Intro. to Machine Learning

Part 1:

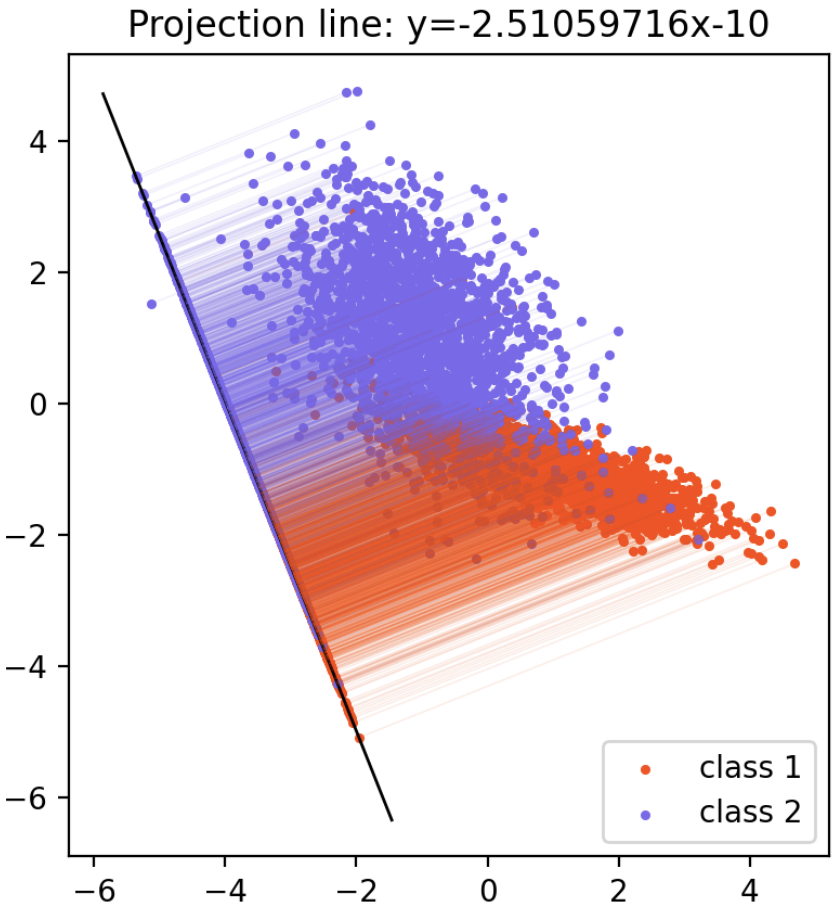
```
mean vector of class 1: [ 0.99253136 -0.99115481] mean vector of class 2: [-0.9888012   1.00522778]


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


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


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


K=1 accuracy rate: 0.8488
K=2 accuracy rate: 0.8704
K=3 accuracy rate: 0.8792
K=4 accuracy rate: 0.8824
K=5 accuracy rate: 0.8912
```



Part 2:

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

Principal component analysis is an unsupervised dimensionality reduction method, it ignores the original class label, capturing the direction of maximum variation of the whole dataset; Linear Discriminant is a supervised reduction method that will consider the label while finding the direction maximizing the separability between groups.

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 K= class number, D= dimension of input, D'= dimensional space that weight vector will project to.
The within-class and between-class covariance matrix for multiclass should be:

$$\mathbf{S}_{\mathrm{W}} = \sum_{k=1}^{K} \mathbf{S}_k \quad \mathbf{S}_{\mathrm{B}} = \sum_{k=1}^{K} N_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^{\mathrm{T}}$$

When D'=1, the objective function:

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}} \mathbf{S}_{\mathrm{B}} \mathbf{w}}{\mathbf{w}^{\mathrm{T}} \mathbf{S}_{\mathrm{W}} \mathbf{w}}$$

When D'>1, the objective function:

$$J(\mathbf{w}) = \mathrm{Tr}\left\{ (\mathbf{W}\mathbf{S}_{\mathrm{W}}\mathbf{W}^{\mathrm{T}})^{-1} (\mathbf{W}\mathbf{S}_{\mathrm{B}}\mathbf{W}^{\mathrm{T}}) \right\}$$

The optimal weight vector is the eigenvector of $S_W^{-1}S_B$ corresponding to the largest eigenvalue, and the dimension of the projected space is at most K-1.

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

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x} \quad \text{Eq (1)}$$

$$\mathbf{m}_1 = \frac{1}{N_1}\sum_{n \in \mathcal{C}_1} \mathbf{x}_n \qquad \mathbf{m}_2 = \frac{1}{N_2}\sum_{n \in \mathcal{C}_2} \mathbf{x}_n \quad \text{Eq (2)}$$

$$m_2 - m_1 = \mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1) \quad \text{Eq (3)}$$

$$m_k = \mathbf{w}^{\mathrm{T}}\mathbf{m}_k \quad \text{Eq (4)}$$

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2 \quad \text{Eq (5)}$$

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \quad \text{Eq (6)}$$

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}} \quad \text{Eq (7)}$$

$$J(w) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} = \frac{\left\| w^T (m_2 - m_1) \right\|^2}{\sum_{n \in C_1} (w^T x_n - m_1)^2 + \sum_{n \in C_2} (w^T x_n - m_2)^2}$$

$$\text{numerator} = \left[w^T (m_2 - m_1) \right] \left[w^T (m_2 - m_1) \right]^T = w^T S_B w$$

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

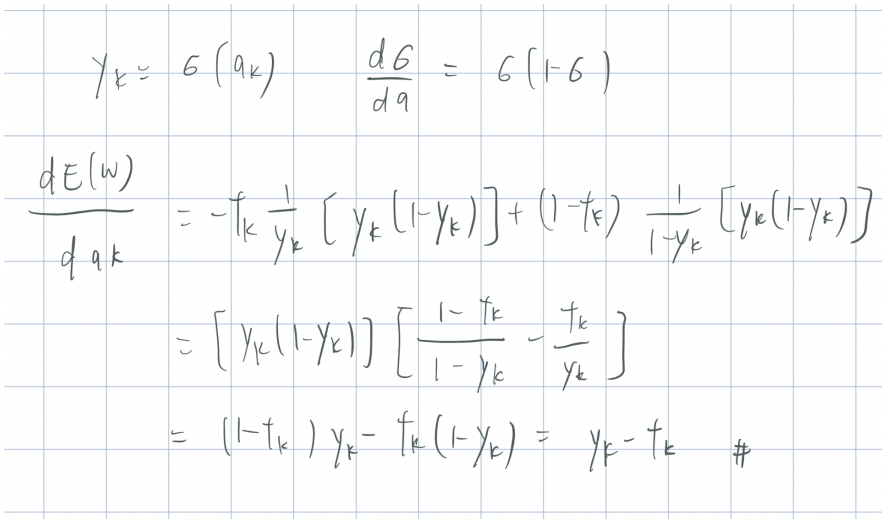
$$= w^T S_{w_1} w + w^T S_{w_2} w = w^T S_w w$$

$$J(w) = \frac{w^T S_B w}{w^T S_w w}$$

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

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\} \quad \text{Eq (8)}$$

$$\frac{\partial E}{\partial a_k} = y_k - t_k \quad \text{Eq (9)}$$



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).

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn} \ln y_k(\mathbf{x}_n, \mathbf{w}) \quad \text{Eq (10)}$$