol{x})d\boldsymbol{x} \tag{4-74}$$

我们的目的是要证明 MSE 解

$$\boldsymbol{a}^* = Y^+\boldsymbol{b} = Y^+\boldsymbol{u}_N$$

使 e^2 达到极小。

设样本未被规范化，则 $J_s(\boldsymbol{a})$ 可以写成

$$J_s(\boldsymbol{a}) = \sum_{n=1}^{N}(\boldsymbol{a}^T\boldsymbol{y}_n - b_n)^2 = \sum_{\boldsymbol{y}\in\mathscr{Y}_1}(\boldsymbol{a}^T\boldsymbol{y} - 1)^2 + \sum_{\boldsymbol{y}\in\mathscr{Y}_2}(\boldsymbol{a}^T\boldsymbol{y} + 1)^2$$

$$= N\cdot\left[\frac{N_1}{N}\cdot\frac{1}{N_1}\sum_{\boldsymbol{y}\in\mathscr{Y}_1}(\boldsymbol{a}^T\boldsymbol{y} - 1)^2 + \frac{N_2}{N}\cdot\frac{1}{N_2}\sum_{\boldsymbol{y}\in\mathscr{Y}_2}(\boldsymbol{a}^T\boldsymbol{y} + 1)^2\right] \tag{4-75}$$

利用大数定律，当 $N\to\infty$ 时，有

$$\frac{N_1}{N} \doteq P(\omega_1), \frac{N_2}{N} \doteq P(\omega_2),$$

$$\frac{1}{N_1}\sum_{\boldsymbol{y}\in\mathscr{Y}_1}(\boldsymbol{a}^T\boldsymbol{y} - 1)^2 \doteq E_1[(\boldsymbol{a}^T\boldsymbol{y} - 1)^2] = \int(\boldsymbol{a}^T\boldsymbol{y} - 1)^2 p(\boldsymbol{x}|\omega_1)d\boldsymbol{x}$$

$$\frac{1}{N_2}\sum_{\boldsymbol{y}\in\mathscr{Y}_2}(\boldsymbol{a}^T\boldsymbol{y} + 1)^2 \doteq E_2[(\boldsymbol{a}^T\boldsymbol{y} + 1)^2] = \int(\boldsymbol{a}^T\boldsymbol{y} + 1)^2 p(\boldsymbol{x}|\omega_2)d\boldsymbol{x}$$

根据以上结果，式(4-75)可写成

$$\frac{1}{N}J_s(\boldsymbol{a}) \doteq J_s'(\boldsymbol{a}) = P(\omega_1)E_1[(\boldsymbol{a}^T\boldsymbol{y} - 1)^2] + P(\omega_2)E_2[(\boldsymbol{a}^T\boldsymbol{y} + 1)^2]$$

$$= \int(\boldsymbol{a}^T\boldsymbol{y} - 1)^2 p(\boldsymbol{x},\omega_1)d\boldsymbol{x} + \int(\boldsymbol{a}^T\boldsymbol{y} + 1)^2 p(\boldsymbol{x},\omega_2)d\boldsymbol{x}$$

$$= \int\Big\{(\boldsymbol{a}^T\boldsymbol{y})^2[p(\boldsymbol{x},\omega_1) + p(\boldsymbol{x},\omega_2)] - 2\boldsymbol{a}^T\boldsymbol{y}\left[\frac{p(\boldsymbol{x},\omega_1) - p(\boldsymbol{x},\omega_2)}{p(\boldsymbol{x})}\right]p(\boldsymbol{x})$$

$$+ [p(\boldsymbol{x},\omega_1) + p(\boldsymbol{x},\omega_2)]\Big\}d\boldsymbol{x} \tag{4-76}$$

若把式(4-73)写成

$$g_0(\boldsymbol{x}) = \frac{p(\boldsymbol{x},\omega_1) - p(\boldsymbol{x},\omega_2)}{p(\boldsymbol{x})},$$

则式(4-76)成为

$$J_s'(\boldsymbol{a}) = \int(\boldsymbol{a}^T\boldsymbol{y})^2 p(\boldsymbol{x})d\boldsymbol{x} - 2\int\boldsymbol{a}^T\boldsymbol{y}g_0(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} + 1$$

$$= \int[\boldsymbol{a}^T\boldsymbol{y} - g_0(\boldsymbol{x})]^2 p(\boldsymbol{x})d\boldsymbol{x} + \left[1 - \int g_0^2(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}\right]$$

$$= e^2 + \left[1 - \int g_0^2(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}\right] \tag{4-77}$$

上式第二项与 $\boldsymbol{a}$ 无关，因此，使 $J_s'(\boldsymbol{a})$ 减至最小（当然也就使 $J_s(\boldsymbol{a})$ 减至最小）的 $\boldsymbol{a}^*$（即 MSE 解），也一定使 e^2 达到最小。这正是我们要证明的。

四、答:

$$\mu = \begin{bmatrix}0\\1\\0\end{bmatrix}$$

$$S_b = \frac{1}{2}\sum_{i=1}^{2}(\mu_i - \mu)(\mu_i - \mu)^T = \frac{1}{4}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T = \frac{1}{16}\begin{bmatrix}1 & 2 & -1\\2 & 4 & -2\\-1 & 2 & 1\end{bmatrix}$$

$$S_w = \frac{1}{2}(S_1 + S_2) = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad S_w^{-1} = \frac{1}{8}\begin{bmatrix} 3 & -1 & 0 \\ -1 & 3 & 0 \\ 0 & 0 & 8 \end{bmatrix}$$

$$S_w^{-1}S_b = \frac{1}{8\times 16}\begin{bmatrix} 1 & 2 & -1 \\ 5 & 10 & -5 \\ -8 & 16 & 8 \end{bmatrix}$$

因为$S_w^{-1}S_b$的秩为 1,所以$S_w^{-1}S_b$只有一个非零本征值,W 是 D×1 矩阵,即 W=w。为求$S_w^{-1}S_b$的本征值应解方程：

$$S_w^{-1}S_b w = \lambda_1 w$$

或

$$\frac{1}{4}S_w^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w = \lambda_1 w$$

$\frac{1}{4}(\mu_1 - \mu_2)^T w$是标量，所以

$$w = S_w^{-1}(\mu_1 - \mu_2) = \begin{bmatrix} \frac{1}{4} \\ \frac{5}{4} \\ -2 \end{bmatrix}$$

五、答：

实际应用中，多元正态分布的更典型情况是，均值 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$ 都未知。这样，参数向量 $\boldsymbol{\theta}$ 就是由这两个成分组成。我们首先考虑单变量的情况，其中参数向量 $\boldsymbol{\theta}$ 的组成成分是：$\theta_1=\mu, \theta_2=\sigma^2$。这样，对于单个训练样本的对数似然函数为：

$$\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k-\theta_1)^2 \tag{12}$$

对上式关于变量 $\boldsymbol{\theta}$ 求导：

$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \ln p(x_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k-\theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2} \end{bmatrix} \tag{13}$$

运用式(7)，我们得到对于全体样本的对数似然函数的极值条件

$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 \tag{14}$$

和

$$-\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k-\hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 \tag{15}$$

其中的 $\hat{\theta}_1, \hat{\theta}_2$ 分别是对于 θ_1, θ_2 的最大似然估计。

把 $\hat{\theta}_1, \hat{\theta}_2$ 用 $\hat{\mu}, \hat{\sigma}^2$ 代替，并进行简单的整理，我们得到下述的对于均值和方差的最大似然估计结果

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{16}$$

和

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2 \tag{17}$$

六、答：

图 8.5

解 我们用样本均值代替总体均值，样本协方差矩阵代替总体协方差矩阵。

$$\mu_1 \doteq m_1 = \frac{1}{4}\begin{bmatrix}3\\1\\1\end{bmatrix}, \qquad \mu_2 \doteq m_2 = \frac{1}{4}\begin{bmatrix}1\\3\\3\end{bmatrix}$$

$$\mu_1 - \mu_2 \doteq m_1 - m_2 = \frac{1}{4}\begin{bmatrix}2\\-2\\-2\end{bmatrix}$$

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

这是两类协方差阵相等的情况。

$$\Sigma^{-1} = \begin{bmatrix}8 & -4 & -4\\-4 & 8 & 4\\-4 & 4 & 8\end{bmatrix}$$

$$\Sigma^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T = \frac{1}{4}\begin{bmatrix}1 & -1 & -1\\-1 & 1 & 1\\-1 & 1 & 1\end{bmatrix}$$

此矩阵之非零本征值及其对应之本征向量为：

$$\lambda = \frac{3}{4}, \qquad v = \begin{bmatrix}-1\\1\\1\end{bmatrix}$$

因此

$$w^T = (-1, 1, 1)$$

对所给样本集做变换：$x^* = w^T x$，可得到下面一维结果(图 8.4)。

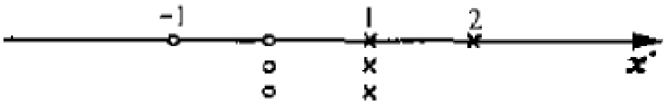


图 8.4　图 8.3 的样本投影到一维的情况

$$\omega_1: x_{11}^* = 0 \qquad \omega_2: x_{21}^* = 1$$

$$x_{12}^* = -1 \qquad x_{22}^* = 1$$

$$x_{13}^* = 0 \qquad x_{23}^* = 2$$

$$x_{14}^* = 0 \qquad x_{24}^* = 1$$

03 年试题

一、答：五种常用决策是

1. 基于最小错误率的贝叶斯决策。利用概率论中的贝叶斯公式，得出使错误率为最小的分类规则。
2. 最小最大决策。使最大可能的风险为最小的决策。
3. 序贯分类方法
4.

我们可以考虑把 d 维空间的样本投影到一条直线上，形成一维空间，即把维数压缩到一维。这在数学上总是容易办到的。然而，即使样本在 d 维空间里形成若干紧凑的互相分得开的集群，若把它们投影到一条任意的直线上，也可能使几类样本混在一起而变得无法识别。但在一般情况下，总可以找到某个方向，使在这个方向的直线上，样本的投影能分开得最好。问题是如何根据实际情况找到这条最好的、最易于分类的投影线。这就是 Fisher 法所要解决的基本问题(见图 4.3)。

5 .

特征选择 从一组特征中挑选出一些最有效的特征以达到降低特征空间维数的目的，这个过程叫特征选择。

二、答:

在实际中有很多模式识别问题并不是线性可分的，这时就需要采用非线性分类器。比如当两类样本分布具有多峰性质并互相交错时，简单的线性判别函数往往会带来较大的分类错误。这时，树分类器作为一种分段线性分类器，常常能有效地应用于这种情况。本章重点

三、答：

两类别问题：判别函数

$$g_1(x) = \lambda_{11}p(\omega_1|x) + \lambda_{12}p(\omega_2|x)$$
$$g_2(x) = \lambda_{21}p(\omega_1|x) + \lambda_{22}p(\omega_2|x)$$

决策面方程：$g_1(x) = g_2(x)$

c 类别问题：判别函数

$$g_i(x) = \sum_{j=1}^{c} \lambda_{ji}p(\omega_j|x), \qquad i = 1, \cdots, c$$

决策面方程：$g_i(x) = g_j(x)$, $i \neq j \quad i = 1, \cdots, c \quad j = 1, \cdots, c$

四、答：

本章的目的是通过样本集推断总体分布 $p(\boldsymbol{x}|\mathscr{X})$，因为已假定总体分布的形式是已知的，所以问题就转化为参数 θ 的估计问题，上面我们已介绍了两种参数估计的方法。此外，我们还可以直接推断总体分布 $p(\boldsymbol{x}|\mathscr{X})$，其思路中的前三步都与求贝叶斯估计量 $\hat{\theta}$ 相同，当第三步利用贝叶斯公式求出 θ 的后验密度 $p(\theta|\mathscr{X})$后，就不再去估计 $\hat{\theta}$，而是直接通过联合密度求类条件概率密度

$$p(\boldsymbol{x}|\mathscr{X}) = \int p(\boldsymbol{x},\theta|\mathscr{X})\mathrm{d}\theta = \int p(\boldsymbol{x}|\theta)p(\theta|\mathscr{X})\mathrm{d}\theta \tag{3-31}$$

式(3-31)的关键问题是求参数 θ 的后验密度 $p(\theta|\mathscr{X})$，根据贝叶斯公式

$$p(\theta|\mathscr{X}) = \frac{p(\mathscr{X}|\theta)p(\theta)}{\int p(\mathscr{X}|\theta)p(\theta)\mathrm{d}\theta} \tag{3-32}$$

在最大似然法的讨论中，我们已经知道，似然函数 $p(\mathscr{X}|\theta)$在 $\theta=\hat{\theta}$ 处很可能有一个尖锐的凸峰。现在参数 θ 是随机变量，如果其先验概率密度 $p(\theta)$在 $\theta=\hat{\theta}$ 处不为零，而且在附近又变化不大，则根据式(3-32)，$p(\theta|\mathscr{X})$在 $\theta=\hat{\theta}$ 处也将有尖锐的凸峰，这样式(3-31)可近似为

$$p(\boldsymbol{x}|\mathscr{X}) \doteq p(\boldsymbol{x}|\hat{\theta}) \tag{3-33}$$

$\hat{\theta}$ 是 θ 的最大似然估计值，也就是说贝叶斯解的结果 $p(\boldsymbol{x}|\mathscr{X})$与最大似然估计的结果近似相等。

五、证：因为各特征分量相互独立，所以

$$p(\chi|\omega_i) = \prod_{j=1}^{D} p(\chi_j|\omega_i)$$

$$J_D = \int_X [p(\chi|\omega_i) - p(\chi|\omega_j)]\ln\frac{p(\chi|\omega_i)}{p(\chi|\omega_j)}d\chi$$

$$= \int_X [p(\chi|\omega_i) - p(\chi|\omega_j)]\ln\frac{\prod_{k=1}^{D} p(\chi_k|\omega_i)}{\prod_{k=1}^{D} p(\chi_k|\omega_i)}d\chi$$

$$= \int_X [p(\chi|\omega_i) - p(\chi|\omega_j)]\left[\ln\frac{p(\chi_1|\omega_i)}{p(\chi_1|\omega_i)} + \ln\frac{p(\chi_1|\omega_i)}{p(\chi_1|\omega_i)} + \cdots + \ln\frac{p(\chi_1|\omega_i)}{p(\chi_1|\omega_i)}\right]d\chi = \sum_{j=1}^{D} J_{Dj}$$

六、答：

$$B^T S_w B = I, \qquad S_w = P_1\Sigma_1 + P_2\Sigma_2$$

$\therefore\ B^T(P_1\Sigma_1 + P_2\Sigma_2)B = I$, 即: $P_1B^T\Sigma_1B + P_2B^T\Sigma_2B = I$

设λ_{1i},x_{1i}是$P_1B^T\Sigma_1B$的本征值和本征向量,则$P_1B^T\Sigma_1Bx_{1i} = \lambda_{1i}x_{1i}$

而$P_2B^T\Sigma_2Bx_{1i} = (I - P_1B^T\Sigma_1B)x_{1i} = (1 - \lambda_{1i})x_{1i}$

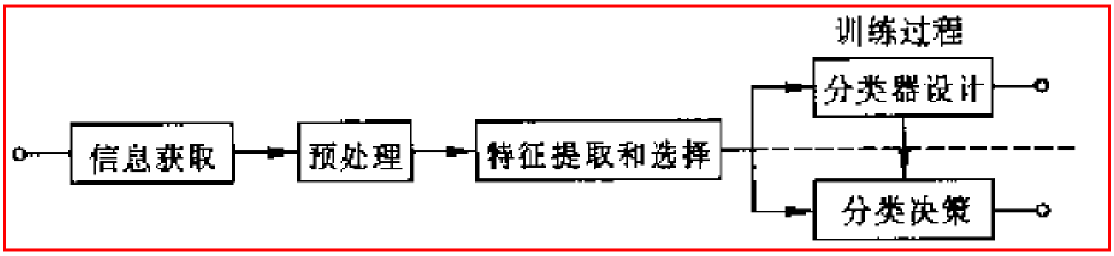
所以$1 - \lambda_{1i}$,x_{1i}是$P_2B^T\Sigma_2B$的本征值和本征向量

因$P_1B^T\Sigma_1B$和$P_2B^T\Sigma_2B$具有相同的本征向量,所以他们所产生的 K-L 坐标轴是相同的.

$$\Lambda_1 = \begin{bmatrix}\lambda_{11} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \lambda_{1n}\end{bmatrix} = I - \begin{bmatrix}1-\lambda_{11} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & 1-\lambda_{1n}\end{bmatrix} = I - \Lambda_2$$

04 年

一、答：



信息获取：通过测量、采样和量化，可以用矩阵或向量表示二维图象或一维波形。这就是信息获取的过程。

预处理：其的目的是去除噪声，加强有用的信息，并对输入测量仪器或其他因素所造成的退化现象进行复原。

特征提取和选择：为了有效地实现分类识别，就要对原始数据进行变换，得到最能反映分类本质的特征。这就是特征提取和选择的过程。

分类决策：就是在特征空间中用统计方法把被识别对象归为某一类别。

二、答：

(1) $w = \Sigma^{-1}(M_1 - M_2) = \begin{bmatrix} -0.889 \\ -0.1481 \end{bmatrix}$

$$x_0 = \frac{1}{2}(M_1 + M_2) - \frac{\ln\frac{P(\omega_1)}{P(\omega_2)}}{(M_1 - M_2)^T\Sigma^{-1}(M_1 - M_2)}(M_1 - M_2) = \begin{bmatrix} 4.391 \\ 6.391 \end{bmatrix}$$

所以决策面为$[-0.889 \quad -0.1481]\left(x - \begin{bmatrix} 4.391 \\ 6.391 \end{bmatrix}\right) = 0$

(2) $g_1(x) = \lambda_{11}p(\omega_1|x) + \lambda_{12}p(\omega_2|x) = 8p(\omega_2|x)$

$g_2(x) = \lambda_{21}p(\omega_1|x) + \lambda_{22}p(\omega_2|x) = p(\omega_1|x)$

所以决策面为$8p(\omega_2|x) = p(\omega_1|x)$

(3) a)

$$p(\omega_1|x) = \frac{p(x|\omega_1)P(\omega_1)}{p(x|\omega_1)P(\omega_1) + p(x|\omega_2)P(\omega_2)} = \frac{3}{7}$$

$$p(\omega_2|x) = \frac{p(x|\omega_2)P(\omega_2)}{p(x|\omega_1)P(\omega_1) + p(x|\omega_2)P(\omega_2)} = \frac{4}{7}$$

所以 x 属于ω_2

b)

$$R(\alpha_1|x) = \lambda_{11}p(\omega_1|x) + \lambda_{12}p(\omega_2|x) = \frac{32}{7}$$

$$R(\alpha_2|x) = \lambda_{21}p(\omega_1|x) + \lambda_{22}p(\omega_2|x) = \frac{3}{7}$$

所以 x 属于ω_2

(4)

三、答：

$$P(c) = \sum_{i=1}^{c} P(x \in R_i|\omega_i)P(\omega_i) = \sum_{i=1}^{c}\int_{R_i} p(x|\omega_i)P(\omega_i)dx$$

$$=\frac{1}{c}\sum_{i=1}^{c}\int_{R_i}p(x|\omega_i)dx=\frac{1}{c}\left[\int_0^{\frac{cr}{c-1}}dx+\sum_{i=1}^{c}\int_i^{i+1-\frac{cr}{c-1}}dx\right]$$

$$=\frac{1}{c}\left[\frac{cr}{c-1}+\sum_{i=1}^{c}1-\frac{cr}{c-1}\right]=1-r$$

四、答：

实际应用中，多元正态分布的更典型情况是，均值 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$ 都未知。这样，参数向量 $\boldsymbol{\theta}$ 就是由这两个成分组成。我们首先考虑单变量的情况，其中参数向量 $\boldsymbol{\theta}$ 的组成成分是：$\theta_1=\mu,\theta_2=\sigma^2$。这样，对于单个训练样本的对数似然函数为：

$$\ln p(x_k|\boldsymbol{\theta})=-\frac{1}{2}\ln 2\pi\theta_2-\frac{1}{2\theta_2}(x_k-\theta_1)^2 \qquad (12)$$

对上式关于变量 $\boldsymbol{\theta}$ 求导：

$$\nabla_{\boldsymbol{\theta}}l=\nabla_{\boldsymbol{\theta}}\ln p(x_k|\boldsymbol{\theta})=\begin{bmatrix}\frac{1}{\theta_2}(x_k-\theta_1)\\ -\frac{1}{2\theta_2}+\frac{(x_k-\theta_1)^2}{2\theta_2^2}\end{bmatrix} \qquad (13)$$

运用式(7)，我们得到对于全体样本的对数似然函数的极值条件

$$\sum_{k=1}^{n}\frac{1}{\hat{\theta}_2}(x_k-\hat{\theta}_1)=0 \qquad (14)$$

和

$$-\sum_{k=1}^{n}\frac{1}{\hat{\theta}_2}+\sum_{k=1}^{n}\frac{(x_k-\hat{\theta}_1)^2}{\hat{\theta}_2^2}=0 \qquad (15)$$

其中的 $\hat{\theta}_1,\hat{\theta}_2$ 分别是对于 θ_1,θ_2 的最大似然估计。

把 $\hat{\theta}_1,\hat{\theta}_2$ 用 $\hat{\mu},\hat{\sigma}^2$ 代替，并进行简单的整理，我们得到下述的对于均值和方差的最大似然估计结果

$$\hat{\mu}=\frac{1}{n}\sum_{k=1}^{n}x_k \qquad (16)$$

和

$$\hat{\sigma}^2=\frac{1}{n}\sum_{k=1}^{n}(x_k-\hat{\mu})^2 \qquad (17)$$

当高斯函数为多元时，最大似然估计的过程也是非常类似的，当然，也将更加复杂(习题6)。对于多元高斯分布的均值 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$ 的最大似然估计结果为：

$$\hat{\boldsymbol{\mu}}=\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k \qquad (18)$$

$$\hat{\boldsymbol{\Sigma}}=\frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k-\hat{\boldsymbol{\mu}})(\mathbf{x}_k-\hat{\boldsymbol{\mu}})^t \qquad (19)$$

五、答：

$$\mu=\begin{bmatrix}0\\1\\0\end{bmatrix}$$

$$S_b=\frac{1}{2}\sum_{i=1}^{2}(\mu_i-\mu)(\mu_i-\mu)^T=\frac{1}{4}(\mu_1-\mu_2)(\mu_1-\mu_2)^T=\frac{1}{16}\begin{bmatrix}1&2&-1\\2&4&-2\\-1&2&1\end{bmatrix}$$

$$S_w = \frac{1}{2}(S_1 + S_2) = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad S_w^{-1} = \frac{1}{8}\begin{bmatrix} 3 & -1 & 0 \\ -1 & 3 & 0 \\ 0 & 0 & 8 \end{bmatrix}$$

$$S_w^{-1}S_b = \frac{1}{8\times 16}\begin{bmatrix} 1 & 2 & -1 \\ 5 & 10 & -5 \\ -8 & 16 & 8 \end{bmatrix}$$

因为$S_w^{-1}S_b$的秩为 1,所以$S_w^{-1}S_b$只有一个非零本征值,W 是 D×1 矩阵,即 W=w。为求$S_w^{-1}S_b$的本征值应解方程：

$$S_w^{-1}S_b w = \lambda_1 w$$

或

$$\frac{1}{4}S_w^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w = \lambda_1 w$$

$\frac{1}{4}(\mu_1 - \mu_2)^T w$是标量，所以

$$w = S_w^{-1}(\mu_1 - \mu_2) = \begin{bmatrix} \frac{1}{4} \\ \frac{5}{4} \\ -2 \end{bmatrix}$$

六、答：

(1) $E[xx^T] = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}$,其对应的本征值和本征向量为：

$$\Lambda = \begin{bmatrix} \frac{1}{4} & & \\ & \frac{1}{4} & \\ & & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 0 & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{bmatrix}$$

降到二维空间取$U = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix}$

对应的坐标 $\omega_i: \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{2}{\sqrt{6}} \\ \frac{-2}{\sqrt{6}} \end{bmatrix}, \begin{bmatrix} \frac{2\sqrt{2}}{\sqrt{6}} \\ \frac{-1}{\sqrt{6}} \end{bmatrix}, \begin{bmatrix} \frac{2\sqrt{2}}{\sqrt{6}} \\ \frac{-1}{\sqrt{6}} \end{bmatrix}$ $\quad \omega_2: \begin{bmatrix} \frac{\sqrt{2}}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix}, \begin{bmatrix} \frac{\sqrt{2}}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix}, \begin{bmatrix} \frac{2\sqrt{2}}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \end{bmatrix}, \begin{bmatrix} \frac{3\sqrt{2}}{\sqrt{6}} \\ 0 \end{bmatrix}$

(2) $m_1 = \begin{bmatrix} \frac{3}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{bmatrix}, \ m_2 = \begin{bmatrix} \frac{1}{4} \\ \frac{3}{4} \\ \frac{3}{4} \end{bmatrix}, \quad S_w = \frac{1}{12}\begin{bmatrix} 3 & 1 & 1 \\ 1 & 3 & -1 \\ 1 & -1 & 3 \end{bmatrix}$

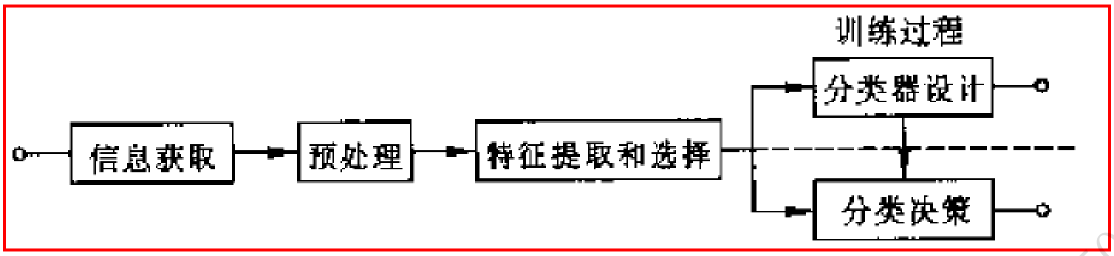
$w^* = S_w^{-1}(m_1 - m_2) = \begin{bmatrix} 6 \\ -6 \\ -6 \end{bmatrix}, \quad y_0 = \frac{w^{*T}m_1 + w^{*T}m_2}{2} = -3$

所以判别函数为：$g(x) = \begin{bmatrix} 6 \\ -6 \\ -6 \end{bmatrix}^T x + 3$

05 年

一、答：

我们把通过对具体的个别事物进行观测所得到的具有时间和空间分布的信息称为模式.



信息获取：通过测量、采样和量化，可以用矩阵或向量表示二维图象或一维波形。这就是信息获取的过程。

预处理：其的目的是去除噪声，加强有用的信息，并对输入测量仪器或其他因素所造成的退化现象进行复原。

特征提取和选择：为了有效地实现分类识别，就要对原始数据进行变换，得到最能反映分类本质的特征。这就是特征提取和选择的过程。

分类决策：就是在特征空间中用统计方法把被识别对象归为某一类别。

二、答:判别函数 g(x)是 x 的线性函数叫线性判别函数.

在实际中有很多模式识别问题并不是线性可分的，这时就需要采用非线性分类器。比如当两类样本分布具有多峰性质并互相交错时，简单的线性判别函数往往会带来较大的分类错误。这时，树分类器作为一种分段线性分类器，常常能有效地应用于这种情况。本章重点

3 个,

三、答:

(1) 判别函数:

$$g_1(x)=p(\omega_1|x)=\frac{p(x|\omega_1)P(\omega_1)}{p(x|\omega_1)P(\omega_1)+p(x|\omega_2)P(\omega_2)}=\frac{0.9p(x|\omega_1)}{0.9p(x|\omega_1)P(\omega_1)+0.1p(x|\omega_2)}$$

$$g_2(x)=p(\omega_2|x)=\frac{p(x|\omega_2)P(\omega_2)}{p(x|\omega_1)P(\omega_1)+p(x|\omega_2)P(\omega_2)}=\frac{0.1p(x|\omega_2)}{0.9p(x|\omega_1)P(\omega_1)+0.1p(x|\omega_2)}$$

决策面方程: $9p(x|\omega_1)=p(x|\omega_2)$

决策:

$$g_1(x)=p(\omega_1|x)=\frac{p(x|\omega_1)P(\omega_1)}{p(x|\omega_1)P(\omega_1)+p(x|\omega_2)P(\omega_2)}=\frac{9}{11}$$

$$g_2(x)=p(\omega_2|x)=\frac{p(x|\omega_2)P(\omega_2)}{p(x|\omega_1)P(\omega_1)+p(x|\omega_2)P(\omega_2)}=\frac{2}{11}$$

所以 x 属于ω_1

(2) 判别函数:

$$g_1(x)=\lambda_{11}p(\omega_1|x)+\lambda_{12}p(\omega_2|x)=6p(\omega_2|x)$$

$$g_2(x)=\lambda_{21}p(\omega_1|x)+\lambda_{22}p(\omega_2|x)=p(\omega_1|x)$$

决策面方程: $6p(\omega_2|x)=p(\omega_1|x)$

决策:

$$g_1(x)=\lambda_{11}p(\omega_1|x)+\lambda_{12}p(\omega_2|x)=6p(\omega_2|x)=\frac{12}{11}$$

$$g_2(x) = \lambda_{21} p(\omega_1|x) + \lambda_{22} p(\omega_2|x) = p(\omega_1|x) = \frac{9}{11}$$

所以 x 属于ω_2

四、答:

$$\mu = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$S_b = \frac{1}{2}\sum_{i=1}^{2}(\mu_i - \mu)(\mu_i - \mu)^T = \frac{1}{4}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T = \frac{1}{16}\begin{bmatrix} 1 & 2 & -1 \\ 2 & 4 & -2 \\ -1 & 2 & 1 \end{bmatrix}$$

$$S_w = \frac{1}{2}(S_1 + S_2) = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad S_w^{-1} = \frac{1}{8}\begin{bmatrix} 3 & -1 & 0 \\ -1 & 3 & 0 \\ 0 & 0 & 8 \end{bmatrix}$$

$$S_w^{-1} S_b = \frac{1}{8 \times 16}\begin{bmatrix} 1 & 2 & -1 \\ 5 & 10 & -5 \\ -8 & 16 & 8 \end{bmatrix}$$

因为$S_w^{-1}S_b$的秩为 1,所以$S_w^{-1}S_b$只有一个非零本征值,W 是 D×1 矩阵,即 W=w。为求$S_w^{-1}S_b$的本征值应解方程:

$$S_w^{-1} S_b w = \lambda_1 w$$

或

$$\frac{1}{4} S_w^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w = \lambda_1 w$$

$\frac{1}{4}(\mu_1 - \mu_2)^T w$是标量，所以

$$w = S_w^{-1}(\mu_1 - \mu_2) = \begin{bmatrix} \frac{1}{4} \\ \frac{5}{4} \\ -2 \end{bmatrix}$$

五、答:

(3) $E[xx^T] = \begin{bmatrix} 25.4 & 25 \\ 25 & 25.4 \end{bmatrix}$,其对应的本征值和本征向量为:

$\Lambda = \begin{bmatrix} 50.4 & \\ & 4 \end{bmatrix}$, $U = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$

降到二维空间取$U = [1 \quad 1]$

对应的坐标 $\omega_i: \{-10, -9, -9, -11, -11\}$ $\omega_2: \{10, 11, 11, 9, 9\}$

六、答:

ISODATA 算法与 C- 均值算法有两点不同:第一,它不是每调整一个样本的类别就重新计算一次各类样本的均值。而是在每次把全部样本都调整完毕之后才重新计算一次样本

的均值，前者一般称为逐个样本修正法，后者称为成批样本修正法。第二，ISODATA 算法不仅能通过调整样本所属类别完成聚类分析，而且还能自动地进行类的“合并”和“分裂”，从而得到类数较为合理的各个聚类。

下面我们给出 ISODATA 算法的程序步骤：

设有 N 个样本组成的样本集为 $\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_N\}$

步骤 1　规定下列控制参数

K——期望得到的聚类数

θ_N——一个聚类中的最少样本数

θ_S——标准偏差参数

θ_C——合并参数

L——每次迭代允许合并的最大聚类对数

I——允许迭代的次数

设一初始的聚类数 c 以及初始的聚类中心 $\boldsymbol{m}_i, i=1,2,\cdots,c$。

步骤 2　按照下列关系

$$\begin{cases} \text{若}\ \|\boldsymbol{y}-\boldsymbol{m}_j\| < \|\boldsymbol{y}-\boldsymbol{m}_i\|\ i=1,2,\cdots,c, i\neq j \\ \text{则}\ \boldsymbol{y}\in\Gamma_j \end{cases}$$

将所有样本分到各个聚类中去。Γ_j 是第 j 个聚类，其中心为 $\boldsymbol{m}_j$

步骤 3　若有任何一个 Γ_j，其基数 $N_j<\theta_N$，则舍去 Γ_j，并令 $c=c-1$。

步骤 4　按照下列关系

$$\boldsymbol{m}_j=\frac{1}{N_j}\sum_{\boldsymbol{y}\in\Gamma_j}\boldsymbol{y}, j=1,2,\cdots,c \tag{10-38}$$

计算更新的均值向量。式中 N_j 是第 j 个聚类的样本数目(基数)

步骤5　计算 Γ_j 中的各个样本离开其中心 $\boldsymbol{m}_j$ 的平均距离 $\bar{\delta}_j$

$$\bar{\delta}_j=\frac{1}{N_j}\sum_{\boldsymbol{y}\in\Gamma_j}\|\boldsymbol{y}-\boldsymbol{m}_j\|, j=1,2,\cdots,c \tag{10-39}$$

步骤 6　计算所有样本离开其相应的聚类中心的平均距离

$$\bar{\delta}=\frac{1}{N}\sum_{j=1}^{c}N_j\bar{\delta}_j \tag{10-40}$$

步骤 7　①若这是最后一次迭代(由参数 I 确定)，则置 $\theta_c=0$，转向步骤 11；

②若 $c\leqslant K/2$，转向步骤 8；

③若是偶数次迭代，或若是 $c\geqslant 2K$，则转向步骤 11。否则，往下进行。

步骤 8　对每一个聚类 j，用下列公式求标准偏差 $\sigma_j=[\sigma_{j1},\sigma_{j2},\cdots,\sigma_{jd}]^T$

$$\sigma_{ji}=\sqrt{\frac{1}{N_j}\sum_{\boldsymbol{y}_k\in\Gamma_j}(y_{ki}-m_{ji})^2} \tag{10-41}$$

式中 y_{ki}是第 k 个样本的第 i 个分量(特征)，$\boldsymbol{y}_k$ 应在 Γ_j 中；m_{ji}是第 j 个聚类中心的第 i 个分量；σ_{ji}是第 j 个聚类第 i 个分量的标准偏差，d 是样本 $\boldsymbol{y}$ 的维数。

步骤 9　对每一个聚类，求出具有最大标准偏差的分量 $\sigma_{j\max}, j=1,2,\cdots,c$。

步骤 10　若对任一个 $\sigma_{j\max}, j=1,2,\cdots,c$ 存在 $\sigma_{j\max}>\theta_s$，并且有

① $\bar{\delta}_j>\bar{\delta}$ 且 $N_j>2(\theta_N+1)$，

或　②$c \leqslant K/2$，
则把 Γ_j 分裂成两个聚类，其中心相应为 $\boldsymbol{m}_j^+$ 和 $\boldsymbol{m}_j^-$，把原来的 $\boldsymbol{m}_j$ 取消，且令 $c=c+1$。$\boldsymbol{m}_j^+$ 和 $\boldsymbol{m}_j^-$ 的计算如下：

① 给定一个 k 值，使 $0<k\leqslant 1$；

② 令 $\boldsymbol{\gamma}_j = k\boldsymbol{\sigma}_j$ 或 $\boldsymbol{\gamma}_j = k[0,\cdots,\sigma_{j\max},\cdots,0]^T$；

③ $\boldsymbol{m}_j^+ = \boldsymbol{m}_j + \boldsymbol{\gamma}_j$，$\boldsymbol{m}_j^- = \boldsymbol{m}_j - \boldsymbol{\gamma}_j$。

式中的 k 值应选得使 Γ_j 中的样本到 $\boldsymbol{m}_j^+$ 和 $\boldsymbol{m}_j^-$ 的距离不同，但又应使 Γ_j 中的样本仍然在这两个新的集合中。

步骤 11　对于所有的聚类中心，计算两两之间的距离

$$\delta_{ij} = \| \boldsymbol{m}_i - \boldsymbol{m}_j \| \quad \begin{array}{l} i = 1,2,\cdots,c-1 \\ j = i+1, i+2,\cdots,c \end{array} \tag{10-42}$$

步骤 12　比较 δ_{ij} 和 θ_c。把小于 θ_c 和 δ_{ij} 按 d_{ij} 的上升次序排列

$$\delta_{i_1j_1} < \delta_{i_2j_2} < \cdots < \delta_{i_lj_l}$$

这里的 l 是步骤 1 所给定的每次迭代可允许合并的最大聚类对的数。

步骤 13　从最小的 $\delta_{i_1j_1}$ 开始，对于每个 $\delta_{i_lj_l}$，相应有两个聚类中心 $\boldsymbol{m}_{i_l}$ 和 $\boldsymbol{m}_{j_l}$，假使在同一次迭代中，还没有把 $\boldsymbol{m}_{i_l}$ 和 $\boldsymbol{m}_{j_l}$ 合并，就把这两者合并为

$$\boldsymbol{m}_l = \frac{1}{N_{i_l} + N_{j_l}} [N_{i_l} \cdot \boldsymbol{m}_{i_l} + N_{j_l} \cdot \boldsymbol{m}_{j_l}] \tag{10-43}$$

并把类数 c 减少一个，即 $c=c-1$。由于每次迭代，一个聚类中心最多只能被合并一次，所以实际的合并对数总是小于或等于 l 的。

步骤 14　若这是最后一次迭代，则程序终止，否则，若根据经验需要改变参数，则转向步骤 1；若不需要改变参数，则转向步骤 2。这两种情况都要使迭代计数器加 1。

06年

一、答：

$$p(\omega_i|x)=\frac{p(x|\omega_i)P(\omega_i)}{\sum_{j=1}^{M}p(x|\omega_j)P(\omega_j)},\quad i=1,\cdots,M$$

$$R(\alpha_i|x)=\sum_{j=1}^{M}\lambda_{ij}p(\omega_j|x),\quad i=1,\cdots,M$$

如果

$$R(\alpha_i|x)=\min_{j=1,\cdots,M}R(\alpha_j|x)$$

则 x 属于ω_i

$$\lambda(\alpha_i,\omega_j)=\begin{cases}0, i=j,\\ 1, i\neq j,\end{cases}\quad i,j=1,2,\cdots,c \tag{2-18}$$

式中假定对于 c 类只有 c 个决策，即不考虑“拒绝”的情况。式(2-18)中 $\lambda(\alpha_i,\omega_j)$ 是对于正确决策(即 $i=j$)没有损失；而对于任何错误决策，其损失均为 1。这样定义的损失函数称为 0-1 损失函数。

根据式(2-15)，条件风险为

$$R(\alpha_i|\boldsymbol{x})=\sum_{j=1}^{c}\lambda(\alpha_i,\omega_j)P(\omega_j|\boldsymbol{x})=\sum_{\substack{j=1\\ j\neq i}}^{c}P(\omega_j|\boldsymbol{x}) \tag{2-19}$$

$\sum_{\substack{j=1\\ j\neq i}}^{c}P(\omega_j|\boldsymbol{x})$表示对 $\boldsymbol{x}$ 采取决策 ω_i 的条件错误概率。所以在 0-1 损失函数时，使

$$R(\alpha_k|\boldsymbol{x})=\min_{i=1,\cdots,c}R(\alpha_i|\boldsymbol{x})$$

的最小风险贝叶斯决策就等价于

$$\sum_{\substack{j=1\\ j\neq k}}^{c}P(\omega_j|\boldsymbol{x})=\min_{i=1,\cdots,c}\sum_{\substack{j=1\\ j\neq i}}^{c}P(\omega_j|\boldsymbol{x})$$

的最小错误率贝叶斯决策。

由此可见，最小错误率贝叶斯决策就是在 0-1 损失函数条件下的最小风险贝叶斯决策。

二、答：基于参数方法：是由已知类别的样本集对总体分布的某些参数进行统计推断。

非参数方法：已知样本所属类别，但未知总体概率密度函数形式。

三、答：

实际应用中，多元正态分布的更典型情况是，均值 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$ 都未知。这样，参数向量 $\boldsymbol{\theta}$ 就是由这两个成分组成。我们首先考虑单变量的情况，其中参数向量 $\boldsymbol{\theta}$ 的组成成分是：$\theta_1=\mu, \theta_2=\sigma^2$。这样，对于单个训练样本的对数似然函数为：

$$\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k-\theta_1)^2 \tag{12}$$

对上式关于变量 $\boldsymbol{\theta}$ 求导：

$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \ln p(x_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k-\theta_1) \\ -\frac{1}{2\theta_2}+\frac{(x_k-\theta_1)^2}{2\theta_2^2} \end{bmatrix} \tag{13}$$

运用式(7)，我们得到对于全体样本的对数似然函数的极值条件

$$\sum_{k=1}^{n}\frac{1}{\hat{\theta}_2}(x_k-\hat{\theta}_1)=0 \tag{14}$$

和

$$-\sum_{k=1}^{n}\frac{1}{\hat{\theta}_2}+\sum_{k=1}^{n}\frac{(x_k-\hat{\theta}_1)^2}{\hat{\theta}_2^2}=0 \tag{15}$$

其中的 $\hat{\theta}_1, \hat{\theta}_2$ 分别是对于 θ_1, θ_2 的最大似然估计。

把 $\hat{\theta}_1, \hat{\theta}_2$ 用 $\hat{\mu}, \hat{\sigma}^2$ 代替，并进行简单的整理，我们得到下述的对于均值和方差的最大似然估计结果

$$\hat{\mu}=\frac{1}{n}\sum_{k=1}^{n}x_k \tag{16}$$

和

$$\hat{\sigma}^2=\frac{1}{n}\sum_{k=1}^{n}(x_k-\hat{\mu})^2 \tag{17}$$

四、答：

Fisher:

$$\mathrm{m}_1=\begin{bmatrix}\frac{2}{3}\\ \frac{2}{3}\end{bmatrix},\ \mathrm{m}_2=\begin{bmatrix}-\frac{1}{3}\\ -\frac{1}{3}\end{bmatrix},\ \Sigma_1=\begin{bmatrix}\frac{1}{3} & -\frac{1}{6}\\ -\frac{1}{6} & \frac{1}{3}\end{bmatrix},\ \ \Sigma_2=\begin{bmatrix}\frac{1}{3} & -\frac{1}{6}\\ -\frac{1}{6} & \frac{1}{3}\end{bmatrix}\ \mathrm{S_w}=\begin{bmatrix}\frac{1}{3} & -\frac{1}{6}\\ -\frac{1}{6} & \frac{1}{3}\end{bmatrix}$$

$$\mathrm{w}^*=\mathrm{S_w^{-1}}(\mathrm{m}_1-\mathrm{m}_2)=\begin{bmatrix}6\\6\end{bmatrix},\quad \mathrm{y}_0=\frac{\widetilde{\mathrm{m}_1}+\widetilde{\mathrm{m}_2}}{2}=\frac{\mathrm{w}^{*\mathrm{T}}\mathrm{m}_1+\mathrm{w}^{*\mathrm{T}}\mathrm{m}_1\mathrm{m}_2}{2}=2$$

所以判别函数为：$\mathrm{g(x)}=\begin{bmatrix}6\\6\end{bmatrix}^{\mathrm{T}}\mathrm{x}-2$

最小平方误差法：

$$Y=\begin{bmatrix}1&0\\0&1\\1&1\\0&0\\1&0\\0&1\end{bmatrix}\qquad Y^{+}=\left(Y^TY\right)^{-1}Y^T=\frac{1}{8}\begin{bmatrix}3 & -1 & 2 & 0 & 3- \\ -1 & 3 & 2 & 0- & 1\end{bmatrix}$$

令 $b=\begin{bmatrix}1 & 1 & 1 & 1 & 1 & 1\end{bmatrix}^T$， $a^*=Y^+b=\begin{bmatrix}\frac{3}{4}\\ \frac{3}{4}\end{bmatrix}$

五、答：

$$\mu=\begin{bmatrix}0\\0\end{bmatrix}$$

$$S_b=\frac{1}{2}\sum_{i=1}^{2}(\mu_i-\mu)(\mu_i-\mu)^T=\frac{1}{4}(\mu_1-\mu_2)(\mu_1-\mu_2)^T=\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}$$

$$S_w=\frac{1}{2}(S_1+S_2)=\frac{1}{2}\begin{bmatrix}2 & 0\\0 & 5\end{bmatrix},\qquad S_w^{-1}=\begin{bmatrix}5 & 0\\0 & 2\end{bmatrix}$$

$$S_w^{-1}S_b=\begin{bmatrix}5 & 5\\2 & 2\end{bmatrix}$$

因为$S_w^{-1}S_b$的秩为 1,所以$S_w^{-1}S_b$只有一个非零本征值,W 是 D×1 矩阵,即 W=w。为求$S_w^{-1}S_b$的本征值应解方程：

$$S_w^{-1}S_bw=\lambda_1w$$

或

$$\frac{1}{4}S_w^{-1}(\mu_1-\mu_2)(\mu_1-\mu_2)^Tw=\lambda_1w$$

$\frac{1}{4}(\mu_1-\mu_2)^Tw$是标量，所以

$$w=S_w^{-1}(\mu_1-\mu_2)=\begin{bmatrix}10\\4\end{bmatrix}$$

六、答：

$X=\emptyset, J(x_1)=0, J(x_2)=3, J(x_3)=3, J(x_4)=4$

所以$X=\{x_4\}, J(X+x_1)=4, J(X+x_2)=5.2785, J(X+x_3)=5.1569$

所以$X=\{x_2x_4\}, J(X+x_1)=5.2785, J(X+x_3)=6.2237$

所以$X=\{x_2x_3x_4\}$

七、答：

第一类：(-1,1), (-1,0), (-1,-1), (0,-1), (0,-2)

第二类：(0,1)

第三类：(1,1), (1,0), (2,0)