R05631027_楊皓文_機器學習_作業#6

Problem 1 (LDA)

(a) Write a general program to calculate the optimal direction $\boldsymbol{v}$ for a linear discriminant analysis based on three-dimensional data

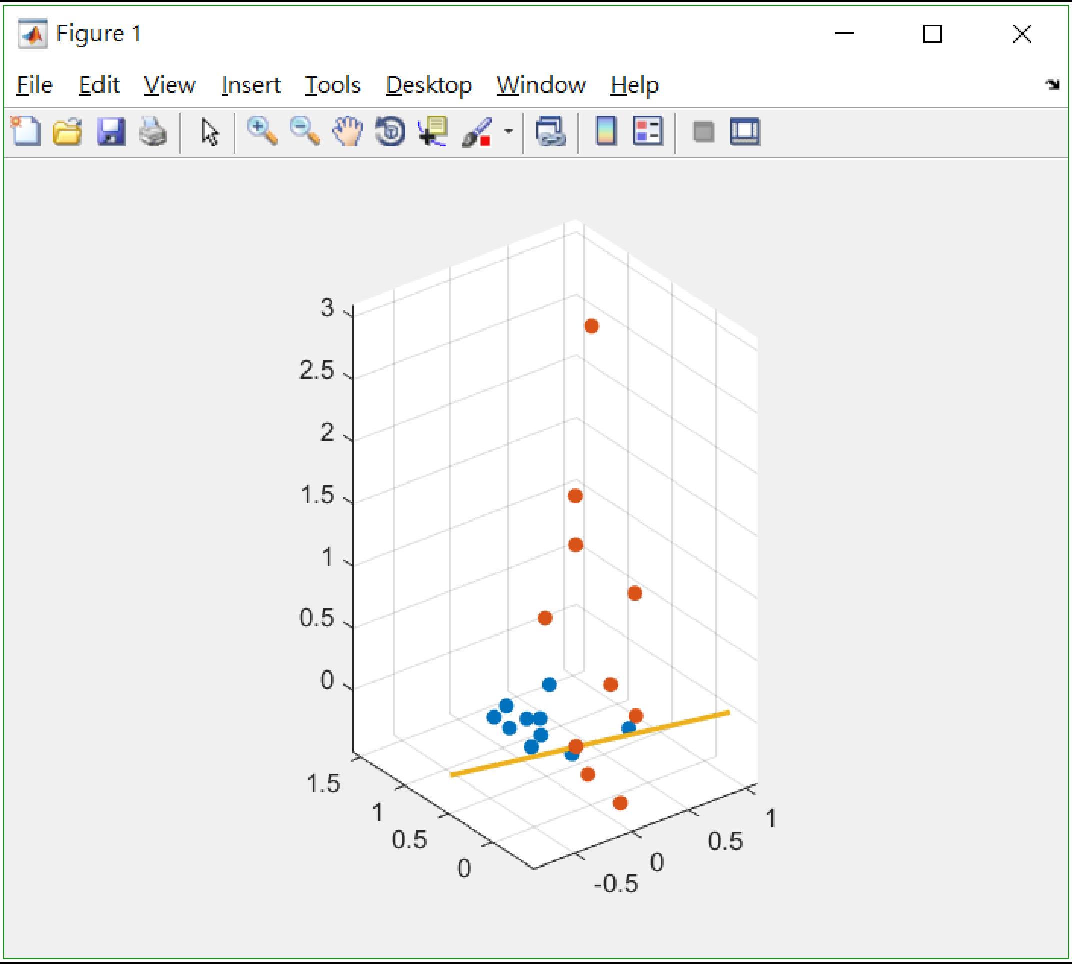
程式碼 ：

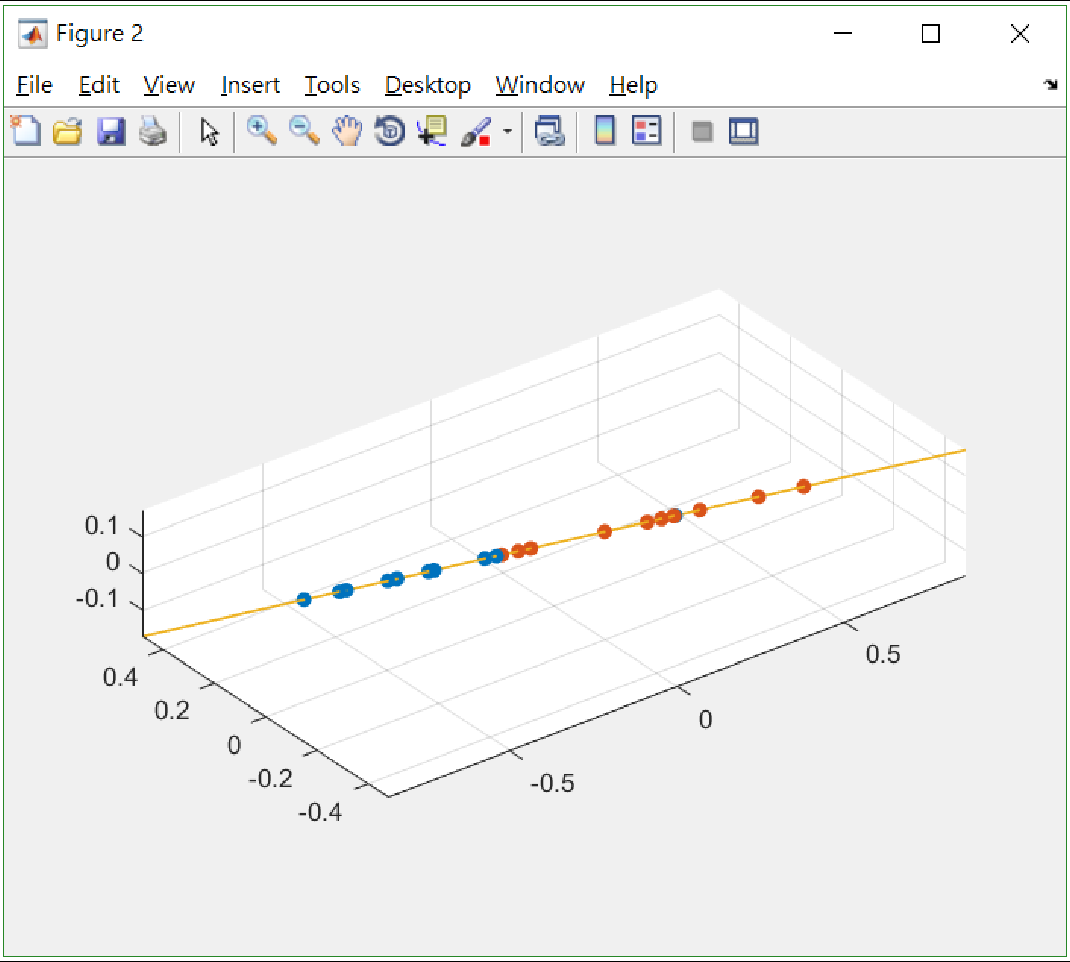
```
clear; clc; close all;
data = load('data.txt');
y1 = data(:, 1:3);
y2 = data(:, 4:6);
y = {y1, y2};

N1 = 10;
N2 = 10;
N = N1 + N2;
m1 = (1/N1)*sum(y1);
m2 = (1/N2)*sum(y2);
m = [m1; m2];
m0 = (N1/N)*m1 + (N2/N)*m2;
% Sb
Sb = 0;
for i = 1:2
    Sb = Sb + (N1/N)*(m(i, :) - m0)'*(m(i, :) - m0);
end

% Sw
Sw = 0;
for i = 1:2
    for j = 1:N1
        Sw = Sw + (N1/N)*( y{i}(j, :) - m(i, :) )'*( ( y{i}(j, :) - m(i, :)) );
    end
end

% (b) Find the optimal v for the data in the table above

[w, ~] = eig( Sb^(1/2)*Sw^(-1)*Sb^(1/2) );
v = inv(Sw)*(m1' - m2');
```

(b) Find the optimal $\boldsymbol{v}$ for the data in the table above

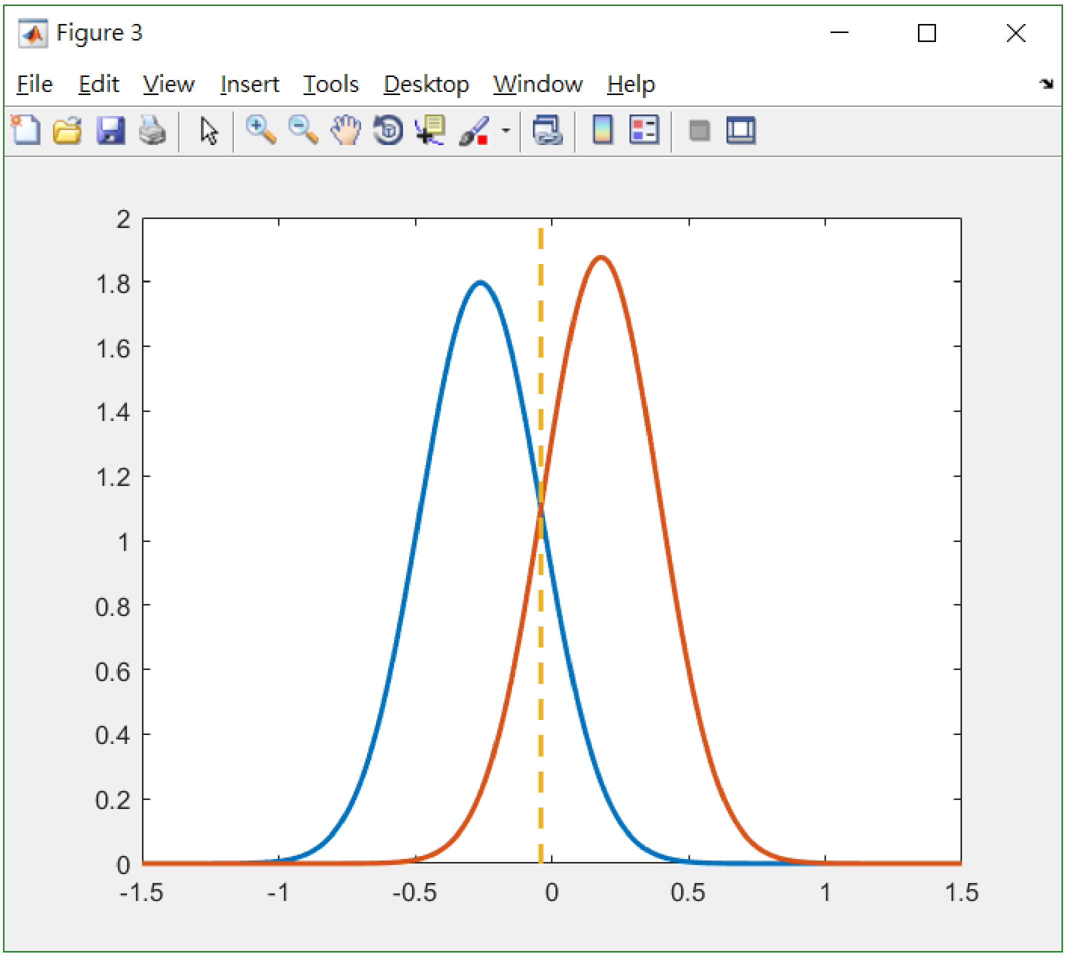
```
optimal vector  :
   -0.7665
    0.4275
   -0.1535
```

(c) Plot a line representing your optimal direction v. Mark on the line the positions of the projected points





(d) Fit each distribution with a (univariate) Gaussian, and find the resulting decision boundary



(e) What is the training error in the optimal subspace you found in (b)?

```
training_error = 20.0 %
```

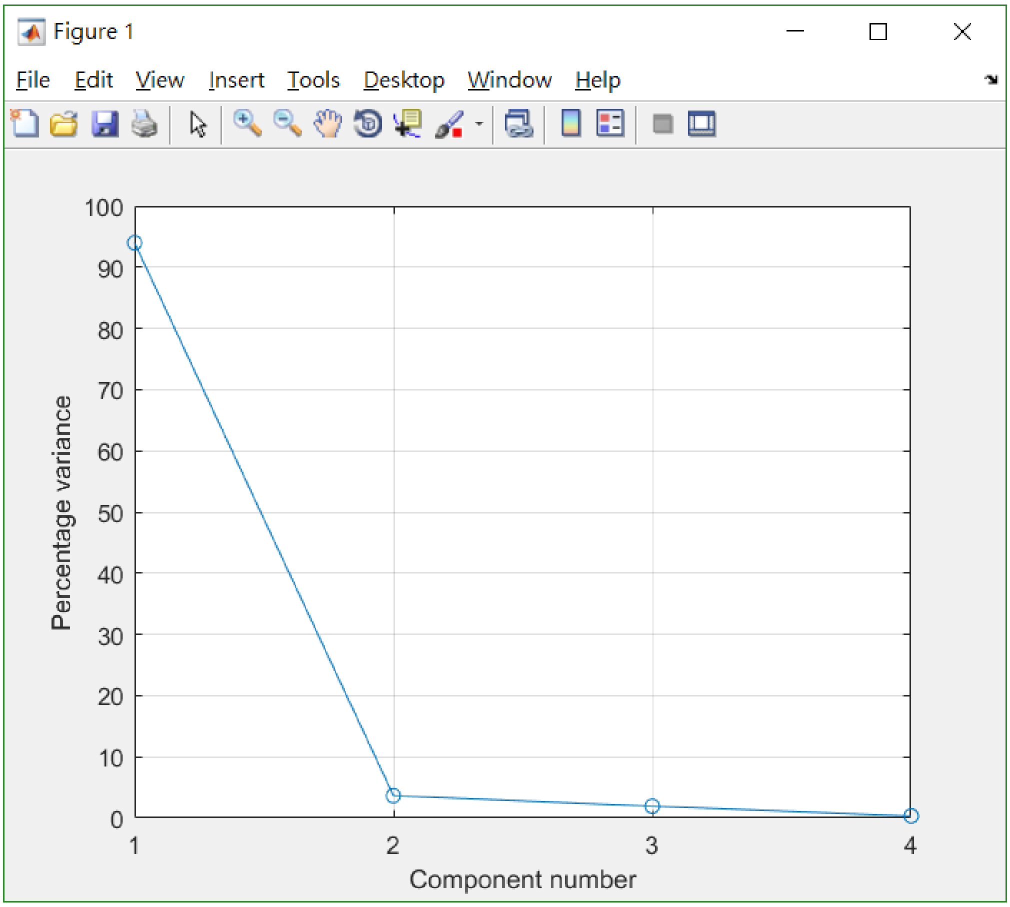
Discussion:

我們可以發現，經由 LDA 後，可以找到一條最佳的投影向量，接著將點投影到該線上，就可以作大概的分類，經由高斯分布得到 decision boundary = -0.04，用來當作分類的標準，這個資料最後的 training error 為 20%

Problem 2 (PCA and LDA)

(a) Perform PCA on the unlabeled points from the Fisher' Iris flower data set provided on ceiba

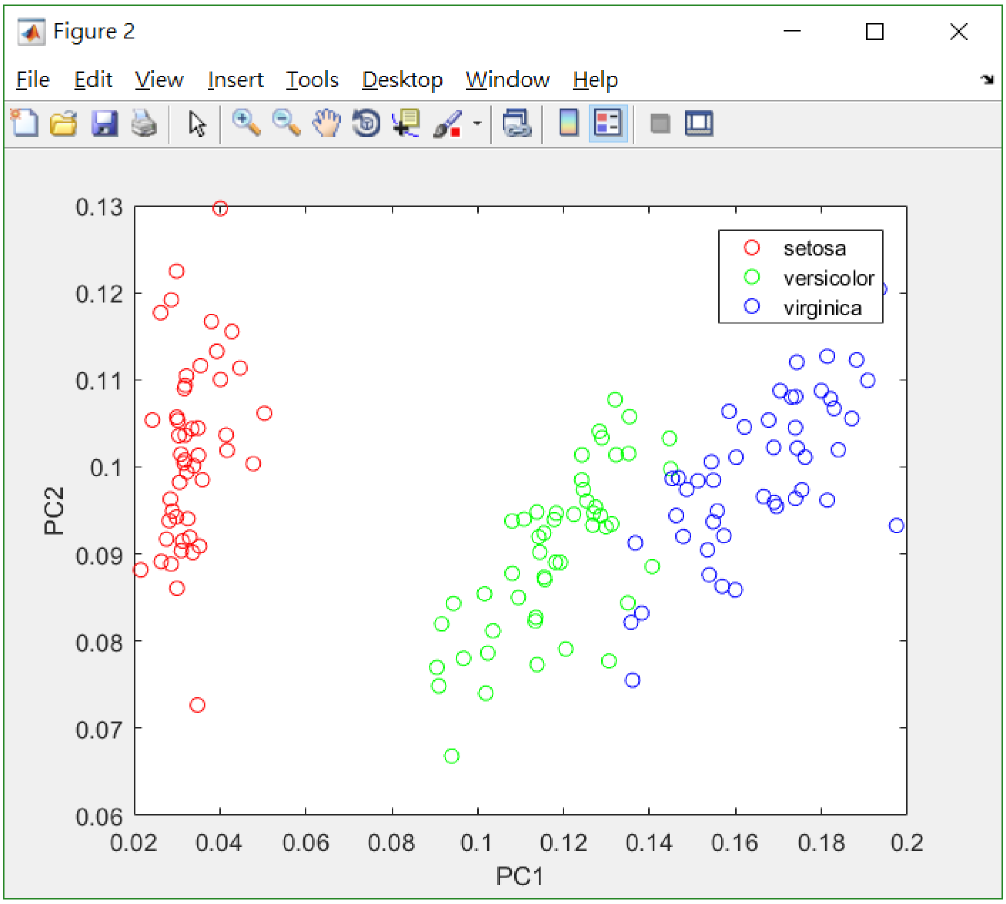
(1) List the principal components explaining 95% of the total variance in the dataset



Discussion:

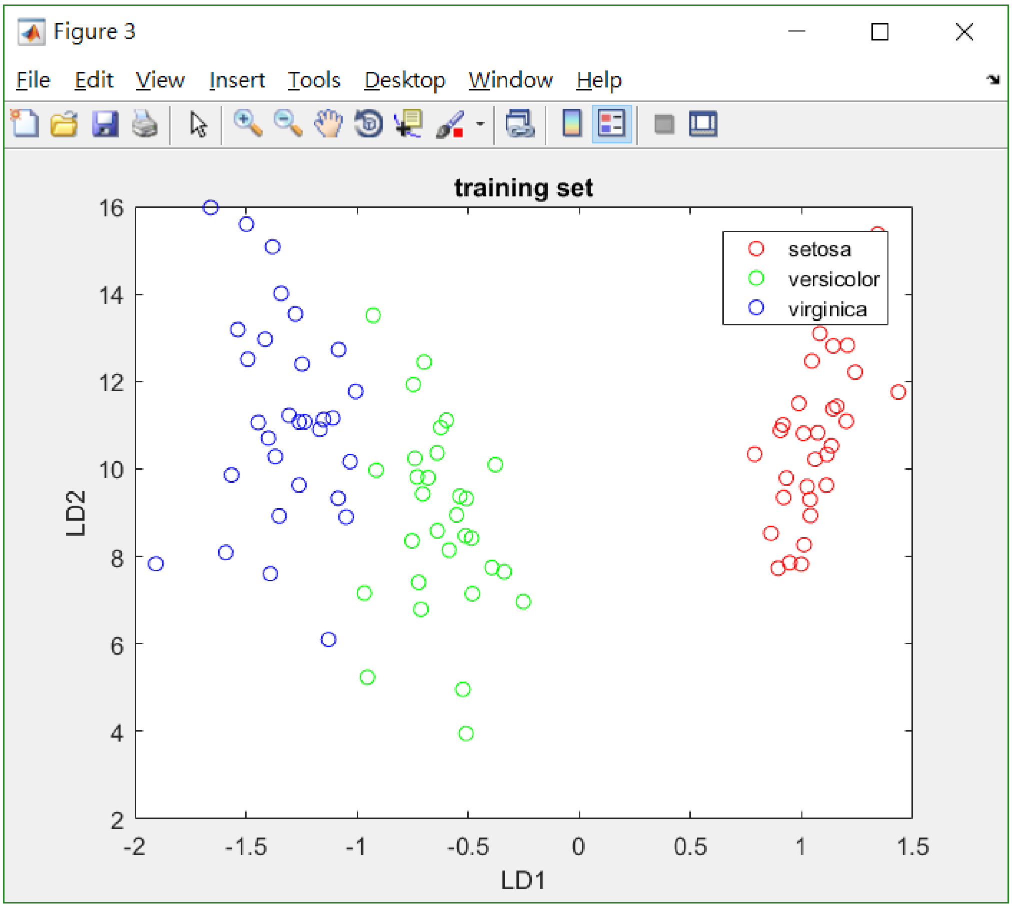
由上圖可以看出，PC1 和 PC2 即可決定超過 95%的總變異數。

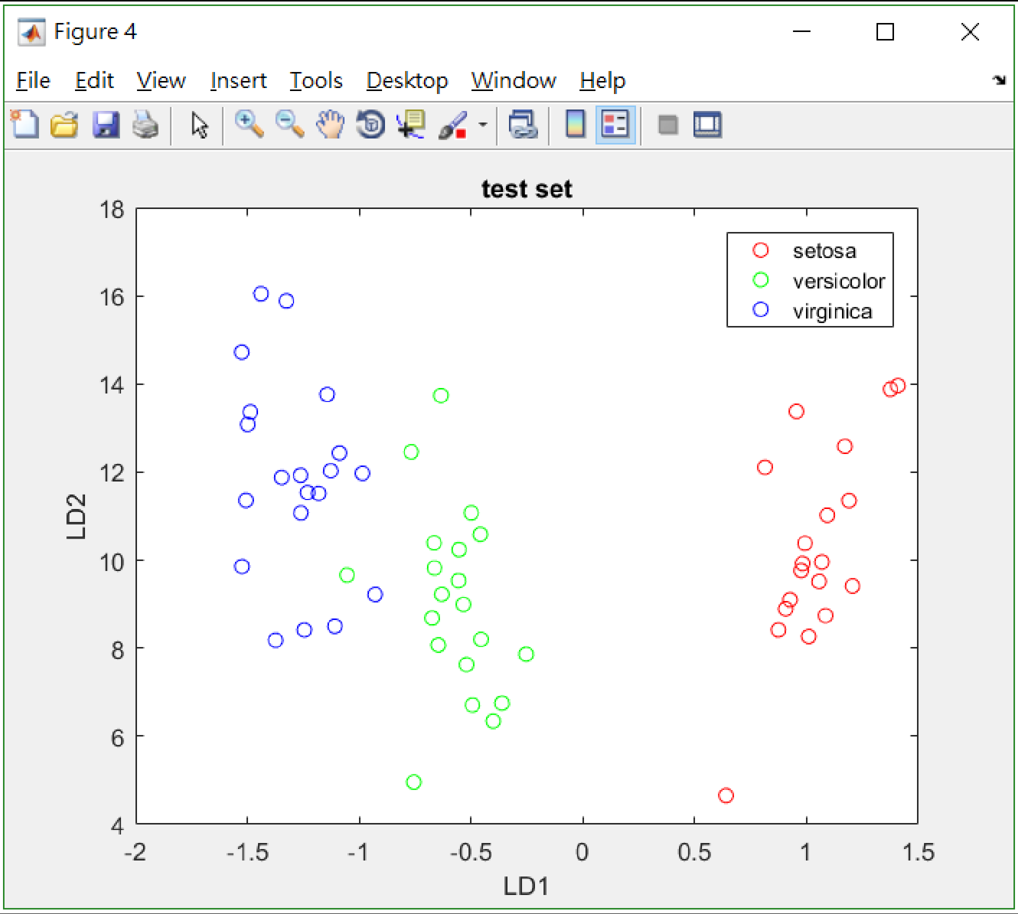
(2) Plot the data points using the first two PCs as axes, distinguishing between the classes using different color or marker



(b) Split the Fisher' Iris flower data set into a training and test set - use the first 30 points from each class for training and the last 20 points from each class for testing

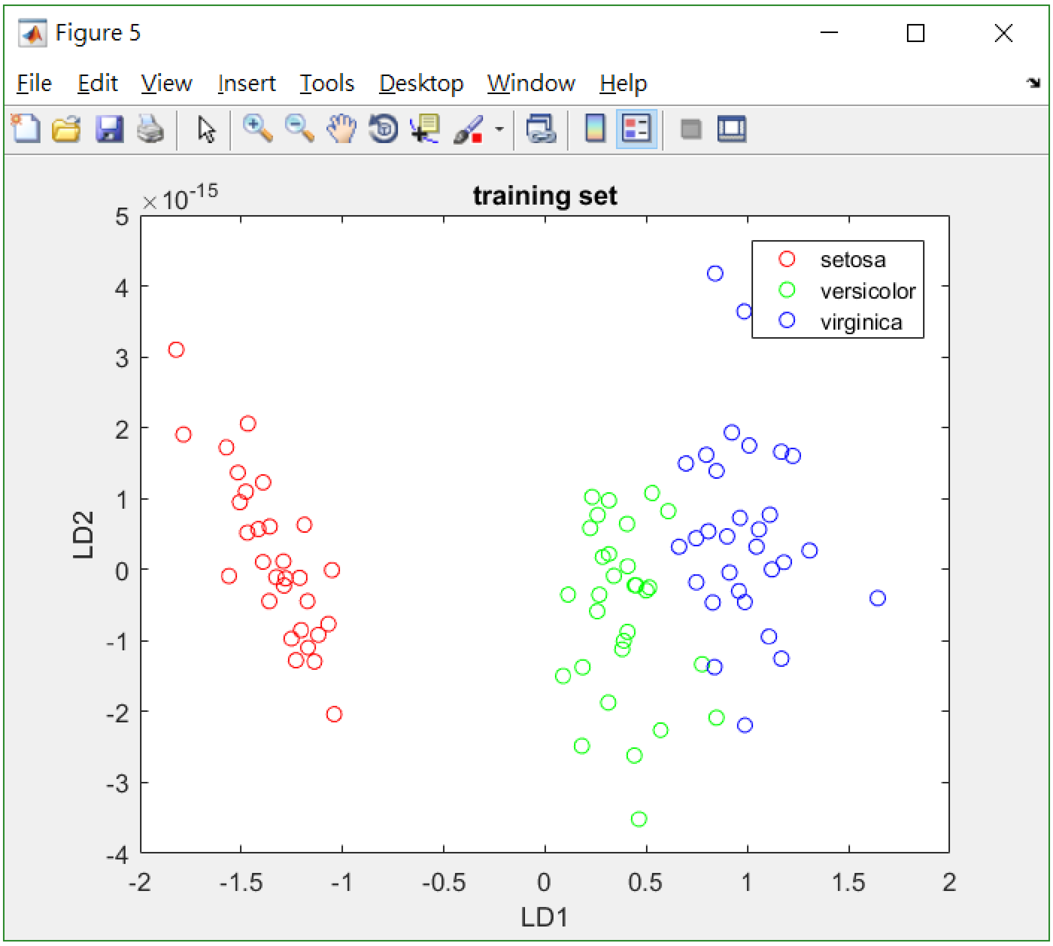
(1) Perform LDA on the original explanatory variables

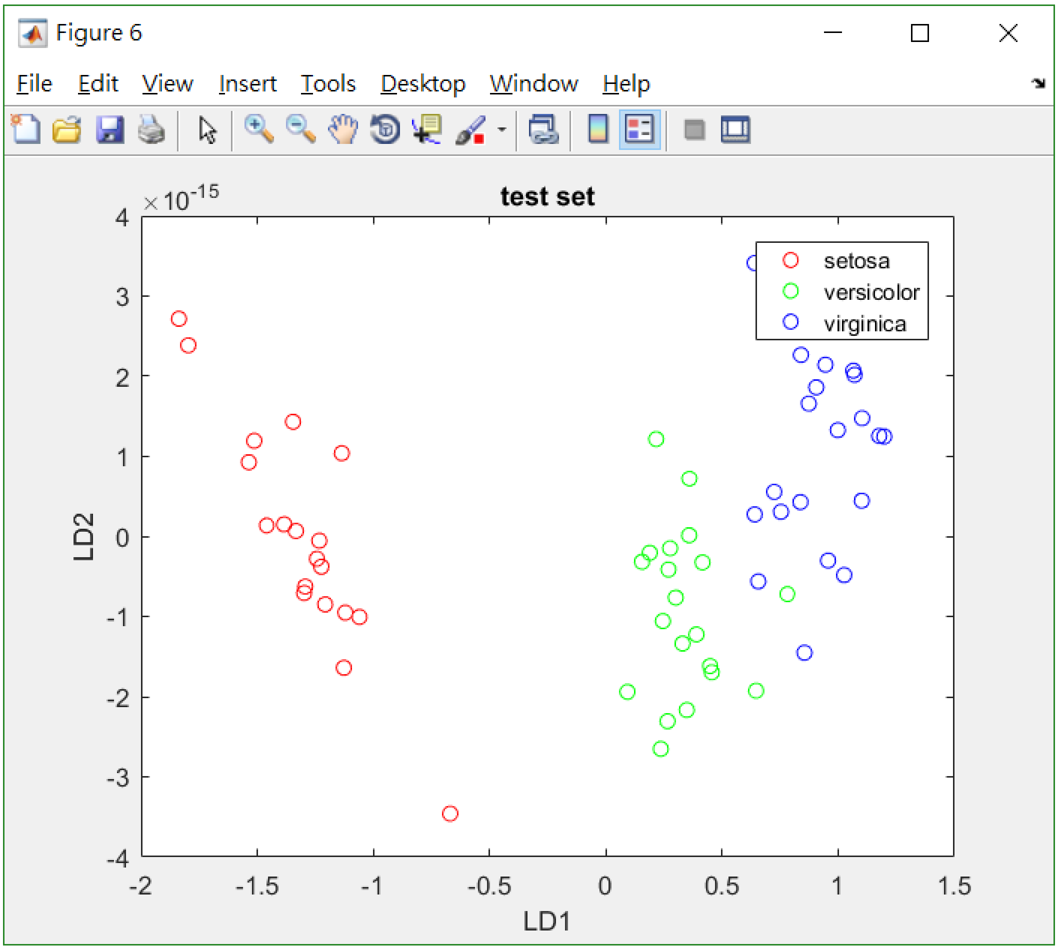




Training error : 約 1.1% ， Test error : 約 2.2%

(2) Perform LDA on the PCs obtained above





Training error：約 3.3% ， Test error：約 3.3%

(3) Compare the results of 1) and 2). Explain the discrepancy
首先，我們看第一個經由 PCA 後的散佈圖，可以發現從 PC1 的角度，versicolor 和 virginica(綠色、藍色)並沒有分得很好，應該是由於 PCA 最主要是由整體的變異數決定方向，而沒有分群標記，因此可以找出全部資料散佈的主要方向　，但並無法作分群。 因此，我們可以看到經過 PCA 再進行 LDA，不同群之間的差異反而縮小，因為已經將 PC3 和 PC4 除去，剩下 PC1 和 PC2，數據遺失掉一些資訊，反而比直接進行 LDA 更難將資料藉由投影分群。