NYCU Introduction to Machine Learning, Homework 2

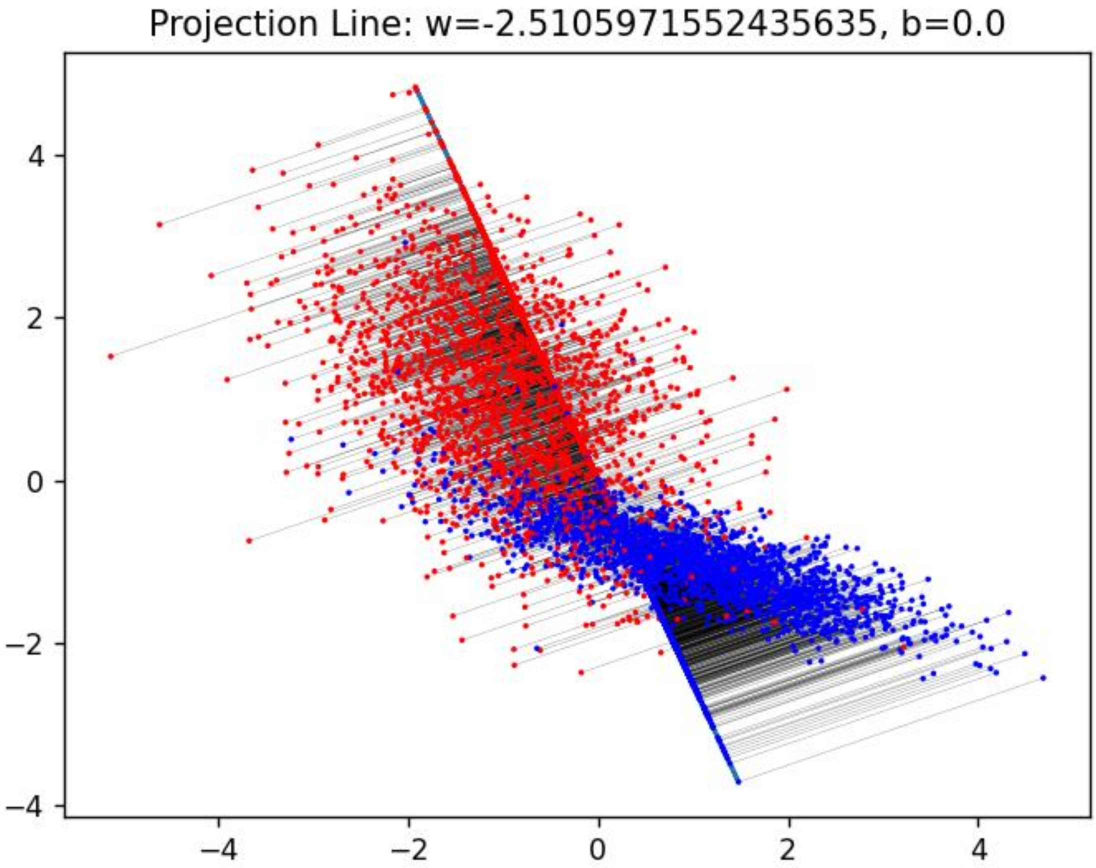
109550136 邱弘竣

Part. 1, Coding (60%):

1. (5%) Compute the mean vectors m_i (i=1, 2) of each 2 classes on <u>training data</u>
2. (5%) Compute the within-class scatter matrix $S_W$on <u>training data</u>
3. (5%) Compute the between-class scatter matrix S_B on <u>training data</u>
4. (5%) Compute the Fisher’ s linear discriminant w on <u>training data</u>
5. (20%) Project the <u>testing data</u> by Fisher’ s linear discriminant to get the class prediction by K-Nearest-Neighbor rule and report the accuracy score on <u>testing data</u> with K values from 1 to 5 (you should get accuracy over 0.88)

```
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]
For K=1, Accuracy of test-set 0.8648
For K=2, Accuracy of test-set 0.8808
For K=3, Accuracy of test-set 0.88
For K=4, Accuracy of test-set 0.8896
For K=5, Accuracy of test-set 0.9
```

6. (20%) Plot the **1) best projection line** on the <u>training data</u> and <u>show the slope and intercept on the title</u> *(you can choose any value of **intercept** for better visualization)*
   2) colorize the data with each class **3)** project all data points on your projection line. Your result should look like the below image (This image is for reference, not the answer)



Part. 2, Questions (40%):

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

LDA is supervised and attempts to find a feature subspace that maximizes class separability. PCA i s unsupervised and is a technique that finds the directions of maximal variance.

(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).

2.

2 class

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

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

$$S_W = \sum_{n \in C_1} (x_n - m_1)(x_n - m_1)^T + \sum_{n \in C_2} (x_n - m_2)(x_n - m_2)^T$$

$$\frac{\partial}{\partial w} J(w) = 0 \Rightarrow w \propto S_W^{-1}(m_2 - m_1)$$

K classes

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

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

$$J(W) = \frac{W^T S_B W}{W^T S_W W}$$

the optimal w is the eigenvector of $S_W^{-1} S_B$ that corresponds to the largest eigenvalue

(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).

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

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

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

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

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

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

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

3.

$$s_k^2 = \sum_{n\in C_k} (y_n - m_k)^2$$
$$= \sum_{n\in C_k} (w^T x_n - w^T m_k)^2$$
$$= \sum_{n\in C_k} w^T (x_n - m_k)(x_n - m_k)^T w$$
$$= w^T S_k w$$
$$s_1^2 + s_2^2 = w^T S_1 w + w^T S_2 w = w^T S_W w \quad \text{— ①}$$

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

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 a_k for an output unit having a logistic sigmoid activation function satisfies Eq (9).

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

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

4.

$$\frac{\partial E}{\partial a_k} = \frac{\partial E}{\partial y}\frac{\partial y}{\partial a}$$
$$= \left(\frac{-t_k}{y_k} + \frac{1 - t_k}{1 - y_k}\right) y_k (1 - y_k)$$
$$= y_k - t_k$$

#

$$\frac{\partial E}{\partial y} = \frac{\partial}{\partial y}(t_n \ln y_n) + \frac{\partial}{\partial y}\left((1 - t_n)\ln(1 - y_n)\right)$$
$$= \frac{-t_k}{y_k} + \frac{1 - t_k}{1 - y_k}$$

$$\frac{\partial y}{\partial a} = \frac{\partial}{\partial a}\left(\frac{1}{1 + e^{-a}}\right) = \frac{-(-e^{-a})}{(1 + e^{-a})^2}$$
$$= \frac{1 + e^{-a}}{(1 + e^{-a})^2} - \frac{1}{(1 + e^{-a})^2}$$
$$= \frac{1}{1 + e^{-a}}\left(1 - \frac{1}{1 + e^{-a}}\right)$$
$$= y(1 - y)$$

(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).

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

5.

2-class

$$p(t|w) = \prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}. \quad \text{where } t=(t_1, \cdots, t_n)^T,\ y_n = p(C_1|\phi_n)$$

multiclass

$$p(T|w_1, \cdots w_K) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(C_k|\phi_n)^{t_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

$$E(w_1 \cdots w_K) = -\ln p(T|w_1, \cdots w_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}$$