NYCU Introduction to Machine Learning,Homework 2

109550164 徐聖哲

Part. 2, Questions (40%):.

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

Ans: Both Principle Component Analysis and Fisher's Linear Discriminant are used to find the best linear combination to explain the data. However Fisher's Linear Discriminant is supervised learning, it uses the label of train data to check the new projected dot is classified correct or not. Principle Component Analysis is unsuupervisied learning, which focus on maximum variance between data.

(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).

Ans: First SW is the sum of all Sk for k = k class, Then SB will multiply with number of dot in the class in each between-class variance. Finally, in order to find w , we use lagrangian function to minimize within-class-covariance , thus w is the eigenvector of (Sw)^-1 * Sb that corresponds to the largest eigenvalue.

2.

① SW

k=2: $S_W = \sum_{n \in C_1} (x_n - m_1)(x_n - m_1)^T + \sum_{n \in C_2} (x_n - m_2)(x_n - m_2)^T$

k=n: $S_W = \sum_{k=1}^{k} S_k$, $S_k = \sum_{n \in C_k} (x_n - m_k)(x_n - m_k)^T$

② SB

k=2: $S_B = (m_2 - m_1)(m_2 - m_1)^T$

k=n: $S_B = \sum_{k=1}^{k} N_k (m_k - m)(m_k - m)^T$

③ W

k=2: $W = S_W^{-1}(m_2 - m_1)$

k=n: use lagrangian function

$L_P = -\frac{1}{2} W^T S_B W + \frac{1}{2}\lambda (W^T S_W W - 1)$

$\Rightarrow S_B W = \lambda S_W W$

$\Rightarrow S_W^{-1} S_B = \lambda W$

$\Rightarrow$ select optimal W

W = eigenvectors of $S_W^{-1} S_B$ with 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. $J(w) = \dfrac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$

$$J_1 = (m_2 - m_1)^2 = (w^T(m_2 - m_1))^2$$

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

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

$$= w^T S_b w$$

$$J_2 = s_1^2 + s_2^2$$

$$= \sum_{n \in C_1}(y_n - m_1)^2 + \sum_{n \in C_2}(y_n - m_2)^2$$

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

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

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

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

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

(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).

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

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

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

4. $y = \sigma(a) = \dfrac{1}{1+e^{-a}}$

☆ $\dfrac{dE}{da_k} = \dfrac{dE}{dy} \cdot \dfrac{dy}{da_k}$

$$\frac{dE}{dy} = \frac{d}{dy}\left(-t_n \ln y_n + (1-t_n)\ln(1-y_n)\right)$$

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

$$\frac{dy}{da_k} = \frac{d}{da_k}\left(\frac{1}{1+e^{-a_k}}\right) = \frac{-(-e^{-a})}{(1+e^{-a_k})^2} = \frac{e^{-a}}{(1+e^{-a})^2} = y_k(1-y_k)$$

$$\Rightarrow \frac{dE}{dk} = \frac{dE}{dy} \cdot \frac{dy}{da_k} = \frac{y_k - t_k}{y_k(1-y_k)} \cdot y_k(1-y_k) = y_k - t_k$$

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

5. $y_k(x, w) = p(t_k = 1 | w)$

$w = \arg\max P(X|w) = \arg\max \prod_{i=1}^{N} \prod_{k=1}^{K} y_k(x, w)^{t_k}$

then we take negative log-likelihood, turn the problem to minimization and log is strictly increased function

$\Rightarrow w = \arg\min(-\log y(x, w)) = \arg\min\left(-\sum_{i=1}^{N} \sum_{k=1}^{K} t_{kn} \ln y_k(x_n, w)\right)$

$\Rightarrow$ maximum likelihood equals minimization of cross entropy error function

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