深圳大学研究生课程：模式识别理论与方法

课程作业实验报告

实验名称：最小错误率贝叶斯分类器

实验编号：Proj02-01

签　　名：

姓　　名：夏荣杰

学　　号：2170269107

截止提交日期：2018 年 3 月 30 日

摘要：贝叶斯决策论是解决模式分类问题的一种基本统计途径。其出发点是利用概率的不同分类决策与相应的决策代价之间的定量折中。贝叶斯分类器是各种分类器中分类错误概率最小或者在预先给定代价的情况下平均风险最小的分类器。它的设计方法是一种最基本的统计分类方法。其分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。本实验旨在编程实现一个可以对两类模式样本进行分类的贝叶斯分类器，分析影响贝叶斯分类器分类效果的因素，以加深对贝叶斯分类器的理解。

一、 背景技术 或 基本原理

1. 贝叶斯公式：$P(\omega_j \mid x)=\dfrac{P(x \mid \omega_j)P(\omega_j)}{P(x)}$。

贝叶斯公式表明，通过观察 x 的值我们可将先验概率 $P(\omega_j)$ 转换为后验概率 $P(\omega_j \mid x)$，即假设特征值 x 已知的条件下类别属于 ω_j 的概率。由于证据(evidence)因子 $P(x)$ 对于所有类别的样本都一样，由贝叶斯公式可知，后验概率主要取决于先验概率和似然函数的乘机所决定的。

2. 最小错误率贝叶斯分类器：

考虑 c 类样本，令分类器函数为：

$$g_j(x)=P(x \mid \omega_j)P(\omega_j)\ j=1,2,\cdots,c \qquad (1)$$

则

$$x \in \omega_j, if \quad i=\arg\max_j \{g_j(x) \mid j=1,2,\cdots,c\} \qquad (2)$$

称上述为最小错误率贝叶斯分类器。

3. 一个贝叶斯分类器的结构可由类条件概率密度 $P(x \mid \omega_j)$ 和先验概率 $P(\omega_j)$ 来决定。如果样本类的条件概率密度未知，可以从可用的训练样本中估计出来。在各种概率密度函数中，常用多元高斯函数，很大程度上是源于它分析的简易性。正因如此，本实验中用到的样本类的条件概率密度服从高斯分布，具体计算公式如下：

若样本 x 为 d 维向量，第 j 类 ω_j 样本的条件概率密度服从均值为 m_j，协方差为 S_j 的多元高斯分布：

$$P(x \mid \omega_j)=\frac{1}{(2\pi)^{d/2}\sqrt{|S_j|}}e^{-\frac{1}{2}(x-m_j)^T S_j^{-1}(x-m_j)} \qquad (3)$$

上式中，$|S_j|$ 是 S_j 的行列式函数，T 是矩阵转置符号，S_j^{-1} 是 S_j 的逆矩阵。上面的高斯概率密度函数值的计算可以使用 matlab 中的函数 mvnpdf()实现。

二、 实验方法 或 算法流程步骤

实验 1 用到的方法和步骤如下：

1. 利用均值和协方差矩阵不同的两个模式各生成为 100 个服从二维高斯分布的随机样本 P1 和 P2，并将它们组合为一组长度为 200（sample_num）的总样本 P；
2. 根据公式（3）分别计算 P1 和 P2 的条件概率密度 p1 和 p2；
3. 根据公式（1）分别计算 P1 和 P2 的分类器函数值 g1(x)和 g2(x)；
4. 根据最小错误率贝叶斯分类器，即公式（2），比较 g1(x)和 g2(x)的大小，对 P1 和 P2 进行分类得到分类结果；

最小错误率贝叶斯分类器的分类规则简化为：

当 $g_1(x) > g_2(x)$ 时，则 $x \in \omega_1$

当 $g_1(x) < g_2(x)$ 时，则 $x \in \omega_2$

当 $g_1(x) = g_2(x)$ 时，则 $x \in unsure$，即该数据样本属于非确定类

贝叶斯分类规则就是当样本出现时，根据比较分类器函数值的大小判断样本最有可能属于的类别。

5. 将最小错误率贝叶斯分类器的分类结果与 P1 和 P2 的原本的标签进行对比，验证分类的正确与否，并统计正确分类的样本数量 correct_num 与错误分类的样本数量 wrong_num；

6. 计算正确分类的百分比：$accuracy = \frac{\texttt{correct\_num}}{\texttt{sample\_num}} = \frac{\texttt{correct\_num}}{200}$；

7. 最后改变两类样本的先验概率 $P(\omega_j)$，对相同的 200 个样本重新进行上述的分类。

实验 2、实验 3 和实验 4：通过改变实验 1 的均值向量或协方差矩阵，进行重复实验。

三、 实验结果

1. 实验 1 结果如下：

(1) 模式 1 的均值矢量 $m_1 = (1,3)^T$，协方差矩阵 $S_1 = (1.5, 0; 0, 1)^T$；模式 2 的均值矢量 $m_1 = (3,1)^T$，协方差矩阵 $S_2 = (1, 0.5; 0.5, 2)^T$，利用两个模式各生成 100 个随机样本（一共 200 个，·表示第一类样本点，▽表示第二类样本点），二维散点分布图如图 1-1 左图。利用贝叶斯分类器（先验概率 $P(\omega_1) = P(\omega_2) = 0.5$），对生成的 200 个随机样本进行分类，我们用不同的符号来区分正确和错误分类（○表示正确分类，×表示错误分类），以下实验都是采用上述记号表示，如图 1-1 右图。求得正确分类的百分比为 92.5%。

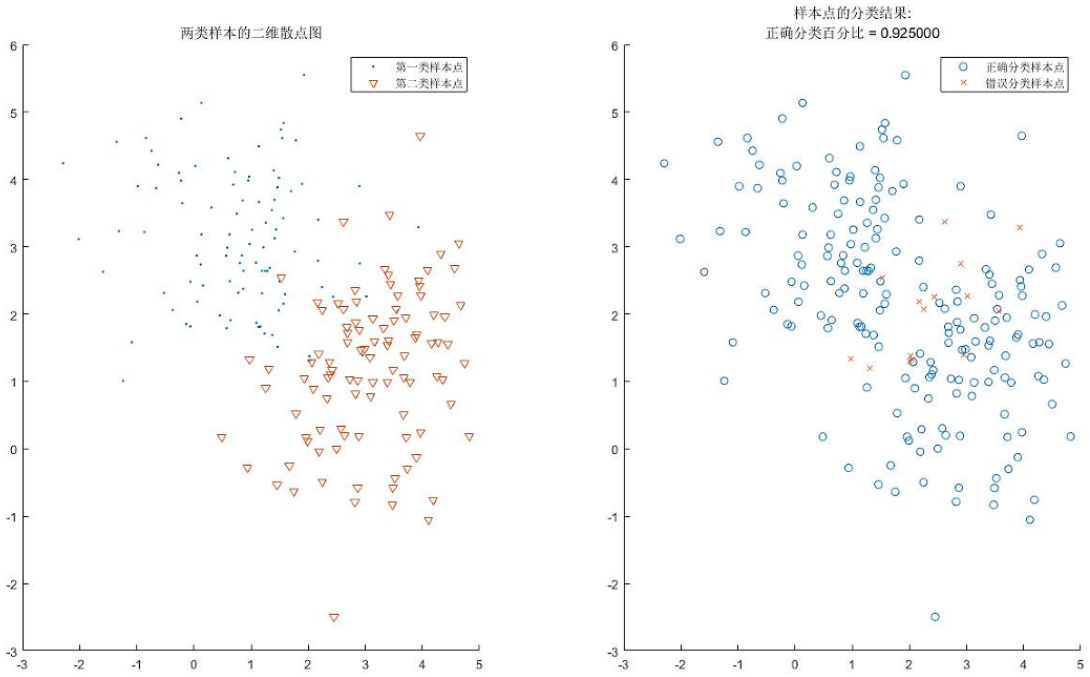

图 1-1

(2) 对于相同的 200 个随机样本，改变先验概率为 $P(\omega_1)=0.4, P(\omega_2)=0.6$，其他保持不变，重新进行实验，得到的正确分类和错误分类的样本如图 1-2 所示。求得正确分类的百分比为 94.0%。

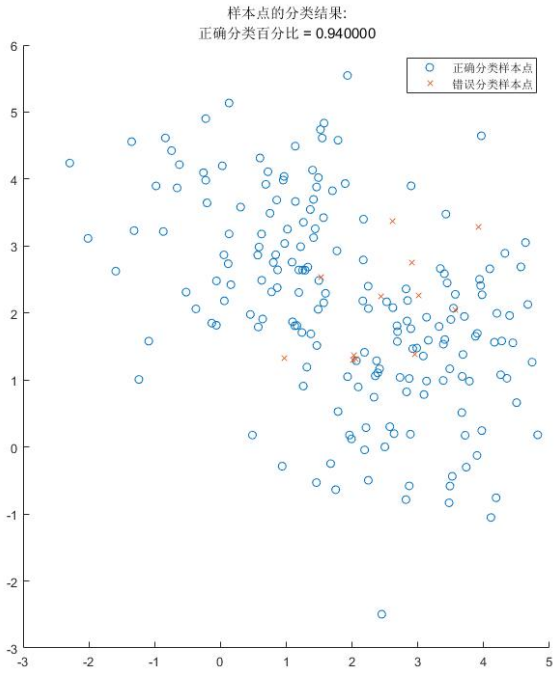

图 1-2

2. 实验 2 结果如下：

(1) 在基本实验 1 中，协方差矩阵不变，但类均值向量分别变为 $m_1=(1,3)^T$，$m_2=(2,2)^T$，重新生成 200 个随机样本，二维散点分布图如图 2-1 左图。利用贝叶斯

分类器（先验概率 $P(\omega_1)=P(\omega_2)=0.5$）进行分类，在二维图中画出正确分类和错误分类的样本如图 2-1 右图。求得正确分类的百分比为 78.5%。

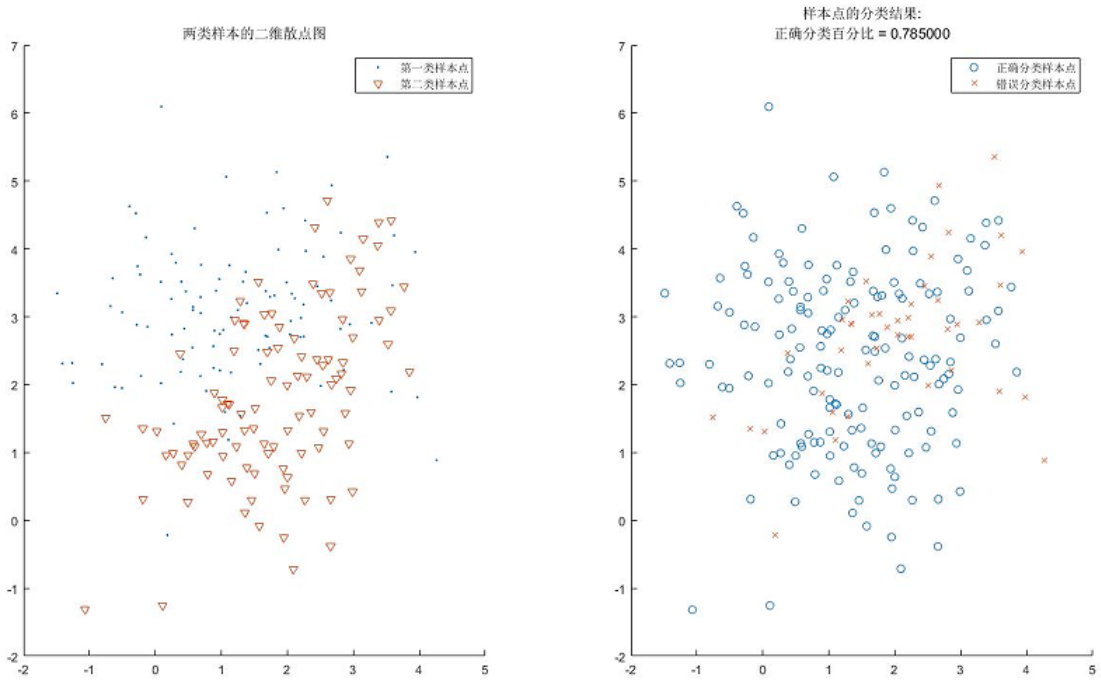

图 2-1

(2) 对于相同的 200 个随机样本，改变先验概率为 $P(\omega_1)=0.4, P(\omega_2)=0.6$，其他保持不变，重新进行实验，得到的正确分类和错误分类的样本如图 2-2 所示。求得正确分类的百分比为 74.5%。

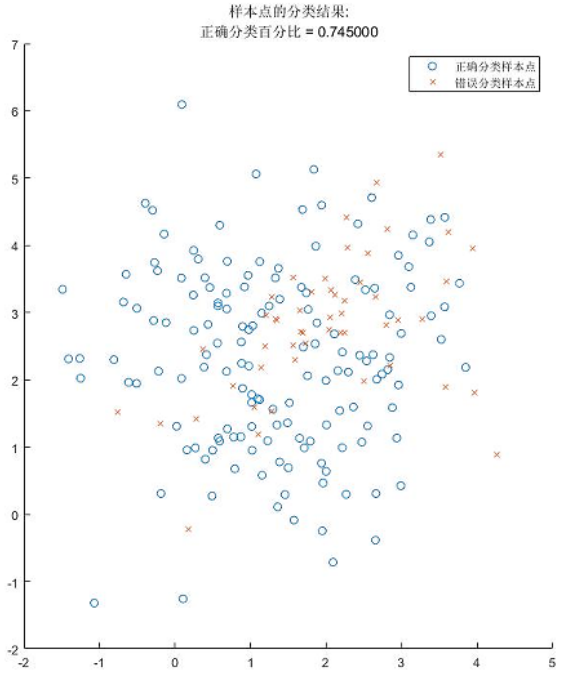

图 2-2

3. 实验 3 结果如下：

(1) 在基本实验 1 中，协方差矩阵不变，但类均值向量分别变为 $m_1=(1,3)^T$，$m_2=(4,0)^T$，重新生成 200 个随机样本，二维散点分布图如图 3-1 左图。利用贝叶斯分类器（先验概率 $P(\omega_1)=P(\omega_2)=0.5$）进行分类，在二维图中画出正确分类和错误分类的样本如图 3-1 右图。求得正确分类的百分比为 99.0%。

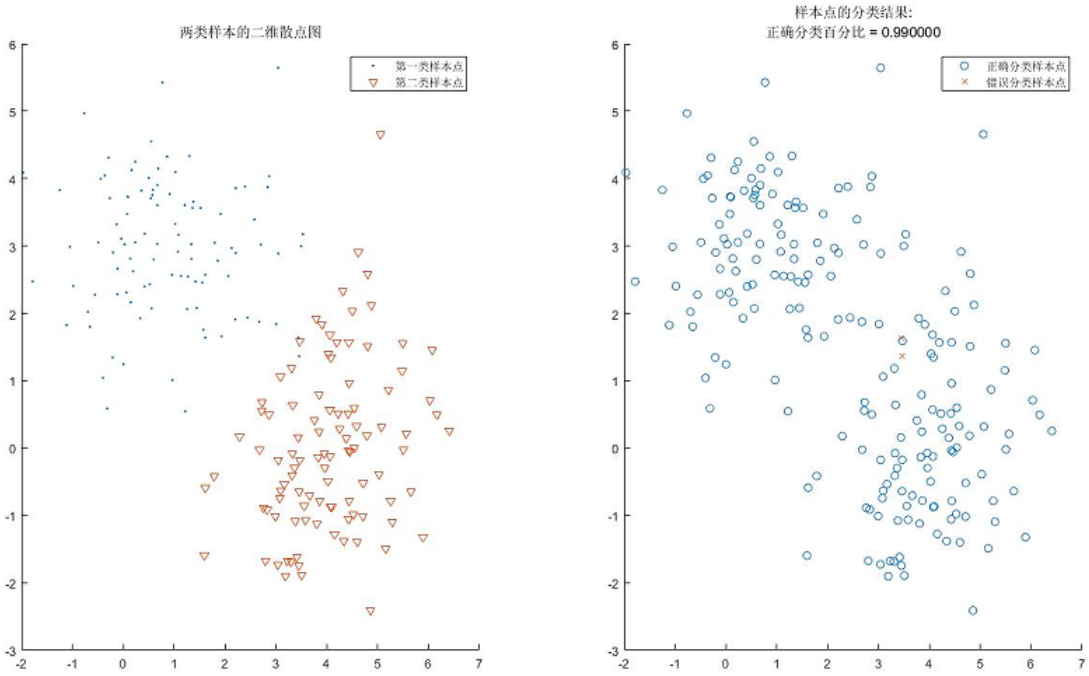

图 3-1

(2) 对于相同的 200 个随机样本，改变先验概率为 $P(\omega_1)=0.4, P(\omega_2)=0.6$，其他保持不变，重新进行实验，得到的正确分类和错误分类的样本如图 3-2 所示。求得正确分类的百分比为 98.5%。

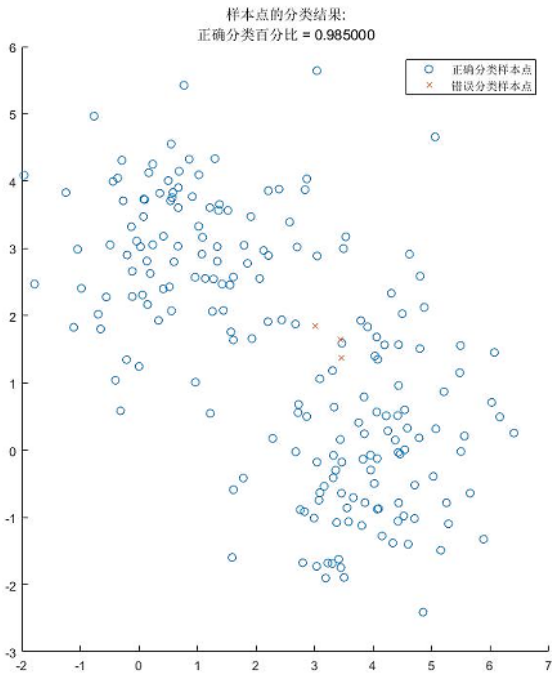

图 3-2

4. 实验 4 结果如下：

(1) 在基本实验 1 中，均值向量不变，但协方差矩阵分别变为 $S_1=(1.5,1;1,1)^T$，$S_2=(1,0.5;0.5,2)^T$ 重新生成 200 个随机样本，二维散点分布图如图 4-1 左图。利用贝叶斯分类器（先验概率 $P(\omega_1)=P(\omega_2)=0.5$）进行分类，在二维图中画出正确分类和错误分类的样本如图 4-1 右图。求得正确分类的百分比为 96.5%。

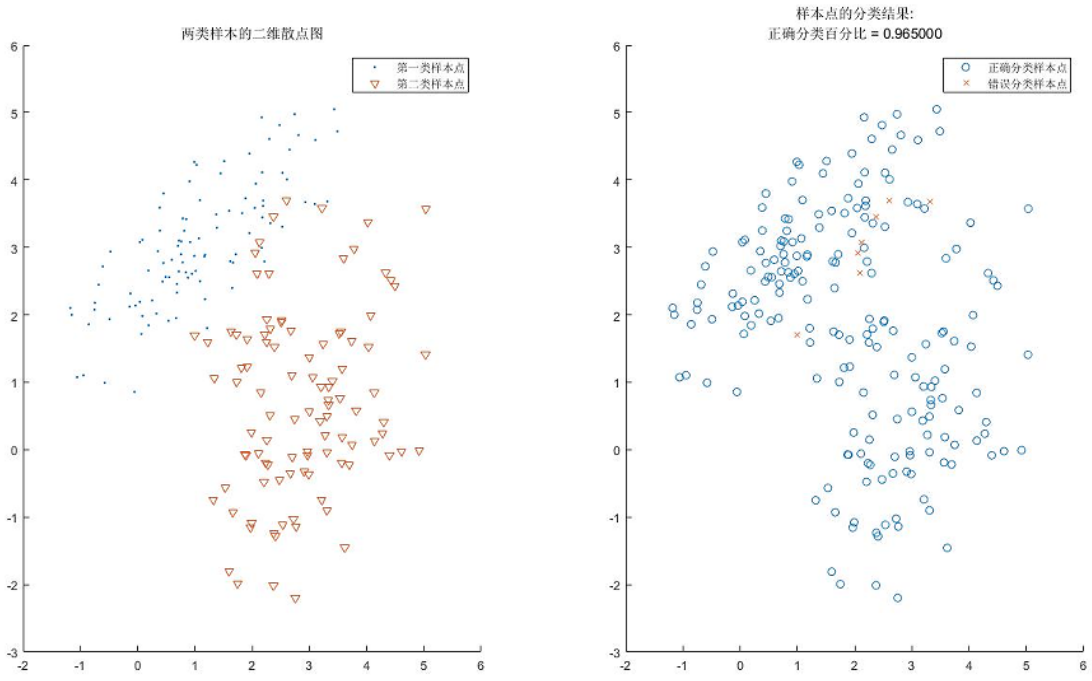

图 4-1

(2) 对于相同的 200 个随机样本，改变先验概率为 $P(\omega_1)=0.4, P(\omega_2)=0.6$，其他保持不变，重新进行实验，得到的正确分类和错误分类的样本如图 4-2 所示。求得正确分类的百分比为 95.5%。

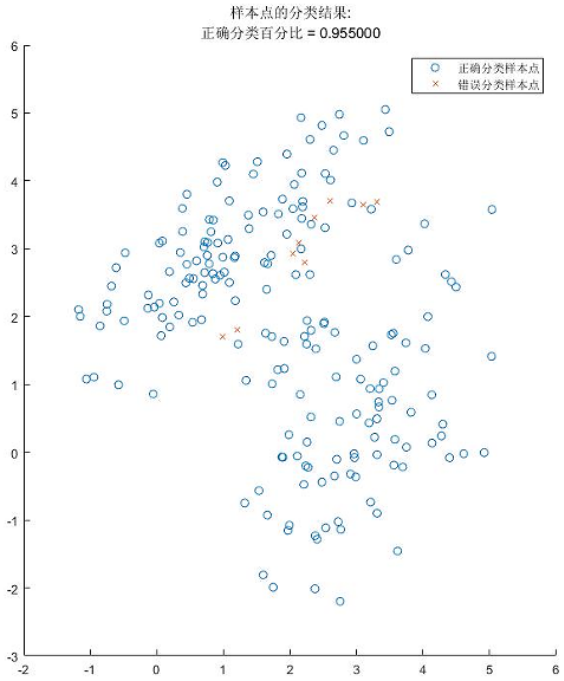

图 4-2

四、 讨论与分析

1. 两个模式类先验概率对分类结果的影响

	$P(\omega_1)=P(\omega_2)=0.5$	$P(\omega_1)=0.4, P(\omega_2)=0.6$
实验 1	92.5%	94.0%
实验 2	78.5%	74.5%
实验 3	99.0%	98.5%
实验 4	96.5%	95.5%

由上述的实验结果可以发现，在两个模式类的均值向量和协方差矩阵不变的情况下，改变两个模式类的先验概率，分类准确率大致相等，即模式类的先验概率对分类结果影响不大。

2. 两个模式类的距离对分类结果的影响

	$m_1=(1,3)^T\ m_2=(2,2)^T$ 欧式距离 $D^2=2$	$m_1=(1,3)^T\ m_2=(3,1)^T$ 欧式距离 $D^2=8$		$m_1=(1,3)^T\ m_2=(4,0)^T$ 欧式距离 $D^2=18$
	实验 2	实验 1	实验 4	实验 3
$P(\omega_1)=P(\omega_2)=0.5$	78.5%	92.5%	96.5%	99.0%
$P(\omega_1)=0.4,$ $P(\omega_2)=0.6$	74.5%	94.0%	95.5%	98.5%

由上述的实验结果可以发现，在两个模式类的先验概率不变的情况下，随着两个模式类的距离的增大，分类准确率增大，即模式类之间的距离越大，越容易进行分类，分类准确率越高。

3. 两个模式类的协方差对分类结果的影响

	$S_1=(1.5,0;0,1)^T\ S_2=(1,0.5;0.5,2)^T$ |S1|=1.5, |S2|=1.75			$S_1=(1.5,1;1,1)^T\ S_2=(1,0.5;0.5,2)^T$ |S1|=0.5, |S2|=1.75
	实验 1	实验 2	实验 3	实验 4
$P(\omega_1)=P(\omega_2)=0.5$	92.5%	78.5%	99.0%	96.5%
$P(\omega_1)=0.4,$ $P(\omega_2)=0.6$	94.0%	74.5%	98.5%	95.5%

由上述的实验结果可以发现，在两个模式类的先验概率不变的情况下，两个模式类的分类准确率不单单受模式类的协方差影响，还受到模式类之间的距离影响。实验 1 和实验 4 的模式类的协方差不同（实验 4 协方差矩阵的模值大于实验 1），由上述的实验结果说明，在两个模式类的先验概率和距离不变的情况下，模式类的协方差矩阵的模值越小，数据聚合越好，越容易进行分类，分类准确率越高。

综上所述，模式类的分类结果受先验概率的影响不大，主要受模式类之间的距离和协方差的影响：模式类之间的距离越大，协方差矩阵的模值越小，越容易进行分类，分类准确率越高。

附录.

实验 1：

```
%%----------------Proj02-01：最小错误率贝叶斯分类器--------------%%
%%----------------Proj02-01-exp1----------------%%
%%为两个模式类各生成 100 个随机样本，并在一幅图中用不同的符号画出这两类样本的二维散点图；
%%利用贝叶斯分类器，对生成的这 200 个样本进行分类，统计正确分类的百分比，并在 2 维图上用不同的颜色画出正确分类和错分的样本；
clear; clc;
m1 = [1; 3]; m2 = [3; 1]; %均值
S1 = [1.5 0; 0 1]; S2 = [1 0.5; 0.5 2]; %协方差矩阵
n = 100; %随机样本数量为 100
P1 = mvnrnd(m1, S1, n); %第一类样本
P2 = mvnrnd(m2, S2, n); %第二类样本
subplot(1, 2, 1);
s1 = scatter(P1(:, 1), P1(:, 2), '.');
hold on; s2 = scatter(P2(:, 1), P2(:, 2), 'v');
title('两类样本的二维散点图');
legend([s1 s2], '第一类样本点', '第二类样本点');

P = [P1; P2]; %两类样本
p1 = mvnpdf(P, m1', S1); %条件概率
p2 = mvnpdf(P, m2', S2);
p_w1 = 0.5; p_w2 = 0.5; %先验概率
g1 = p1 .* p_w1; %分类器函数
g2 = p2 .* p_w2;

%%————————最小错误率贝叶斯分类器————————%%
g = g1 - g2;
D = zeros(size(P, 1), 1);
D(find(g > 0)) = 1;
D(find(g < 0)) = 2;
D(find(g == 0)) = inf;
%%————————————————————————————————%%

label = [ones(size(P1, 1), 1); 2 * ones(size(P2, 1), 1)]; %200 个样本的标签
correct_id = find(D - label == 0);
wrong_id = find(abs(D - label) == 1);
unsure_id = find(D - label == inf);

correct = P(correct_id, :);
wrong = P(wrong_id, :);
```

```
unsure = P(unsure_id, :);

accuracy = size(correct, 1) / size(P, 1);

subplot(1, 2, 2);
c = scatter(correct(:, 1), correct(:, 2), 'o');
hold on; w = scatter(wrong(:, 1), wrong(:, 2), 'x');
acc = sprintf('正确分类百分比 = %f', accuracy);
st = ['样本点的分类结果:', string(acc)];
title(st);
legend([c w], '正确分类样本点', '错误分类样本点');
if ~isempty(unsure)
    u = scatter(unsure(:, 1), unsure(:, 2), '*');
    legend([c w u], '正确分类样本点', '错误分类样本点', '随机分类样本点');
end

%%
%%改变先验概率，利用贝叶斯分类器对相同样本进行重新分类
p_w11 = 0.4; p_w22 = 0.6; %先验概率
g11 = p1 .* p_w11; %分类器函数
g22 = p2 .* p_w22;

%%——————————最小错误率贝叶斯分类器——————————%%
gg = g11 - g22;
DD = zeros(size(P, 1), 1);
DD(find(gg > 0)) = 1;
DD(find(gg < 0)) = 2;
DD(find(gg == 0)) = inf;
%%————————————————————————————————————%%

label = [ones(size(P1, 1), 1); 2 * ones(size(P2, 1), 1)]; %200 个样本的
标签
correct_id1 = find(DD - label == 0);
wrong_id1 = find(abs(DD - label) == 1);
unsure_id1 = find(DD - label == inf);

correct1 = P(correct_id1, :);
wrong1 = P(wrong_id1, :);
unsure1 = P(unsure_id1, :);

accuracy1 = size(correct1, 1) / size(P, 1);

figure;
cc = scatter(correct1(:, 1), correct1(:, 2), 'o');
```

```
hold on; ww = scatter(wrong1(:, 1), wrong1(:, 2), 'x');
acc1 = sprintf('正确分类百分比 = %f', accuracy1);
st1 = ['样本点的分类结果:', string(acc1)];
title(st1);
legend([cc ww], '正确分类样本点', '错误分类样本点');
if ~isempty(unsure1)
    uu = scatter(unsure1(:, 1), unsure1(:, 2), '*');
    legend([cc ww uu], '正确分类样本点', '错误分类样本点', '随机分类样本点');
end
```