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 $\boldsymbol{\mu}_k$, 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

$$
\begin{aligned}
\frac{\partial}{\partial \pi} \ell(\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) &= \sum_{n=1}^{N} t_n \frac{1}{\pi} - \sum_{n=1}^{N} (1 - t_n) \frac{1}{1 - \pi} \\
&= 0.
\end{aligned}
$$

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

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

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 $\mathbf{0}$,

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

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 $\boldsymbol{\mu}_k$,

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

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,

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

where we used the following properties

$$\mathrm{Tr}(\mathbf{ABC}) = \mathrm{Tr}(\mathbf{BCA})$$
$$\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}(\mathbf{XA}) = \mathbf{A}^{\mathrm{T}}$$
$$\frac{\partial}{\partial \mathbf{X}} \ln |\mathbf{X}| = \mathbf{X}^{-\mathrm{T}}.$$

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

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

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

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

where in the second last step we used the fact

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

Hence, we obtain

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

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

4.3 Probabilistic Discriminative Models

(4.88)

This is easy to be verified using the chain rule.

$$\begin{aligned}
\frac{\mathrm{d}\sigma}{\mathrm{d}a} &= \frac{\mathrm{d}}{\mathrm{d}a}\frac{1}{1+\exp(-a)} \\
&= -\frac{1}{(1+\exp(-a))^2}\cdot 1\cdot \exp(-a)\cdot(-1) \\
&= \frac{1}{1+\exp(-a)}\left(1-\frac{1}{1+\exp(-a)}\right) \\
&= \sigma(1-\sigma).
\end{aligned}$$

(4.89)

This can be interpreted as *under the assumption that the probability of $\boldsymbol{\phi}_n$ belonging to class $\mathcal{C}_1$ is y_n, what is the chance of the given dataset coming into existence.*

(4.91)

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}} y_n &= \frac{\partial}{\partial \mathbf{w}}\sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n) \\
&= y_n(1-y_n)\boldsymbol{\phi}_n.
\end{aligned}$$

Using this conclusion, we can compute the gradient of the error function with respect to $\mathbf{w}$, giving

$$\begin{aligned}
\nabla_{\mathbf{w}} E(\mathbf{w}) &= -\nabla_{\mathbf{w}} \sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} \\
&= -\sum_{n=1}^{N}\left\{t_n\frac{1}{y_n}y_n(1-y_n)\boldsymbol{\phi}_n - (1-t_n)\frac{1}{1-y_n}y_n(1-y_n)\boldsymbol{\phi}_n\right\} \\
&= \sum_{n=1}^{N}(y_n-t_n)\boldsymbol{\phi}_n.
\end{aligned}$$

(4.97)

For any vector $\mathbf{u}$ that is not perpendicular to all the feature vectors, since $0 < y_n < 1$, we have

$$\begin{aligned}
\mathbf{u}^{\mathrm{T}}\mathbf{H}\mathbf{u} &= \mathbf{u}^{\mathrm{T}}\left(\sum_{n=1}^{N} y_n(1-y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}\right)\mathbf{u} \\
&= \sum_{n=1}^{N} y_n(1-y_n)(\mathbf{u}^{\mathrm{T}}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}\mathbf{u}) \\
&= \sum_{n=1}^{N} y_n(1-y_n)(\boldsymbol{\phi}_n^{\mathrm{T}}\mathbf{u})^2 \\
&> 0.
\end{aligned}$$

Hence, the Hessian is positive definite, which implies that the error function is convex and has a unique minimum.

(4.106)

When $j = k$,

$$\begin{aligned}
\frac{\partial}{\partial a_j} y_k &= \frac{\exp(a_k) \sum_j \exp(a_j) - \exp(a_k)^2}{\left(\sum_j \exp(a_j) \right)^2} \\
&= \frac{\exp(a_k)}{\sum_j \exp(a_j)} \frac{\sum_j \exp(a_j) - \exp(a_k)}{\sum_j \exp(a_j)} \\
&= y_k(1 - y_k).
\end{aligned}$$

When $j \neq k$,

$$\begin{aligned}
\frac{\partial}{\partial a_j} y_k &= -\frac{\exp(a_j) \exp(a_k)}{\left(\sum_i \exp(a_i) \right)^2} \\
&= -y_k y_j \\
&= y_k(0 - y_j).
\end{aligned}$$

Combining the two cases, we obtain

$$\frac{\partial}{\partial a_j} y_k = y_k(I_{kj} - y_j),$$

where I_{kj} are the elements of the identity matrix.

(4.109)

$$\begin{aligned}
\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) &= \nabla_{\mathbf{w}_j} \left(-\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk} \right) \\
&= -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \frac{1}{y_{nk}} \frac{\partial y_{nk}}{\partial a_j} \frac{\partial a_j}{\partial w_j} \\
&= -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} (I_{kj} - y_{nj}) \boldsymbol{\phi}_n \\
&= \sum_{n=1}^{N} \left(\sum_{k=1}^{K} t_{nk} \right) y_{nj} \boldsymbol{\phi}_n - \sum_{n=1}^{N} \left(\sum_{k=1}^{K} t_{nk} I_{kj} \right) \boldsymbol{\phi}_n \\
&= \sum_{n=1}^{N} y_{nj} \boldsymbol{\phi}_n - \sum_{n=1}^{N} t_{nj} \boldsymbol{\phi}_n \\
&= \sum_{n=1}^{N} (y_{nj} - t_{nj}) \boldsymbol{\phi}_n.
\end{aligned}$$

(4.110)

Using the result of (4.109), this equation should be

$$\nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \nabla_{\mathbf{w}_k} \sum_{n=1}^{N} y_{nj} (I_{kj} - y_{nk}) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^{\mathrm{T}}.$$

(4.116)

$$
\begin{aligned}
\Phi(a) &= \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)\,\mathrm{d}\theta \\
&= \frac{1}{2} + \int_{0}^{a} \mathcal{N}(\theta|0,1)\,\mathrm{d}\theta \\
&= \frac{1}{2} + \int_{0}^{a} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{\theta^2}{2}\right)\mathrm{d}\theta \\
&= \frac{1}{2} + \frac{1}{(2\pi)^{1/2}} \int_{0}^{\frac{a}{\sqrt{2}}} \sqrt{2}\exp(-u^2)\,\mathrm{d}u \\
&= \frac{1}{2} + \frac{1}{\sqrt{\pi}} \int_{0}^{\frac{a}{\sqrt{2}}} \exp(-u^2)\,\mathrm{d}u \\
&= \frac{1}{2}\left\{1 + \mathrm{erf}\left(\frac{1}{\sqrt{2}}\right)\right\}.
\end{aligned}
$$

where in the forth step we made change of variable $u = \theta/\sqrt{2}$.

(4.119)

Recall (4.118)

$$
p(t|\eta, s) = \frac{1}{s} h\left(\frac{t}{s}\right) g(\eta) \exp\left\{\frac{\eta t}{s}\right\}.
$$

Taking advantage of the fact that the integral of $p(t|\eta, s)$ equals to 1, we have

$$
\int \frac{1}{s} h\left(\frac{t}{s}\right) g(\eta) \exp\left\{\frac{\eta t}{s}\right\} \mathrm{d}t = 1. \tag{*}
$$

Differentiating both sides with respect to η, we obtain

$$
\left(\frac{\mathrm{d}}{\mathrm{d}\eta} g(\eta)\right) \int \frac{1}{s} h\left(\frac{t}{s}\right) g(\eta) \exp\left\{\frac{\eta t}{s}\right\} \mathrm{d}t + \frac{1}{s} g(\eta) \int \frac{1}{s} h\left(\frac{t}{s}\right) g(\eta) \exp\left\{\frac{\eta t}{s}\right\} t\,\mathrm{d}t = 0.
$$

Then, by making use of (*), the equation can be reduced to

$$
\frac{\mathrm{d}}{\mathrm{d}\eta} g(\eta) + \frac{1}{s} g(\eta) \mathbb{E}[t|\eta] = 0,
$$

which implies that

$$
\begin{aligned}
\mathbb{E}[t|\eta] &= -s \frac{1}{g(\eta)} \frac{\mathrm{d}}{\mathrm{d}\eta} g(\eta) \\
&= -s \frac{\mathrm{d}}{\mathrm{d}\eta} \ln g(\eta).
\end{aligned}
$$