統計學習初論（105-2）

Self-Assessment Homework 3A

國立台灣大學資管系

Question 1

Consider the problem of performing least square for classification problem with K classes. Using 1-of-K binary coding scheme, each outcome is represented as $t_i = (0, 0, .., 1, ..0)$, where t_i is a vector of length K and the "one" is located at the position that corresponds to the outcome class. The feature of point i is $x_i = (1, x_{i1}, x_{i2}, \dots, x_{iD})^T$. We predict the outcome by comparing $y_k(x_i) = w_k^T x_i$ for $k = 1, 2, \dots, K$, and assign to the class a with $y_a(x_i) > y_b(x_i)$ for all $b \neq a$.

To estimate w_k, $k = 1, 2, \dots, K$, we adopted a vector notation:

$$T = \begin{bmatrix} t_1^T \\ t_2^T \\ \vdots \\ t_N^T \end{bmatrix}, X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix},$$

Here T is a $N \times K$ matrix and X is a $N \times (D + 1)$ matrix. The prediction for the N training data points are XW, where $W = [w_1\ w_2 \dots w_K]$ are the parameters for $y_i(\cdot)$. Define $\hat{Y} = XW$, then $\hat{Y}$ is a $N \times K$ matrix. The row i is the predicted score for data point i, and we assign row i to the class that has the largest score. To train this model, we minimize the difference between the true 1-of-K representation and the predicted score for all training data point. Define $E_D(W) = \frac{1}{2} tr\,\{(XW - T)^T(XW - T)\}$. Show that the solution that minimize $E_D(W)$ is $W = (X^T X)^{-1} X^T T$. Use the derivative equations in "Matrix Cookbook" (cf. http://www.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf) to help you solve the problem.

Question 2 (Exercise 4.8)

4.8 ($\star$) Using (4.57) and (4.58), derive the result (4.65) for the posterior class probability in the two-class generative model with Gaussian densities, and verify the results (4.66) and (4.67) for the parameters $\mathbf{w}$ and w_0.

Question 3 (Exercise 4.9)

4.9 ($\star$) **www** Consider a generative classification model for K classes defined by prior class probabilities $p(\mathcal{C}_k) = \pi_k$ and general class-conditional densities $p(\phi|\mathcal{C}_k)$ where ϕ is the input feature vector. Suppose we are given a training data set $\{\phi_n, \mathbf{t}_n\}$ where $n = 1, \ldots, N$, and $\mathbf{t}_n$ is a binary target vector of length K that uses the 1-of-K coding scheme, so that it has components $t_{nj} = I_{jk}$ if pattern n is from class $\mathcal{C}_k$. Assuming that the data points are drawn independently from this model, show that the maximum-likelihood solution for the prior probabilities is given by

$$\pi_k = \frac{N_k}{N} \tag{4.159}$$

where N_k is the number of data points assigned to class $\mathcal{C}_k$.

Question 4 (Exercise 4.10)

4.10 ($\star\star$) Consider the classification model of Exercise 4.9 and now suppose that the class-conditional densities are given by Gaussian distributions with a shared covariance matrix, so that

$$p(\phi|\mathcal{C}_k) = \mathcal{N}(\phi|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}). \tag{4.160}$$

Show that the maximum likelihood solution for the mean of the Gaussian distribution for class $\mathcal{C}_k$ is given by

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N} t_{nk}\phi_n \tag{4.161}$$

which represents the mean of those feature vectors assigned to class $\mathcal{C}_k$. Similarly, show that the maximum likelihood solution for the shared covariance matrix is given by

$$\boldsymbol{\Sigma} = \sum_{k=1}^{K} \frac{N_k}{N}\mathbf{S}_k \tag{4.162}$$

where

$$\mathbf{S}_k = \frac{1}{N_k}\sum_{n=1}^{N} t_{nk}(\phi_n - \boldsymbol{\mu}_k)(\phi_n - \boldsymbol{\mu}_k)^{\mathrm{T}}. \tag{4.163}$$

Thus $\boldsymbol{\Sigma}$ is given by a weighted average of the covariances of the data associated with each class, in which the weighting coefficients are given by the prior probabilities of the classes.