吉林大学自编教材

模式识别基础实验指导书

林 琳 编

吉林大学教材发行与调配中心
2020 年 10 月

目　　录

实验课概况

课程名称：**模式识别基础**

适应专业：信息工程、通信工程、电子工程

实验学时：8

开科学期：5 学期

一、 实验的性质、任务和基本要求

（一） 实验课的性质

《模式识别基础》实验课是一门非独立的实验课，是同学对模式识别理论内容进行充分的理解的基础上，根据相应的原理，设计实验内容，完成实验任务，是理论知识实践化的方式，利于学生更好的吸收，领悟模式识别的原理与应用，培养学生的动手实践的能力。

（二） 实验课的基本要求

1、理解模式识别的基本概念

2、掌握各种算法的流程，以及相应的优缺点。

3、会使用相应的模式识别分类器等算法处理实验问题。

二、 实验的分配情况

序号	实验内容	学时	选作
1	Bayes 分类器算法	2	必做
2	总体概率密度分布的非参数估计	2	必做
3	Fisher 线性分类器设计	2	二选一
4	基于感知器准则线性分类器设计	2	
5	近邻法分类器设计	2	必做

实验一 Bayes 分类器设计

1.1 实验目的

本实验旨在让学生对模式识别有一个初步的理解，能够根据自己的设计对贝叶斯决策理论算法有一个深刻地认识，理解二类分类器的设计原理。

1.2 实验条件

PC 机一台、Matlab 仿真软件

1.3 实验原理

1.3.1 最小错误率贝叶斯决策

最小错误率贝叶斯决策就是利用贝叶斯公式，按照尽量减少分类错误的原则而得出的分类规则，可以使得分类的错误率最小。

最小错误贝叶斯决策可按下列步骤进行：

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

$$P(\omega_i|\mathbf{x})=\frac{P(\omega_i,\mathbf{x})}{p(\mathbf{x})}=\frac{P(\omega_i)p(\mathbf{x}|\omega_i)}{\sum_{j=1}^{c}P(\omega_j)p(\mathbf{x}|\omega_j)} \tag{1.1}$$

(2)利用式(1.1)计算出待识别样本 $\mathbf{x}$ 属于每类的后验概率 $P(\omega_i|\mathbf{x})$，i=1,…，c，找到最大后验概率对应的类别，将 $\mathbf{x}$ 归属到那一类中。其分类准则可以表示为：

两类情况：
$$\begin{array}{l}\text{若}P(\omega_1|\mathbf{x})>P(\omega_2|\mathbf{x}),\ \text{则}\mathbf{x}\in\omega_1\\ \text{若}P(\omega_2|\mathbf{x})>P(\omega_1|\mathbf{x}),\ \text{则}\mathbf{x}\in\omega_2\end{array} \tag{1.2}$$

多类情况：
$$\text{若}j=\arg\max_i P(\omega_i|\mathbf{x}),\ \text{则}\mathbf{x}\in\omega_j \tag{1.3}$$

在式(1.1)中，比较样本 **x** 分属各个类别的后验概率大小不需要计算 $p(\mathbf{x})$，该判别依据有下列几种等价形式

规则 2： $p(\mathbf{x}\mid\omega_i)P(\omega_i)=\underset{j=1,2}{\operatorname{argmax}}\{p(\mathbf{x}\mid\omega_j)P(\omega_j)\}$，则$x\in\omega_i$ (1.4)

规则 3： 似然比$l(\mathbf{x})=\dfrac{p(\mathbf{x}\mid\omega_i)}{p(\mathbf{x}\mid\omega_j)}>\dfrac{P(\omega_j)}{P(\omega_i)}$ $\mathbf{x}\in\begin{cases}\omega_1\\ \omega_2\end{cases}$ (1.5)

规则 4: $\ln p(\mathbf{x}\mid\omega_i)P(\omega_i)=\ln p(\mathbf{x}\mid\omega_i)+\ln P(\omega_i)$

$$\ln p(\mathbf{x}\mid\omega_i)+\ln P(\omega_i)=\underset{j=1,2}{\operatorname{argmax}}\{\ln p(\mathbf{x}\mid\omega_j)+\ln P(\omega_j)\},\ \text{则}x\in\omega_i \tag{1.6}$$

1.3.2 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$P(\omega_i)$，$p(\mathbf{x}\mid\omega_i)$，$i$=1,…，$c$ 及给出待识别的 **x** 的情况下，利用式(1.1)根据贝叶斯公式计算出后验概率；

(2) 利用式(1.1)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

$$R(\alpha_i\mid\mathbf{x})=\sum_{j=1}^{c}\lambda(\alpha_i,\omega_j)P(\omega_j\mid\mathbf{x}),\ \ i=1,2,...a \tag{1.7}$$

(3) 对式(1.7)中得到的 a 个条件风险值$R(\alpha_i\mid\mathbf{x}),\ \ i=1,...,a$,进行比较，找出使其条件风险最小的决策$a_k$，即

$$R(\alpha_k\mid\mathbf{x})=\min_{j=1,..,c}R(\alpha_j\mid\mathbf{x}),\ \ k=1,..,a \tag{1.8}$$

则a_k就是最小风险贝叶斯决策。

1.4 实验内容

假定某个局部区域细胞识别中正常（ω_1）和非正常（ω_2）两类先验概率分别为

正常状态：P（ω_1）=0.9；

异常状态：P（ω_2）=0.1。

现有一系列待观察的细胞，其观察值为 x：

-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	2.8792	-0.9780	0.7932	1.1882	3.0682
-1.5799	-1.4885	-0.7431	-0.4221	-1.1186	4.2532

已知两类的类条件概率密度 $p(\mathbf{x}|\omega_1)$、 $p(\mathbf{x}|\omega_2)$ 的曲线如下图所示。

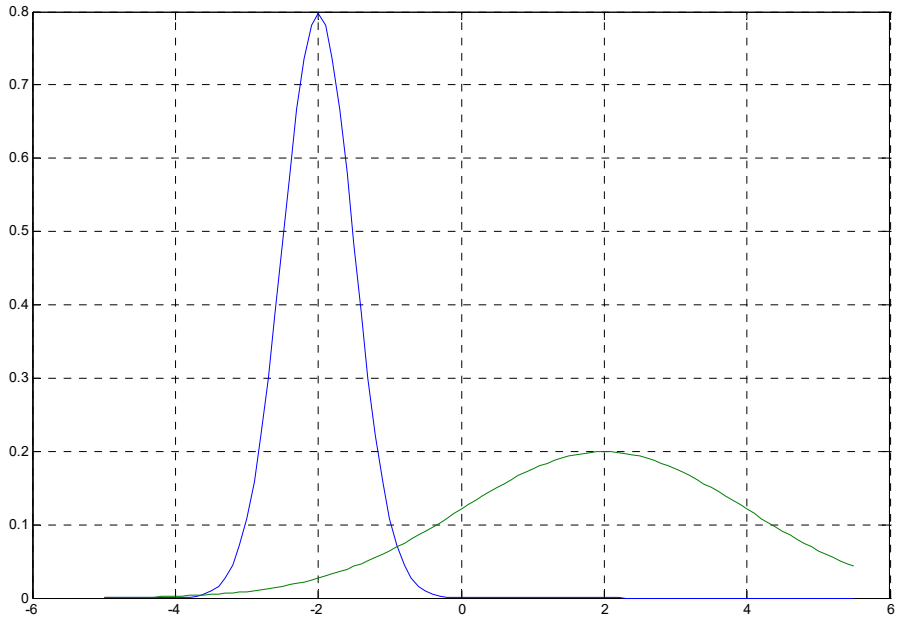

图 1 两类样本的类条件概率密度曲线

$p(\mathbf{x}|\omega_1)$、 $p(\mathbf{x}|\omega_2)$ 类条件概率分布分别是（-2，0.25）（2,4）的正态分布，试对观察的结果进行分类。

1.5 实验要求

1.根据最小错误率贝叶斯决策，利用 Matlab 完成分类器的设计。

1）写出相应程序语句的文字说明；

2）程序设计过程中，要求有子程序的调用。

3）根据上述例题中的数据，画出后验概率的分布曲线以及分类的结果示意图。

2.根据最小风险贝叶斯决策，决策表如下。

表 1 最小风险贝叶斯决策表

状态 / 决策	ω_1	ω_2
α_1	0	6
α_2	2	0

1）请重新设计程序，画出相应的条件风险的分布曲线和分类结果,并比较两个结果。

2）在损失矩阵为 0-1 损失函数时，比较最小错误贝叶斯决策和最小风险决策的结果是否一致。

实验二 总体概率密度分布的非参数估计

2.1 实验目的

本实验旨在让学生对了解基于 Parzen 窗的概率密度估计方法，加深非参数估计基本思想的认识和理解。

2.2 实验条件

PC 机一台、Matlab 仿真软件

2.3 实验原理

设区域 R_N 是一个 d 维的超立方体，并设 h_N 是超立方体的棱长，则超立方体的体积为

$$V_N = h_d{}^N \tag{2.1}$$

定义窗函数 $\varphi(u)$ 为

$$\varphi(u) = \begin{cases} 1, & \left|u_j \le \frac{1}{2}\right|; j = 1,2,\cdots,d \\ 0, & \text{其他} \end{cases} \tag{2.2}$$

由于 $\varphi(u)$ 是以原点为中心的一个超立方体，所以当 x_i 落入到以 x 为中心，体积为 V_N 的超立方体时，$\varphi(u) = \varphi[(x - x_i)/h_N] = 1$，否则 $\varphi(u) = 0$，因此，落入该超立方体内的样本数为

$$k_N = \sum_{i=1}^{N} \varphi\left(\frac{(x - x_i)}{h_N}\right) \tag{2.3}$$

$\hat{p}_N(x)$ 是 $p_N(x)$ 的第 N 次估计，则有

$$\hat{p}_N(x) = \frac{k_N / N}{V_N} \tag{2.4}$$

由上面两式可得

$$\hat{p}_N(x)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{V}\varphi(\frac{x-x_i}{h_N}) \tag{2.5}$$

上式就是 Parzen 窗法的基本公式。当然，窗函数不限于超立方体窗函数，还可以有更一般的形式。式(2.5)表示 $p(x)$ 的估计可以看作是 x 和 x_i 的函数的一种平均。实际上，窗函数的作用是内插，每一样本对估计所起的作用依赖于它到 x 的距离。

2.4 实验内容

1. 理解利用 Parzen 窗法进行概率密度的估计过程。

2. 假设已知待估计概率密度函数 $p(x)$ 是一个均值为零，方差为 1 的正态概率密度函数，设 N 为样本个数，h_1 为窗宽。若采用 Matlab 编程，利用正态窗实现概率密度估计，并观察参数对估计结果的影响。

本实验采用正态窗进行 Parzen 窗法的设计实现，实验步骤如下：

1) 利用 Matlab 随机产生 N 个学习样本 x_i 的样本集，作为仿真数据；

2)利用公式(2.5)，选取正态窗函数，编写利用 Parzen 窗函数求取 $\hat{p}_N(x)$ 的仿真程序，并设 $h_N=\frac{h_1}{\sqrt{N}}$。其中，正态窗函数

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2} \tag{2.6}$$

3)绘制不同参数 h_1 和不同样本数 N 下所获得的估计概率密度曲线，并进行比较；

4)根据仿真结果，分析 Parzen 窗法进行概率密度估计的非参数估计的特点。

2.5 实验报告要求

1．完成基于正态窗函数的 Parzen 窗法概率密度函数估计的 Matlab 程序编写，并写出相应程序语句的文字说明；

2．选取 h_1=0.25, 1.0, 4.0，分别在样本数 N=1, 16, 256, 1024, 4096 时画出原始概率密度曲线和不同参数下估计的概率密度曲线。分析所得到概率密度曲线的变化

情况，说明 N、h_1 对概率密度函数估计的影响。

3.分析程序运行和实验中遇到的困难。

实验三 基于 Fisher 准则线性分类器设计

3.1 实验类型

设计型：线性分类器设计（Fisher 准则）

3.2 实验目的

本实验旨在让学生进一步了解分类器的设计概念，能够根据自己的设计对线性分类器有更深刻地认识，理解 Fisher 准则方法确定最佳线性分界面方法的原理，以及 Lagrande 乘子求解的原理。

3.3 实验条件

PC 机一台、Matlab 软件

3.4 实验原理

线性判别函数的一般形式可表示成

$$g(\mathbf{x})=\mathbf{w}^T\mathbf{x}+w_0 \tag{3.1}$$

其中

$$\mathbf{x}=\begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_d \end{pmatrix} \qquad \mathbf{w}=\begin{pmatrix} w_1 \\ w_2 \\ \cdots \\ w_d \end{pmatrix}$$

根据 Fisher 选择投影方向 **w** 的原则，即使原样本向量在该方向上的投影能兼顾类间分布尽可能分开，类内样本投影尽可能密集的要求，用以评价投影方向 **w** 的准则函数为

$$J_F(\mathbf{w})=\frac{(\tilde{m}_1-\tilde{m}_2)^2}{\tilde{S}_1^{\,2}+\tilde{S}_2^{\,2}} \tag{3.2}$$

其中，$\tilde{m}_1$、$\tilde{m}_2$ 分别是一维投影上两类样本的均值，$\tilde{S}_1^{\,2}$、$\tilde{S}_2^{\,2}$ 分别是一维投影上两类样本的类内离散度，可以由下面两个公式计算得到。

$$\widetilde{m}_i = \frac{1}{N_i}\sum_{y\in Y_i} y = \frac{1}{N_i}\sum_{\mathbf{x}\in \mathbf{X}_i} \mathbf{w}^T\mathbf{x} = \mathbf{w}^T\mathbf{m}_i \quad , \quad i=1,2 \tag{3.3}$$

$$\widetilde{S}_i^{\,2} = \sum_{y\in Y_i}(y-\widetilde{m}_i)^2 = \sum_{\mathbf{x}\in \mathbf{X}_i}(\mathbf{w}^T\mathbf{x}-\mathbf{w}^T\mathbf{m}_i) = \mathbf{w}^T\mathbf{S}_i\mathbf{w} \ , i=1,2 \tag{3.4}$$

其中 $\mathbf{m}_i$ 是 d 维样本向量空间第 i 类的均值向量，$\mathbf{S}_i$ 是第 i 类的类内离散度矩阵。

使得 $J_F(\mathbf{w})$ 为最大值的 $\mathbf{w}$ 就是要求的最佳投影向量 $\mathbf{w}^*$，其表达式如下

$$\mathbf{w}^* = \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \tag{3.5}$$

式(3.5)是使用 Fisher 准则求最佳法线向量的解，利用求得的 $\mathbf{w}^*$ 对样本向量进行的如式(3.1)所示变换形式，称为线性变换。在式(3.5)中，$\mathbf{m}_1 - \mathbf{m}_2$ 是 d 维样本向量空间两类均值向量的差，结果是一个 d 维向量；$\mathbf{S}_w^{-1}$ 是 $\mathbf{S}_w$ 的逆矩阵，其中 $\mathbf{S}_w$ 是 d 维样本向量空间总的类内离散度矩阵，是一个 $d\times d$ 维矩阵，$\mathbf{S}_w^{-1}$ 也是 $d\times d$ 维矩阵，这样当采用式(3.5)计算最佳投影方向 $\mathbf{w}^*$ 时，得到的 $\mathbf{w}^*$ 也是一个 d 维的向量。类内总离散度矩阵 $\mathbf{S}_w=\mathbf{S}_1+\mathbf{S}_2$，其中

$$\mathbf{S}_i = \sum_{\mathbf{x}\in \mathbf{X}_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T, \quad i=1,2 \tag{3.6}$$

向量 $\mathbf{w}^*$ 是使 Fisher 准则函数 $J_F(\mathbf{w})$ 达到极大值的解，也就是按照 Fisher 准则将 d 维样本向量 $\mathbf{x}$ 投影到一维 Y 空间的最佳投影方向，$\mathbf{w}^*$ 的各个分量是对原 d 维特征向量求加权和的权值。

以上讨论了线性判别函数加权向量 $\mathbf{w}$ 的确定方法，并讨论了使 Fisher 准则函数极大的 d 维向量 $\mathbf{w}^*$ 的计算方法。但是判别函数中的另一项 w_0 尚未确定，一般可采用以下几种方法确定 w_0 ，即

$$w_0 = -\frac{\widetilde{m}_1 + \widetilde{m}_2}{2} \tag{3.7}$$

$$w_0 = -\frac{N_1\widetilde{m}_1 + N_2\widetilde{m}_2}{N_1+N_2} = -\widetilde{m} \tag{3.8}$$

当 $P(w_1)$、$P(w_2)$ 已知时，

$$w_0 = -\left[\frac{\widetilde{m}_1 + \widetilde{m}_2}{2} + \frac{\ln(P(w_1)/P(w_2))}{N_1 + N_2 - 2}\right] \tag{3.9}$$

这里阈值点 $y_0 = -w_0$。式(3.7)根据投影后两类样本均值之间的平均距离来确定阈值点的。式(3.8)既考虑了投影后样本均值之间的平均距离，又考虑了两类样本的容量大小作为阈值位置的偏移修正。式(3.9)使用了先验概率 $P(w_i)$，这样可以使分类误差尽可能小。

当阈值点 y_0 确定之后，则可按以下规则进行分类

$$\begin{cases} y = \mathbf{w}^T\mathbf{x} > -w_0(\text{即}y_0) \rightarrow \mathbf{x} \in \omega_1 \\ y = \mathbf{w}^T\mathbf{x} > -w_0(\text{即}y_0) \rightarrow \mathbf{x} \in \omega_2 \end{cases} \tag{3.10}$$

Fisher 准则进行分类器的设计步骤：

1）选择训练样本向量，根据 Fisher 准则计算得到最佳投影方向 $\mathbf{w}^*$；

2）对输入的训练样本向量进行线性映射，并设定阈值点 y_0；

3）计算未知样本的投影，根据式(3.9)对样本的类别进行判断，实现分类。

3.5 实验内容：

已知两类分类问题，类别用 ω_1 和 ω_2 表示，每类的先验概率已知，$P(w_1)$=0.6，$P(w_2)$=0.4。这里样本向量的维数是 3 维。

ω_1 中数据向量 **xx1**=[$x1, y1, z1$]^T，其数据点的坐标对应如下。

$x1$ =

0.2331	1.5207	0.6499	0.7757	1.0524	1.1974
0.2908	0.2518	0.6682	0.5622	0.9023	0.1333
-0.5431	0.9407	-0.2126	0.0507	-0.0810	0.7315
0.3345	1.0650	-0.0247	0.1043	0.3122	0.6655

0.5838	1.1653	1.2653	0.8137	-0.3399	0.5152
0.7226	-0.2015	0.4070	-0.1717	-1.0573	-0.2099

*y*1=

2.3385	2.1946	1.6730	1.6365	1.7844	2.0155
2.0681	2.1213	2.4797	1.5118	1.9692	1.8340
1.8704	2.2948	1.7714	2.3939	1.5648	1.9329
2.2027	2.4568	1.7523	1.6991	2.4883	1.7259
2.0466	2.0226	2.3757	1.7987	2.0828	2.0798
1.9449	2.3801	2.2373	2.1614	1.9235	2.2604

*z*1=

0.5338	0.8514	1.0831	0.4164	1.1176	0.5536
0.6071	0.4439	0.4928	0.5901	1.0927	1.0756
1.0072	0.4272	0.4353	0.9869	0.4841	1.0992
1.0299	0.7127	1.0124	0.4576	0.8544	1.1275
0.7705	0.4129	1.0085	0.7676	0.8418	0.8784
0.9751	0.7840	0.4158	1.0315	0.7533	0.9548

ω_2 中数据向量 **xx2**=[*x*2, *y*2, *z*2]$^{\mathrm{T}}$，数据点的对应的三维坐标为

x2 =

1.4010	1.2301	2.0814	1.1655	1.3740	1.1829
1.7632	1.9739	2.4152	2.5890	2.8472	1.9539
1.2500	1.2864	1.2614	2.0071	2.1831	1.7909
1.3322	1.1466	1.7087	1.5920	2.9353	1.4664
2.9313	1.8349	1.8340	2.5096	2.7198	2.3148
2.0353	2.6030	1.2327	2.1465	1.5673	2.9414

y2 =

1.0298	0.9611	0.9154	1.4901	0.8200	0.9399
1.1405	1.0678	0.8050	1.2889	1.4601	1.4334
0.7091	1.2942	1.3744	0.9387	1.2266	1.1833

0.8798	0.5592	0.5150	0.9983	0.9120	0.7126
1.2833	1.1029	1.2680	0.7140	1.2446	1.3392
1.1808	0.5503	1.4708	1.1435	0.7679	1.1288

z2 =

0.6210	1.3656	0.5498	0.6708	0.8932	1.4342
0.9508	0.7324	0.5784	1.4943	1.0915	0.7644
1.2159	1.3049	1.1408	0.9398	0.6197	0.6603
1.3928	1.4084	0.6909	0.8400	0.5381	1.3729
0.7731	0.7319	1.3439	0.8142	0.9586	0.7379
0.7548	0.7393	0.6739	0.8651	1.3699	1.1458

数据的样本点分布如下图所示。

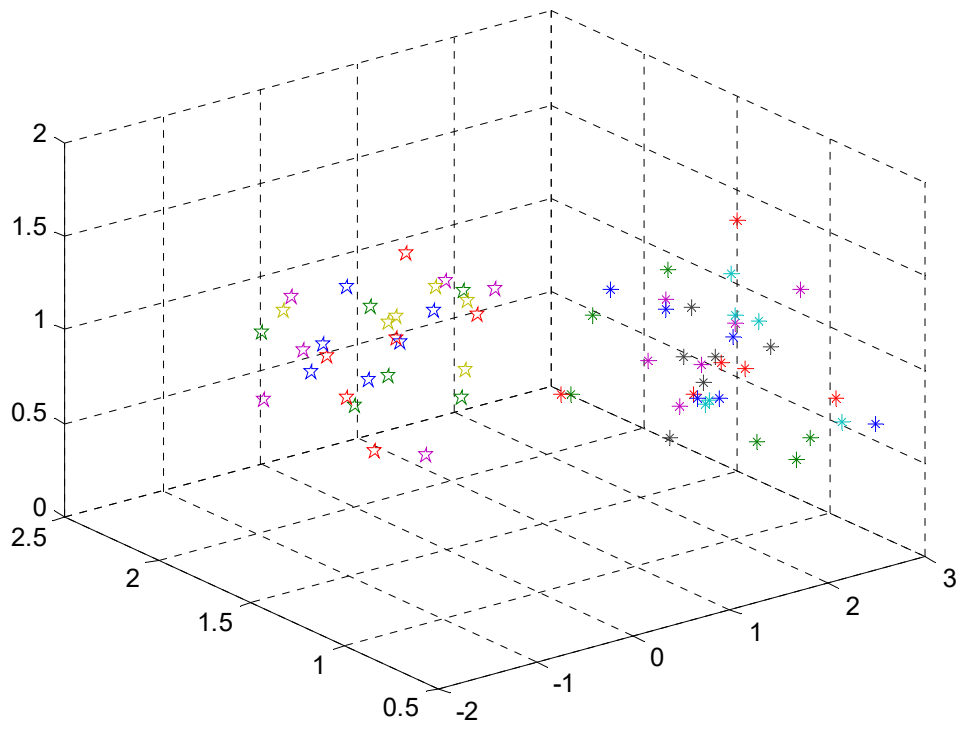

图 3-1 数据样本点分布图

3.6 实验要求：

1）把数据作为样本，根据 Fisher 准则选择投影方向 $\mathbf{w}$ 的原则，使原样本向量在该方向上的投影能兼顾类间分布尽可能分开，类内样本投影尽可能密集的要求，求出使 $J_F(\mathbf{w})$ 取极大值的 $\mathbf{w}^*$，并在图中表示出来。

2）用 Matlab 完成 Fisher 线性分类器的设计，程序的语句要求有注释。

3）根据上述的结果并判断（1，1.5，0.6）(1.2，1.0，0.55)，(2.0，0.9，0.68)，(1.2，1.5，0.89)，（0.23，2.33，1.43），属于哪个类别，并画出数据分类相应的结果图，要求画出其在 $\mathbf{w}$ 上的投影。

4）分析 $\mathbf{w}$ 的比例因子对于 Fisher 判别函数是否有影响，给出分析过程。

实验四　基于感知函数准则线性分类器设计

4.1　实验类型：

设计型：线性分类器设计（感知函数准则）

4.2　实验目的：

本实验旨在让同学理解感知准则函数的原理，通过软件编程模拟线性分类器，理解感知函数准则的确定过程，掌握梯度下降算法求增广权向量，进一步深刻认识线性分类器。

4.3　实验条件：

PC 机一台、Matlab 软件

4.4　实验原理：

感知准则函数是五十年代由 Rosenblatt 提出的一种自学习判别函数生成方法，由于 Rosenblatt 企图将其用于脑模型感知器，因此，被称为感知准则函数。其特点是随意确定的判别函数初始值，在对样本分类训练过程中逐步修正直至最终确定。

感知器（perceptioon）是一种神经网络模型。对于两类线性可分的样本模式类ω_1和ω_2，首先对样本进行性规范化处理，即将ω_2类的全部样本乘以(-1),这样对于所有的样本，判别函数都满足下式

$$g(\mathbf{x})=\mathbf{a}^T\mathbf{x}>0 \tag{4.1}$$

其中$\mathbf{a}=[a_1,a_2,...,a_d]^T$，$\mathbf{x}=[x_1,x_2,...,x_d]^T$。

感知准则函数通过对已知类别的训练样本集学习，利用梯度下降算法求满足上式的增广权向量。其步骤如下：

1）选择 N 个分属于ω_1和ω_2类的模式样本构成训练样本集，将训练样本写成增广向量的形式，并进行规范化处理。

2）采用单一样本修正法，用全部训练样本进行一轮迭代。每输入一个样本，计算判别函数$g(\mathbf{x})=\mathbf{a}^T\mathbf{x}$，根据判别函数分类结果的正误修正权向量，此时迭代次数

加 1。当任意给定增广权向量初始值$\mathbf{a}(1)$，第 k+1 次迭代时的权向量$\mathbf{a}(k+1)$等于第 k 次的权向量加$\mathbf{a}(k)$上被错分类的所有样本之和与ρ_k的乘积。假设第 k 次迭代，输入样本为$\mathbf{x}_k$

$$\begin{cases}\mathbf{a}(k+1)=\mathbf{a}(k) & \mathbf{a}^T(k)\mathbf{x}_k>0\\ \mathbf{a}(k+1)=\mathbf{a}(k)+\rho_k\mathbf{x}_k & \mathbf{a}^T(k)\mathbf{x}_k\le 0\end{cases} \tag{4.2}$$

可以证明，对于线性可分的样本集，经过有限次修正，一定可以找到一个解向量$\mathbf{a}^*$，即算法能在有限步内收敛。其收敛速度的快慢取决于初始权向量$\mathbf{a}(1)$和系数ρ_k。

4.5 实验内容

已知有两个样本空间ω_1和ω_2，这些点对应的横纵坐标的分布情况是：

x1=[1,2,4,1,5]; y1=[2,1,-1,-3,-3];

x2=[-2.5,-2.5,-1.5,-4,-5,-3];y2=[1,-1,5,1,-4,0];

在二维空间样本分布图形如下所示。（plot(x1,y1,x2,y2)）

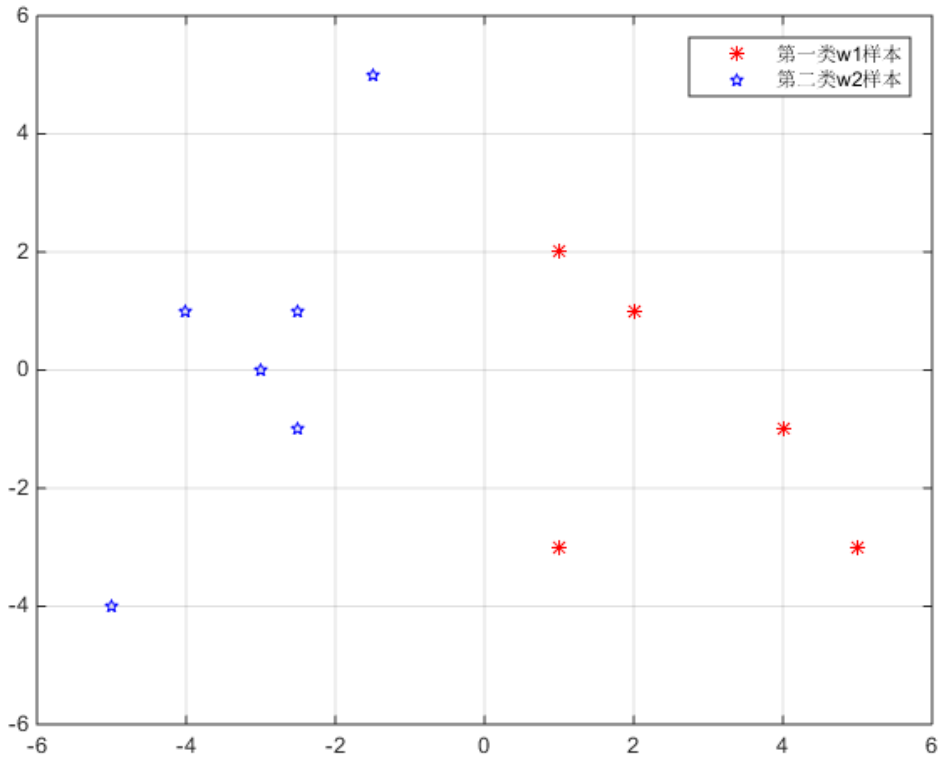

图 4-1 样本分布图

4.6 实验任务：

1、用 matlab 完成感知准则函数确定程序的设计。

2、请确定

sample=[(0,-3),(1,3),(-1,5),(-1,1),(0.5,6),(-3,-1),(2,-1),(0,1),(1,1),(-0.5,-0.5),(0.5,-0.5)];属于哪个样本空间,根据数据画出分类的结果。

3、请分析一下 ρ_k 和对 **a(1)** 于感知函数准则确定的影响，并确定当 ρ_k=1/2/3 时，相应的 k 的值，以及 **a(1)** 不同时，k 值得变化情况。

4、根据实验结果请说明感知准则函数是否是唯一的，为什么？

实验五 近邻法分类器设计

5.1 实验类型：

设计型：近邻法分类器设计

5.2 实验目的：

本实验旨在让同学理解近邻法的原理，通过软件编程，理解 k-近邻法和剪辑近邻的设计过程，掌握影响 k-近邻法错误率的估算因素等。

5.3 实验条件：

PC 机一台、Matlab 软件

5.4 实验原理：

1. k-近邻法

最近邻法可以扩展成找测试样本的 k 个最近样本作决策依据的方法。其基本规则是，在所有 N 个样本中找到与测试样本的 k 个最近邻者，其中各类别所占个数表示成 k_i，$i=1,\ldots,c$ 则决策规划是：

$$\text{如果 } g_j(\mathbf{x})=\max_{i=1,\ldots c} g_i(\mathbf{x}) \qquad \text{则决策 } \mathbf{x}\in\omega_j \tag{5-1}$$

k 近邻一般采用 k 为奇数，跟投票表决一样，避免因两种票数相等而难以决策。

2. 剪辑近邻法

剪辑近邻法的基本思想是从这样一个现象出发的，即当不同类别的样本在分布上有交迭部分的，分类的错误率主要来自处于交迭区中的样本。当我们得到一个作为识别用的参考样本集时，由于不同类别交迭区域中不同类别的样本彼此穿插，导致用近邻法分类出错。因此如果能将不同类别交界处的样本以适当方式筛选，可以实现既减少样本数又提高正确识别率的双重目的。为此，可以利用现有样本集对其自身进行剪辑。下面以两类别问题为例说明这种方法的原理。

假设现有一个样本集，样本数量为 N。我们将此样本集分成两个互相独立的样

本子集。一个被当作考试集 a^{NT}，另一个作为参考集a^{NR}，数量分别为N_T与N_R，$N_T + N_R = N$。将a^{NT}中的样本表示成$X_i,(i=1,\cdots,N_T)$，而在a^{NR}中的样本表示为$Y_j,(j=1,\cdots,N_R)$。

将一个样本集分成两个相互独立的样本子集是指，分完以后的两个子集具有相同的分布例如将一个样本集分成两个相互独立的对等子集，则在每个特征空间的子区域，两个子集都有相同的比例，或说各类数量近似相等。要注意指出的是每个子区域(从大空间到小空间)实际做时要用从总的集合中随机抽取的方式进行。

剪辑的过程是： 首先对a^{NT}中每一个 $\mathbf{x}_i$ 在a^{NR}中找到其最近邻的样本 $Y_i(\mathbf{x}_i)$，用 $Y_i(\mathbf{x}_i)$表示 Y_i 是 $\mathbf{x}_i$ 的最近邻参考样本。如果 Y_i 与 $\mathbf{x}_i$ 不属于同一类别，则将 $\mathbf{x}_i$ 从a^{NT}中删除，最后从a^{NT}中得到一个经过剪辑的样本集，称为剪辑样本集a^{NTE}。a^{NTE}可用来取代原样本集a^{N}，作为参考样本集对待识别样本进行分类。

a^{NT}经过剪辑后，要作为新的训练样本集，则a^{NR}是对其性能进行测试的样本，如发现a^{NT}中的某个训练样本对分类不利，就要把它剪辑掉。

实际上剪辑样本的过程也可以用 *k*-近邻法进行，即对a^{NT}中的每个样本 $\mathbf{x}_i$，找到在a^{NR}中的 *k* 个近邻，用 *k*-近邻法判断 $\mathbf{x}_i$ 是否被错分类。从而决定其取舍，其它过程与前述方法完全一样。

剪辑近邻法也可用到多类别情况。剪辑过程也可不止一次。重复多次的称为重复剪辑近邻法。

5.5 实验内容

有两个类别的样本 **x** 和 **y**，两类样本的分布规律服从正态分布，其均值和方差分别为（2，2），(-2，4)，每个类别里面分别有 100 个样本。可以利用下面阐述的 Matlab 程序产生上述数据。每一类的数据都是二维数据形式，若设数据格式为第一行为横坐标，相应的下一行对应的是纵坐标，对应的数据分布图如图 5-1 所示。

```
clear all
close all
```

mu1=2；sigma1=2；

mu2=-6，sigma2=4；

x = mu1 + sqrt(sigma1) * randn(2,100)

y = mu2 + sqrt(sigma2) * randn(2,100)

%画图

plot(x(1,:),x(2,:),'ro');

hold on

plot(y(1,:),y(2,:),' b*');

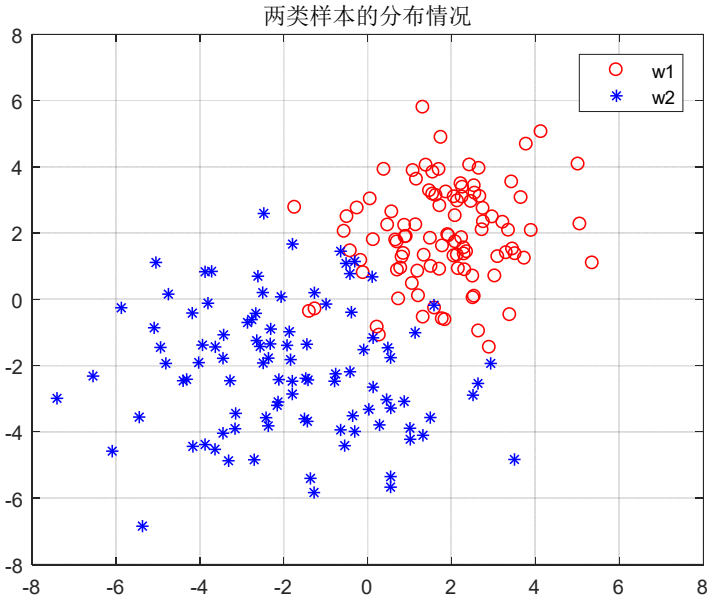

图 5-1 两类样本的分布图

请使用 *k*-近邻法判断下列 sample 中样本的分类情况

（-0.7303，2.1624），（1.4445，-0.1649），（-1.2587，0.9187），（1.2617，-0.2086），（0.7302，1.6587）

5.6 实验要求：

1、要求用 Matlab 编程，来确定分类的情况，并以图形的方式表示出来。附 Matlab 程序以及对应程序说明。

2、分析 *k* 值的不同对分类的情况是否有影响，并把结果用图形的方式表示出来。

3、请根据剪辑方法近邻的原理，对样本的空间进行剪辑，再确定上述样本点的分类情况。并对两种分类结果进行分析（选作）。