Part 1. Coding

Q1:

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

Q2:

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

Q3:

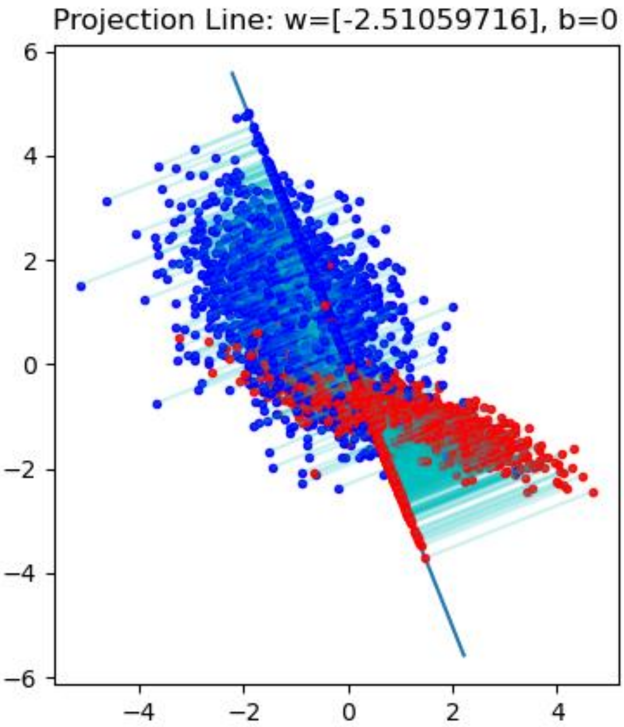
```
Between-class scatter matrix SB:
[[ 3.92567873 -3.95549783]
 [-3.95549783  3.98554344]]
```

Q4:

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

Q5:

```
Accuracy for K=1: 0.8488
Accuracy for K=2: 0.8704
Accuracy for K=3: 0.8792
Accuracy for K=4: 0.8824
Accuracy for K=5: 0.8912
```



Q6:

Part 2. Questions

Q1: Principle Component Analysis only try to maximize the variance between all data points, while Fisher's Linear Discriminant maximizes the variance between classes and minimizes the variance within class at the same time.

Q2:

First, calculate the mean vector for each class $m_k = \frac{1}{N_k}\sum_{n\in C_k} x_n$.

Within-class covariance matrix Sw changes from sum of two classes to all the classes: $S_w = \sum_{n\in C_1}(x_n - m_1)(x_n - m_1)^T + \sum_{n\in C_2}(x_n - m_2)(x_n - m_2)^T$
$\Rightarrow S_w = \sum_{k=1}^{K}\sum_{n\in C_k}(x_n - m_k)(x_n - m_k)^T$.

Between-class covariance matrix extended from $S_B = (m_2 - m_1)(m_2 - m_1)^T$ to $S_B = \sum_{k=1}^{K} N_k(m_k - m)(m_k - m)^T$,where $m = \frac{1}{N}\sum_{n=1}^{N} x_n$.

Fisher's linear discriminants W equals to the eigenvectors corresponding to the most D' eigenvalues when we want to project to a D'-dimensional space. $W = max_{D'}\left(eig(S_W^{-1}S_B)\right)$

Q3:

$$J(w) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

$$= \frac{(\mathbf{w^T m_2} - \mathbf{w^T m_1})^2}{\sum_{n\in C_1}(y_n - m_1)^2 + \sum_{n\in C_2}(y_n - m_2)^2}$$ by eq(3), eq(5)

$$= \frac{\left(\mathbf{w^T(m_2 - m_1)}\right)^2}{\sum_{n\in C_1}(\mathbf{w^T(x_n - m_1)})^2 + \sum_{n\in C_2}(\mathbf{w^T(x_n - m_2)})^2}$$ by eq(1), eq(4)

$$= \frac{\mathbf{w^T(m_2 - m_1)^2 w}}{\sum_{n\in C_1}(\mathbf{w^T(x_n - m_1)^2 w}) + \sum_{n\in C_2}(\mathbf{w^T(x_n - m_2)^2 w})}$$

$$= \frac{\mathbf{w^T S_B w}}{\mathbf{w^T}(\sum_{n\in C_1}(\mathbf{x_n - m_1})^2 + \sum_{n\in C_2}(\mathbf{x_n - m_2})^2)\mathbf{w}}$$ by Sb definition

$$= \frac{\mathbf{w^T S_B w}}{\mathbf{w^T S_W w}} \blacksquare$$ by Sw definition

Q4:

$$E(\mathbf{w}) = -\sum_{n=1}^{N} t_n \ln y_n + (1 - t_n)\ln(1 - y_n)$$

Because when n not equal to k, the derivative is 0, so:

$$\frac{\partial E}{\partial a_k} = -\frac{\partial}{\partial a_k} t_k \ln y_k + (1 - t_k)\ln(1 - y_k)$$

$$= -\frac{t_k}{y_k} y_k(1 - y_k) + \frac{1 - t_k}{1 - y_k} y_k(1 - y_k)$$

$$= -t_k(1 - y_k) + (1 - t_k)y_k$$

$$= -t_k + t_k y_k + y_k - t_k y_k$$

$$= y_k - t_k \blacksquare$$

Q5:

$$L = p(\mathbf{t}|\mathbf{x}) = \prod_{k=1}^{K} p(t_k = 1|\mathbf{x})^{t_k}$$

Take negative log on L:

$$E(\mathbf{w}) := -\ln \prod_{n=1}^{N} p(\mathbf{t}_n|\mathbf{x}_n)$$

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

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

Maximizing L also minimizing -log L, which is equivalent to minimizing the cross entropy error E(w).