模式识别导论上机题7-特征选择与提取

薛犇 1500012752

1. 程序实现说明

本次实验采用Matlab作为编程语言，使用的版本为2016b。

Part I : PCA

PCA希望求得一个线性变换，使得样本变换后的协方差尽量大。

$$max_A tr(A^T \Sigma A)$$

$$= max_A tr(\sum_{i=1}^{r} a_i^T \Sigma a_i)$$

$$s.t. a_i^T a_i = 1, a_i^T \Sigma a_j = 0, i \neq j$$

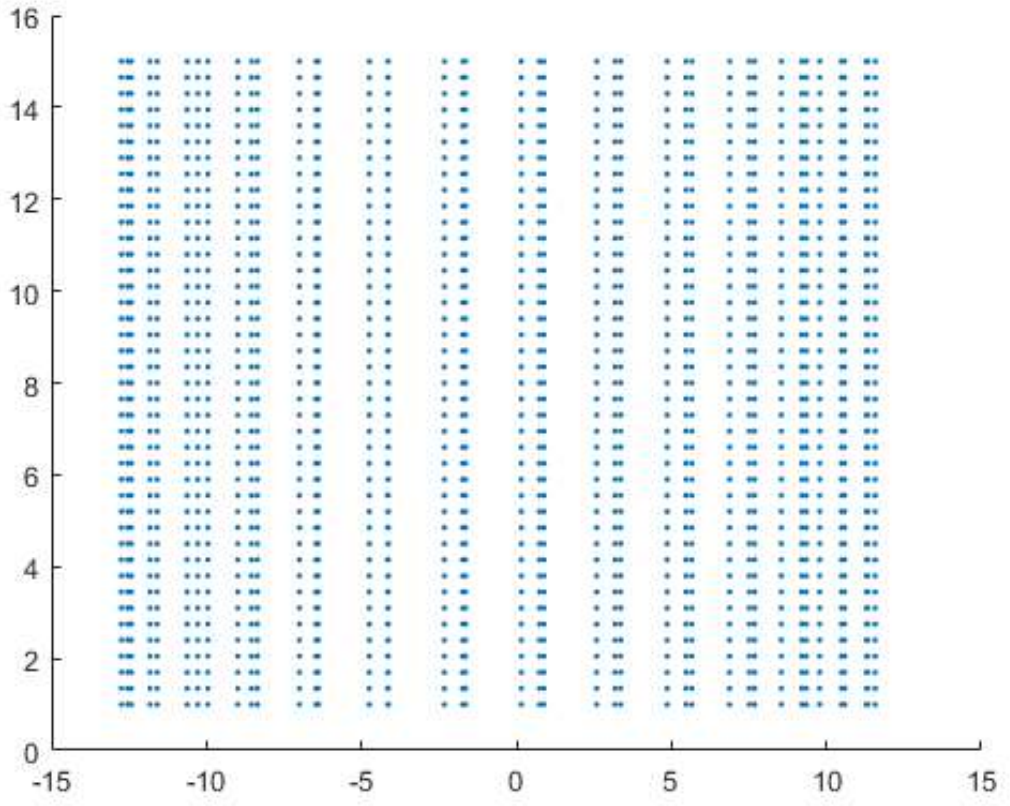
得到一组线性无关的变量。具体的求法是利用样本协方差矩阵的前r个最大的特征值对应的单位特征向量，来组成A。题目中要求投影到二维平面上，所以取最大的两个特征值。

```
% PCA
cov_x = cov(x);
[V, D] = eig(cov_x);
d = diag(D);
[sort_d, idx] = sort(d,'descend');
a1 = V(:,idx(1));
a2 = V(:,idx(2));
A = [a1,a2];
```

随后再把A作用在原来的样本上，得到样本在两个主维上的投影：

```
PCA_res = x * A;
```

最后将样本投影出来，可得：



Part II: LDA

LDA是结合了样本类别的特性，求一个使得变换后的类间离散度尽量大，类内离散度尽量小的线性变换

$$max_A tr(A^T S_b A)$$

$$s.t. A^T S_w A = I_r$$

$$S_w = \sum_{j=1}^{K} \sum_{x_i \in C_j} (x_i - m_j)(x_i - m_j)^T$$

$$S_b = \sum_{j=1}^{K} n_j (m_j - m)(m_j - m)^T$$

可以转化成求$S_w^{-1} S_b$的前r个最大的特征值所对应的单位特征向量

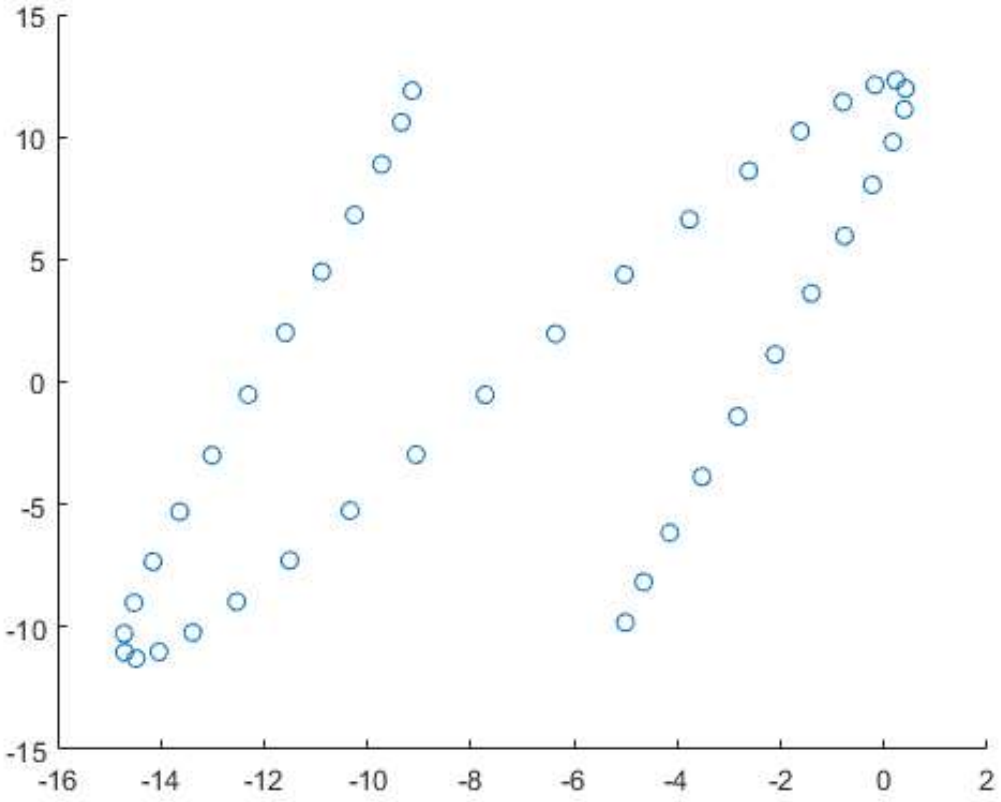
```
% LDA
x1 = x(label==1,:);
x2 = x(label==2,:);
x3 = x(label==3,:);

[n1,d] = size(x1);
[n2,d] = size(x2);
[n3,d] = size(x3);

m1 = mean(x1);
m2 = mean(x2);
```

```
m3 = mean(x3);
m = mean(x);
s_w = cov(x1)+cov(x2)+cov(x3);
s_b = n1.*(m1-m)'*(m1-m) + n2.*(m2-m)'*(m2-m) + n3.*(m3-m)'*(m3-m);
[V, D] = eig(s_b/s_w);
d = diag(D);
[sort_d,idx] = sort(d,'descend');
a1 = V(:,idx(1));
a2 = V(:,idx(2));
A = [a1, a2];
```

然后得出LDA的结果并画图：



2. 实验结果分析

可以看到，结合了类别信息的LDA能够更好得分离出与类别相关的信息，得到比较美观的图形，而PCA确实也能反应出样本的统计信息。