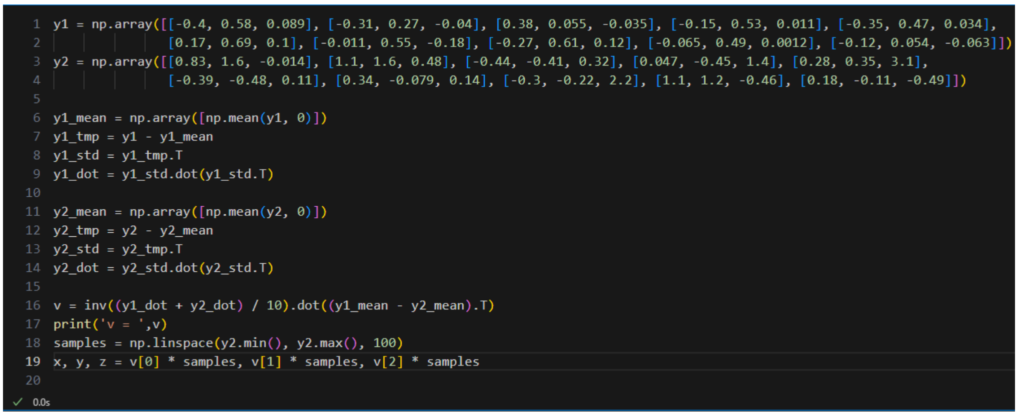
HW6 r12631001

Problem 1

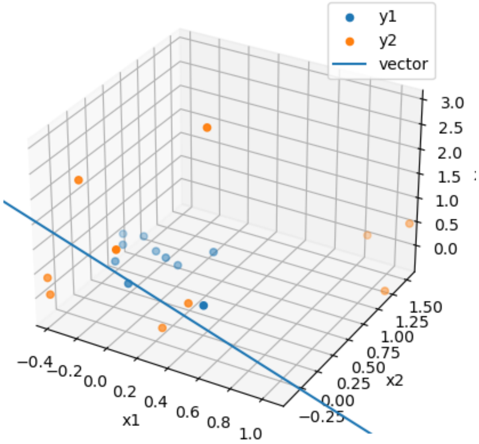
(a) Write a general program to calculate the optimal direction $\boldsymbol{v}$ for a linear discriminant analysis based on three-dimensional data.



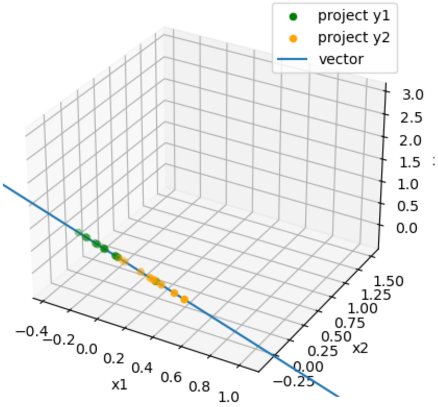
```
y1 = np.array([[-0.4, 0.58, 0.089], [-0.31, 0.27, -0.04], [0.38, 0.055, -0.035], [-0.15, 0.53, 0.011], [-0.35, 0.47, 0.034],
                [0.17, 0.69, 0.1], [-0.011, 0.55, -0.18], [-0.27, 0.61, 0.12], [-0.065, 0.49, 0.0012], [-0.12, 0.054, -0.063]])
y2 = np.array([[0.83, 1.6, -0.014], [1.1, 1.6, 0.48], [-0.44, -0.41, 0.32], [0.047, -0.45, 1.4], [0.28, 0.35, 3.1],
                [-0.39, -0.48, 0.11], [0.34, -0.079, 0.14], [-0.3, -0.22, 2.2], [1.1, 1.2, -0.46], [0.18, -0.11, -0.49]])

y1_mean = np.array([np.mean(y1, 0)])
y1_tmp = y1 - y1_mean
y1_std = y1_tmp.T
y1_dot = y1_std.dot(y1_std.T)

y2_mean = np.array([np.mean(y2, 0)])
y2_tmp = y2 - y2_mean
y2_std = y2_tmp.T
y2_dot = y2_std.dot(y2_std.T)

v = inv((y1_dot + y2_dot) / 10).dot((y1_mean - y2_mean).T)
print('v = ',v)
samples = np.linspace(y2.min(), y2.max(), 100)
x, y, z = v[0] * samples, v[1] * samples, v[2] * samples
```

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

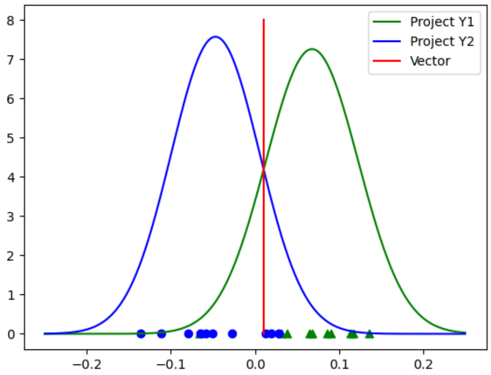
$\boldsymbol{v}$ = [[-3.83246075], [2.1374852], [-0.76736865]]



(c) Plot a line representing your optimal direction $\boldsymbol{v}$. 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)?

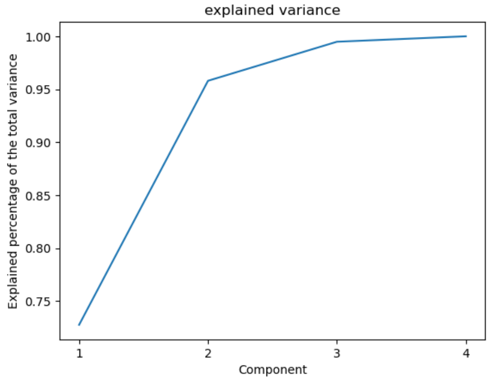
Training error = 0.2

Problem 2

(a)

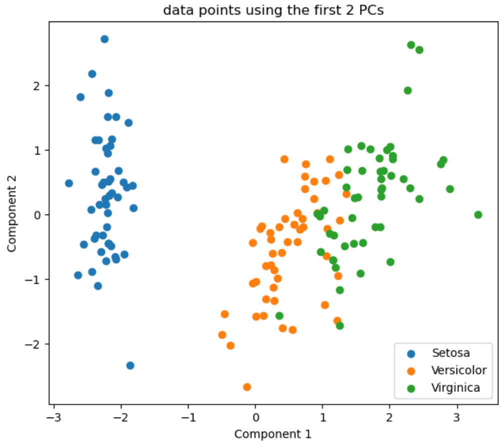
(1) List the principal components explaining 95% of the total variance in the dataset.

Explained variance= [0.728 0.958, 0.995, 1]，前兩個特徵佔了 95%以上的比例。



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

Versicolor 和 virginica 分布較為接近

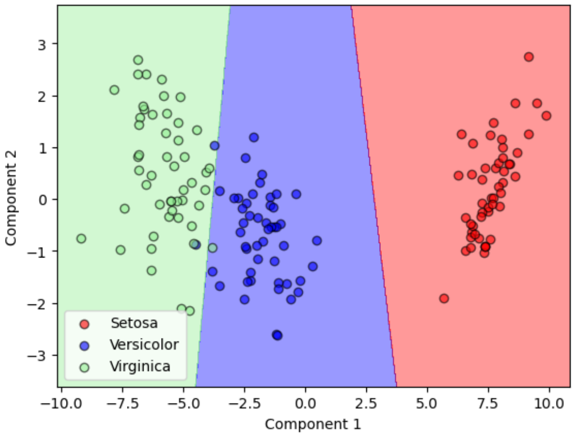


(b)

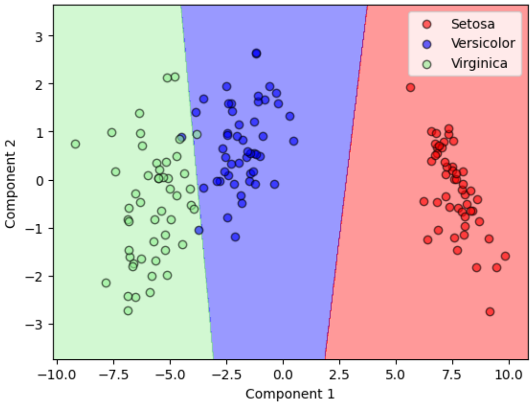
(1) Perform LDA on the original explanatory variables. The dataset should be normalized before performing the LDA. Plot the training and test sets using the first two linear discriminant axes. These can be on separate plots. Calculate the training and test errors.

Training error = 0.011111111111111112

Test error = 0.022222222222222223



(2) Perform LDA on the PCs obtained above. Plot the training and test set using the first two linear discriminant axes. These can be on separate plots. Calculate the training and test error.



(3) Compare the results of 1) and 2). Explain the discrepancy.

test 和 traning 的誤差皆相當相近，但和所繪製的結果不相同，可能是因為在 Iris dataset 上，利用 PCA+LDA 或許並沒有太大的效果，並且計算速度上也沒有提升或是降低太多