PRML Solutions

Kaikai Zhao
Email: kkai_zhao@yeah.net

First draft: January 24, 2023 Last update: September 9, 2023

Contents

Notations

$$\binom{\alpha}{k} = \frac{\alpha(\alpha-1)(\alpha-2)\cdots(\alpha-k+1)}{k!} \quad (k = 1, 2, \cdots)$$

$$\binom{\alpha}{0} = 1$$

where α is a nonzero real number. Note that in combinatorics α is usually a positive integer n, i.e., $\binom{n}{k}$ which is also denoted as C_n^k with $\mathrm{C}_n^0 = 1$. In this case, we have

$$\mathrm{C}_n^k = \frac{n!}{k!\,(n-k)!}$$

1 Introduction

1.1 Exercises

Exercise 1.1

Consider the sum-of-squares error function given by (1.2),

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 \tag{1}$$

in which the function $y(x, \mathbf{w})$ is given by the polynomial (1.1),

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j. \tag{2}$$

Show that the coefficients $\mathbf{w} = \{w_i\}$ that minimize this error function are given by the solution to the following set of linear equations

$$\sum_{j=0}^{M} A_{ij} w_j = T_i \tag{3}$$

where

$$A_{ij} = \sum_{n=1}^{N} (x_n)^{i+j}, \qquad T_i = \sum_{n=1}^{N} (x_n)^i t_n. \tag{4}$$

Here a suffix i or j denotes the index of a component, whereas $(x)^i$ denotes x raised to the power of i.

Proof. Since $E(\mathbf{w})$ is a quadratic function, it follows that $E(\mathbf{w})$ is convex with respect to $\mathbf{w}$. Additionally, $E(\mathbf{w})$ is lower bounded by 0 and its feasible set is the entire space $\mathbb{R}^M$, which is convex as well. Hence, the minimum of $E(\mathbf{w})$ can be achieved at its stationary points $\mathbf{w}^*$, i.e. $\nabla_{\mathbf{w}^*}(E(\mathbf{w}^*)) = \mathbf{0}$.

We will denote by $\mathbf{x}$ and $\mathbf{t}$ the column vectors $(1, x, x^2, \ldots, x^M)^T$ and $(t_1, t_2, \ldots, t_N)^T$, respectively. Furthermore, we can combine the observations $\{\mathbf{x}_n\}$ into a data matrix $\mathbf{X}$ in which the n^{th} row of $\mathbf{X}$ corresponds to the row vector $\mathbf{x}_n^T$. Then, we can get the following compact formulations,

$$y(x, \mathbf{w}) = \mathbf{w}^T\mathbf{x}, \qquad E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{\mathbf{w}^T\mathbf{x}_n - t_n\}^2 = \frac{1}{2}(\mathbf{X}\mathbf{w} - \mathbf{t})^T(\mathbf{X}\mathbf{w} - \mathbf{t}) \tag{5}$$

Taking gradients of $E(\mathbf{w})$ w.r.t. $\mathbf{w}$ gives

$$\nabla_{\mathbf{w}}(E(\mathbf{w})) = \mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{t}) = \mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{t} \tag{6}$$

Setting $\nabla_{\mathbf{w}}(E(\mathbf{w})) = \mathbf{0}$ yields $\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{t}$. By expanding this compact result, we get

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^M & x_2^M & \cdots & x_n^M \end{pmatrix} \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^M \\ 1 & x_2 & x_2^2 & \cdots & x_2^M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^M \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^M & x_2^M & \cdots & x_n^M \end{pmatrix} \begin{pmatrix} t_0 \\ t_1 \\ \vdots \\ t_N \end{pmatrix} \tag{7}$$

$$\Downarrow$$

$$\begin{pmatrix} \sum_{n=1}^{N} 1^{0+0} & \sum_{n=1}^{N} x_n^{0+1} & \cdots & \sum_{n=1}^{N} x_n^{0+M} \\ \sum_{n=1}^{N} x_n^{1+0} & \sum_{n=1}^{N} x_n^{1+1} & \cdots & \sum_{n=1}^{N} x_n^{1+M} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{n=1}^{N} x_n^{M+0} & \sum_{n=1}^{N} x_n^{M+1} & \cdots & \sum_{n=1}^{N} x_n^{M+M} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{pmatrix} = \begin{pmatrix} \sum_{j=0}^{M} \sum_{n=0}^{N} x_n^{0+j} w_j \\ \sum_{j=0}^{M} \sum_{n=0}^{N} x_n^{1+j} w_j \\ \vdots \\ \sum_{j=0}^{M} \sum_{n=0}^{N} x_n^{M+j} w_j \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^{N} t_n \\ \sum_{n=1}^{N} x_n t_n \\ \vdots \\ \sum_{n=1}^{N} x_n^{M} t_n \end{pmatrix} \tag{8}$$

$$\Downarrow$$

$$\mathbf{A}\mathbf{w} = \mathbf{T} \tag{9}$$

where $A_{ij} = \sum_{n=1}^{N} x_n^{i+j}$ and $\mathbf{T}$ is a column vector with elements $T_i = \sum_{n=1}^{N} x_n^i t_n$ for $i, j = 0, 1, \cdots, M$. Note that we omitted the brackets around x_n for notational brevity. This completes the proof. □

Exercise 1.2

Write down the set of coupled linear equations, analogous to (3) ((1.122) in PRML), satisfied by the coefficients w_i which minimize the regularized sum-of-squares error function given by ((1.4) in PRML)

$$\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2. \tag{10}$$

Solution: Taking gradients of $\tilde{E}(\mathbf{w})$ w.r.t. $\mathbf{w}$ gives

$$\nabla_{\mathbf{w}}(\tilde{E}(\mathbf{w})) = \mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{t}) + \lambda\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w} - \mathbf{X}^T\mathbf{t}. \tag{11}$$

Setting $\tilde{E}(\mathbf{w}) = \mathbf{0}$ yields $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w} = \mathbf{X}^T\mathbf{t}$. Thus, we get $\tilde{\mathbf{A}}\mathbf{w} = \mathbf{T}$ where $T_i = \sum_{n=1}^{N} x_n^i t_n$ is identical to the counterpart in Exercise 1.1. Following a similar argument, we obtain $\tilde{A}_{ij} = \sum_{n=1}^{N} (x_n^{i+j} + \lambda\delta_{ij})$ where $\delta_{ij} = 1$ when $i = j$ otherwise $\delta_{ij} = 0$. □

Exercise 1.3

Suppose that you have three coloured boxes r (red), b (blue), and g (green). Box r contains 3 apples, 4 oranges, and 3 limes, box b contains 1 apple, 1 orange, and 0 limes, and box g contains 3 apples, 3 oranges, and 4 limes. If a box is chosen at random with probabilities $p(r) = 0.2$, $p(b) = 0.2$, $p(g) = 0.6$, and a piece of fruit is removed from the box with equal probability of selecting any of the items in the box), then what is the probability of selecting an apple? If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?

Solution: The probabilities of selecting an apple from the red, the blue, or the green box are given by

$$p(F = a|r) = \frac{3}{3+4+3} = 0.3 \tag{12}$$

$$p(F = a|b) = \frac{1}{1+1} = 0.5 \tag{13}$$

$$p(F = a|g) = \frac{3}{3+3+4} = 0.3 \tag{14}$$

respectively. We use the sum and product rules of probability to evaluate the probability of selecting an apple.

$$p(F=a) = p(F=a|r)p(r) + p(F=a|b)p(b) + p(F=a|g)p(g) \tag{15}$$
$$= 0.3 \times 0.2 + 0.5 \times 0.2 + 0.3 \times 0.6 = 0.06 + 0.1 + 0.18 \tag{16}$$
$$= 0.34 \tag{17}$$

By the Bayes' Theorem, the probability of a selected orange that came from the green box is

$$p(g|F=o) = \frac{p(F=o|g)p(g)}{p(F=o)} \tag{18}$$
$$= \frac{p(F=o|g)p(g)}{p(F=o|r)p(r) + p(F=o|b)p(b) + p(F=o|g)p(g)} \tag{19}$$
$$= \frac{0.3 \times 0.6}{0.4 \times 0.2 + 0.5 \times 0.2 + 0.3 \times 0.6} = \frac{0.18}{0.08 + 0.1 + 0.18} \tag{20}$$
$$= 0.5 \tag{21}$$

□

Exercise 1.4

Consider a probability density $p_x(x)$ defined over a continuous variable x, and suppose that we make a nonlinear change of variable using $x = g(y)$, so that the density transforms according to ((1.27) in PRML book)

$$p_y(y) = p_x(x)|g'(y)|. \tag{22}$$

By differentiating (22), show that the location $\hat{y}$ of the maximum of the density in y is not in general related to the location $\hat{x}$ of the maximum of the density over x by the simple functional relation $\hat{x} = g(\hat{y})$ as a consequence of the Jacobian factor. This shows that the maximum of a probability density (in contrast to a simple function) is dependent on the change of variable. Verify that, in the case of a linear transformation, the location of the maximum transforms in the same way as the variable itself.

The proof below follows the same logic as the official solution.

Proof. Given a funtion $f(x)$ and the relation $x = g(y)$, we can get a new function

$$\tilde{f}(y) = f(g(y)). \tag{23}$$

Suppose $f(x)$ achieves its maximum at $\hat{x}$ so that $f'(\hat{x}) = 0$. The corresponding maximum $\tilde{f}(\hat{y})$ will be obtained by differentiating both sides of (23) w.r.t y

$$\tilde{f}'(\hat{y}) = f'(g(\hat{y}))g'(\hat{y}) = 0. \tag{24}$$

Assuming $g'(\hat{y}) \neq 0$ at the maximum $\tilde{f}(\hat{y})$, then $\tilde{f}'(g(\hat{y})) = 0$. Since $f'(\hat{x}) = 0$, we see that the locations of the maximum are related by $\hat{x} = g(\hat{y})$. Thus, finding a maximum w.r.t x is equivalent to first transforming to y, and then find a maximum w.r.t y, and then transforming back to x.

Now consider the behavior of a probability density $p_x(x)$ under the change of variables $x = g(y)$. According to (22), the new density is given by

$$p_y(y) = p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = p_x(g(y))|g'(y)|. \tag{25}$$

Let $|g'(y)| = sg'(y)$ where $s \in \{-1, 1\}$, then

$$p_y(y) = sp_x(g(y))g'(y). \tag{26}$$

Differentiating both sides w.r.t y yields

$$p'_y(y) = sp'_x(g(y))(g'(y))^2 + sp_x(g(y))g''(y). \tag{27}$$

Due to the presence of the second term on the right hand side of (27), the result $\hat{x} = g(\hat{y})$ no longer holds. This implies that we can not get the maximum of $p_x(x)$ by simply transforming to $p_y(y)$ then maximizing w.r.t y and then transforming back to x. In other words, maxima of densities are dependent on the choice of variables. From the above analyses, we see that this is exactly the consequence of the Jacobian factor $|g'(y)|$.

In the case of linear transformation, $g''(y)$ vanishes and $g'(y)$ is a constant denoted c, then we have

$$p'_y(y) = sc^2 p'_x(g(y)). \tag{28}$$

which implies $p'_y(\hat{y}) = p'_x(g(\hat{y})) = p'_x(\hat{x}) = 0$ at the stationarity $\hat{y}$. Thus, the location of the maximum transforms according to $\hat{x} = g(\hat{y})$. This completes the proof. □

Exercise 1.5

Using the definition ((1.38) in PRML book)

$$\text{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] \tag{29}$$

show that $\text{var}[f(x)]$ satisfies ((1.39) in PRML book)

$$\text{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2. \tag{30}$$

Proof. Expanding the right hand side of (29) gives

$$\begin{aligned}
\text{var}[f] &= \mathbb{E}\left[f(x)^2 - 2\mathbb{E}[f(x)]f(x) + \mathbb{E}[f(x)]^2\right] && (31)\\
&= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2 && (32)\\
&= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]^2 + \mathbb{E}[f(x)]^2 && (33)\\
&= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 && (34)
\end{aligned}$$

as desired. □

Exercise 1.6

Show that if two variables x and y are independent, then their covariance is zero.

Proof. By the definition of covariance, we have

$$\begin{aligned}
\text{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] && (35)\\
&= \mathbb{E}_{x,y}\left[xy - \mathbb{E}[x]y - \mathbb{E}[y]x + \mathbb{E}[x]\mathbb{E}[y]\right] && (36)\\
&= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] - \mathbb{E}[y]\mathbb{E}[x] + \mathbb{E}[x]\mathbb{E}[y] && (37)\\
&= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] && (38)\\
&= \iint xyp(x, y)\mathrm{d}x\mathrm{d}y - \mathbb{E}[x]\mathbb{E}[y] && (39)
\end{aligned}$$

$$= \iint xyp(x)p(y)\mathrm{d}x\mathrm{d}y - \mathbb{E}[x]\mathbb{E}[y] \tag{40}$$

$$= \int xp(x)\mathrm{d}x \int yp(y)\mathrm{d}y - \mathbb{E}[x]\mathbb{E}[y] \tag{41}$$

$$= \mathbb{E}[x]\mathbb{E}[y] - \mathbb{E}[x]\mathbb{E}[y] = 0 \tag{42}$$

as desired. □

Exercise 1.7

In this exercise, we prove the normalization condition ((1.48) in PRML book)

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) = 1 \tag{43}$$

for the univariate Gaussian. To do this consider, the integral

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right)\mathrm{d}x \tag{44}$$

which we can evaluate by first writing its square in the form

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}y^2\right)\mathrm{d}x\mathrm{d}y. \tag{45}$$

Now make the transformation from Cartesian coordinates (x, y) to polar coordinates (r, θ) and then substitute $u = r^2$. Show that, by performing the integrals over θ and u, and then taking the square root of both sides, we obtain

$$I = (2\pi\sigma^2)^{1/2}. \tag{46}$$

Finally, use this result to show that the Gaussian distribution $\mathcal{N}(x|\mu, \sigma^2)$ is normalized.

Proof. By making the transformation from Cartesian coordinates (x, y) to polar coordinates (r, θ) and then substitute $u = r^2$, we have

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)\mathrm{d}x\mathrm{d}y \tag{47}$$

$$= \int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{r^2}{2\sigma^2}\right) r\mathrm{d}r\mathrm{d}\theta \tag{48}$$

$$= 2\pi\sigma^2 \int_0^{\infty} \exp\left(-\frac{r^2}{2\sigma^2}\right)\mathrm{d}\frac{r^2}{2\sigma^2} \tag{49}$$

$$= 2\pi\sigma^2 \int_0^{\infty} \exp\left(-\frac{u}{2\sigma^2}\right)\mathrm{d}\frac{u}{2\sigma^2} \tag{50}$$

$$= -2\pi\sigma^2 \exp\left(-\frac{u}{2\sigma^2}\right)|_0^{+\infty} \tag{51}$$

$$= -2\pi\sigma^2(0-1) = 2\pi\sigma^2. \tag{52}$$

Thus,

$$I = (2\pi\sigma^2)^{1/2}. \tag{53}$$

Furthermore,

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)\mathrm{d}x = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x = \frac{1}{\sqrt{2\pi}\sigma} I = 1. \tag{54}$$

This completes our proof. □

Exercise 1.8

By using a change of variables, verify that the univariate Gaussian distribution given by ((1.46) in PRML book)

$$\mathcal{N}(x \mid \mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{55}$$

satisfies ((1.49) in PRML book)

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu,\sigma^2) x \mathrm{d}x = \mu. \tag{56}$$

Next, by differentiating both sides of the normalization condition

$$\int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu,\sigma^2)\mathrm{d}x = 1 \tag{57}$$

with respect to σ^2, verify that the Gaussian satisfies ((1.50) in PRML book)

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu,\sigma^2) x^2 \mathrm{d}x = \mu^2 + \sigma^2. \tag{58}$$

Finally, show that ((1.51) in PRML book)

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 \tag{59}$$

holds.

Proof. Let's first verify (56).

$$\begin{aligned}
\mathbb{E}[x] &= \int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} x\mathrm{d}x && (60)\\
&= \int_{-\infty}^{\infty} \frac{y+\mu}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{y^2}{2\sigma^2}\right\} \mathrm{d}y \qquad (y = x-\mu) && (61)\\
&= \int_{-\infty}^{\infty} \frac{y}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{y^2}{2\sigma^2}\right\} \mathrm{d}y + \mu \underbrace{\int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{y^2}{2\sigma^2}\right\} \mathrm{d}y}_{=1} && (62)\\
&= 0 + \mu = \mu && (63)
\end{aligned}$$

where the first term of the second last line vanishes since the integrand is an odd function with respect to y and the region of integration is symmetric about 0.

Next, to derive (58), we first substitute the standard form of Gaussian distribution into (43) and make some rearrangements.

$$\int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x = \sqrt{2\pi\sigma^2}. \tag{64}$$

Before doing differentiation on both sides, we need to explain why we can swap the differentiation and the integration. Define $f(x,\sigma^2)=\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ and $I(\sigma^2)=\int_{-\infty}^{\infty}f(x,\sigma^2)\mathrm{d}x$, then $f'_{\sigma^2}(x,\sigma^2)$ is given by

$$f'_{\sigma^2}(x,\sigma^2)=\frac{(x-\mu)^2}{2(\sigma^2)^2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{65}$$

It is easy to see that $f(x,\sigma^2)$ and $f'_{\sigma^2}(x,\sigma^2)$ are continuous on $(-\infty,+\infty)\times(0,+\infty)$, and for every $\sigma^2\in(0,+\infty)$, $I(\sigma^2)$ converges to $\sqrt{2\pi\sigma^2}$[1]. The last thing we need to check is if $\int_{-\infty}^{\infty}f'_{\sigma^2}(x,\sigma^2)\mathrm{d}x$ is uniformly convergent for $\sigma^2\in(0,+\infty)$. Actually, since $(0,+\infty)$ is open, we only need to see if $\int_{-\infty}^{\infty}f'_{\sigma^2}(x,\sigma^2)\mathrm{d}x$ is uniformly convergent on any closed subset of $(0,+\infty)$. To do this, let $z=(x-\mu)/\sigma^2$, then

$$\begin{aligned}
\int_{-\infty}^{\infty}f'_{\sigma^2}(x,\sigma^2)\mathrm{d}x&=\int_{-\infty}^{\infty}\frac{(x-\mu)^2}{2(\sigma^2)^2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x &(66)\\
&=\frac{4\sigma^2}{2(\sigma^2)^2}\int_{-\infty}^{\infty}\frac{(x-\mu)^2}{4\sigma^2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x &(67)\\
&<\frac{2}{\sigma^2}\int_{-\infty}^{\infty}\exp\left(\frac{(x-\mu)^2}{4\sigma^2}\right)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x &(68)\\
&=\frac{2}{\sigma^2}\int_{-\infty}^{\infty}\exp\left(-\frac{(x-\mu)^2}{4\sigma^2}\right)\mathrm{d}x &(69)\\
&=\frac{2}{\sigma^2}\sqrt{2\pi(\sqrt{2}\sigma)^2}=\frac{4}{\sigma^2}\sqrt{\pi\sigma^2} &(70)
\end{aligned}$$

where the inequality in the third line follows from $x<e^x$ for any $x\in\mathbb{R}$. Since $f'_{\sigma^2}(x,\sigma^2)\geq 0$, according to Weiestrass's test for absolute uniform convergence, the above derivation shows that $\int_{-\infty}^{\infty}f'_{\sigma^2}(x,\sigma^2)\mathrm{d}x$ is uniformly convergent. Thus, we can interchange the differentiation and integral safely.

$$I'(\sigma^2)=\int_{-\infty}^{\infty}f'_{\sigma^2}(x,\sigma^2)\mathrm{d}x=\frac{\mathrm{d}}{\mathrm{d}\sigma^2}\sqrt{2\pi\sigma^2}=\frac{\sqrt{2\pi}}{2\sqrt{\sigma^2}} \tag{71}$$

We can rewrite (67) as

$$\begin{aligned}
&\frac{4\sigma^2}{2(\sigma^2)^2}\int_{-\infty}^{\infty}\frac{(x-\mu)^2}{4\sigma^2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x &(72)\\
=&\frac{4\sigma^2\sqrt{2\pi\sigma^2}}{2(\sigma^2)^2}\int_{-\infty}^{\infty}\frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}4\sigma^2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x &(73)\\
=&\frac{\sqrt{2\pi\sigma^2}}{2(\sigma^2)^2}\int_{-\infty}^{\infty}\frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x &(74)\\
=&\frac{\sqrt{2\pi\sigma^2}}{2(\sigma^2)^2}\mathrm{var}[x]. &(75)
\end{aligned}$$

Combining (71) and (72) yields

$$\frac{\sqrt{2\pi\sigma^2}}{2(\sigma^2)^2}\mathrm{var}[x]=\frac{\sqrt{2\pi}}{2\sqrt{\sigma^2}}\Longleftrightarrow\mathrm{var}[x]=\sigma^2 \tag{76}$$

Furthermore, by the definition of variance,

$$\mathrm{var}[x]=\sigma^2=\int_{-\infty}^{\infty}\frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \tag{77}$$

[1] http://homepages.math.uic.edu/~jyang06/stat411/handouts/InterchangeDiffandIntegral.pdf

$$= \int_{-\infty}^{\infty} \frac{x^2 - 2\mu x + \mu^2}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x \tag{78}$$

$$= \mathbb{E}[x^2] - 2\mu\mathbb{E}[x] + \mu^2 = \mathbb{E}[x^2] - \mu^2 \tag{79}$$

$$\iff \mathbb{E}[x^2] = \mu^2 + \sigma^2. \tag{80}$$

The last two claims have been proved together. □

Exercise 1.9

Show that the mode (i.e. maximum) of the Gaussian distribution ((1.46) in PRML book)

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{81}$$

is given by μ. Similarly, show that the mode of the multivariate Gaussian ((1.52) in PRML book)

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{82}$$

is given by $\boldsymbol{\mu}$. Here, $\mathbf{x}$ is a D-dimensional vector of continuous variables.

Proof. For the univariate case, differentiating the Gaussian density function with respect to x gives

$$\frac{\partial \mathcal{N}(x \mid \mu, \sigma^2)}{\partial x} = -\frac{x-\mu}{\sigma^2} \cdot \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{83}$$

Setting this to 0 yields $x = \mu$. So $x = \mu$ is the only stationary point. Since $x \in \mathbb{R}$ and $\lim_{x\to\infty} \mathcal{N}(x \mid \mu, \sigma^2) = 0$, then the mode of $\mathcal{N}(x \mid \mu, \sigma^2)$ is given by μ.

Similarly, for the multivariate case, according to the result $\nabla_{\mathbf{x}} \mathbf{x}^T \mathbf{A}\mathbf{x} = 2\mathbf{A}\mathbf{x}$ where $\mathbf{A}$ is a symmetric matrix, differentiating the multivariate Gaussian with respect to $\mathbf{x}$ gives

$$\frac{\partial \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{x}} = -\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{84}$$

Setting this to $\mathbf{0}$ and left-multiplying by $\boldsymbol{\Sigma}$ yield $\mathbf{x} = \boldsymbol{\mu}$. The same argument is applicable here. □

Exercise 1.10

Suppose that the two variables x and z are statistically independent. Show that the mean and variance of their sum satisfies

$$\mathbb{E}[x+z] = \mathbb{E}[x] + \mathbb{E}[z] \tag{85}$$

$$\mathrm{Var}[x+z] = \mathrm{Var}[x] + \mathrm{Var}[z]. \tag{86}$$

Proof. We first consider the case when x and z are continuous.

$$\mathbb{E}[x+z] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+z)p(x,z)\mathrm{d}x\mathrm{d}z \qquad \text{(Definition of mean)} \tag{87}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+z)p(x)p(z)\mathrm{d}x\mathrm{d}z \qquad (x \text{ and } z \text{ are independent}) \tag{88}$$

$$= \int_{-\infty}^{\infty} xp(x)\mathrm{d}x + \int_{-\infty}^{\infty} zp(z)\mathrm{d}z \tag{89}$$

$$= \mathbb{E}[x] + \mathbb{E}[z] \qquad \text{(Definition of mean)} \tag{90}$$

For the variances, since x and z are independent,

$$(x + z - \mathbb{E}(x + z))^2 = \tag{91}$$

$$\begin{aligned}
\mathrm{Var}[x+z] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x + z - \mathbb{E}(x+z))^2 \, p(x,z)\mathrm{d}x\mathrm{d}z && \text{(Definition of variance)} && (92)\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} ((x - \mathbb{E}[x])^2 + (z - \mathbb{E}[z])^2 && && (93)\\
&\quad -2(x - \mathbb{E}[x])(z - \mathbb{E}[z])) \, p(x)p(z)\mathrm{d}x\mathrm{d}z && (x \text{ and } z \text{ are independent}) && (94)\\
&= \int_{-\infty}^{\infty} (x - \mathbb{E}[x])^2 p(x)\mathrm{d}x + \int_{-\infty}^{\infty} (z - \mathbb{E}[z])^2 p(z)\mathrm{d}z && && (95)\\
&\quad - 2\int_{-\infty}^{\infty} (x - \mathbb{E}(x))p(x)\mathrm{d}x \int_{-\infty}^{\infty} (z - \mathbb{E}(z))p(z)\mathrm{d}z && && (96)\\
&= \int_{-\infty}^{\infty} (x - \mathbb{E}[x])^2 p(x)\mathrm{d}x + \int_{-\infty}^{\infty} (z - \mathbb{E}[z])^2 p(z)\mathrm{d}z && && (97)\\
&= \mathrm{Var}[x] + \mathrm{Var}[z] && \text{(Definition of variance)} && (98)
\end{aligned}$$

□

Exercise 1.11

By setting the derivatives of the log likelihood function ((1.54) in PRML book)

$$\ln p(\mathbf{x} \mid \mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) \tag{99}$$

with respect to μ and σ^2 equal to zero, verify the results

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{100}$$

and

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2. \tag{101}$$

Proof.

$$\frac{\partial \ln p(\mathbf{x} \mid \mu, \sigma^2)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu) = 0 \tag{102}$$

$$\Downarrow \tag{103}$$

$$\sum_{n=1}^{N}(x_n - \mu) = 0 \Rightarrow \sum_{n=1}^{N} x_n - N\mu) = 0 \Rightarrow \mu = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{104}$$

Thus, $\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$. Now we plug μ_{ML} into (99) and then take derivatives with respect to σ^2.

$$\frac{\partial \ln p(\mathbf{x} \mid \mu, \sigma^2)}{\partial \sigma^2} = \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 - \frac{N}{2\sigma^2} = 0 \tag{105}$$

$$\Downarrow \tag{106}$$

$$\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2-N=0\Rightarrow N\sigma^2=\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2\Rightarrow\sigma^2=\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2. \tag{107}$$

Hence, $\sigma^2_{\mathrm{ML}}=\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$. This completes the verification. □

Exercise 1.12

Using the results in PRML book, i.e. (1.49)

$$\mathbb{E}[x]=\int_{-\infty}^{\infty}\mathcal{N}(x\mid\mu,\sigma^2)x\mathrm{d}x=\mu \tag{108}$$

and (1.50)

$$\mathbb{E}[x^2]=\int_{-\infty}^{\infty}\mathcal{N}(x\mid\mu,\sigma^2)x^2\mathrm{d}x=\mu^2+\sigma^2, \tag{109}$$

show that

$$\mathbb{E}[x_nx_m]=\mu^2+I_{nm}\sigma^2 \tag{110}$$

where x_n and x_m denote data points sampled from a Gaussian distribution with mean μ and variance σ^2, and I_{nm} satisfies $I_{nm}=1$ if $n=m$ and $I_{nm}=0$ otherwise. Hence prove that the results (1.57) and (1.58) in PRML book as follows.

$$\mathbb{E}[\mu_{\mathrm{ML}}]=\mu \tag{111}$$

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}]=\left(\frac{N-1}{N}\right)\sigma^2. \tag{112}$$

Proof. When $n=m$, $\mathbb{E}[x_nx_m]=\mathbb{E}[x_n^2]=\mu^2+\sigma^2$. However, if $n\neq m$, since x_n and x_m are independent, $\mathbb{E}[x_nx_m]=\mathbb{E}[x_n]\mathbb{E}[x_m]=\mu^2$. Thus, $\mathbb{E}[x_nx_m]=\mu^2+I_{nm}\sigma^2$ holds.

$$\mathbb{E}[\mu_{\mathrm{ML}}]=\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}x_n\right]=\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n\right]=\frac{N\mu}{N}=\mu \tag{113}$$

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}]=\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2\right] \tag{114}$$

$$=\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2-2x_n\mu_{\mathrm{ML}}+\mu^2_{\mathrm{ML}}\right] \tag{115}$$

$$=\frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{E}[x_n^2]-2\mathbb{E}[x_n\mu_{\mathrm{ML}}]+\mathbb{E}[\mu^2_{\mathrm{ML}}]\right) \tag{116}$$

$$=\frac{1}{N}\sum_{n=1}^{N}\left(\mu^2+\sigma^2-2\mathbb{E}\left[x_n\frac{1}{N}\sum_{n=1}^{N}x_n\right]+\mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right]\right) \tag{117}$$

$$=\frac{1}{N}\sum_{n=1}^{N}\left(\mu^2+\sigma^2-\frac{2}{N}\mathbb{E}\left[x_n^2+\sum_{i\neq n}^{N}x_nx_i\right]+\frac{1}{N^2}\mathbb{E}\left[\sum_{i=1}^{N}x_i^2+\sum_{i\neq j}^{N}x_ix_j\right]\right) \tag{118}$$

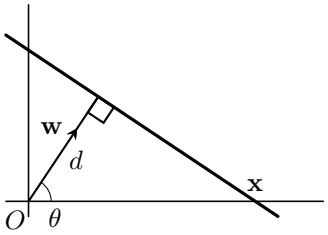

Figure 1: Illustration of the geometry of a linear discriminant function in two dimensions. The direction of $\mathbf{w}$ depends on the form of the decision surface, shown in thick, $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$. This makes θ in range $[0, \pi]$.

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2 + \sigma^2 - \frac{2}{N}\left(\mathbb{E}[x_n^2] + \sum_{i\neq n}^{N}\mathbb{E}[x_n x_i]\right) + \frac{1}{N^2}\left(\sum_{n=1}^{N}\mathbb{E}[x_n^2] + \sum_{i\neq j}^{N}\mathbb{E}[x_i x_j]\right)\right) \quad (119)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2 + \sigma^2 - \frac{2}{N}\left(\mu^2 + \sigma^2 + (N-1)\mu^2\right) + \frac{1}{N^2}\left(N(\mu^2+\sigma^2) + N(N-1)\mu^2\right)\right) \quad (120)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2 + \sigma^2 - \frac{1}{N}(\sigma^2 + N\mu^2)\right) = \left(\frac{N-1}{N}\right)\sigma^2 \quad (121)$$

which completes the proof. □

2 Chapter 4 Linear Models for Classification

2.1 Discriminant Functions

2.1.1 The derivation of Equation (4.5)

Equation (4.5) gives the normal distance from the origin to the decision surface $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = 0$. Here is a thorough derivation with an illustration shown in Figure 1.

$$\mathbf{w}^T\mathbf{x} = \|\mathbf{w}\|\|\mathbf{x}\|\cos\theta = \|\mathbf{w}\|\cdot d \Longrightarrow \frac{\mathbf{w}^T\mathbf{x}}{\|\mathbf{w}\|} = d. \quad (122)$$

Since any point $\mathbf{x}$ on the decision surface satisfies $\mathbf{w}^T\mathbf{x} + w_0 = 0$, we have

$$\mathbf{w}^T\mathbf{x} = -w_0 \Longrightarrow \frac{\mathbf{w}^T\mathbf{x}}{\|\mathbf{w}\|} = d = \frac{-w_0}{\|\mathbf{w}\|}. \quad (123)$$

You may have noticed Equation (4.7), i.e. $r = \frac{y(\mathbf{x})}{\|\mathbf{w}\|}$, on Page 182 of the PRML textbook. When the point $\mathbf{x}$ lies at the origin, namely $\mathbf{x} = \mathbf{0}$, it is easy to get

$$r = \frac{y(\mathbf{0})}{\|\mathbf{w}\|} = \frac{w_0}{\|\mathbf{w}\|}. \quad (124)$$

which does not contradict the form of d, though both represent the distance between the origin and the decision surface. Mathematically speaking, they are signed distances. We can follow the wording from the textbook to interpret d as the normal distance *from the origin to the decision surface* and r as the perpendicular (orthogonal) distance *from the decision surface to the point* $\mathbf{x}$.

2.2 Exercises

Exercise 4.1

Given a set of data points $\mathbf{x}_n$, we can define the *convex hull* to be the set of all points $\mathbf{x}$ given by

$$\mathbf{x} = \sum_n a_n \mathbf{x}_n$$

where $a_n \geq 0$ and $\sum_n a_n = 1$. Consider a second set of points $\mathbf{y}_n$ together with their corresponding convex hull. By definition, the two sets of points will be linearly separable if there exists a vector $\widehat{\mathbf{w}}$ and a scalar w_0 such that $\widehat{\mathbf{w}}^T\mathbf{x}_n + w_0 > 0$ for all $\mathbf{x}_n$, and $\widehat{\mathbf{w}}^T\mathbf{y}_n + w_0 < 0$ for all $\mathbf{y}_n$. Show that if their convex hulls intersect, the two sets of points cannot be linearly separable, and conversely that if they are linearly separable, their convex hulls do not intersect.

Proof. We prove the first part by contradiction. Given the convex hulls of the two sets of points intersect, we need to show that the two sets cannot be linearly separable. Since the two convex hulls intersect, there exists at least one point $\mathbf{z} = \sum_n a_n \mathbf{x}_n = \sum_n a_n \mathbf{y}_n$ that lies in their intersection, where $a_n, b_n \geq 0$ and $\sum_n a_n = \sum_n b_n = 1$. For the sake of contradiction, suppose that the two sets of points are linearly separable, then there exists a vector $\widehat{\mathbf{w}}$ and a scalar w_0 such that $\widehat{\mathbf{w}}^T\mathbf{x}_n + w_0 > 0$ for all $\mathbf{x}_n$, and $\widehat{\mathbf{w}}^T\mathbf{y}_n + w_0 < 0$ for all $\mathbf{y}_n$. Furthermore,

$$\begin{aligned}
\widehat{\mathbf{w}}^T\mathbf{z} = \widehat{\mathbf{w}}^T\left(\sum_n a_n\mathbf{x}_n\right) = \sum_n a_n\widehat{\mathbf{w}}^T\mathbf{x}_n &= \sum_n a_n(\widehat{\mathbf{w}}^T\mathbf{x}_n + w_0 - w_0) \\
&= \sum_n a_n(\widehat{\mathbf{w}}^T\mathbf{x}_n + w_0) - w_0\overbrace{\sum_n a_n}^{=1} \\
&= \sum_n a_n \underbrace{(\widehat{\mathbf{w}}^T\mathbf{x}_n + w_0)}_{>0} - w_0 \\
&> -w_0
\end{aligned}$$

Similarly,

$$\begin{aligned}
\widehat{\mathbf{w}}^T\mathbf{z} = \widehat{\mathbf{w}}^T\left(\sum_n b_n\mathbf{y}_n\right) = \sum_n b_n\widehat{\mathbf{w}}^T\mathbf{y}_n &= \sum_n b_n(\widehat{\mathbf{w}}^T\mathbf{y}_n + w_0 - w_0) \\
&= \sum_n b_n(\widehat{\mathbf{w}}^T\mathbf{y}_n + w_0) - w_0\overbrace{\sum_n b_n}^{=1} \\
&= \sum_n b_n \underbrace{(\widehat{\mathbf{w}}^T\mathbf{y}_n + w_0)}_{<0} - w_0. \\
&< -w_0
\end{aligned}$$

which gives an obvious contradiction. This implies that the assumption does not hold. In other words, the two sets of points cannot be linearly separable.

Now we show the second half still by contradiction. Given the two sets are linearly separable, then there exists a vector $\widehat{\mathbf{w}}$ and a scalar w_0 such that $\widehat{\mathbf{w}}^T\mathbf{x}_n + w_0 > 0$ for all $\mathbf{x}_n$, and $\widehat{\mathbf{w}}^T\mathbf{y}_n + w_0 < 0$ for all $\mathbf{y}_n$. Suppose their convex hulls intersect, then there exists at least one point $\mathbf{z}$ such that $\mathbf{z} = \sum_n a_n = \sum_n b_n$ with $\sum_n a_n = \sum_n b_n = 1$ and $a_n, b_n \geq 0$. For any $a_n > 0$, we have

$$\widehat{\mathbf{w}}^T\mathbf{x}_n + w_0 > 0 \Longrightarrow a_n\widehat{\mathbf{w}}^T\mathbf{x}_n + a_n w_0 > 0$$

$$\Longrightarrow \widehat{\mathbf{w}}^T a_n \mathbf{x}_n + a_n w_0 > 0.$$

For $a_n = 0$, $\widehat{\mathbf{w}}^T a_n \mathbf{x}_n + a_n w_0 = 0$. Summing over n, we get

$$\sum_n \widehat{\mathbf{w}}^T a_n \mathbf{x}_n + a_n w_0 > 0 \Longrightarrow \widehat{\mathbf{w}}^T \underbrace{\sum_n (a_n \mathbf{x}_n)}_{=\mathbf{z}} + w_0 \underbrace{\sum_n a_n}_{=1} > 0 \Longrightarrow \widehat{\mathbf{w}}^T \mathbf{z} + w_0 > 0.$$

Likewise,

$$\sum_n \widehat{\mathbf{w}}^T b_n \mathbf{y}_n + b_n w_0 < 0 \Longrightarrow \widehat{\mathbf{w}}^T \underbrace{\sum_n (b_n \mathbf{y}_n)}_{=\mathbf{z}} + w_0 \underbrace{\sum_n b_n}_{=1} < 0 \Longrightarrow \widehat{\mathbf{w}}^T \mathbf{z} + w_0 < 0.$$

which leads to a contradiction. This shows their convex hulls do not intersect. This completes the proof.

□