

Chapter 4 Linear Models for Classification

Yue Yu

4.1 Discriminant Functions

Skipped reading.

4.2 Probabilistic Generative Models

(4.57)

$$\begin{aligned} p(\mathcal{C}_1|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x})} \\ &= \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{\sum_{k=1}^K p(\mathbf{x}, \mathcal{C}_k)} \\ &= \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \\ &= \frac{1}{1 + \frac{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}} \\ &= \frac{1}{1 + \exp(-a)}. \end{aligned}$$

(4.65) – (4.67)

We can readily derive (4.65) by noticing that all terms will be canceled out except for those containing μ_k , provided (4.66) and (4.67).

(4.73)

As given by (4.72), the terms in the log likelihood depending on π are

$$\sum_{n=1}^N \{t_n \ln \pi + (1 - t_n) \ln(1 - \pi)\}.$$

Setting the derivative of the log likelihood function with respect to π to 0, we have

$$\begin{aligned} \frac{\partial}{\partial \pi} \ell(\pi, \mu_1, \mu_2, \Sigma) &= \sum_{n=1}^N t_n \frac{1}{\pi} - \sum_{n=1}^N (1 - t_n) \frac{1}{1 - \pi} \\ &= 0. \end{aligned}$$

Solving for π while denoting the total number of data points in class \mathcal{C}_1 by N_1 , we obtain

$$\pi = \frac{N_1}{N},$$

which is the fraction of points in class \mathcal{C}_1 .

This can be generalized to $K > 2$ classes where \mathbf{t}_n is a one hot vector of length K such that $t_{nj} = I_{jk}$. Then, the likelihood function can be written as

$$p(\mathbf{X}, \mathbf{T} | \pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) = \prod_{n=1}^N \prod_{k=1}^K (\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}))^{t_{nk}}.$$

The corresponding log likelihood function is

$$\ell(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) = \sum_{n=1}^N \sum_{k=1}^K t_{nk} (\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma})).$$

Here, we are only interested in the terms depending on π_k , namely,

$$\sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln \pi_k.$$

To find π_k , we construct the Lagrangian using the constraint $\sum_{k=1}^K \pi_k = 1$, given by

$$\mathcal{L}(\pi_k, \lambda) = \sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln \pi_k + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right).$$

Setting the derivative with respect to π_k to 0, we have

$$\begin{aligned} \frac{\partial}{\partial \pi_k} \mathcal{L}(\pi_k, \lambda) &= \sum_{n=1}^N t_{nk} \frac{1}{\pi_k} + \lambda \\ &= 0. \end{aligned}$$

Solving for π_k , we obtain

$$\pi_k = -\frac{1}{\lambda} \sum_{n=1}^N t_{nk} = -\frac{1}{\lambda} N_k. \quad (*)$$

Summing over k on both sides, we have

$$\sum_{k=1}^K \pi_k = -\frac{N}{\lambda} = 1,$$

which implies that

$$\lambda = -N.$$

Substituting back into (*), we obtain

$$\pi_k = \frac{N_k}{N},$$

which is the fraction of points in class \mathcal{C}_k .

(4.75) – (4.76)

To find $\boldsymbol{\mu}_1$, we set the derivative of the log likelihood with respect to $\boldsymbol{\mu}_1$ to $\mathbf{0}$,

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\mu}_1} \ell(\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) &= \frac{\partial}{\partial \boldsymbol{\mu}_1} \sum_{n=1}^N t_n \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \\ &= -\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\mu}_1} \sum_{n=1}^N t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) \\ &= \sum_{n=1}^N t_n (-\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}^{-1} \mathbf{x}_n) \\ &= \mathbf{0}. \end{aligned}$$

Solving for $\boldsymbol{\mu}_1$, we obtain

$$\boldsymbol{\mu}_1 = \frac{1}{N_1} \sum_{n=1}^N t_n \mathbf{x}_n,$$

where we denote $N_1 = \sum_{n=1}^N t_n$ as the number of data points assigned to class \mathcal{C}_1 . Similarly,

$$\boldsymbol{\mu}_2 = \frac{1}{N_2} \sum_{n=1}^N (1 - t_n) \mathbf{x}_n,$$

where we denote $N_2 = \sum_{n=1}^N (1 - t_n)$ as the number of data points assigned to class \mathcal{C}_2 .

This can be generalized to $K > 2$ classes with the same settings as the derivation of (4.73). The log likelihood function is

$$\ell(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) = \sum_{n=1}^N \sum_{k=1}^K t_{nk} (\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma})).$$

Here we are only interested in $\boldsymbol{\mu}_k$. Setting the derivative with respect to $\boldsymbol{\mu}_k$,

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\mu}_k} \ell(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) &= -\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_{n=1}^N \sum_{k=1}^K t_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \\ &= \sum_{n=1}^N t_{nk} (-\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \boldsymbol{\Sigma}^{-1} \mathbf{x}_n) \\ &= \mathbf{0}. \end{aligned}$$

Solving for $\boldsymbol{\mu}_k$, we obtain

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^N t_{nk} \mathbf{x}_n,$$

where $N_k = \sum_{n=1}^N t_{nk}$, representing the number of data points that are assigned to class \mathcal{C}_k .

(4.77) – (4.80)

To find $\boldsymbol{\Sigma}$, we set the derivative of the log likelihood function with respect to $\boldsymbol{\Sigma}^{-1}$ to $\mathbf{0}$. Specifically,

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \ell(\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) &= \frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \left(-\frac{1}{2} \sum_{n=1}^N t_n \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{n=1}^N t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) \right. \\ &\quad \left. - \frac{1}{2} \sum_{n=1}^N (1 - t_n) \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{n=1}^N (1 - t_n) (\mathbf{x}_n - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_2) \right) \\ &= \frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \left(-\frac{N}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{n=1}^N t_n \text{Tr}\{(\mathbf{x}_n - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1)\} \right. \\ &\quad \left. - \frac{1}{2} \sum_{n=1}^N (1 - t_n) \text{Tr}\{(\mathbf{x}_n - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_2)\} \right) \\ &= \frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}} \left(\frac{N}{2} \ln |\boldsymbol{\Sigma}^{-1}| - \frac{1}{2} \sum_{n=1}^N t_n \text{Tr}\{\boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) (\mathbf{x}_n - \boldsymbol{\mu}_1)^T\} \right. \\ &\quad \left. - \frac{1}{2} \sum_{n=1}^N (1 - t_n) \text{Tr}\{\boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_2) (\mathbf{x}_n - \boldsymbol{\mu}_2)^T\} \right) \\ &= \frac{N}{2} \boldsymbol{\Sigma} - \frac{1}{2} \sum_{n=1}^N t_n (\mathbf{x}_n - \boldsymbol{\mu}_1) (\mathbf{x}_n - \boldsymbol{\mu}_1)^T - \frac{1}{2} \sum_{n=1}^N (1 - t_n) (\mathbf{x}_n - \boldsymbol{\mu}_2) (\mathbf{x}_n - \boldsymbol{\mu}_2)^T \\ &= \mathbf{0}, \end{aligned}$$

where we used the following properties

$$\begin{aligned}\text{Tr}(\mathbf{ABC}) &= \text{Tr}(\mathbf{BCA}) \\ \frac{\partial}{\partial \mathbf{X}} \text{Tr}(\mathbf{XA}) &= \mathbf{A}^T \\ \frac{\partial}{\partial \mathbf{X}} \ln |\mathbf{X}| &= \mathbf{X}^{-T}.\end{aligned}$$

Solving for $\mathbf{\Sigma}$, we obtain

$$\mathbf{\Sigma} = \frac{1}{N} \sum_{n=1}^N \{t_n(\mathbf{x}_n - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_1)^T + (1 - t_n)(\mathbf{x}_n - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_2)^T\},$$

which is equivalent to (4.78) to (4.80).

A generalization to $K > 2$ classes can be derived using the same techniques. Consider the log likelihood function

$$\ell(\pi_k, \boldsymbol{\mu}_k, \mathbf{\Sigma}) = \sum_{n=1}^N \sum_{k=1}^K t_{nk} (\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \mathbf{\Sigma})).$$

Setting the derivative with respect to $\mathbf{\Sigma}^{-1}$ to $\mathbf{0}$ while taking advantage of the above properties, that is

$$\begin{aligned}\frac{\partial}{\partial \mathbf{\Sigma}^{-1}} \ell(\pi_k, \boldsymbol{\mu}_k, \mathbf{\Sigma}) &= \frac{\partial}{\partial \mathbf{\Sigma}^{-1}} \left(-\frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln |\mathbf{\Sigma}| - \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K t_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \right) \\ &= \frac{\partial}{\partial \mathbf{\Sigma}^{-1}} \left(\frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln |\mathbf{\Sigma}^{-1}| - \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K t_{nk} \text{Tr}\{\mathbf{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T\} \right) \\ &= \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K t_{nk} \mathbf{\Sigma} - \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K t_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \\ &= \frac{N}{2} \mathbf{\Sigma} - \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K t_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \\ &= \mathbf{0},\end{aligned}$$

where in the second last step we used the fact

$$\sum_{n=1}^N \sum_{k=1}^K t_{nk} = N.$$

Hence, we obtain

$$\mathbf{\Sigma} = \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K t_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T,$$

which is a weighted average of the covariances of the data points assigned to each class.