模式识别上机实验 2：参数估计及两分类问题

给定 2 维样本 500 个，存放在文件“500 样本.txt”中，其中前 300 个是属于第一类的样本，接着 200 个是属于第二类的样本（第一列为样本的类别）。假设两类样本均来自正态总体，试分别估计其参数，求出决策函数和决策规则并对如下五个未知类别的样本进行分类。

类别	x_1	x_2
	-1.0221	3.2155
	5.0000	10.000
	2.4344	4.3210
	3.1932	8.7089
	-0.6212	1.8253

决策结果为：

用马氏距离得到的结果

x_1	x_2	到样本一的马氏距离	到样本二的马氏距离	马氏距离决策结果
-1.0221	3.2155	1.7808	1.9772	属于样本一
5.0000	10.000	1.9367	2.9772	属于样本一
2.4344	4.3210	0.446	2.9741	属于样本一
3.1932	8.7089	1.9068	1.731	属于样本二
-0.6212	1.8253	0.9918	2.9054	属于样本一

用最小贝叶斯决策结果

x_1	x_2	$g(x)$	决策结果
-1.0221	3.2155	0.1359	属于样本一
5.0000	10.000	2.3234	属于样本一
2.4344	4.3210	4.0899	属于样本一
3.1932	8.7089	-0.553	属于样本二
-0.6212	1.8253	3.4958	属于样本一

参数估计及两分类问题

姓名：寸正雄

学号：20081910073

1.问题分析

该实验目的要通过也知道的 300 个一类和 200 个二类样本，由参数估计得到两类的正态函数，通过正态分布统计决策设计出分类器将实验中的五个数据进行分类。

1.1． 多元正态分布参数估计

多元正态分布的概率密度定义如下：

$$f(x)=\frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma|}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{1.1}$$

其中，$x=[x_1,x_2,\cdots,x_d]^{\mathrm{T}}$ 是 d 维向量，$\mu=[\mu_1,\mu_2,\cdots,\mu_d]^{\mathrm{T}}$ 是 d 维均值向量，Σ 是 $d\times d$ 维的协方差矩阵，Σ^{-1} 是 Σ 的逆矩阵，$|\Sigma|$ 是 Σ 的行列式。在其密度函数中有 μ 和 Σ 两组参数。

而多元正态分布对于每一个 x_i 得边缘分布都是一个一元的正态分布 $\mathrm{N}(\mu_i,\sigma_i)$，其密度函数为

$$f_{x_i}(x_i)=(2\pi)^{-\frac{1}{2}}e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}},x_i\in(-\infty,+\infty) \tag{1.2}$$

由一元的正态分布参数估计可知

$$\mu_i=\mathrm{E}X_i=\overline{X_i},\sigma_i^2=DX_i=\frac{1}{n}\Sigma(X_i-\overline{X_i})^2=m_2' \tag{1.3}$$

这样可以得到哥分布函数，多元正态分布函数中参数 Σ 由 σ_i^2 和各分布函数所有任意两两变量的协方差组成

$$\Sigma_{ij}=\begin{cases}\sigma_i^2,i=j\\ \mathrm{cov}(x_i,x_j)=\mathrm{E}(x_i-\mathrm{EX}_i)(x_j-\mathrm{EX}_j),i\neq j\end{cases} \tag{1.4}$$

可表示成

$$\Sigma=E((x-\mu)(x-u)^T) \tag{1.5}$$

1.2． Mahalanobis 距离

马氏距离是由印度统计学家马哈拉诺比斯(P. C. Mahalanobis)提出的，表示数据的协方差距离。它是一种有效的计算两个未知样本集的相似度的方法。对给定的两个样本 $\mathrm{X}=[x_1,x_2,\cdots,x_d]^{\mathrm{T}}$ 和 $\mathrm{Y}=[y_1,y_2,\cdots,y_d]^{\mathrm{T}}$，$\Sigma$ 为 X 和 Y 的协方差矩阵，马氏距离定义如

下

$$R(\mathrm{X},\mathrm{Y})=\sqrt{(\mathrm{X}-\mathrm{Y})^{\mathrm{T}}\Sigma^{-1}(\mathrm{X}-\mathrm{Y})} \qquad (1.6)$$

表示是 X 到 Y 的马氏距离。如图 1.1 所示，在椭圆中椭圆边上任意一点到中心的马氏距离是相等的。

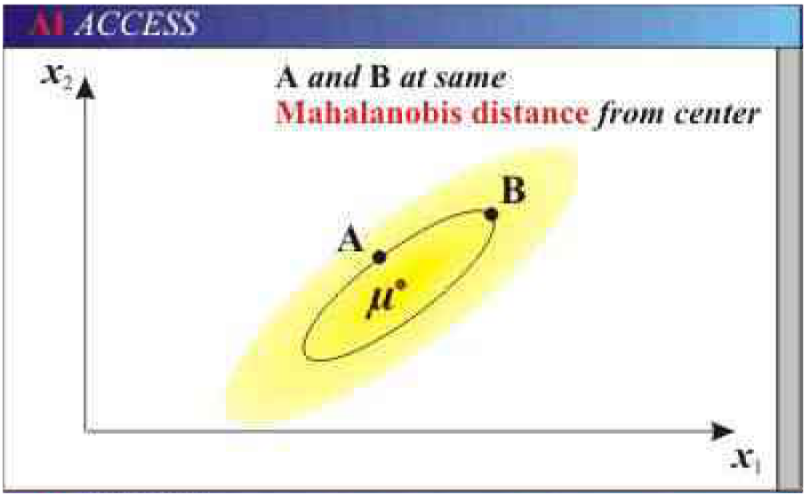

图 1.1

1.3. 多元正态概率型下的最小错误率贝叶斯判别

最小错误率贝叶斯决策规则常有四种方法：

（1） $P(w_1 \mid x)=\max\limits_{j=1,2} P(w_j \mid x)$,则$x\in w_i$

（2） $p(x \mid w_i)P(w_i)=\max\limits_{j=1,2} p(x \mid w_j)P(w_j)$,则$x\in w_i$

（3） $l(x)=\dfrac{p(x \mid w_1)}{p(x \mid w_2)} > or < \dfrac{P(w_2)}{P(w_1)}$,则$x\in \begin{cases} w_1 \\ w_2 \end{cases}$

（4） $\ln p(x \mid w_1)+\ln P(w_1) > or < \ln p(x \mid w_2)+\ln P(w_2)$,则$x\in \begin{cases} w_1 \\ w_2 \end{cases}$

在多元正态函数中采用上述中方法（4），$p(x \mid w_i)\sim \mathrm{N}(\mu_i,\sigma_i)$，可得到多元正态型下的最小错误率贝叶斯判别函数为

$$g_i(x)=-\frac{1}{2}R^2(x,\mu_i)-\frac{d}{2}\ln(2\pi)-\frac{1}{2}\ln|\Sigma_i|+\ln P(w_i) \qquad (1.7)$$

其中$R(x,\mu_i)$表示x到μ_i的马氏距离，Σ_i是协方差矩阵，$P(w_i)$为先验概率。

决策面方程为 $g_i(x)=g_j(x)$

决策函数为 $g(x)=g_i(x)-g_j(x)$

$$= -\frac{1}{2}[R^2(x,\mu_i) - R^2(x,\mu_j)] - \frac{1}{2}\ln\frac{|\Sigma_i|}{|\Sigma_j|} + \ln\frac{P(w_i)}{P(w_j)} \quad （1.8）$$

决策结果为 $$x \in \begin{cases} w_i, g(x) > 0 \\ w_j, g(x) \le 0 \end{cases} \quad (1.9)$$

2. 问题求解

2.1. 参数估计

该题为二元正态分布， $x = [x, y]$，其密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \quad （2.1）$$

即 $$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]} \quad （2.2）$$

由二元正态参数估计可知

$$\mu_1 = \overline{X}, \mu_2 = \overline{Y}, \sigma_1 = DX, \sigma_2 = DY \quad （2.3）$$

$$\rho = \mathrm{cov}(X,Y) = E(X - EX)(Y - EY) \quad (2.4)$$

$$\Sigma = \begin{bmatrix} \sigma_1 & \rho \\ \rho & \sigma_2 \end{bmatrix} \quad （2.5）$$

题所给的两个样本可得到各样本的密度函数参数见表 2.1

表 2.1

	μ_1	μ_1	Σ
样本一	1.6473	3.9287	5.7140 6.5198 6.5198 10.2668
样本二	0.6823	6.8952	2.1154 1.6827 1.6827 3.4679

求出各参数就可以得到两个样本的分布密度函数，把其绘制成二维图像如下，其中*号代表的是第二类样本点，o 号代表第一类样本点。

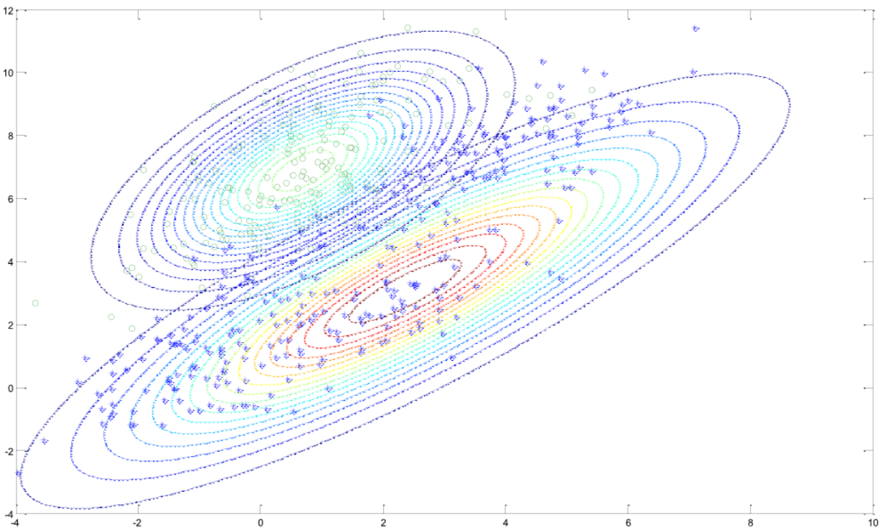

图 2.1 两个样本的二维图像

2.2．求未知类别的样本到各样本的$(\mu_1,\mu_2)^T$的马氏距离

由上面 1.2 中公式(1.6)可以得到未知类别的样本到 2.1 中所求的两个样本的(μ_1,μ_2)的马氏距离如下表，并用马氏距离的比较，到哪个样本的马氏距离小将其分为该类。

表 **2.2**

x_1	x_2	到样本一的马氏距离	到样本二的马氏距离	马氏距离决策结果
-1.0221	3.2155	1.7808	1.9772	属于样本一
5.0000	10.000	1.9367	2.9772	属于样本一
2.4344	4.3210	0.446	2.9741	属于样本一
3.1932	8.7089	1.9068	1.731	属于样本二
-0.6212	1.8253	0.9918	2.9054	属于样本一

2.3．最小错误率贝叶斯判别

设w_1,w_2分别表示两个类别，$P(w_1)$，$P(w_2)$分别表示两类的先验概率，由上面 1.3 中决策函数（1.8）式，和决策结果（1.9）氏，就可以得到决策结果，但在（1.8）式中还有$P(w_1)$，$P(w_2)$是未知的，这里就假设他所选的两个样本与实际相符，两类的先验概率就假设为$P(w_1)=\frac{3}{5}=0.6$，$P(w_2)=\frac{2}{5}=0.4$，做此假设后并可求出$g(x)$值，再更具决策结果（1.9）式得到决策表如下

表 2.3

x_1	x_2	$g(x)$	决策结果
-1.0221	3.2155	0.1359	属于样本一
5.0000	10.000	2.3234	属于样本一
2.4344	4.3210	4.0899	属于样本一
3.1932	8.7089	-0.553	属于样本二
-0.6212	1.8253	3.4958	属于样本一

3. 程序代码

实验结果在上面个表中，实验中所需的 MATLAB 代码如下：

3.1.样本 yangbenzhi.m

```
yangben=[1.0000    2.1839    3.0859
          ……
          2.0000   -0.0694    8.1524];
yangben1=yangben(1:300,:,:);
yangben2=yangben(301:500,:,:);
```

3.2.参数求解 canshu.m

```
function [U S]=canshu(YB)
X=YB(:,2);Y=YB(:,3);
U=[mean(X);mean(Y)];
S=cov(X,Y);
```

3.3.正态密度函数 f.m

```
function z=f(X,S,U)
d=size(S);d=d(1);
z=(2*pi)^(d/2)*sqrt(det(S))*exp(-1/2*(X-U)'*S^-1*(X-U));
```

3.4.马氏距离 mashijuli.m

```
function R=mashijuli(X,U,S)
R=sqrt((X-U)'*S^-1*(X-U));
```

3.5.用马氏距离判别 MSJL.m

```
YB=[-1.0221 3.2155
5.0000  10.000
2.4344  4.3210
3.1932  8.7089
```

```
-0.6212 1.8253];
yangbenzhi;
[U1 S1]=canshu(yangben1);
[U2 S2]=canshu(yangben2);
R=[];JG=[];
for i=1:5
    R1=mashijuli(YB(i,:)',U1,S1);
    R2=mashijuli(YB(i,:)',U2,S2);
    R=[R;R1 R2];
    if R1>R2
        JG=[JG;'属于样本二 '];
    else
        JG=[JG;'属于样本一'];
    end
end
disp('马氏距离为：');disp(R);
disp('马氏距离决策结果为：');disp(JG);
```

3.6.绘制样本二维图像 **HZTX.m**

```
yangbenzhi;
X1=yangben1(:,2);Y1=yangben1(:,3);
X2=yangben2(:,2);Y2=yangben2(:,3);
plot(X1,Y1,'*',X2,Y2,'o');
[U1 S1]=canshu(yangben1);[U2 S2]=canshu(yangben2);
x=linspace(-5,9,200);y=linspace(-5,12,200);
for i=1:200
    for j=1:200
        zf1(i,j)=f([x(i) y(j)]',S1,U1);
        zf2(i,j)=f([x(j) y(i)]',S2,U2);
    end
end
hold on;
contour(x,y,zf1,16);contour(x,y,zf2,16); hold off;
```

3.7.最小错误率贝叶斯决策 **ZXBYS.m**

```
YB=[-1.0221 3.2155
5.0000  10.000
2.4344  4.3210
3.1932  8.7089
-0.6212 1.8253];
PW1=0.6;PW2=0.4;
yangbenzhi;
[U1 S1]=canshu(yangben1);
[U2 S2]=canshu(yangben2);
```

```
JG=[];
for i=1:5
    R1=mashijuli(YB(i,:)',U1,S1);
    R2=mashijuli(YB(i,:)',U2,S2);
    G(i)=-1/2*(R1^2-R2^2)-1/2*log(det(S1)/det(S2))+log(PW1/PW2);
    if G(i)>0
        JG=[JG;'属于样本一'];
    else
        JG=[JG;'属于样本二'];
    end
end
disp('G(X)为：');disp(G');
disp('最小贝叶斯错误率决策结果为：');disp(JG);
```