NYCU Introduction to Machine Learning,Homework 2

陳存佩 109550171

Part. 1, Coding (60%):

1. (5%) Compute the mean vectors m_i (i=1, 2) of each 2 classes on training data

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

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

2. (5%) Compute the within-class scatter matrix S_W on training data

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

3. (5%) Compute the between-class scatter matrix S_B on training data

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

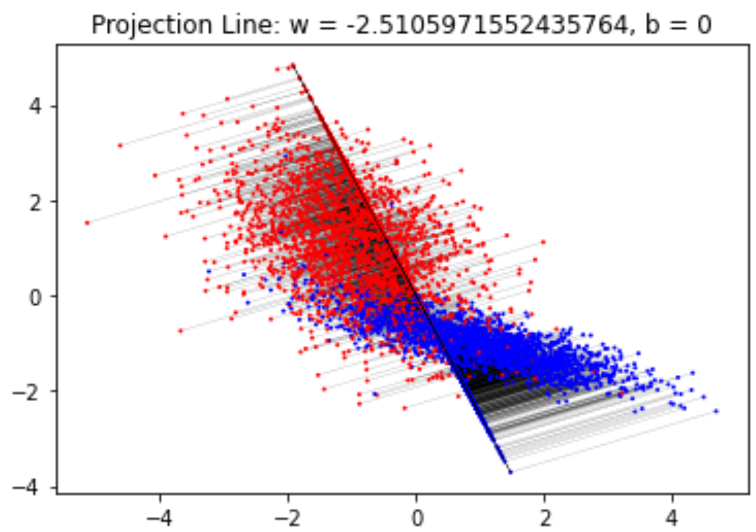
4. (5%) Compute the Fisher's linear discriminant $_w$ on training data

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

5. (20%) Project the testing data

```
For K=1, Accuracy of test-set 0.8496
For K=2, Accuracy of test-set 0.88
For K=3, Accuracy of test-set 0.8832
For K=4, Accuracy of test-set 0.9
For K=5, Accuracy of test-set 0.8864
```

6. (20%) Plot



Part. 2, Questions (40%):

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

Principle Component Analysis	Fisher's Linear Discriminant
Unsupervised dimensionality reduction	**Supervised** dimensionality reduction
Goal: Find the direction of maximum **variation** in the data set.	Goal: Find a feature subspace that maximizes the **variance/separability** between different groups and minimizes the variance within the class.

(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).

Sw 公式更新如下：

$$S_W = \sum_{k=1}^{K} S_k,\ \text{where}\ S_k = \sum_{n\in C_k} \underline{(x_n - m_k)(x_n - m_k)'},\ m_k = \frac{1}{N_k}\sum_{n\in C_k} x_n$$

共有K class,所以相加

概念和2-class相同

Sb 公式更新如下：

Goal:每個class都和平均拉越遠越好

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

multi-class才有
每類別個數不同
數量越多,權重越大

所有資料平均

用更新後的 Sb 和 Sw 去計算 J(w)，有條件的最佳化可用 Lagrangian function 去處理

(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).

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

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

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

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

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

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

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

$$S_k^2 = \sum_{n \in C_k} (y_n - m_k)^2$$
$$= \sum_{n \in C_k} (W^T x_n - W^T m_k)^2$$
$$= \sum_{n \in C_k} W^T (x_n - m_k)(x_n - m_k)^T W$$
$$= W^T S_k W$$
$$S_1^2 + S_2^2 = W^T S_1 W + W^T S_2 W = W^T S_W W \text{ — ①}$$

$$m_2 - m_1 = W^T (m_2 - m_1)$$
$$(m_2 - m_1)^2 = W^T (m_2 - m_1)(m_2 - m_1)^T W$$
$$= W^T S_B W \text{ — ②}$$

from ①, ②:

$$J(W) = \frac{(m_2 - m_1)^2}{S_1^2 + S_2^2} = \frac{W^T S_B W}{W^T S_W W}$$

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

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

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

$y_k = \sigma(a_k)$, $\sigma(\cdot)$ represents the logistic sigmoid function

$$\frac{\partial \sigma}{\partial a} = \sigma(1 - \sigma)$$

$$\frac{\partial E}{\partial a_k} = -t_k \frac{1}{y_k}[y_k(1 - y_k)] + (1 - t_k)\frac{1}{1 - y_k}[y_k(1 - y_k)]$$
$$= \left[\frac{1 - t_k}{1 - y_k} - \frac{t_k}{y_k}\right][y_k(1 - y_k)]$$
$$= (1 - t_k) y_k - t_k (1 - y_k)$$
$$= y_k - t_k \ \#$$

(7%) 5. Show that maximizing likelihood for a multiclass neural network model in which the network outputs have the interpretation $\boldsymbol{y_k(x, w) = p(t_k = 1 \mid x)}$ is equivalent to the minimization of the cross-entropy error function Eq (10).

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

2 class:

$$p(t|w) = \prod_{n=1}^{N} y_n^{t_n} (1-y_n)^{1-t_n}$$

$$t = (t_1, t_2, \ldots, t_N)^T \text{ and } y_n = p(C_1|\phi_n)$$

extend to multi-class:

$$p(\underset{n \times K}{\underset{\downarrow}{T}}|w_1, w_2, \ldots, w_K) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(C_k|\phi_n)^{t_{kn}} = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{kn}}$$

$$E(W) = -\ln p(T|w_1, w_2, \ldots, w_k) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn} \ln y_k(x_n, w)$$