Part. 1, Coding

1. mean vector of each 2 classes

```
mean vector of class 1: [ 0.99253136 -0.99115481]  mean vector of class 2: [-0.9888012   1.00522778]
```

2. the within-class scatter matrix SW

```
Within-class scatter matrix SW: [[ 4337.38546493 -1795.55656547]
 [-1795.55656547  2834.75834886]]
```

3. the between-class scatter matrix SB

```
Between-class scatter matrix SB: [[ 3.92567873 -3.95549783]
 [-3.95549783  3.98554344]]
```

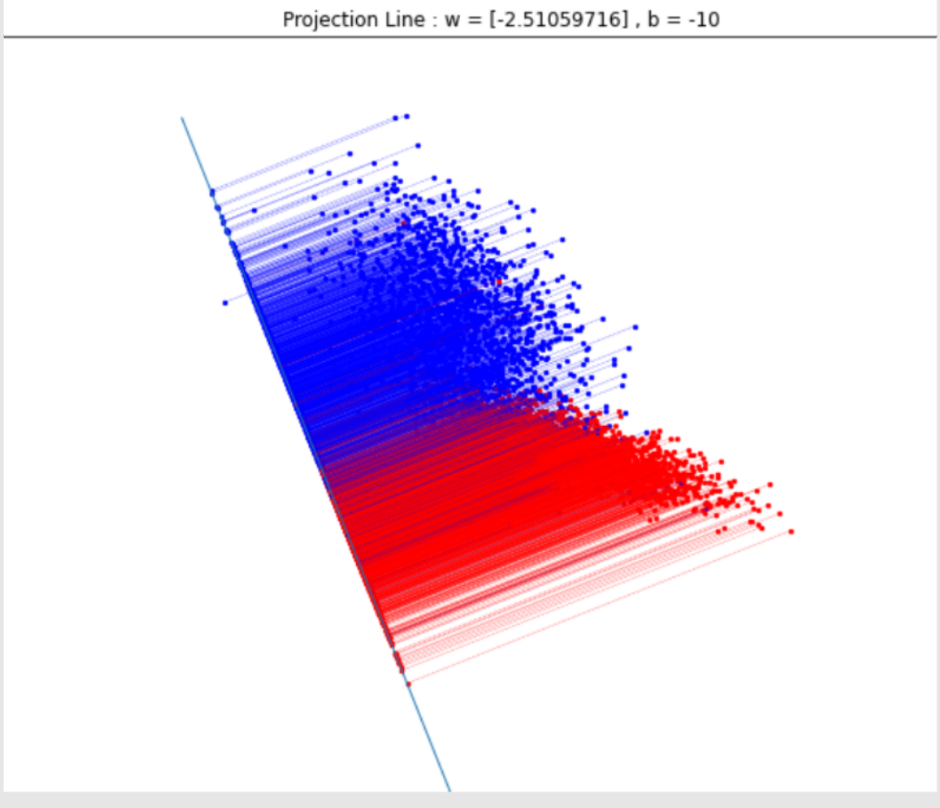
4. The Fisher's Linear Discriminant W

```
Fisher’ s linear discriminant: [-0.37003809  0.92901658]
```

5. Accuracy of KNN on testing data

```
Accuracy of test-set 0.8488 when k = 1
Accuracy of test-set 0.8488 when k = 2
Accuracy of test-set 0.8792 when k = 3
Accuracy of test-set 0.8744 when k = 4
Accuracy of test-set 0.8912 when k = 5
```

6. Plot the projection line and projections of training data poin



Part. 2, Questions (40%):

(10%) 1. What's the difference between the Principle Component Analysis and Fisher's Linear Discriminant?

FLD 和PCA都是嘗試將高維數據投影到某個軸上以進行降維，而這2種降維方式的核心思想不同：

① FLD考慮到要最小差異化組內數據+最大差異化組間數據，所以是屬於supervised的過程；PCA旨在使得降維後(不要考慮label)的數據的variance最大，不考慮原資料label，所以屬於unsupervised的過程。又因為PCA只是考慮到表示數據的方便性，所以通常只是用來做數據處理的手段，需要和其他學習模型組合使用。

② FLD降維最多能降到k-1維(k為組別總數)，而PCA則無此限制。

(10%) 2. Please explain in detail how to extend the 2-class FLD into multi-class FLD (the number of classes is greater than two).

Assume the dimension of input x is D, which is greater than K ($K>2$, K為類別總數).

Introdue a matrix W, and columns of W is the weight vectors (ex: weight vectors w_k belongs to the k-th class), and $y = W^T x$ projects each data point x to a D'-dimensional space (D'是weight vector的個數)

Define within-class covariance matrix to be

$$S_w = \sum_{k=1}^{K} S_k \text{ where } S_k = \sum_{n \in C_k} (x_n - m_k)(x_n - m_k)^T \text{ and}$$

$$m_k = \frac{1}{N_k} \sum_{n \in C_k} x_n$$

(m_k為第k類data的組內平均)

Define between-class covariance matrix to be

$$S_B = \sum_{k=1}^{K} N_k \cdot (m_k - m)(m_k - m)^T \text{ and } m = \frac{1}{N} \sum_{n=1}^{N} x_n$$

(m為所有資料點的平均)

Consider the objective function J_{cw}, we cannot directly extend the J_{cw} which project data to 1-D to the case that is multidimensional.

One criteria is:

$J(w) = Tr\{(WS_W W^T)^{-1}(WS_B W^T)\}$, this criteria also meets the concept of maxmizing $J(w)$ through $(WS_B W^T)$ to be larger (組間差異提升) or $(wS_W W^T)^{-1}$ to be lesser (組內差異下降) to make the trace larger.

*補充: why 紅筆部份:

$\because J(w) = \frac{w^T S_B w}{w^T S_W w}$ in 1-D: $\frac{\text{constant}}{\text{constant}}$

$J_{cw} = \frac{W^T S_B W}{W^T S_W W}$ in multi-D: $\frac{\text{matrix}}{\text{matrix}}$

(6%) 3. By making use of Eq (1) ~ Eq (5), show that the Fisher criterion Eq (6) can be written in the form Eq (7).

*黑筆都仍為表粗體字的m

分子=

$$(m_2-m_1)^2 = (m_2-m_1)(m_2-m_1)^T$$

$$= \left[w^T(m_2-m_1)\right]\left[w^T(m_2-m_1)\right]^T \quad \text{Eq(3)}$$

$$= w^T(m_2-m_1)(m_2-m_1)^T w$$

$$= w^T S_B w \ \#$$

分母:

$$S_1^2+S_2^2 = \sum_{n\in C_1}(y_n-m_1)^2 + \sum_{n\in C_2}(y_n-m_2)^2 \quad \text{Eq(5)}$$

$$= \sum_{n\in C_1}(w^Tx_n-w^Tm_1)(w^Tx_n-w^Tm_1)^T + \sum_{n\in C_2}(w^Tx_n-w^Tm_2)(w^Tx_n-w^Tm_2)^T$$

$$\text{Eq(1)+(4)}$$

$$= w^T\cdot\left\{\sum_{n\in C_1}(x_n-m_1)(x_n-m_1)^T + \sum_{n\in C_2}(x_n-m_2)(x_n-m_2)^T\right\}\cdot w$$

$$= w^T\cdot S_W\cdot w \ \#$$

(7%) 4. Show the derivative of the error function Eq (8) with respect to the activation for an output unit having a logistic sigmoid activation function satisfies Eq (9).

$$\frac{\partial E}{\partial a_n} = -\left\{\frac{t_n}{y_n}\cdot y_n' + \frac{1-t_n}{1-y_n}\cdot -y_n'\right\}$$

(此時Eq(8)的n為k)

$$= -\left\{\frac{t_n}{y_n} y_n(1-y_n) - \frac{1-t_n}{1-y_n} y_n(1-y_n)\right\}$$

$$= -(t_n - y_n t_n - y_n + y_n t_n)$$

$$= y_n - t_n \# \Rightarrow \frac{\partial E}{\partial a_k} = y_k - t_k \#$$

$$* \; y_n' = \sigma'(a_n) = \frac{\partial \sigma}{\partial a_n} = \frac{-(-e^{-a_n})}{(1+e^{-a_n})^2}$$

$$= \frac{1}{1+e^{-a_n}} \times \left(1 - \frac{1}{1+e^{-a_n}}\right)$$

$$= \sigma(a_n)(1-\sigma(a_n))$$

$$= y_n(1-y_n)$$

(7%) 5. Show that maximizing likelihood for a multiclass neural network model in which the network outputs have the interpretation

$y_k(x, w) = p(t_k = 1 \mid x)$

is equivalent to the minimization of the cross-entropy error function Eq (10).

This is under the assumption that we use 1-of-K coding $(0, 0, \dots 1, \dots 0)^T$ to represent each target label vector. $1 \times K$

Suppose we have training data of size N, where these training data are all i.i.d and we have K class :

第n筆training data用x_n記

$$p(T \mid w_1, w_2 \dots w_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(t_k = 1 \mid x_n)^{t_{kn}}$$

$$= \prod_{n=1}^{N} \prod_{k=1}^{K} y_k(x_n, w)^{t_{kn}}$$

$T_{N \times K}$，每列對應到這N筆資料各自的1-of-K coding結果

t_{kn}表在第n筆資料的1-of-K coding中，第k行的值(i.e.是否被label)

成第k類)

$\Rightarrow$ 想要maximize $p(T|w_1, w_2 \cdots w_k)$

等價於想要minimize $-\ln(P(T|w_1, \cdots w_k))$

$\Rightarrow$ equivalent to minimize

$$-\ln(P(T|w_1, w_2 \cdots w_k))$$

$$= -\ln\left(\prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}}\right)$$

$$= -\ln\left(\prod_{n=1}^{N} y_{n1}^{t_{n1}} \cdot y_{n2}^{t_{n2}} \cdots y_{nk}^{t_{nk}}\right)$$

$$= -\ln\left((y_{11}^{t_{11}} y_{12}^{t_{12}} \cdots y_{1K}^{t_{1K}})(y_{21}^{t_{21}} y_{22}^{t_{22}} \cdots y_{2K}^{t_{2K}})\right.$$

$$\left. \cdots (y_{N1}^{t_{N1}} y_{N2}^{t_{N2}} \cdots y_{NK}^{t_{NK}})\right)$$

$$= -\left((t_{11}\ln y_{11} + \cdots + t_{1K}\ln y_{1K}) + \cdots (t_{N1}\ln y_{N1} + \cdots + t_{NK}\ln y_{NK})\right)$$

$$= -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \cdot \ln(\underbrace{y_{nk}}_{=y_k(x_n, w)})$$

$= E(w)$