Experiment 2 Synthetical Design of Bayesian Classifier

一、 Principle and Theory

The essence of the Bayesian approach is to provide a mathematical rule explaining how you should change your existing beliefs in the light of new evidence. In other words, it allows us to combine new data with their existing knowledge or expertise. In terms of classification, the Bayesian theorem allows us to combine prior probabilities, along with observed evidence to arrive at the posterior probability. More or less, conditional probabilities represent the probability of an event occurring given evidence. According to the Bayesian Theorem, if $P(\omega i)$ ， $P(X|\omega i)$ ， i= 1,2,3,..., c,and X are known or given, the posterior probability can be derived as follows

$$P(\omega_i|X)=\frac{P(X|\omega_i)P(\omega_i)}{\sum_{j=1}^{c}P(X|\omega_j)P(\omega_j)} \qquad \text{i=1,…， c} \qquad (1)$$

Let the series of decision actions as {a1,a2,a3,...,ac}, the conditional risk of decision action ai can be computed by

$$R(a_i|X)=\sum_{j=1,j\neq i}^{c}\lambda(a_i,\omega_j)P(\omega_j|X), \qquad \text{i=1,…， c} \qquad (2)$$

Thus the minimum risk Bayesian decision can be found as

$$a_k^* = \text{Arg min}_i\ R(a_i|X), \qquad \text{i=1,…， c} \qquad (3)$$

二，**Objective**

The goals of the experiment are as follows:

(1) To understand the computation of likelihood of a class, given a sample.

(2) To understand the use of density/distribution functions to model a class.

(3) To understand the effect of prior probabilities in Bayesian classification.

(4) To understand how two (or more) density functions interact in the feature space to decide a decision boundary between classes.

(5) To understand how the decision boundary varies based on the nature of density functions.

三，**Contents and Procedure**

（1）Design a Bayesian classifier for the classification of two classes of patterns which are subjected to Gaussian normal distribution and compile the corresponding programme codes.

最小错误贝叶斯决策可按下列步骤进行：

1)在已知 $P(\omega_i)$， $P(X|\omega_i)$，i=1,…，c 及给出待识别的 X 的情况下，根据贝叶斯公式计算出后验概率：

$$P(\omega_i|X)=\frac{P(X|\omega_i)P(\omega_i)}{\sum_{j=1}^{c}P(X|\omega_i)P(\omega_i)} \qquad \text{j=1,…，x}$$

2)将不同的后验概率进行比较，找出使其错误率最小，即后验概率最大的决策。

%计算后验概率，其中 pw1 是先验概率，normpdf(x(i),e1,a1))是满足均值为 e1，标准差为 a1 的高斯分布的类条件概率，x 是样本矩阵，i 是第 i 个样本。

```
pw1_x(i)=(pw1*normpdf(x(i),e1,a1))/(pw1*normpdf(x(i),e1,a1)+pw2*normpdf(x(i),e2,a2)) ;
%比较两类后验概率
if pw1_x(i)>pw2_x(i)
        result(i)=0;            %第一类
    else
        result(i)=1;            %第二类
    end
```

（2）In view of the normal cell class ω𝟏, the corresponding data of sample features are extracted as Ω1 ={ -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}，and the sample features of abnormal cell class ω𝟐 are listed as Ω2 = { 2.8792，0.7932，1.1882，3.0682，4.2532， 0.3271, 0.9846, 2.7648, 2.6588}.The prior probabilities of both ω𝟏 and ω𝟐 are known as P(W1) = 0.9, P(W2) = 0.1

The loss parameters for different decision action are given as table 1

Table 1 the loss parameters for different decision

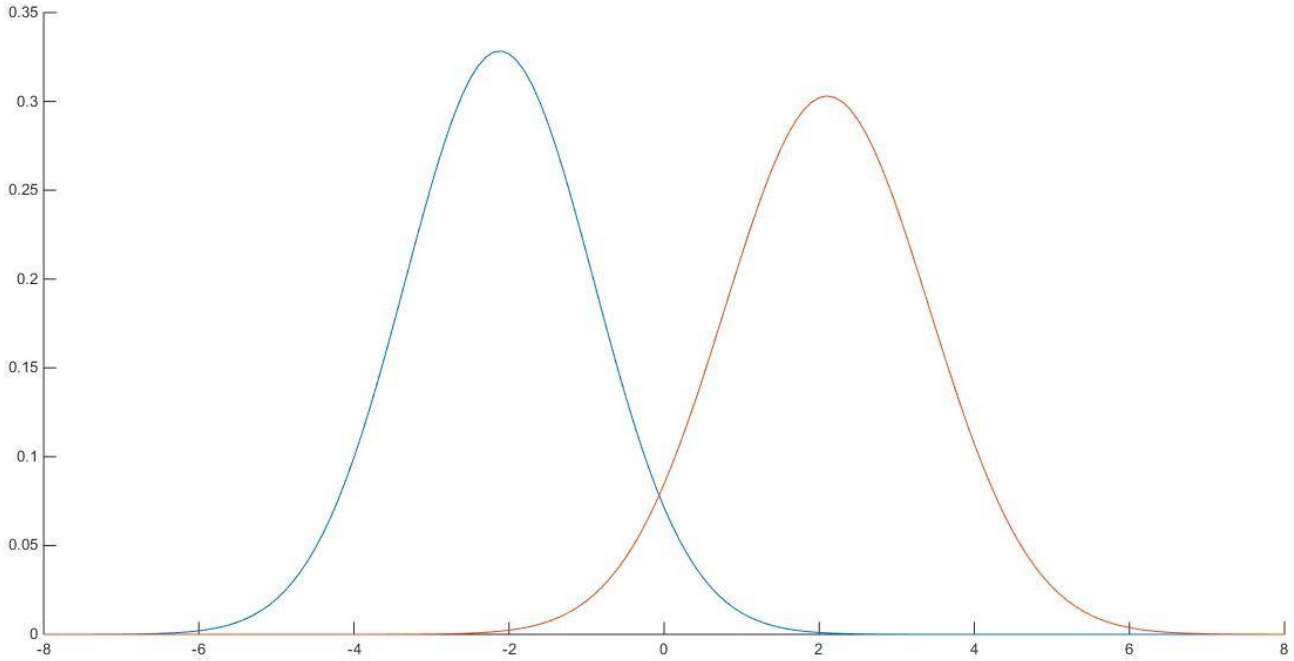
real class		ω_1	ω_2	
decision action	a_1	0	1	loss parameters
	a_2	6	0	

Suppose the conditional probability distributions are Gaussian, find the conditional probability density functions P(X|W1) and P(X|W2)and complete the design of Bayessian classifier with minimum risk, and then give a comparative analysis with the situation without considering decision loss. Draw the curves of prior and posterior probability density functions, P(X|W1) ,P(X|W2), P(W1|X), and P(W2|X) , give the classifying decision boundary function and illustration of classification result.

根据训练集得到满足高斯分布的类条件概率曲线：

```
x=-8:0.1:8;
a=[ -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];
b=[2.8792,0.7932,1.1882,3.0682,4.2532, 0.3271, 0.9846, 2.7648, 2.6588];
amean=mean(a);
astd=std(a);
bmean=mean(b);
bstd=std(b);
y=normpdf(x,amean,astd);
y1=normpdf(x,bmean,bstd);
hold on
plot(x,y)
plot(x,y1)
hold off
```

类条件概率曲线如图：



P(X|W1) 类条件概率分布正态分布分别为（-2.1226 ， 1.4757），

P(X|W2) 类条件概率分布正态分布分别为（2.1019 ,1.7326）。

1. 最小错误率贝叶斯决策

最小错误贝叶斯决策可按下列步骤进行：

1)在已知 $P(\omega_i)$， $P(X|\omega_i)$ ，i=1,…，c 及给出待识别的 X 的情况下，根据贝叶斯公式计算出后验概率：

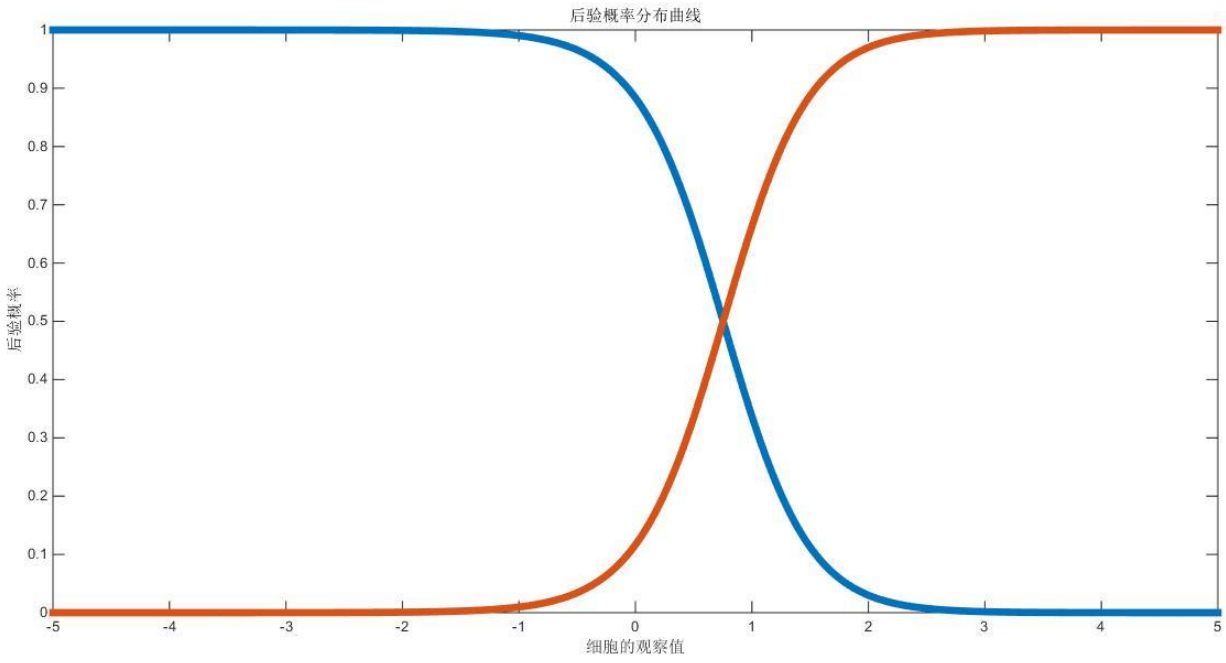
$$P(\omega_i|X)=\frac{P(X|\omega_i)P(\omega_i)}{\sum_{j=1}^{c}P(X|\omega_i)P(\omega_i)} \qquad j=1,\cdots,\ x$$

2)将不同的后验概率进行比较，找出使其错误率最小，即后验概率最大的决策。

```
%计算在 w1,w2 下的后验概率
pw1_x(i)=(pw1*normpdf(x(i),e1,a1))/(pw1*normpdf(x(i),e1,a1)+pw2*normpdf(x(i),e
2,a2)) ;
pw2_x(i)=(pw2*normpdf(x(i),e2,a2))/(pw1*normpdf(x(i),e1,a1)+pw2*normpdf(x(i),e
2,a2)) ;
end


%比较两类后验概率
for i = 1:m
    if pw1_x(i)>pw2_x(i)
        result(i)=0;          %正常细胞
    else
        result(i)=1;          %异常细胞
    end
end
```

后验概率曲线图：



蓝色曲线表示正常细胞的后验概率曲线，红色曲线表示异常细胞的后验概率曲线。可知当 X≤0.752 时，正常细胞的后验概率大于异常细胞的后验概率，故对于最小错误贝叶斯决策，观测值在此区间内的都认为是正常细胞，否则都被看作异常细胞。

2.最小风险贝叶斯决策

最小风险贝叶斯决策可按下列步骤进行：

1)在已知 $P(\omega_i)$，$P(X|\omega_i)$，i=1,…，c 及给出待识别的 X 的情况下，根据贝叶斯公式计算出后验概率：

$$P(\omega_i|X)=\frac{P(X|\omega_i)P(\omega_i)}{\sum_{j=1}^{c}P(X|\omega_i)P(\omega_i)} \qquad j=1,\cdots,\ x$$

2)利用计算出的后验概率及决策表，按下面的公式计算出采取 a_i，i=1,…，a 的条件风险

$$R(a_i|X)=\sum_{j=1}^{c}\lambda(a_i,\omega_j)P(\omega_j|X),\ i=1,2,\cdots,a$$

3)对(2)中得到的 a 个条件风险值 $R(a_i|X)$，i=1,…，a 进行比较，找出使其条件风险最小的决策 a_k，即 $R(a_k|x)=\min_{i=1,\perp,a}R(a_i|x)$

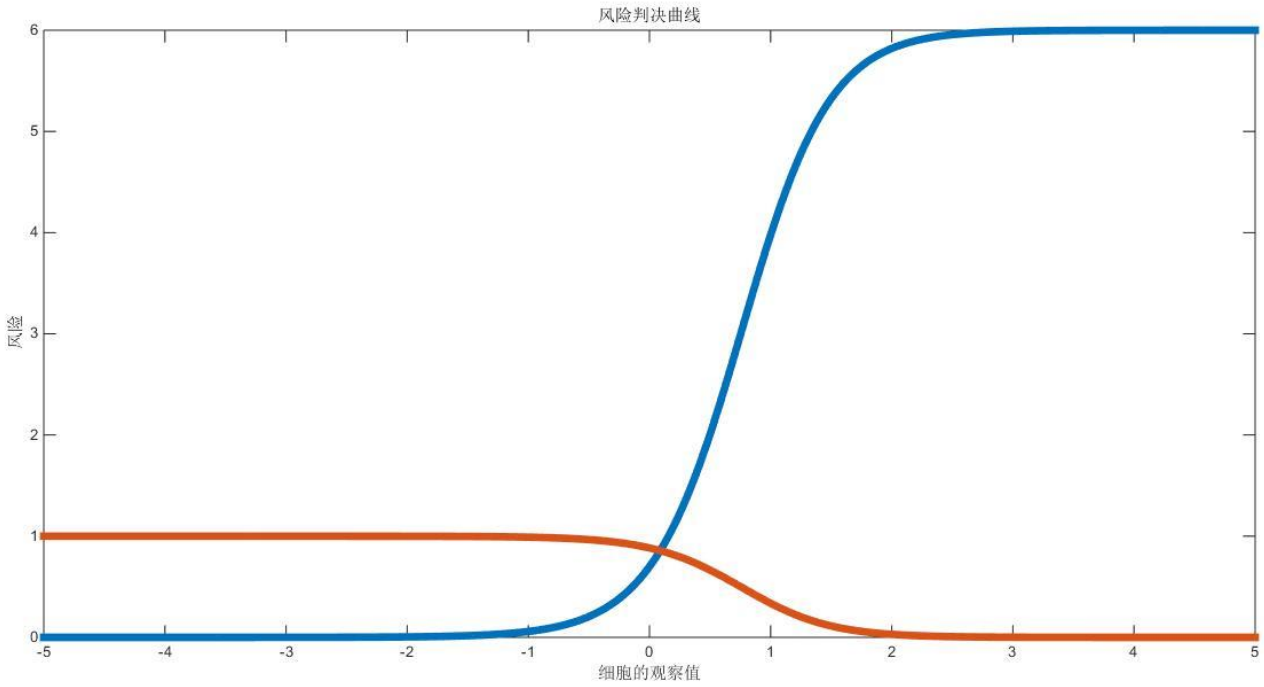
则 a_k 就是最小风险贝叶斯决策。

```
%风险决策表
r11=0;r12=1;
r21=6;r22=0;

%计算两类风险值
R1_x(i)=r11*pw1*normpdf(x(i),e1,a1)/(pw1*normpdf(x(i),e1,a1)+pw2*normpdf(x(i)
,e2,a2))+r21*pw2*normpdf(x(i),e2,a2)/(pw1*normpdf(x(i),e1,a1)+pw2*normpdf(x(i)
,e2,a2));

R2_x(i)=r12*pw1*normpdf(x(i),e1,a1)/(pw1*normpdf(x(i),e1,a1)+pw2*normpdf(x(i)
,e2,a2))+r22*pw2*normpdf(x(i),e2,a2)/(pw1*normpdf(x(i),e1,a1)+pw2*normpdf(x(i)
,e2,a2));
%比较条件风险值
for i=1:m
    if R2_x(i)>R1_x(i)   %第二类比第一类风险大
        result(i)=0;      %判为正常细胞（损失较小），用 0 表示
    else
        result(i)=1;      %判为异常细胞，用 1 表示
    end
```

风险判决曲线：



蓝色曲线表示正常细胞的风险判决曲线，红色曲线表示异常细胞的风险判决曲线。可知当 X≤0.06 时正常细胞的风险小于异常细胞的风险，观测值在此区间内的都认为是正常细胞，否则都被看作异常细胞。同时，我们可以看到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，即前者是后者的特例。

（3）Create a pattern dataset of multiple classes and high dimension with more than 50 samples for each class. Then design a Bayesian classifier and complete the corresponding experiments and comparative analysis. Think and analyse the intrinsic relationship between the classifier of two classes and the one of multiple classes, give your comments.

本实验中，我们采用有名的 Iris 数据集。Iris 也称鸢尾花卉数据集，是一类多重变量分析的数据集。数据集包含 150 个数据集，分为 3 类，每类 50 个数据，每个数据包含 4 个属性。可通过花萼长度，花萼宽度，花瓣长度，花瓣宽度 4 个属性预测鸢尾花卉属于（Setosa，Versicolour，Virginica）三个种类中的哪一类。

三类数据集的贝叶斯分类器实验原理：

对于具有多个特征参数的样本，其正态分布的概率密度函数可定义为

$$p(\mathbf{x})=\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})^{T}\right\}$$

式中，$\mathbf{x}=\left[x_1,x_2,\mathrm{L}\ ,x_d\right]$ 是 d 维行向量，$\boldsymbol{\mu}=\left[\mu_1,\mu_2,\mathrm{L}\ ,\mu_d\right]$ 是 d 维行向量，Σ 是 $d\times d$ 维协方差矩阵，Σ^{-1} 是 Σ 的逆矩阵，$|\Sigma|$ 是 Σ 的行列式。

本实验我们采用最小错误率的贝叶斯决策，使用如下的函数作为判别函数

$$g_i(\mathbf{x})=p(\mathbf{x}\,|\,\omega_i)P(\omega_i),\qquad i=1,2,3\text{（3 个类别）}$$

其中 $P(\omega_i)$ 为类别 ω_i 发生的先验概率，$p(\mathbf{x}\,|\,\omega_i)$ 为类别 ω_i 的类条件概率密度函数。

由其判决规则，如果使 $g_i(\mathbf{x})>g_j(\mathbf{x})$ 对一切 $j\neq i$ 成立，则将 $\mathbf{x}$ 归为 ω_i 类。

我们根据假设：类别 ω_i，i=1,2,……,N 的类条件概率密度函数 $p(\mathbf{x}\,|\,\omega_i)$，i=1,2,……,N 服从正态分布，即有 $p(\mathbf{x}\,|\,\omega_i)\sim N(\boldsymbol{\mu}_i,\Sigma_i)$，那么上式就可以写为

$$g_i(\mathbf{x})=\frac{P(\omega_i)}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})^{T}\right\},\quad i=1,2,3$$

对上式右端取对数，可得

$$g_i(\mathbf{x})=-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{i})\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{i})^{T}+\ln P(\omega_i)-\frac{1}{2}\ln|\Sigma_i|-\frac{d}{2}\ln(2\pi)$$

上式中的第二项与样本所属类别无关，将其从判别函数中消去，不会改变分类结果。则判别函数 $g_i(\mathbf{x})$ 可简化为以下形式：

$$g_i(\mathbf{x})=-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{i})\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{i})^{T}+\ln P(\omega_i)-\frac{1}{2}\ln|\Sigma_i|$$

实验步骤：

（1）每一类样本抽出前 40 个，分别求其均值，公式如下

$$\boldsymbol{\mu}^{\omega_i} = \frac{1}{N_i}\sum_{\mathbf{x}\in\omega_i}\mathbf{x}^{\omega_i} \qquad i=1,2,3$$

```
%求第一类样本均值
for i = 1:N
    for j = 1:4
        w1(i,j) = iris(i,j+1);
    end
end
sumx1 = sum(w1,1);
for i=1:4
    meanx1(1,i)=sumx1(1,i)/N;
end
```

（2）求每一类样本的协方差矩阵、逆矩阵 Σ_i^{-1} 以及协方差矩阵的行列式 $|\Sigma_i|$，协方差矩阵计算公式如下

$$\sigma_{jk}^{i} = \frac{1}{N_i - 1}\sum_{l=1}^{N_i}(x_{lj}-\mu_j^{\omega_i})(x_{lk}-\mu_k^{\omega_i}) \qquad j,k=1,2,3,4$$

```
%求第一类样本协方差矩阵
z1(4,4) = 0;
var1(4,4) = 0;
for i=1:4
for j=1:4
   for k=1:N
     z1(i,j)=z1(i,j)+(w1(k,i)-
meanx1(1,i))*(w1(k,j)-meanx1(1,j));
   end
   var1(i,j) = z1(i,j) / (N-1);
 end
end
```

（3）对三个类别，分别取每组剩下的 10 个样本，每两组进行分类。由于每一类样本都相等，且每一类选取用作训练的样本也相等，在每两组进行分类时，待分类样本的类先验概率 $P(\omega_i)=0.5$。将各个样本代入判别函数

$$g_i(\mathbf{x})=-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)^T+\ln P(\omega_i)-\frac{1}{2}\ln|\Sigma_i|$$

根据判决规则，如果使 $g_i(\mathbf{x})>g_j(\mathbf{x})$ 对一切 $j\neq i$ 成立，则将 $\mathbf{x}$ 归为 ω_i 类。

1）取第一类样本的后 10 个数据，按 ω_1 、ω_2 分类，若 g1>g2，则属于 ω_1 。按 ω_1 、ω_3 分类，若 g1>g3，则属于 ω_1 。

2）取第二类样本的后 10 个数据，按 ω_1 、ω_2 分类，若 g2>g1，则属于 ω_2 。按 ω_2 、ω_3 分类，若 g2>g3，则属于 ω_2 。

3）取第三类样本的后 10 个数据，按 ω_1 、ω_3 分类，由 g3>g1，则属于 ω_2 。按 ω_2 、ω_3 分类，由 g3>g2，则属于 ω_3 。

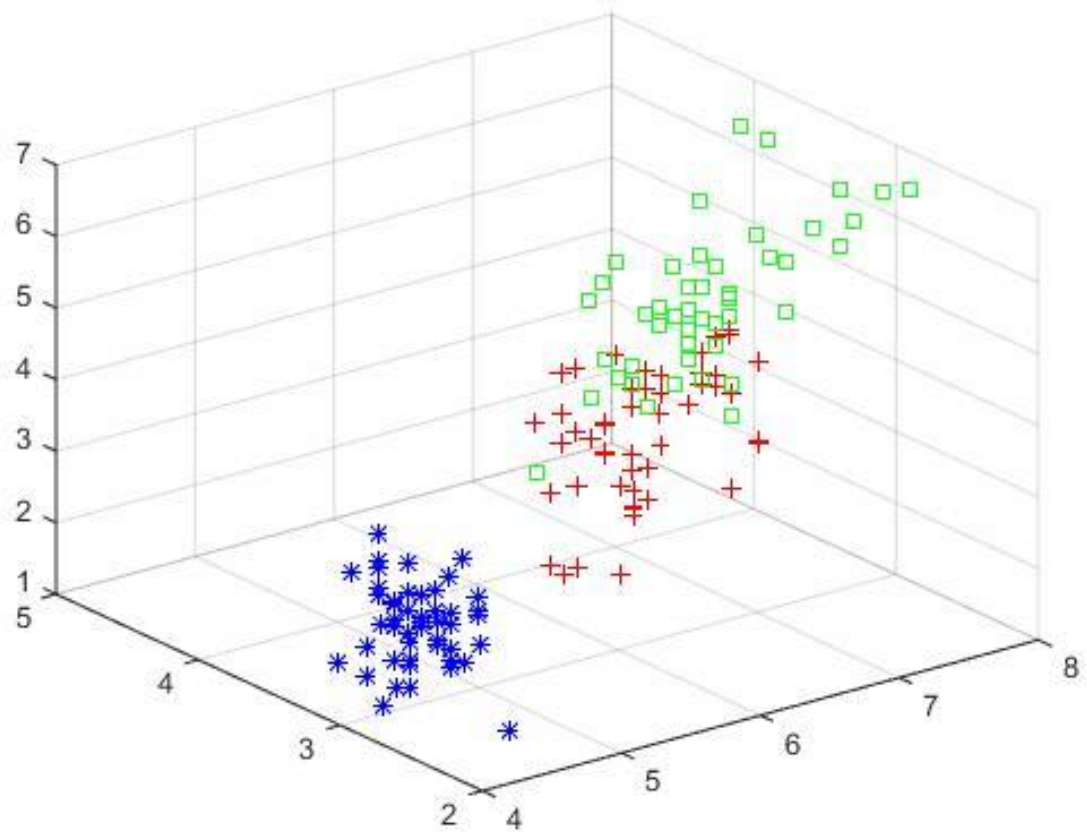
若取第一类后 10 个数据按 ω_1 、ω_2 分类进行分类，代码如下：

```
g1 = (-0.5)*([x,y,z,h]-meanx1)*var1_inv*([x,y,z,h]'-meanx1') -0.5*log(abs(var1_det))
+log(0.5);
g2 = (-0.5)*([x,y,z,h]-meanx2)*var2_inv*([x,y,z,h]'-meanx2') -0.5*log(abs(var2_det))
+log(0.5);
if g1>g2
t1=t1+1; %若 g1>g2,则属于第一类，否则属于第二类
else
t2=t2+1;
end
```

同理第二类和第三类、第一类和第三类可进行分类。

实验结果：第一类后 10 个数据属于第一类；
第二类后 10 个数据属于第二类；
第三类后 10 个数据属于第三类。

Iris 数据集共有四个特征属性，我们可以选取前三个特征属性的值，做出三维空间的点分布如图：



蓝色的星号表示：Iris Setosa（山鸢尾）

红色的十字表示：Iris Versicolour（杂色鸢尾）

绿色的方块表示：Iris Virginica（维吉尼亚鸢尾）

三个坐标轴表示：花萼长度、花萼宽度、花瓣长度共三个特征。

总结：

经过贝叶斯两类与多类分类器的实验，我认识到，两类分类是分类器工作的基础，多类分类只是两类分类的扩展。