模式识别(第二版)习题解答

目录

§1 绪论

略

§2 贝叶斯决策理论

- 2.1 如果只知道各类的先验概率，最小错误率贝叶斯决策规则应如何表示？

  解：设一个有C类，每一类的先验概率为$P(w_i)$，$i = 1, ..., C$。此时最小错误率贝叶斯决策规则为：如果$i^* = \max\limits_i P(w_i)$，则$x \in w_i$。

- 2.2 利用概率论中的乘法定理和全概率公式证明贝叶斯公式（教材中下面的公式有错误）

$$P(w_i|x) = \frac{p(x|w_i)P(w_i)}{p(x)}.$$

  证明：

$$\begin{aligned} P(w_i|x) &= \frac{P(w_i, x)}{p(x)} \\ &= \frac{p(x|w_i)P(w_i)}{p(x)} \end{aligned}$$

- 2.3 证明：在两类情况下$P(w_i|x) + P(w_2|x) = 1$。

  证明：

$$\begin{aligned} P(w_1|x) + P(w_2|x) &= \frac{P(w_1, x)}{p(x)} + \frac{P(w_2, x)}{p(x)} \\ &= \frac{P(w_1, x) + P(w_2, x)}{p(x)} \\ &= \frac{p(x)}{p(x)} \\ &= 1 \end{aligned}$$

- 2.4 分别写出在以下两种情况

  1. $P(x|w_1) = P(x|w_2)$
  2. $P(w_1) = P(w_2)$

  下的最小错误率贝叶斯决策规则。

  解：当$P(x|w_1) = P(x|w_2)$时，如果$P(w_1) > P(w_2)$，则$x \in w_1$，否则$x \in w_2$。

  当$P(w_1) = P(w_2)$时，如果$P(x|w_1) > P(x|w_2)$，则$x \in w_1$，否则$x \in w_2$。

- 2.5

  1. 对c类情况推广最小错误率率贝叶斯决策规则；
  2. 指出此时使错误率最小等价于后验概率最大，即$P(w_i|x) > P(w_j|x)$ 对一切$j \neq i$成立时，$x \in w_i$。

解：对于c类情况，最小错误率贝叶斯决策规则为:

如果 $P(w_i|x) = \max\limits_{j=1,\ldots,c} P(w_j|x)$，则$x \in w_i$。利用贝叶斯定理可以将其写成先验概率和类条件概率相联系的形式，即

如果 $p(x|w_i)P(w_i) = \max\limits_{j=1,\ldots,c} p(x|w_j)P(w_j)$，则$x \in w_i$。

- 2.6 对两类问题，证明最小风险贝叶斯决策规则可表示为，若

$$\frac{p(x|w_1)}{p(x|w_2)} > \frac{(\lambda_{12}-\lambda_{22})P(w_2)}{(\lambda_{21}-\lambda_{11})P(w_1)},$$

则$x \in w_1$，反之则属于w_2。

解：计算条件风险

$$\begin{aligned} R(\alpha_1|x) &= \sum_{j=1}^{2} \lambda_{1j} P(w_j|x) \\ &= \lambda_{11}P(w_1|x) + \lambda_{12}P(w_2|x) \end{aligned}$$

$$\begin{aligned} R(\alpha_2|x) &= \sum_{j=1}^{2} \lambda_{2j} P(w_j|x) \\ &= \lambda_{21}P(w_1|x) + \lambda_{22}P(w_2|x) \end{aligned}$$

如果$R(\alpha_1|x) < R(\alpha_2|x)$，则$x \in w_1$。

$$\begin{aligned} \lambda_{11}P(w_1|x) + \lambda_{12}P(w_2|x) &< \lambda_{21}P(w_1|x) + \lambda_{22}P(w_2|x) \\ (\lambda_{21}-\lambda_{11})P(w_1|x) &> (\lambda_{12}-\lambda_{22})P(w_2|x) \\ (\lambda_{21}-\lambda_{11})P(w_1)p(x|w_1) &> (\lambda_{12}-\lambda_{22})P(w_2)p(x|w_2) \\ \frac{p(x|w_1)}{p(x|w_2)} &> \frac{(\lambda_{12}-\lambda_{22})P(w_2)}{(\lambda_{21}-\lambda_{11})P(w_1)} \end{aligned}$$

所以，如果$\dfrac{p(x|w_1)}{p(x|w_2)} > \dfrac{(\lambda_{12}-\lambda_{22})P(w_2)}{(\lambda_{21}-\lambda_{11})P(w_1)}$，则$x \in w_1$。反之则$x \in w_2$。

- 2.7 若$\lambda_{11} = \lambda_{22} = 0, \lambda_{12} = \lambda_{21}$，证明此时最小最大决策面是来自两类的错误率相等。

解： 最小最大决策时满足

$$(\lambda_{11}-\lambda_{22}) + (\lambda_{21}-\lambda_{11})\int_{\mathscr{R}_2} p(x|w_1)dx - (\lambda_{12}-\lambda_{22})\int_{\mathscr{R}_1} p(x|w_2)dx = 0$$

容易得到

$$\int_{\mathscr{R}_1} p(x|w_2)dx = \int_{\mathscr{R}_2} p(x|w_1)dx$$

所以此时最小最大决策面使得$P_1(e) = P_2(e)$

- 2.8 对于同一个决策规则判别函数可定义成不同形式，从而有不同的决策面方程，指出决策区域是不变的。

解： 对于同一决策规则（如最小错误率贝叶斯决策规则），它的判别函数可以是$j^* = \max\limits_{j=1,\dots,c} P(w_j|x)$，则$x \in w_{j^*}$。另外一种形式为$j^* = \max\limits_{j=1,\dots,c} p(x|w_j)P(w_j)$，则$x \in w_{j^*}$。考虑两类问题的分类决策面为：$P(w_1|x) = P(w_2|x)$，与$p(x|w_1)P(w_1) = p(x|w_2)P(w_2)$是相同的。

- 2.9 写出两类和多类情况下最小风险贝叶斯决策判别函数和决策面方程。
- 2.10 随机变量$l(x)$定义为$l(x) = \dfrac{p(x|w_1)}{p(x|w_2)}$，$l(x)$又称为似然比，试证明
  - (1) $E\{l^n(x)|w_1\} = E\{l^{n+1}(x)|w_2\}$
  - (2) $E\{l(x)|w_2\} = 1$
  - (3) $E\{l(x)|w_1\} - E^2\{l(x)|w_2\} = \text{var}\{l(x)|w_2\}$（教材中题目有问题）

  证明： 对于(1)，$E\{l^n(x)|w_1\} = \displaystyle\int l^n(x)p(x|w_1)\mathrm{d}x = \int \frac{(p(x|w_1))^{n+1}}{(p(x|w_2))^n}\mathrm{d}x$ 又$E\{l^{n+1}(x)|w_2\} = \displaystyle\int l^{n+1}p(x|w_2)\mathrm{d}x = \int \frac{(p(x|w_1))^{n+1}}{(p(x|w_2))^n}\mathrm{d}x$ 所以，$E\{l^n(x)|w_1\} = E\{l^{n+1}(x)|w_2\}$

  对于(2)，$E\{l(x)|w_2\} = \displaystyle\int l(x)p(x|w_2)\mathrm{d}x = \int p(x|w_1)\mathrm{d}x = 1$

  对于(3)，$E\{l(x)|w_1\} - E^2\{l(x)|w_2\} = E\{l^2(x)|w_2\} - E^2\{l(x)|w_2\} = \text{var}\{l(x)|w_2\}$
- 2.11 $x_j(j = 1, 2, ..., n)$为n个独立随机变量，有$E[x_j|w_i] = ij\eta$，$\text{var}[x_j|w_i] = i^2j^2\sigma^2$，计算在$\lambda_{11} = \lambda_{22} = 0$及$\lambda_{12} = \lambda_{21} = 1$的情况下，由贝叶斯决策引起的错误率。（中心极限定理）

  解： 在$0-1$损失下，最小风险贝叶斯决策与最小错误率贝叶斯决策等价。
- 2.12 写出离散形式的贝叶斯公式。

  解：

$$P(w_i|x) = \frac{P(x|w_i)P(x)}{\sum_{j=1}^{c} P(x|w_i)P(w_i)}$$

- 2.13 把连续情况的最小错误率贝叶斯决策推广到离散情况，并写出其判别函数。
- 2.14 写出离散情况条件风险$R(a_i|x)$的定义，并指出其决策规则。

  解：

$$\begin{aligned} R(a_i|x) &= \sum_{j=1}^{c} \lambda_{ij}P(w_j|x) \\ &= \sum_{j=1}^{c} \lambda_{ij}p(x|w_j)P(w_j)////\text{omit the same part } p(x) \end{aligned}$$

  $R(a_k|x) = \min\limits_{j=1,2,\dots,N} R(a_j|x)$，则$a_k$就是最小风险贝叶斯决策。
- 2.15 证明多元正态分布的等密度点轨迹是一个超椭球面，且其主轴方向由Σ的特征向量决定，轴长度由Σ的特征值决定。

  证明： 多元正态分布的等密度点满足：$x^T\Sigma^{-1}x = C$，C为常数。

- 2.16 证明$Mahalanobis$距离r符合距离定义三定理，即
  - (1) $r(a,b)=r(b,a)$
  - (2) 当且仅当$a=b$时，$r(a,b)=0$
  - (3) $r(a,c)\leq r(a,b)+r(b,c)$

  证明：

  (1) $r(a,b)=(a-b)^T\Sigma^{-1}(a-b)=(b-a)^T\Sigma^{-1}(b-a)=r(b,a)$

  (2) Σ为半正定矩阵所以$r(a,b)=(a-b)^T\Sigma^{-1}(a-b)\geq 0$，只有当$a=b$时，才有$r(a,b)=0$。

  (3) Σ^{-1}可对角化，$\Sigma^{-1}=P\Lambda P^T$

- 2.17 若将Σ^{-1}矩阵写为：$\Sigma^{-1}=\begin{bmatrix}h_{11} & h_{12} & \cdots & h_{1d}\\ h_{12} & h_{22} & \cdots & h_{2d}\\ \vdots & \vdots & \ddots & \vdots\\ h_{1d} & h_{2d} & \cdots & h_{dd}\end{bmatrix}$，证明$Mahalanobis$距离平方为

$$\gamma^2=\sum_{i=1}^{d}\sum_{j=1}^{d}h_{ij}(x_i-u_i)(x_j-u_j)$$

  证明：

$$\begin{aligned}\gamma^2&=(x-u)^T\begin{bmatrix}h_{11} & h_{12} & \cdots & h_{1d}\\ h_{12} & h_{22} & \cdots & h_{2d}\\ \vdots & \vdots & \ddots & \vdots\\ h_{1d} & h_{2d} & \cdots & h_{dd}\end{bmatrix}(x-u)\\&=\sum_{i=1}^{d}\sum_{j=1}^{d}h_{ij}(x_i-u_i)(x_j-u_j)\end{aligned}$$

- 2.18 分别对于$d=2,d=3$证明对应与Mahalanobis距离γ的超椭球体积是$V=V_d|\Sigma|^{\frac{1}{2}}\gamma^d$
- 2.19 假定x和m是两个随机变量，并设在给定m时，x的条件密度为

$$p(x|m)=(2\pi)^{\frac{1}{2}}\sigma^{-1}\exp\left\{-\frac{1}{2}(x-m)^2/\sigma^2\right\}$$

  再假设m的边缘分布是正态分布，期望值是m_0，方差是σ_m^2，证明

$$p(m|x)=\frac{(\sigma^3+\sigma_m)^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}\sigma\sigma_m}\exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_m^2}{\sigma^2\sigma_m^2}\left(m-\frac{\sigma_m^2x+m_0\sigma^2}{\sigma^2+\sigma_m^2}\right)^2\right]$$

证明：

$$\begin{aligned}
p(m|x) &= \frac{p(x|m)p(m)}{p(x)} \\
&= \frac{p(x|m)p(m)}{\int p(x|m)p(m)\mathrm{d}m} \\
&= \frac{(2\pi)^{\frac{1}{2}}\sigma^{-1}\exp\left\{-\frac{1}{2}(x-m)^2/\sigma^2\right\}(2\pi)^{\frac{1}{2}}\sigma_m^{-1}\exp\left\{-\frac{1}{2}(m-m_0)^2/\sigma_m^2\right\}}{\int(2\pi)^{\frac{1}{2}}\sigma^{-1}\exp\left\{-\frac{1}{2}(x-m)^2/\sigma^2\right\}(2\pi)^{\frac{1}{2}}\sigma_m^{-1}\exp\left\{-\frac{1}{2}(m-m_0)^2/\sigma_m^2\right\}\mathrm{d}m} \\
&= \frac{(\sigma^3+\sigma_m)^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}\sigma\sigma_m}\exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_m^2}{\sigma^2\sigma_m^2}\left(m-\frac{\sigma_m^2x+m_0\sigma^2}{\sigma^2+\sigma_m^2}\right)^2\right]
\end{aligned}$$

- 2.20 对$\Sigma_i=\sigma^2I$的特殊情况，证明
  - (1) 若$P(w_i)\neq P(w_j)$，则超平面靠近先验概率较小的类；
  - (2) 在甚么情况下，先验概率对超平面的位置影响不大。

  证明： (1)当$P(w_i)=P(w_j)$时，超平面经过$x_0=\frac{1}{2}(u_i+u_j)$，则对于先验概率较小的类属于它的区域会减少，所以超平面经过的点会靠近先验概率较小的类。 (可以这样理解，具体证明也很简单)

  (2)？不知道这是什么问题，先验概率不管在什么时候都很重要！

- 2.21 对$\Sigma_i=\Sigma$的特殊情况，指出在先验概率不等时，决策面沿u_i点与u_j点连线向先验概率小的方向移动。

  证明： 同上面一题解释一样。

- 2.24 似然比决策准则为：若

- 2.23 二维正态分布，$u_1=(-1,0)^T,u_2=(1,0)^T,\Sigma_1=\Sigma_2=I,P(w_1)=P(w_2)$。试写出对数似然比决策规则。

  解：

$$\begin{aligned}
h(x) &= -\ln\left[l(x)\right] \\
&= -\ln p(x|w_1)+\ln p(x|w_2) \\
&= \frac{1}{2}(x_1-u_1)^T\Sigma_1^{-1}(x_1-u_1)-\frac{1}{2}(x_2-u_2)^T\Sigma_2^{-1}(x_2-u_2)+\frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|} \\
&= \frac{1}{2}\left[(x-u_1)^T(x-u_1)-(x-u_2)^T(x-u_2)\right]
\end{aligned}$$

  而，$\ln\left[\frac{P(w_1)}{P(w_2)}\right]=0$。所以判别规则为当$(x-u_1)^T(x-u_1)>(x-u_2)^T(x-u_2)$则$x\in w_1$,反之则$s\in w_2$。即将$x$判给离它最近的$u_i$的那个类。

- 2.24 在习题2.23中若$\Sigma_1\neq\Sigma_2$，$\Sigma_1=\begin{bmatrix}1 & \frac{1}{2}\\ \frac{1}{2} & 1\end{bmatrix}$，$\Sigma_2=\begin{bmatrix}1 & -\frac{1}{2}\\ -\frac{1}{2} & 1\end{bmatrix}$，写出负对数似然比决策规则。

解：

$$
\begin{aligned}
h(x) &= -\ln\left[l(x)\right] \\
&= -\ln p(x|w_1) + \ln p(x|w_2) \\
&= \frac{1}{2}(x_1-u_1)^T\Sigma_1^{-1}(x_1-u_1) - \frac{1}{2}(x_2-u_2)^T\Sigma_2^{-1}(x_2-u_2) + \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|} \\
&= \frac{1}{2}x^T(\Sigma_1^{-1}-\Sigma_2^{-1})x - (\Sigma_1^{-1}u_i - \Sigma_2^{-1}u_j)^T x + \\
&= \frac{1}{2}(u_1^T\Sigma_1^{-1}u_1 - u_2^T\Sigma_2^{-1}u_2 + \ln\frac{|\Sigma_1|}{|\Sigma_2|}) \\
&= -\frac{4}{3}x_1x_2 + \frac{4}{3}x_1
\end{aligned}
$$

而，$\ln\left[\frac{P(w_1)}{P(w_2)}\right] = 0$。决策面为$x_1(x_2-1)=0$，如图1所示

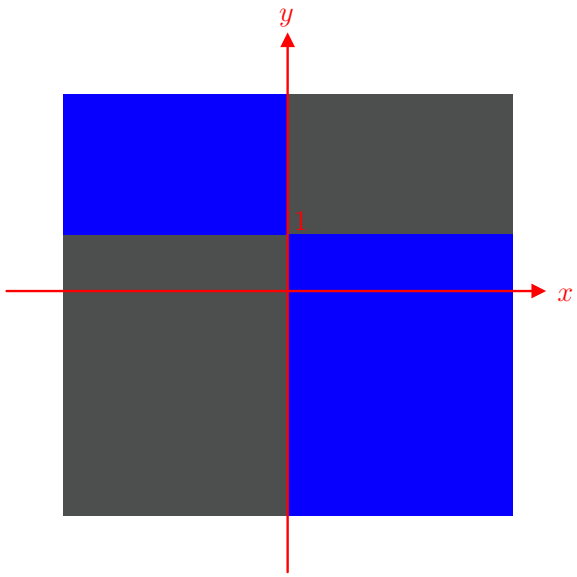

图 1: 分类决策面

- 2.25 在习题2.24的情况下，若考虑损失函数$\lambda_{11}=\lambda_{22}=0, \lambda_{12}=\lambda_{21}$，画出似然比阈值与错误率之间的关系。
  - (1)求出$P(e)=0.05$时完成Neyman-Pearson决策时总的错误率；（$P(e)$应该为$P(e_1)$或者$P(e_2)$）
  - (2)求出最小最大决策的域值和总的错误率。

解：

（1）损失函数在0-1损失函数条件下的最小风险贝叶斯决策等价于最小错误率贝叶斯决策。似然比等于0的情况下错误率最小。当$P(e_1)=0.05$时，

（2）最小最大决策时，$(\lambda_{11}-\lambda_{22})+(\lambda_{21}-\lambda_{11})\int_{\mathscr{R}_2} p(x|w_1)\mathrm{d}x-(\lambda_{12}-\lambda_{22})\int_{\mathscr{R}_1} p(x|w_2)\mathrm{d}m=0$ 可以得到，$\int_{\mathscr{R}_2} p(x|w_1)\mathrm{d}x=\int_{\mathscr{R}_1} p(x|w_2)\mathrm{d}m$，所以$\mathscr{R}_1=\{(x_1,x_2)|x_1(x_2-1)>0\}$，$\mathscr{R}_2=\{(x_1,x_2)|x_1(x_2-1)<0\}$

§3 概率密度函数的估计

- 3.1 设总体分布密度为$N(u,1)$，$-\infty<u<+\infty$，并设$\mathscr{X}=\{x_1,x_2,...,x_N\}$，分别用最大似然估计和贝叶斯估计计算$\hat{u}$。已知$u$的先验分布$p(u)\sim N(0,1)$。

  解：似然函数为：

$$L(u)=\ln p(\mathscr{X}|u)=\sum_{i=1}^{N}\ln p(x_i|u)=-\frac{1}{2}\sum_{i=1}^{N}(x_i-u)^2+C$$

  似然函数u求导

$$\frac{\partial L(u)}{\partial u}=\sum_{i=1}^{N}x_i-Nu=0$$

  所以u的最大似然估计：$\hat{u}=\frac{1}{N}\sum_{i=1}^{N}x_i$

  贝叶斯估计：MAP(maximum a posterior)

$$\begin{aligned}
p(u|\mathscr{X}) &= \frac{p(\mathscr{X}|u)p(u)}{\int p(\mathscr{X}|u)p(u)\mathrm{d}u} \\
&= \alpha\prod_{i=1}^{N}p(x_i|u)p(u) \\
&= \alpha\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_i-u)^2}{2\sigma^2}\right]\cdot\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left[-\frac{(u-u_0)^2}{2\sigma_0^2}\right] \\
&= \alpha'\exp\left[-\frac{1}{2}\left(\sum_{i=1}^{N}\left(\frac{u-x_i}{\sigma}\right)^2+\left(\frac{u-u_0}{\sigma_0}\right)^2\right)\right] \\
&= \alpha''\exp\left[-\frac{1}{2}\left[\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)u^2-2\left(\frac{1}{\sigma^2}\sum_{i=1}^{N}x_i+\frac{u_0}{\sigma_0^2}\right)u\right]\right]
\end{aligned}$$

  将$p(u|\mathscr{X})$写成$N(u_n,\sigma_n^2)$的形式，利用待定系数法，可以求得：

$$\frac{1}{\sigma_n^2}=\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}$$

$$\frac{u_n}{\sigma_n^2}=\frac{1}{\sigma^2}\sum_{i=1}^{N}x_i+\frac{u_0}{\sigma_0^2}$$

进一步求得u_n和σ_n^2

$$u_n = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}m_N + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}u_0$$
$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}$$

其中，$m_N = \frac{1}{N}\sum_{i=1}^{N} x_i$，$u_n$就是贝叶斯估计。

- 3.3 设$\mathscr{X} = \{x_1, x_2, ..., x_N\}$为来自点二项式分布的样本集，即$f(x,P) = P^x Q^{(1-x)}, x = 0, 1, 0 \leq P \leq 1, Q = 1-P$，试求参数$P$的最大似然估计。

  解：似然函数为：

$$\begin{aligned} L(P) &= \sum_{i=1}^{N} \ln\left(P^{x_i}(1-P)^{(1-x_i)}\right) \\ &= \sum_{i=1}^{N} x_i \ln P + N\ln(1-P) + \sum_{i=1}^{N} x_i \ln P \end{aligned}$$

  两边对P求导可得

$$\frac{\partial L}{\partial P} = \frac{\sum_{i=1}^{N} x_i}{P} - \frac{N}{1-P} + \frac{\sum_{i=1}^{N} x_i}{1-P} = 0$$

  所以P得最大似然估计为：$\hat{P} = \frac{1}{N}\sum_{i=1}^{N} x_i$。

- 3.4 假设损失函数为二次函数$\lambda(\hat{P}, P) = (\hat{P}-P)^2$，以及$P$的先验密度为均匀分布$f(P) = 1, 0 \leq P \leq 1$。在这样的假设条件下，求3.3题的贝叶斯估计$\hat{P}$。

  解：

$$P(\mathscr{X}|P) = \prod_{i=1}^{N} P^{x_i}(1-P)^{1-x_i}$$

  利用贝叶斯公式求出P的后验概率

$$\begin{aligned} P(P|\mathscr{X}) &= \frac{P(\mathscr{X}|P)f(P)}{\int P(\mathscr{X}|P)f(P)\mathrm{d}P} \\ &= \frac{\prod_{i=1}^{N} P^{x_i}(1-P)^{1-x_i}}{\int_0^1 \prod_{i=1}^{N} P^{x_i}(1-P)^{1-x_i}\mathrm{d}P} \end{aligned}$$

- 3.7 设$\mathcal{X} = \{x_1, x_2, ..., x_N\}$是来自$p(x|\theta)$的随机样本，其中$0 \leqslant x \leqslant \theta$时，$p(x|\theta) = \frac{1}{\theta}$，否则为0。证明$\theta$的最大似然估计是$\max_k x_k$。

  最大似然估计的目标是最大化对数似然函数，对数似然函数为：

$$\begin{aligned} \mathcal{L}(\theta) &= \ln p(\mathcal{X}|\theta) &(1)\\ &= -N\ln\theta &(2) \end{aligned}$$

要使得$\mathcal{L}(\theta)$最大，同时要满足$\theta \geqslant x$。那么此时θ的最大似然估计为:

$$\theta^* = \max_k x_k \tag{3}$$

- 3.8 利用矩阵恒等式

$$(A^{-1}+B^{-1})^{-1} = A(A+B)^{-1}B = B(A+B)^{-1}A$$

证明:

$$\begin{aligned}(A^{-1}+B^{-1})A(A+B)^{-1}B &= (I+B^{-1}A)(A+B)^{-1}B \\ &= B^{-1}(B+A)(A+B)^{-1}B \\ &= B^{-1}B \\ &= I\end{aligned}$$

所以: $(A^{-1}+B^{-1})^{-1} = A(A+B)^{-1}B$ 同理证明$(A^{-1}+B^{-1})^{-1} = B(A+B)^{-1}A$

- 3.15 设$p(x) \sim N(u,\sigma^2)$，窗函数$\varphi(x) \sim N(0,1)$，指出Parzen窗估计

$$\hat{p}_N(x) = \frac{1}{Nh_N}\sum_{i=1}^{N}\varphi\left(\frac{x-x_i}{h_N}\right)$$

对于小的h_N，有如下性质:

  - (1) $E\left[\hat{p}_N(x)\right] \sim N(u,\sigma^2+h_N^2)$
  - (2) $Var\left[\hat{p}_N(x)\right] = \frac{1}{Nh_N 2\sqrt{\pi}}p(x)$

证明:

(1)$E[\hat{p}_N(x)] = \int \hat{p}_N(x)p(x)\mathrm{d}x$

8.1 S_w表示类内离散度矩阵，S_b表示类间离散度矩阵

§4 线性判别函数

- 4.1 （1）指出从x到超平面$g(x) = w^Tx + w_0 = 0$的距离$r = \frac{|g(x)|}{||w||}$是在$g(x_q) = 0$的约束条件下，使$||x-x_q||^2$达到极小解;

（2）指出在超平面上的投影是$x_p = x - \frac{g(x)}{||w||^2}w$

解： （1）设x在超平面的正侧$g(x) > 0$，x_q是x在超平面上的投影点，则$w^Tx_q + w_0 = 0$。设x到平面的距离为r，则$x - x_p = r\frac{w}{||w||}$，所以$w^Tx - w^Tx_p = r||w||$，得到$r = \frac{w^Tx+w_0}{||w||} = \frac{g(x)}{||w||}$。

x在超平面负侧时$g(x)<0$，得$r=\dfrac{-w^Tx-w_0}{||w||}=\dfrac{-g(x)}{||w||}$。

所以$r=\dfrac{|g(x)|}{||w||}$

（2）x在超平面正侧时，$x-x_p=r\dfrac{w}{||w||}=\dfrac{g(x)}{||w||^2}w$，所以$x_p=x-\dfrac{g(x)}{||w||^2}w$；当$x$在超平面的负侧时，$x-x_p=-r\dfrac{w}{||w||}=\dfrac{g(x)}{||w||^2}w$，所以$x_p=x-\dfrac{g(x)}{||w||^2}w$。

- 4.3 设有一维空间二次判别函数$g(x)=5+7x+9x^2$
  - (1)试映射成广义齐次线性判别函数；
  - (2)总结把高次函数映射成齐次线性函数的方法。

解：（1）设$y=[y_1,y_2,y_3]^T=[1,x,x^2]^T$，$a=[5,7,9]^T$，则广义齐次线性判别函数为：$g(x)=a^Ty$

（2）对于n次函数$g(x)=c_0+c_1x+c_2x^2+...+c_nx^n$，令$y=[y_1,y_2,...,y_{n+1}]^T=[1,x,...,x^n]^T$，$a=[c_0,c_1,...,c_n]^T$，则$g(x)=a^Ty$。

- 4.3 （1）通过映射把一维二次判别函数$g(x)=a_1+a_2x+a_3x^2$映射成三维广义线性判别函数；

  （2）若x在一维空间具有分布密度$p(x)$，说明三维空间中的分布退化成只在一条曲线上有值，且曲线上值无穷大。

  解：（1） $y=[1,x,x^2]^T$，$a=[a_1,a_2,a_3]^T$，则$g(x)=a^Ty$。

  （2）映射$y=[1,x,x^2]^T$把一条直线映射为三维空间中的一条抛物线。

- 4.4 对于二维线性判别函数$g(x)=x_1+2x_2-2$
  - （1）将判别函数写成$g(x)=w^Tx+w_0$的形式，并画出$g(x)=0$的几何图形；
  - （2）映射成广义齐次线性函数$g(x)=a^Ty$；
  - （3）指出上述X空间实际是Y空间的一个子空间，且$a^Ty=0$对于X子空间的划分和原空间中$w^T+w_0=0$对原X空间的划分相同，并在图上表示出来。

  解：（1）$w=[1,2]^T,x=[x_1,x_2]^T,w_0=-2$，则$g(x)=w^Tx+w_0$，$g(x)=0$的图形如下图2：

  （2）$y=[y_1,y_2,y_3]^T=[1,x_1,x_2]^T$，$a=[-2,1,2]^T$，则$g(x)=a^Ty$。

  （3）$y_1=1,y_2=x_1,y_3=x_2$，在所以所有的样本在Y空间中的一个平面$y_1=1$上。

- 4.5 指出在Fisher线性判别中，w的比例因子对Fisher判别结果无影响。

  解： 假设w乘一比例因子α，αw，经过投影后得到$y=\alpha w^Tx$。相当于对所有样本乘以一个比例因子，所以对判别结果没有影响。

- 4.6 证明两向量外积组成的矩阵一般是奇异的。

  证明： 设两向量$a,b\in\mathbb{R}^n$，它们的外积为：$A=ab^T$，因为ab^T与b^Ta有相同的非零特征值，容易得到 A的特征值为$b^Ta,\underbrace{0,0,...,0}_{n-1}$。有零特征值肯定是奇异的，除非$n=1$。

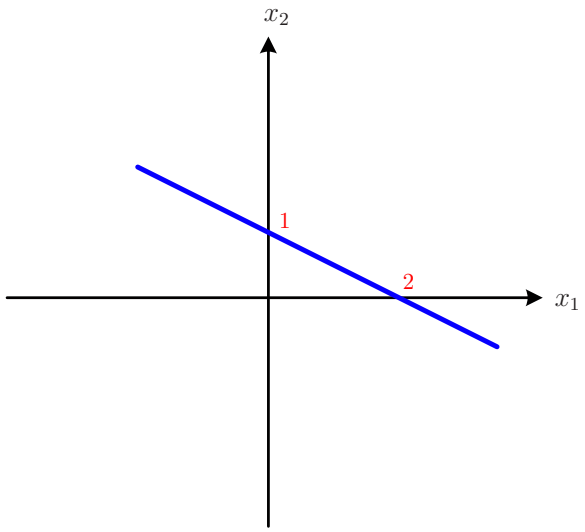

图 2: $g(x)=0$的几何图形

- 4.8 证明在正态等方差条件下，Fisher线性判别等价于贝叶斯判别。

  证明： 在正态等方差的条件下，判别函数$g(x)=w^Tx+w_0$中$w=\Sigma^{-1}(u_1-u_2)$，在Fisher线性判别中最优投影方向为：$w=\Sigma^{-1}(u_1-u_2)$。

- 4.9 证明
  - （1）引入余量b以后的解区($a^Ty_i\geq b$)位于原来的解区($a^Ty_i>0$)之中；
  - （2）与原解区边界之间的距离为$\frac{b}{||y_i||}$。

  解：(1) 设$a*$满足$a^Ty_i\geq b$，则它一定也满足$a^Ty_i>0$，所以引入余量后的解区位于原来的解区$a^Ty>0$之中。(2) $a^Ty_i\geq b$解区边界为：$a^Ty_i=b$，$a^Ty_i>0$解区边界为：$a^Ty_i=0$，$a^Ty_i=b$到$a^Ty_i=0$的距离为$\frac{b}{||y_i||}$。

- 4.10 证明，在几何上，感知器准则函数正比于被错分类样本到决策面的距离之和。

  证明： 感知器准则函数为$J(a)=\sum_{y\in\mathscr{Y}}(-a^Ty)$。决策面方程为：$a^Ty=0$。当$y$为错分类样本时，有$a^Ty\leq 0$，到决策面的距离为$-a^Ty$。所有错分类样本到决策面的距离之和为$\sum_{y\in\mathscr{Y}}(-a^Ty)$，就是感知器准则函数。

- 4.12 写出Widrow-Hoff法程序框图。

  解： 平方误差准则函数$J(a)=||Ya-b||^2=\sum_{n=1}^{N}(a^Ty_n-b_n)^2$，它的最小二乘解，伪逆解或MSE解为：$a^*=(Y^TY)^{-1}Y^Tb$，采用梯度下降法来求解$a^*$。$J(a)$的梯度为$\nabla J(a)=2Y^T(Ya-b)$，则梯度下降法可以写成$\begin{cases} a(1) \\ a(k+1)=a(k)-\rho_kY^T(Ya-b) \end{cases}$，选择$\rho_k=\frac{\rho_1}{k}$，式中$\rho_1$为任意正常数。

为了进一步减小计算量和存储量，可以将上述算法修改为（单样本修正）

$$\begin{cases} a(1) \\ a(k+1) = a(k) - \rho_k(a(k)^T y_k - b_k)y^k \end{cases}$$

让ρ_k随着k的增加而逐渐减小，以确保算法收敛。一般选择$\rho_k = \frac{\rho_1}{k}$，还有$y^k$和前面感知器准则函数中的单样本修正法一样，是在无限重复序列中的错分类样本。

- 4.13
  - （1）证明矩阵恒等式$(A + xx^T)^{-1} = A^{-1} - \frac{A^{-1}xx^TA^{-1}}{1 + x^TA^{-1}x}$
  - （2）利用上试结果证明式（4-98）。

证明： （1）

$$\begin{aligned}(A + xx^T)\left(A^{-1} - \frac{A^{-1}xx^TA^{-1}}{1 + x^TA^{-1}x}\right) &= (A + xx^T)\left(I - \frac{A^{-1}xx^T}{1 + x^TA^{-1}x}\right)A^{-1} \\ &= \left(A + xx^T - \frac{xx^T}{1 + x^TA^{-1}x} - \frac{xx^TA^{-1}xx^T}{1 + x^TA^{-1}x}\right)A^{-1} \\ &= AA^{-1} \\ &= I\end{aligned}$$

所以$(A + xx^T)^{-1} = A^{-1} - \frac{A^{-1}xx^TA^{-1}}{1 + x^TA^{-1}x}$

（2）$R(k+1)^{-1} = R(k)^{-1} + y_k y_k^T$，利用上面的结果可以得到：$R(k+1) = R(k) - \frac{R(k)y_k y_k^T R(k)}{1 + y_k^T R(k) y_k}$

- 4.14 考虑准则函数

$$J(a) = \sum_{y \in \mathscr{Y}(a)} (a^T y - b)^2$$

其中$\mathscr{Y}(a)$是使$a^T y \le b$的样本集合。设y_1是$\mathscr{Y}(a)$中的唯一样本，则$J(a)$的梯度为$\nabla J(a) = 2(a_k^T y_1 - b)y_1$，二阶偏导数矩阵$D = 2y_1 y_1^T$。据此证明，若最优步长选择为$\rho_k = \frac{||\nabla J(a)||^2}{\nabla J^T(a) D \nabla J(a)}$时，梯度下降法的迭代公式为：

$$a_{k+1} = a_k + \frac{b - a_k^T y_1}{||y_1||^2} y_1$$

证明： y_1是$\mathscr{Y}(a)$中的唯一样本，则准则函数为$J(a) = \sum_{y \in \mathscr{Y}(a)} (a^T y - b)^2 = (a^T y_1 - b)^2$，所以$\nabla J(a) = 2(a^T y_1 - b)y_1$，二阶偏导数矩阵为$D = 2y_1 y_1^T$。

梯度下降的迭代公式为：$a_{k+1} = a_k - \rho_k \nabla J(a_k)$，$\rho_k = \frac{4(a_k^T y_1 - b)^2 ||y_1||^2}{8(a_k^T y_1 - b)^2 y_1^T y_1 y_1^T y_1} = \frac{1}{2||y_1||^2}$，将$\rho_k$代入梯度下降的迭代公式：$a_{k+1} = a_k + \frac{b - a_k^T y_1}{||y_1||^2} y_1$

- 4.15 证明：当取

$$b=\left[\underbrace{\frac{N}{N_1},...,\frac{N}{N_1}}_{N_1},\underbrace{\frac{N}{N_2},...,\frac{N}{N_2}}_{N_2}\right]$$

MSE解等价于Fisher解。

证明： $Y=\begin{bmatrix}y_1^T\\y_2^T\\\vdots\\y_N^T\end{bmatrix}=\begin{bmatrix}\mathbf{1}_1 & \mathbf{X}_1\\-\mathbf{1}_2 & -\mathbf{X}_2\end{bmatrix}$，$a=[w_0,\mathbf{w}]^T$则$Y^TYa=Y^Tb$，化为：

$$\begin{bmatrix}\mathbf{1}_1^T & -\mathbf{1}_2^T\\\mathbf{X}_1^T & -\mathbf{X}_2^T\end{bmatrix}\begin{bmatrix}\mathbf{1}_1 & \mathbf{X}_1\\-\mathbf{1}_2 & -\mathbf{X}_2\end{bmatrix}\begin{bmatrix}w_0\\\mathbf{w}\end{bmatrix}=\begin{bmatrix}\mathbf{1}_1^T & -\mathbf{1}_2^T\\\mathbf{X}_1^T & -\mathbf{X}_2^T\end{bmatrix}\begin{bmatrix}\frac{N}{N_1}\mathbf{1}_1\\\frac{N}{N_1}\mathbf{1}_2\end{bmatrix}$$

设$m_1=\frac{1}{N_1}\sum_{i\in C_1}x_i, m_2=\frac{1}{N_2}\sum_{i\in C_2}x_i$，上式可化为：

$$\begin{bmatrix}N & (N_1m_1+N_2m_2)^T\\(N_1m_1+N_2m_2) & S_w+N_1m_1m_1^T+N_2m_2m_2^T\end{bmatrix}\begin{bmatrix}w_0\\\mathbf{w}\end{bmatrix}=\begin{bmatrix}0\\N(m_1-m_2)\end{bmatrix}$$

式中，$S_w=\sum_{i=1}^{2}\sum_{j\in C_i}(x_j-m_i)(x_j-m_i)^T$，且$(N_1m_1+N_2m_2)^T=Nm^T$，$m=\sum_{i=1}^{N}x_i$，上面的等式可以分解出两个等式，第一个得到$w_0=-m^T\mathbf{w}$，将$w_0$代入第二个等式可以得到

$$\left[-\frac{1}{N}(N_1m_1+N_2m_2)(N_1m_1+N_2m_2)^T+S_w+N_1m_1m_1^T+N_2m_2m_2^T\right]\mathbf{w}=N(m_1-m_2)$$

$$\left[\frac{1}{N}S_w+\frac{N_1N_2}{N^2}(m_1-m_2)(m_1-m_2)^T\right]\mathbf{w}=m_1-m_2$$

注意因为$\frac{N_1N_2}{N}(m_1-m_2)(m_1-m_2)^T\mathbf{w}$在$m_1-m_2$的方向上，所以上式可以化为：

$$S_w\mathbf{w}=\alpha(m_1-m_2)$$

与Fisher的解相同。

- 4.16 证明：
  - (1)式(4-113)表示的向量$y-\frac{a^Ty}{||w||^2}\begin{bmatrix}0\\\mathbf{w}\end{bmatrix}$表示$y$到X空间中超平面的投影。
  - (2)该投影正交于X空间的超平面。

证明： （1）先证明这个向量在X空间中的超平面上，再证明$y-\left(y-\frac{a^Ty}{||w||^2}\begin{bmatrix}0\\\mathbf{w}\end{bmatrix}\right)$的向量为X空间中超平面的法向量。 X空间中的超平面的方程为：$g(x)=\mathbf{w}^Tx+$

$x_0 = [1, \mathbf{w}^T]\begin{bmatrix} x_0 \\ x \end{bmatrix} = a^T y = 0$，将向量代入$g(x)$，得 $a^T y - \frac{a^T y}{||\mathbf{w}||^2} a^T \begin{bmatrix} 0 \\ \mathbf{w} \end{bmatrix} = a^T y - \frac{a^T y}{||\mathbf{w}||^2}||\mathbf{w}||^2 = 0$，又因为$y - \left(y - \frac{a^T y}{||w||^2}\begin{bmatrix} 0 \\ \mathbf{w} \end{bmatrix}\right) = \frac{a^T y}{||w||^2}\begin{bmatrix} 0 \\ \mathbf{w} \end{bmatrix}$

- 4.17 在多类问题中，如果一组样本可被一线性机全部正确分类，则称这组样本是线性可分的。对任意w_i类，如果能用一超平面把w_i类的样本同其他样本分开来，则称总体线性可分。举例说明，总体线性可分必定线性可分，但反之不然。

  解：

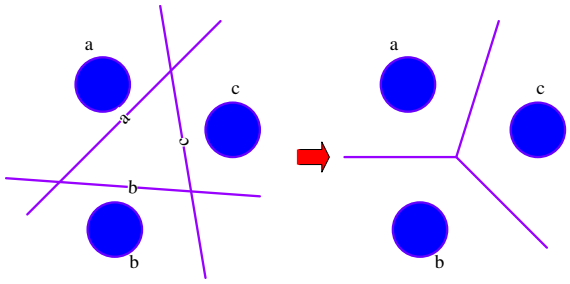

图 3: 总体线性可分必定线性可分

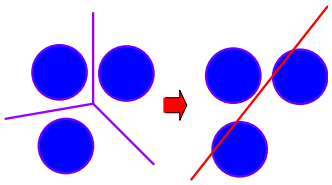

图 4: 线性可分未必总体线性可分

- 4.18 设有一组样本。若存在$c(c-1)/2$个超平面H_{ij}，使H_{ij}把属于w_i类的样本同属于w_j类的样本分开，则称这组样本是成对线性可分的。举列说明，成对线性可分的样本不一定线性可分。

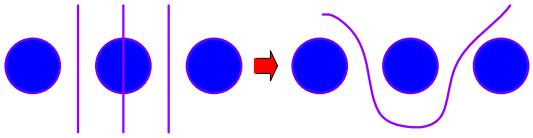

图 5: 成对线性可分不一定定线性可分

§5 非线性判别函数

- 5.1 举例说明分段线性分界面可以逼近贝叶斯判别函数确定的超曲面。

  解： 分段线性函数是一类特殊的非线性函数，它确定的决策面由若干个平面段组成，所以它可以逼近各种形状的超曲面。

- 5.2 已知两类问题如图6所示，其中``×"表示$w_1$类训练样本集合的原型，``◯ "表示w_2类训练样本集的原型。

  – (1)找出紧互对原型集合$\mathscr{P}$;

  – (2)找出与紧互对行集相联系的超平面集$\mathscr{H}$;

  – (3)假设训练集样本与原型完全相同，找出由超平面集$\mathscr{H}$产生的$z(x)$。

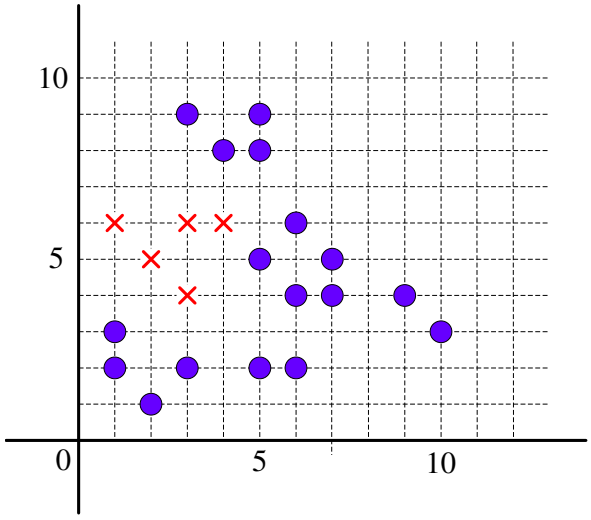

图 6: 一个两类问题的原型分布

解:（1）用坐标来表示样本w_1中的样本$(4,6)$与w_2中的样本$(5,5)$是紧互对原型，$(3,4)$与$(3,2)$是，$(2,5)$与$(1,3)$也是。如图 7所示 （2）如图8所示

§6 近邻法

- 6.1 举例说明最近邻决策面是分段线性的。

  解： 分段线性函数的决策面由若干个超平面组成。由于它的基本组成仍然是超平面，因此，与一般超平面

- 6.2 证明式$(6-14) \sim (6-18)$。

  证明： 记

$$\sum_{i=1}^{c} P^2(w_i|x) = P^2(w_m|x) + \sum_{i \neq m} P^2(w_i|x)$$

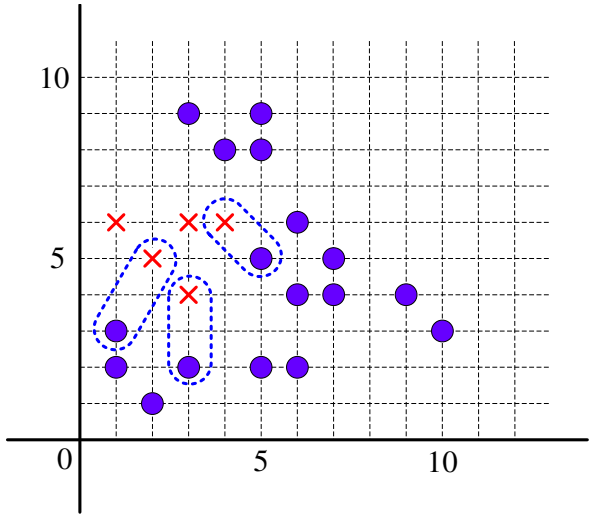

图 7: 紧互对原型

- 6.3 在什么情况下，最近邻平均误差P达到其上界
- 6.5 有7个二维向量：$x_1 = (1,0)^T, x_2 = (0,1)^T, x_3 = (0,-1)^T, x_4 = (0,0)^T, x_5 = (0,2)^T, x_6 = (0,-2)^T, x_7 = (-2,0)^T$，假定前三个为$w_1$类，后四个为$w_2$类。
  - (1) 画出最近邻法决策面；
  - (2) 求样本均值m_1, m_2，若按离样本均值距离的大小进行分类，试画出决策面。

  解： 第一首先要明确什么是“最近邻法”？它实际是一种分段的线性判别函数。第二根据离样本均值的距离来分类，首先求出两类的样本均值，分类决策面就是样本均值的垂直平分线。 (1)如图9所示。 (2) w_1类的均值为$m_1 = (\frac{1}{2}, 0)^T$，w_2类的均值为$m_2 = (-1,0)^T$，决策面如图10所示。
- 6.6 画出k-近邻法得程序框图。

  解：取未知样本x的k近邻，看这k近邻中多数属于哪一类，就把x归为那一类。
- 6.7 对于有限样本，重复剪辑是否比两分剪辑的特性要好。
- 6.8 证明如果$B + D(x_i, M_p) < D(x, M_p)$，其中$x_i \in \mathscr{X}_p$，则$x_i$不是$x$的近邻。

  证明：有三角不等式

$$D(x, x_i) + D(x_i, M_p) > D(x, M_p) \Rightarrow D(x, x_i) > D(x, M_p) - D(x_i, M_p)$$

  所以如果当前近邻距离x的距离为B，$D(x, x_i) > D(x, M_p) - D(x_i, M_p) > B$，即当

$$B + D(x_i, M_p) < D(x, M_p)$$

  时，x的近邻一定不在$\mathscr{X}_p$中。

  知识点：近邻法的快速算法。

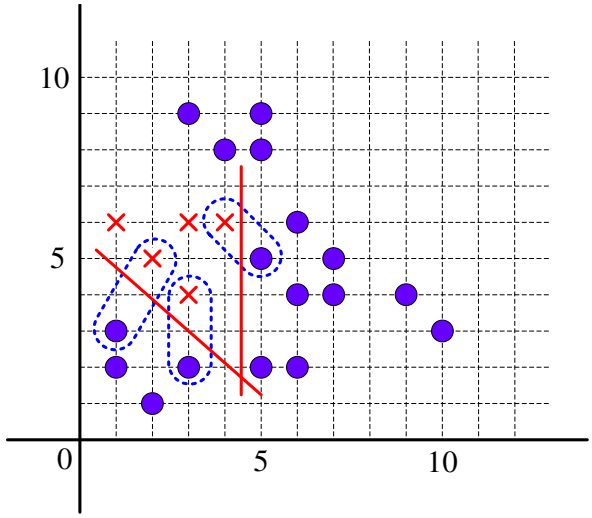

图 8: 利用紧互对原型设计的分段线性分类器

§7 经验风险最小化和有序风险最小化方法

§8 特征的选取和提取

- 8.1 三类w_1, w_2, w_3，求S_w, S_b。

  解：对应w_1类的样本点$\{(1,0)^T, (2,0)^T, (1,1)^T\}$，
  w_2类的样本点$\{(0,1)^T, (-1,0)^T, (-1,1)^T\}$，
  w_3类的样本点$\{(0,-1)^T, (-1,-1)^T, (0,-2)^T\}$。
  w_1类的均值$u_1 = (\frac{4}{3}, \frac{1}{3})^T$，协方差矩阵为：

$$\begin{aligned}\Sigma_1 &= \frac{1}{3}\sum_{x_i \in w_1}(x_i - u_1)(x_i - u_1)^T \\ &= \frac{1}{9}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\end{aligned}$$

  w_2类的均值$u_2 = (-\frac{2}{3}, \frac{2}{3})^T$，协方差矩阵为：

$$\begin{aligned}\Sigma_2 &= \frac{1}{3}\sum_{x_i \in w_2}(x_i - u_2)(x_i - u_2)^T \\ &= \frac{1}{9}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\end{aligned}$$

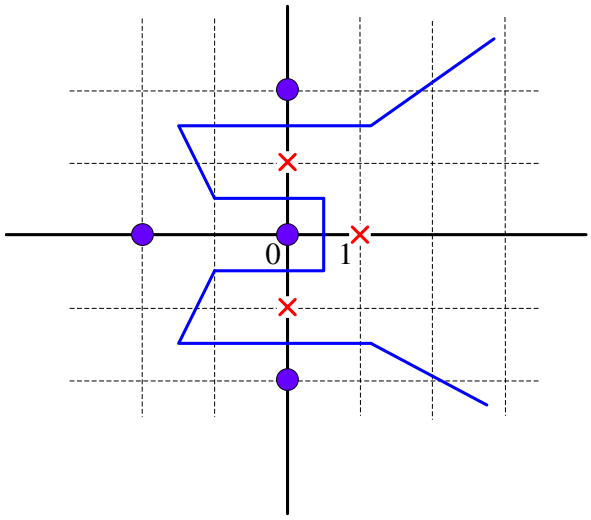

图 9: 最近邻法分类决策面

w_3类的均值$u_3 = (-\frac{1}{3}, -\frac{4}{3})^T$，协方差矩阵为：

$$\begin{aligned}\Sigma_2 &= \frac{1}{3}\sum_{x_i \in w_3}(x_i - u_3)(x_i - u_3)^T \\ &= \frac{1}{9}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\end{aligned}$$

所以类内散布度矩阵为：

$$\begin{aligned}S_w &= \frac{1}{3}(\Sigma_1 + \Sigma_2 + \Sigma_3) \\ &= \frac{1}{27}\begin{pmatrix} 6 & -1 \\ -1 & 6 \end{pmatrix}\end{aligned}$$

总体均值为$u = \frac{1}{3}\sum_{i=1}^{3} u_i = (\frac{1}{9}, -\frac{1}{9})^T$，所以类间散布度矩阵为：

$$\begin{aligned}S_b &= \frac{1}{3}\sum_{i=1}^{3}(u_i - u)(u_i - u)^T \\ &= \frac{1}{81}\begin{pmatrix} 62 & 13 \\ 13 & 62 \end{pmatrix}\end{aligned}$$

- 8.2 设有两个正态分布的样本集，它们的期望及方差矩阵分别等于上题中w_1及w_2的均值向量及协方差矩阵，计算w_1和w_2的散度及Bhattacharyya距离。

  解：

$$p(x|w_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}}\exp[-\frac{1}{2}(x - u_i)^T\Sigma_i^{-1}(x - u_i)]$$

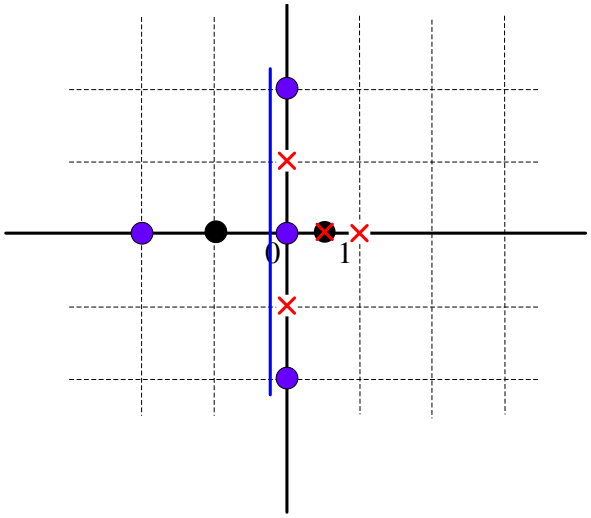

图 10: 按样本均值分类的决策面

$$
\begin{aligned}
I_{12} &= \int p(x|w_1)\ln\frac{p(x|w_1)}{p(x|w_2}\mathrm{d}x \\
&= \int\left\{\frac{1}{2}\ln\frac{|\Sigma_2|}{|\Sigma_1|}-\frac{1}{2}\mathrm{tr}[\Sigma_1^{-1}(x-u_1)(x-u_1)^T]+\frac{1}{2}\mathrm{tr}[\Sigma_2^{-1}(x-u_2)(x-u_2)^T]\right\}p(w|x_1)\mathrm{d}x
\end{aligned}
$$

w_1和w_2的散度为:

$$
\begin{aligned}
J_D =& I_{ij}+I_{ji} \\
=& \int[p(x|w_i)-p(x|w_j)]\ln\frac{p(x|w_i)}{p(x|w_j)} \\
=& \frac{1}{2}tr[\Sigma_1^{-1}\Sigma_2+\Sigma_2^{-1}\Sigma_1-2I]+\frac{1}{2}(u_1-u_2)^T(\Sigma_1^{-1}+\Sigma_2^{-1})(u_1-u_2)
\end{aligned}
$$

§9 基于K-L展开式的特征提取

- 9.1 若有下列两类样本集:

$$
\begin{array}{cc}
w_1 & w_2 \\
x_1=(0,0,0)^T & y_1=(0,0,1)^T \\
x_2=(1,0,0)^T & y_2=(0,1,0)^T \\
x_3=(1,0,1)^T & y_3=(0,1,1)^T \\
x_4=(1,1,0)^T & y_4=(1,1,1)^T
\end{array}
$$

用K-L变换，分别把特征空间维数降到$d=2$和$d=1$并用图画出样本在该特征空间中的位置。

解： w_1和w_2的协方差矩阵分别为:

$$
\Sigma_1=\frac{1}{4}\begin{bmatrix}0.75 & 0.25 & 0.25\\ 0.25 & 0.75 & -0.25\\ 0.25 & -0.25 & 0.75\end{bmatrix}
$$

$$\Sigma_2 = \frac{1}{4}\begin{bmatrix} 0.75 & 0.25 & 0.25 \\ 0.25 & 0.75 & -0.25 \\ 0.25 & -0.25 & 0.75 \end{bmatrix}$$

则总类内散布度矩阵S_w为:

$$S_w = \frac{1}{2}(\Sigma_1 + \Sigma_2) = \frac{1}{8}\begin{bmatrix} 1.5 & 0.5 & 0.5 \\ 0.5 & 1.5 & -0.5 \\ 0.5 & -0.5 & 1.5 \end{bmatrix}$$

它的特征值矩阵和特征向量分别为:

$$\Lambda = \frac{1}{4}\begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}, U = \begin{bmatrix} -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{42}} & \frac{3}{\sqrt{14}} \\ -\frac{1}{\sqrt{3}} & -\frac{5}{\sqrt{42}} & \frac{1}{\sqrt{14}} \\ \frac{1}{\sqrt{3}} & \frac{4}{\sqrt{42}} & \frac{2}{\sqrt{14}} \end{bmatrix}$$

所以，降到$d = 2$维的变换矩阵$U = \begin{bmatrix} -\frac{1}{\sqrt{42}} & \frac{3}{\sqrt{14}} \\ -\frac{5}{\sqrt{42}} & \frac{1}{\sqrt{14}} \\ \frac{4}{\sqrt{42}} & \frac{2}{\sqrt{14}} \end{bmatrix}$

又S_b为

$$S_b = \begin{bmatrix} \frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} \\ -\frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ -\frac{1}{8} & \frac{1}{8} & \frac{1}{8} \end{bmatrix}$$

所以

$$J(x_1) = \frac{u_1^T S_b u_1}{\lambda_1} = 0.375$$
$$J(x_2) = \frac{u_2^T S_b u_2}{\lambda_2} = 0$$
$$J(x_3) = \frac{u_3^T S_b u_3}{\lambda_3} = 0$$

- 9.3 令Σ_i和P_i分别为类$w_i (i = 1, 2)$的协方差矩阵和先验概率。假定对数据进行白化变换，即使得$B^T S_w B = I$，这里 $S_w = \sum_i P_i \Sigma_i$，$I$是单位矩阵。证明矩阵$P_1 B^T \Sigma_1 B$和矩阵$P_2 B^T \Sigma_2 B$所产生的K-L坐标轴是相同的，若用$\Lambda_i$表示矩阵$P_i B^T \Sigma_i B$的特征值矩阵，求证:

$$\Lambda_1 = I - \Lambda_2$$

证明： 因为:

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

设$P_1 B^T \Sigma_1 B u = \lambda u$，则

$$(I - P_2 B^T \Sigma_2 B)u = \lambda u$$
$$P_2 B^T \Sigma_2 B u = (1 - \lambda)u$$

可见矩阵$P_1 B^T \Sigma_1 B$和矩阵$P_2 B^T \Sigma_2 B$具有相同的特征向量。产生的K-L坐标轴相同。再由上面的推倒，设$P_1 B^T \Sigma_1 B$的特征值为λ_1则$P_2 B^T \Sigma_2 B$有一个特征值$\lambda_2 = 1 - \lambda_1$ 容易得到特征值矩阵满足$\Lambda_1 = I - \Lambda_2$

§10 非监督学习方法

- 10.1 令$\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_N$是$d$维样本，$\Sigma$是任一非奇异$d \times d$矩阵，证明使

$$\sum_{k=1}^{N}(\mathrm{x}_k - \mathrm{x})^T\Sigma^{-1}(\mathrm{x}_k - \mathrm{x})$$

最小的向量x是样本的均值$\mu = \dfrac{1}{N}\sum_{i=1}^{N}\mathrm{x}_i$

证明：设

$$g(\mathrm{x}) = \sum_{k=1}^{N}(\mathrm{x}_k - \mathrm{x})^T\Sigma^{-1}(\mathrm{x}_k - \mathrm{x})$$

上式对x求导得

$$\frac{\partial g(\mathrm{x})}{\partial \mathrm{x}} = 2\sum_{k=1}^{N}\Sigma^{-1}(\mathrm{x} - \mathrm{x}_k)$$

导数为0得到极值，易得

$$\mathrm{x} = \frac{1}{N}\sum_{k}^{N}\mathrm{x}_k$$

- 10.2 令$s(\mathrm{x}, \mathrm{x}') = \dfrac{\mathrm{x}^T\mathrm{x}'}{||\mathrm{x}|| \cdot ||\mathrm{x}'||}$。若x的$d$个特征只取$+1$和$-1$二值，即当x具有第$i$个特征时，$\mathrm{x}_i = 1$而当x没有这个特征时$\mathrm{x}_i = -1$，说明$s$是一个相似性度量。证明对于这种情况

$$||\mathrm{x} - \mathrm{x}'||^2 = 2d(1 - s(\mathrm{x}, \mathrm{x}'))$$

证明：

$$\begin{aligned}
||\mathrm{x} - \mathrm{x}'||^2 &= (\mathrm{x} - \mathrm{x}')^T(\mathrm{x} - \mathrm{x}') \\
&= \mathrm{x}^T\mathrm{x} - 2\mathrm{x}^T\mathrm{x}' + \mathrm{x}'^T\mathrm{x}' \\
&= 2d - 2\mathrm{x}^T\mathrm{x}' \\
&= 2d(1 - \frac{\mathrm{x}^T\mathrm{x}'}{\sqrt{d}\sqrt{d}}) \\
&= 2d(1 - \frac{\mathrm{x}^T\mathrm{x}'}{||\mathrm{x}|| \cdot ||\mathrm{x}'||}) \\
&= 2d(1 - s(\mathrm{x}, \mathrm{x}'))
\end{aligned}$$

- 10.3 假使一个有N个样本的集合$\mathscr{X}$划分为c个不相交的子集$\mathscr{X}_1, ..., \mathscr{X}_c$，假使$\mathscr{X}_i$是空集，则$\mathscr{X}_i$中样本的均值$m_i$不定义。在这种情况下，误差平方和只和非空子集有关：

$$J_e = \sum_{i}\sum_{x \in \mathscr{X}_i}||x - m_i||^2$$

这里i是不包含空子集的子集标号。假定$N \geqslant c$,证明使得J_e最小的划分中没有空子集。

证明：假设存在一个空子集$\mathscr{X}_k, 1 \leqslant k \leqslant c$使得$J_e$最小，容易证明可以找到所有子集都不为空的划分使得$J_e$更小。