PRML Solutions

Youming Zhao
Email: youming0.zhao@gmail.com

First draft: January 24, 2023   Last update: November 26, 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),

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 \tag{1.1}$$

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

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j. \tag{1.2}$$

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

$$\sum_{j=0}^{M} A_{ij} w_j = T_i \tag{1.3}$$

where

$$A_{ij} = \sum_{n=1}^{N}(x_n)^{i+j}, \qquad T_i = \sum_{n=1}^{N}(x_n)^i t_n. \tag{1.4}$$

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,

$$y(x, \mathbf{w}) = \mathbf{w}^T\mathbf{x}, \qquad E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{\mathbf{w}^T\mathbf{x}_n - t_n\}^2 = \frac{1}{2}(\mathbf{X}\mathbf{w} - \mathbf{t})^T(\mathbf{X}\mathbf{w} - \mathbf{t}) \tag{1.5}$$

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

$$\nabla_{\mathbf{w}}(E(\mathbf{w})) = \mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{t}) = \mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{t} \tag{1.6}$$

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

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^M & x_2^M & \cdots & x_n^M \end{pmatrix} \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^M \\ 1 & x_2 & x_2^2 & \cdots & x_2^M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^M \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^M & x_2^M & \cdots & x_n^M \end{pmatrix} \begin{pmatrix} t_0 \\ t_1 \\ \vdots \\ t_N \end{pmatrix} \tag{1.7}$$

$$\Downarrow$$

$$\begin{pmatrix} \sum_{n=1}^N 1^{0+0} & \sum_{n=1}^N x_n^{0+1} & \cdots & \sum_{n=1}^N x_n^{0+M} \\ \sum_{n=1}^N x_n^{1+0} & \sum_{n=1}^N x_n^{1+1} & \cdots & \sum_{n=1}^N x_n^{1+M} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{n=1}^N x_n^{M+0} & \sum_{n=1}^N x_n^{M+1} & \cdots & \sum_{n=1}^N x_n^{M+M} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{pmatrix} = \begin{pmatrix} \sum_{j=0}^M \sum_{n=0}^N x_n^{0+j} w_j \\ \sum_{j=0}^M \sum_{n=0}^N x_n^{1+j} w_j \\ \vdots \\ \sum_{j=0}^M \sum_{n=0}^N x_n^{M+j} w_j \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^N t_n \\ \sum_{n=1}^N x_n t_n \\ \vdots \\ \sum_{n=1}^N x_n^M t_n \end{pmatrix} \tag{1.8}$$

$$\Downarrow$$

$$\mathbf{A}\mathbf{w} = \mathbf{T} \tag{1.9}$$

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 (1.3) ((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)

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2. \tag{1.10}$$

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

$$\nabla_{\mathbf{w}}(\tilde{E}(\mathbf{w})) = \mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{t}) + \lambda\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w} - \mathbf{X}^T\mathbf{t}. \tag{1.11}$$

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

$$p(F = a|r) = \frac{3}{3+4+3} = 0.3 \tag{1.12}$$

$$p(F = a|b) = \frac{1}{1+1} = 0.5 \tag{1.13}$$

$$p(F = a|g) = \frac{3}{3+3+4} = 0.3 \tag{1.14}$$

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

$$\begin{aligned} p(F=a) &= p(F=a|r)p(r) + p(F=a|b)p(b) + p(F=a|g)p(g) && (1.15)\\ &= 0.3\times 0.2 + 0.5\times 0.2 + 0.3\times 0.6 = 0.06 + 0.1 + 0.18 && (1.16)\\ &= 0.34 && (1.17)\end{aligned}$$

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

$$\begin{aligned} p(g|F=o) &= \frac{p(F=o|g)p(g)}{p(F=o)} && (1.18)\\ &= \frac{p(F=o|g)p(g)}{p(F=o|r)p(r)+p(F=o|b)p(b)+p(F=o|g)p(g)} && (1.19)\\ &= \frac{0.3\times 0.6}{0.4\times 0.2+0.5\times 0.2+0.3\times 0.6} = \frac{0.18}{0.08+0.1+0.18} && (1.20)\\ &= 0.5 && (1.21)\end{aligned}$$

□

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)

$$p_y(y) = p_x(x)|g'(y)|. \tag{1.22}$$

By differentiating (1.22), 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

$$\tilde{f}(y) = f(g(y)). \tag{1.23}$$

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 (1.23) w.r.t y

$$\tilde{f}'(\hat{y}) = f'(g(\hat{y}))g'(\hat{y}) = 0. \tag{1.24}$$

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 (1.22), the new density is given by

$$p_y(y) = p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = p_x(g(y))|g'(y)|. \tag{1.25}$$

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

$$p_y(y) = sp_x(g(y))g'(y). \tag{1.26}$$

Differentiating both sides w.r.t y yields

$$p'_y(y) = sp'_x(g(y))(g'(y))^2 + sp_x(g(y))g''(y). \tag{1.27}$$

Due to the presence of the second term on the right hand side of (1.27), 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

$$p'_y(y) = sc^2 p'_x(g(y)). \tag{1.28}$$

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)

$$\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] \tag{1.29}$$

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

$$\mathrm{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2. \tag{1.30}$$

Proof. Expanding the right hand side of (1.29) gives

$$\begin{aligned}
\mathrm{var}[f] &= \mathbb{E}\left[f(x)^2 - 2\mathbb{E}[f(x)]f(x) + \mathbb{E}[f(x)]^2\right] && (1.31)\\
&= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2 && (1.32)\\
&= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]^2 + \mathbb{E}[f(x)]^2 && (1.33)\\
&= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 && (1.34)
\end{aligned}$$

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

$$\begin{aligned}
\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] && (1.35)\\
&= \mathbb{E}_{x,y}\left[xy - \mathbb{E}[x]y - \mathbb{E}[y]x + \mathbb{E}[x]\mathbb{E}[y]\right] && (1.36)\\
&= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] - \mathbb{E}[y]\mathbb{E}[x] + \mathbb{E}[x]\mathbb{E}[y] && (1.37)\\
&= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] && (1.38)\\
&= \iint xyp(x, y)\mathrm{d}x\mathrm{d}y - \mathbb{E}[x]\mathbb{E}[y] && (1.39)
\end{aligned}$$

$$
\begin{aligned}
&= \iint xyp(x)p(y)\mathrm{d}x\mathrm{d}y - \mathbb{E}[x]\mathbb{E}[y] && (1.40)\\
&= \int xp(x)\mathrm{d}x \int yp(y)\mathrm{d}y - \mathbb{E}[x]\mathbb{E}[y] && (1.41)\\
&= \mathbb{E}[x]\mathbb{E}[y] - \mathbb{E}[x]\mathbb{E}[y] = 0 && (1.42)
\end{aligned}
$$

43 as desired. □

Exercise 1.7

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

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) = 1 \tag{1.43}$$

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

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right)\mathrm{d}x \tag{1.44}$$

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

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}y^2\right)\mathrm{d}x\mathrm{d}y. \tag{1.45}$$

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

$$I = (2\pi\sigma^2)^{1/2}. \tag{1.46}$$

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

44

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

$$
\begin{aligned}
I^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)\mathrm{d}x\mathrm{d}y && (1.47)\\
&= \int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{r^2}{2\sigma^2}\right) r\mathrm{d}r\mathrm{d}\theta && (1.48)\\
&= 2\pi\sigma^2 \int_0^{\infty} \exp\left(-\frac{r^2}{2\sigma^2}\right)\mathrm{d}\frac{r^2}{2\sigma^2} && (1.49)\\
&= 2\pi\sigma^2 \int_0^{\infty} \exp\left(-\frac{u}{2\sigma^2}\right)\mathrm{d}\frac{u}{2\sigma^2} && (1.50)\\
&= -2\pi\sigma^2 \exp\left(-\frac{u}{2\sigma^2}\right)\Big|_0^{+\infty} && (1.51)\\
&= -2\pi\sigma^2(0-1) = 2\pi\sigma^2. && (1.52)
\end{aligned}
$$

Thus,

$$I = (2\pi\sigma^2)^{1/2}. \tag{1.53}$$

Furthermore,

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)\mathrm{d}x = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right)\mathrm{d}x = \frac{1}{\sqrt{2\pi}\sigma} I = 1. \tag{1.54}$$

This completes our proof. $\square$

Exercise 1.8

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

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{1.55}$$

satisfies ((1.49) in PRML book)

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu, \sigma^2) x \mathrm{d}x = \mu. \tag{1.56}$$

Next, by differentiating both sides of the normalization condition

$$\int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu, \sigma^2)\mathrm{d}x = 1 \tag{1.57}$$

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

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu, \sigma^2) x^2 \mathrm{d}x = \mu^2 + \sigma^2. \tag{1.58}$$

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

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 \tag{1.59}$$

holds.

Proof. Let's first verify (1.56).

$$\begin{aligned}
\mathbb{E}[x] &= \int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} x \mathrm{d}x && (1.60)\\
&= \int_{-\infty}^{\infty} \frac{y+\mu}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{y^2}{2\sigma^2}\right\} \mathrm{d}y \qquad (y = x-\mu) && (1.61)\\
&= \int_{-\infty}^{\infty} \frac{y}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{y^2}{2\sigma^2}\right\} \mathrm{d}y + \mu \underbrace{\int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{y^2}{2\sigma^2}\right\} \mathrm{d}y}_{=1} && (1.62)\\
&= 0 + \mu = \mu && (1.63)
\end{aligned}$$

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 (1.58), we first substitute the standard form of Gaussian distribution into (1.43) and make some rearrangements.

$$\int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x = \sqrt{2\pi\sigma^2}. \tag{1.64}$$

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

$$f'_{\sigma^2}(x,\sigma^2) = \frac{(x-\mu)^2}{2(\sigma^2)^2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{1.65}$$

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

$$\int_{-\infty}^{\infty} f'_{\sigma^2}(x,\sigma^2)\mathrm{d}x = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{2(\sigma^2)^2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \tag{1.66}$$

$$= \frac{4\sigma^2}{2(\sigma^2)^2}\int_{-\infty}^{\infty} \frac{(x-\mu)^2}{4\sigma^2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \tag{1.67}$$

$$< \frac{2}{\sigma^2}\int_{-\infty}^{\infty} \exp\left(\frac{(x-\mu)^2}{4\sigma^2}\right)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \tag{1.68}$$

$$= \frac{2}{\sigma^2}\int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{4\sigma^2}\right)\mathrm{d}x \tag{1.69}$$

$$= \frac{2}{\sigma^2}\sqrt{2\pi(\sqrt{2}\sigma)^2} = \frac{4}{\sigma^2}\sqrt{\pi\sigma^2} \tag{1.70}$$

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.

$$I'(\sigma^2) = \int_{-\infty}^{\infty} f'_{\sigma^2}(x,\sigma^2)\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\sigma^2}\sqrt{2\pi\sigma^2} = \frac{\sqrt{2\pi}}{2\sqrt{\sigma^2}} \tag{1.71}$$

We can rewrite (1.67) as

$$\frac{4\sigma^2}{2(\sigma^2)^2}\int_{-\infty}^{\infty} \frac{(x-\mu)^2}{4\sigma^2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \tag{1.72}$$

$$=\frac{4\sigma^2\sqrt{2\pi\sigma^2}}{2(\sigma^2)^2}\int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}4\sigma^2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \tag{1.73}$$

$$=\frac{\sqrt{2\pi\sigma^2}}{2(\sigma^2)^2}\int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \tag{1.74}$$

$$=\frac{\sqrt{2\pi\sigma^2}}{2(\sigma^2)^2}\mathrm{var}[x]. \tag{1.75}$$

Combining (1.71) and (1.72) yields

$$\frac{\sqrt{2\pi\sigma^2}}{2(\sigma^2)^2}\mathrm{var}[x] = \frac{\sqrt{2\pi}}{2\sqrt{\sigma^2}} \Longleftrightarrow \mathrm{var}[x] = \sigma^2 \tag{1.76}$$

Furthermore, by the definition of variance,

$$\mathrm{var}[x] = \sigma^2 = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \tag{1.77}$$

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

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} \frac{x^2 - 2\mu x + \mu^2}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x && (1.78)\\
&= \mathbb{E}[x^2] - 2\mu\mathbb{E}[x] + \mu^2 = \mathbb{E}[x^2] - \mu^2 && (1.79)\\
\iff \mathbb{E}[x^2] &= \mu^2 + \sigma^2. && (1.80)
\end{aligned}
$$

49 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)

$$
\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{1.81}
$$

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

$$
\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\} \tag{1.82}
$$

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

50

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

$$
\frac{\partial \mathcal{N}(x \mid \mu, \sigma^2)}{\partial x} = -\frac{x-\mu}{\sigma^2} \cdot \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{1.83}
$$

51 Setting this to 0 yields $x = \mu$. So $x = \mu$ is the only stationary point. Since $x \in \mathbb{R}$ and $\lim_{x\to\infty} \mathcal{N}(x \mid$
52 $\mu, \sigma^2) = 0$, 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

$$
\frac{\partial \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{x}} = -\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{1.84}
$$

53 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

$$
\begin{aligned}
\mathbb{E}[x+z] &= \mathbb{E}[x] + \mathbb{E}[z] && (1.85)\\
\operatorname{Var}[x+z] &= \operatorname{Var}[x] + \operatorname{Var}[z]. && (1.86)
\end{aligned}
$$

54

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

$$
\begin{aligned}
\mathbb{E}[x+z] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+z)p(x,z)\mathrm{d}x\mathrm{d}z && \text{(Definition of mean)} && (1.87)\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+z)p(x)p(z)\mathrm{d}x\mathrm{d}z && (x \text{ and } z \text{ are independent}) && (1.88)\\
&= \int_{-\infty}^{\infty} xp(x)\mathrm{d}x + \int_{-\infty}^{\infty} zp(z)\mathrm{d}z && && (1.89)
\end{aligned}
$$

$$= \mathbb{E}[x] + \mathbb{E}[z] \qquad \text{(Definition of mean)} \tag{1.90}$$

For the variances, since x and z are independent,

$$(x + z - \mathbb{E}(x+z))^2 = \tag{1.91}$$

$$\operatorname{Var}[x+z] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x + z - \mathbb{E}(x+z))^2 \, p(x,z)\mathrm{d}x\mathrm{d}z \qquad \text{(Definition of variance)} \tag{1.92}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} ((x - \mathbb{E}[x])^2 + (z - \mathbb{E}[z])^2 \tag{1.93}$$

$$-2(x - \mathbb{E}[x])(z - \mathbb{E}[z]))\, p(x)p(z)\mathrm{d}x\mathrm{d}z \qquad (x \text{ and } z \text{ are independent}) \tag{1.94}$$

$$= \int_{-\infty}^{\infty} (x - \mathbb{E}[x])^2 p(x)\mathrm{d}x + \int_{-\infty}^{\infty} (z - \mathbb{E}[z])^2 p(z)\mathrm{d}z \tag{1.95}$$

$$-\, 2\int_{-\infty}^{\infty} (x - \mathbb{E}(x))p(x)\mathrm{d}x \int_{-\infty}^{\infty} (z - \mathbb{E}(z))p(z)\mathrm{d}z \tag{1.96}$$

$$= \int_{-\infty}^{\infty} (x - \mathbb{E}[x])^2 p(x)\mathrm{d}x + \int_{-\infty}^{\infty} (z - \mathbb{E}[z])^2 p(z)\mathrm{d}z \tag{1.97}$$

$$= \operatorname{Var}[x] + \operatorname{Var}[z] \qquad \text{(Definition of variance)} \tag{1.98}$$

55 □

Exercise 1.11

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

$$\ln p(\mathbf{x} \mid \mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) \tag{1.99}$$

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

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{1.100}$$

and

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2. \tag{1.101}$$

56

Proof.

$$\frac{\partial \ln p(\mathbf{x} \mid \mu, \sigma^2)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu) = 0 \tag{1.102}$$

$$\Downarrow \tag{1.103}$$

$$\sum_{n=1}^{N}(x_n - \mu) = 0 \Rightarrow \sum_{n=1}^{N} x_n - N\mu) = 0 \Rightarrow \mu = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{1.104}$$

Thus, $\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$. Now we plug μ_{ML} into (1.99) and then take derivatives with respect to σ^2.

$$\frac{\partial \ln p(\mathbf{x} \mid \mu, \sigma^2)}{\partial \sigma^2} = \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 - \frac{N}{2\sigma^2} = 0 \tag{1.105}$$

$$\Downarrow \tag{1.106}$$

$$\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2 - N = 0 \Rightarrow N\sigma^2 = \sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2 \Rightarrow \sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2. \tag{1.107}$$

57 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)

$$\mathbb{E}[x] = \int_{-\infty}^{\infty}\mathcal{N}(x \mid \mu, \sigma^2)x\mathrm{d}x = \mu \tag{1.108}$$

and (1.50)

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty}\mathcal{N}(x \mid \mu, \sigma^2)x^2\mathrm{d}x = \mu^2+\sigma^2, \tag{1.109}$$

show that

$$\mathbb{E}[x_nx_m] = \mu^2 + I_{nm}\sigma^2 \tag{1.110}$$

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.

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu \tag{1.111}$$

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2. \tag{1.112}$$

58

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.

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}x_n\right] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n\right] = \frac{N\mu}{N} = \mu \tag{1.113}$$

$$\begin{aligned}
\mathbb{E}[\sigma^2_{\mathrm{ML}}] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2\right] && (1.114)\\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2 - 2x_n\mu_{\mathrm{ML}} + \mu_{\mathrm{ML}}^2\right] && (1.115)\\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{E}[x_n^2] - 2\mathbb{E}[x_n\mu_{\mathrm{ML}}] + \mathbb{E}[\mu_{\mathrm{ML}}^2]\right) && (1.116)\\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2+\sigma^2 - 2\mathbb{E}\left[x_n\frac{1}{N}\sum_{n=1}^{N}x_n\right] + \mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right]\right) && (1.117)\\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2+\sigma^2 - \frac{2}{N}\mathbb{E}\left[x_n^2 + \sum_{i\neq n}^{N}x_nx_i\right] + \frac{1}{N^2}\mathbb{E}\left[\sum_{i=1}^{N}x_i^2 + \sum_{i\neq j}^{N}x_ix_j\right]\right) && (1.118)
\end{aligned}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2 + \sigma^2 - \frac{2}{N}\left(\mathbb{E}[x_n^2] + \sum_{i\neq n}^{N}\mathbb{E}[x_n x_i]\right) + \frac{1}{N^2}\left(\sum_{n=1}^{N}\mathbb{E}[x_n^2] + \sum_{i\neq j}^{N}\mathbb{E}[x_i x_j]\right)\right) \tag{1.119}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2 + \sigma^2 - \frac{2}{N}\left(\mu^2 + \sigma^2 + (N-1)\mu^2\right) + \frac{1}{N^2}\left(N(\mu^2 + \sigma^2) + N(N-1)\mu^2\right)\right) \tag{1.120}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2 + \sigma^2 - \frac{1}{N}(\sigma^2 + N\mu^2)\right) = \left(\frac{N-1}{N}\right)\sigma^2 \tag{1.121}$$

59 which completes the proof. □

Exercise 1.13

Suppose that the variance of a Gaussian is estimated using σ^2_{ML} but with the maximum likelihood estimate μ_{ML} replaced with the true value μ of the mean. Show that this estimator has the property that its expectation is given by the true variance σ^2.

60

Proof. By replacing μ_{ML} in the proof of Exercise 1.12 with μ, we get

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2\right] \tag{1.122}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2 - 2x_n\mu + \mu^2\right] \tag{1.123}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{E}[x_n^2] - 2\mu\mathbb{E}[x_n] + \mathbb{E}[\mu^2]\right) \tag{1.124}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2 + \sigma^2 - 2\mu^2 + \mu^2\right) \tag{1.125}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sigma^2 = \frac{N\sigma^2}{N} = \sigma^2 \tag{1.126}$$

61 □

Exercise 1.14

Show that an arbitrary square matrix with the elements w_{ij} can be written in the form $w_{ij} = w_{ij}^{\mathrm{S}} + w_{ij}^{\mathrm{A}}$ where w_{ij}^{S} and w_{ij}^{A} are symmetric and anti-symmetric matrices, respectively, satisfying $w_{ij}^{\mathrm{S}} = w_{ji}^{\mathrm{S}}$ and $w_{ij}^{\mathrm{A}} = -w_{ji}^{\mathrm{A}}$ for all i and j. Now consider the second order term in a higher order polynomial in D dimensions, given by

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_ix_j. \tag{1.127}$$

Show that

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_ix_j = \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\mathrm{S}}x_ix_j \tag{1.128}$$

so that the contribution from the anti-symmetric matrix vanishes. We therefore see that, without loss of generality, the matrix of coefficients w_{ij} can be chosen to be symmetric, and so not all of the D^2 elements of this matrix can be chosen independently. Show that the number of independent parameters in the matrix w_{ij}^{S} is given by $D(D+1)/2$.

62 *Proof.* [2] Given an arbitrary square matrix $\mathbf{W}$, let $\mathbf{S} = (\mathbf{W} + \mathbf{W}^T)/2$ and $\mathbf{A} = (\mathbf{W} - \mathbf{W}^T)/2$, then we have $\mathbf{W} = \mathbf{S} + \mathbf{A}$, namely, $w_{ij} = w_{ij}^{\mathrm{S}} + w_{ij}^{\mathrm{A}}$. Since $\mathbf{S} = \mathbf{S}^T$ and $\mathbf{A}^T = -\mathbf{A}$, $\mathbf{S}$ and $\mathbf{A}$ are symmetric and anti-symmetric matrices, respectively. With this, we have

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\mathrm{A}}x_ix_j = \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}(w_{ij} - w_{ji})x_ix_j \tag{1.129}$$

$$= \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_ix_j - \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ji}x_ix_j \tag{1.130}$$

$$= \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_ix_j - \frac{1}{2}\sum_{j=1}^{D}\sum_{i=1}^{D} w_{ij}x_jx_i \tag{1.131}$$

$$= \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_ix_j - \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_ix_j = 0 \tag{1.132}$$

It follows that

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_ix_j = \sum_{i=1}^{D}\sum_{j=1}^{D}(w_{ij}^{\mathrm{S}} + w_{ij}^{\mathrm{A}})x_ix_j \tag{1.133}$$

$$= \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\mathrm{S}}x_ix_j + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\mathrm{A}}x_ix_j \tag{1.134}$$

$$= \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\mathrm{S}}x_ix_j \tag{1.135}$$

For a symmetric matrix, the parameters of only the upper triangle part or the lower triangle part are independent. Therefore, the number of independent parameters can be calculated as follows.

$$1 + 2 + \ldots + (D-1) + D = \frac{D(D+1)}{2}. \tag{1.136}$$

[2] When working on the second part of the proof, I referenced the official solution manual.

63 This completes the proof. $\square$

Exercise 1.15

In this exercise and the next, we explore how the number of independent parameters in a polynomial grows with the order M of the polynomial and with the dimension D of the input space. We start by writing down the M^{th} order term for a polynomial in D dimensions in the form

$$\sum_{i_1=1}^{D}\sum_{i_2=1}^{D}\cdots\sum_{i_M=1}^{D} w_{i_1 i_2 \cdots i_M} x_{i_1} x_{i_2} \cdots x_{i_M}. \tag{1.137}$$

The coefficients $w_{i_1 i_2 \cdots i_M}$ comprise D^M elements, but the number of independent parameters is significantly fewer due to the many interchange symmetries of the factor $x_{i_1} x_{i_2} \cdots x_{i_M}$. Begin by showing that the redundancy in the coefficients can be removed by rewriting this M^{th} order term in the form

$$\sum_{i_1=1}^{D}\sum_{i_2=1}^{i_1}\cdots\sum_{i_M=1}^{i_{M-1}} \tilde{w}_{i_1 i_2 \cdots i_M} x_{i_1} x_{i_2} \cdots x_{i_M}. \tag{1.138}$$

Note that the precise relationship between the $\tilde{w}$ coefficients and w coefficients need not be made explicit. Use this result to show that the number of independent parameters $n(D, M)$, which appear at order M, satisfies the following recursion relation

$$n(D, M) = \sum_{i=1}^{D} n(i, M-1). \tag{1.139}$$

Next use proof by induction to show that the following result holds

$$\sum_{i=1}^{D} \frac{(i+M-2)!}{(i-1)!\,(M-1)!} = \frac{(D+M-1)!}{(D-1)!\,M!} \tag{1.140}$$

which can be done by first proving the result for $D = 1$ and arbitrary M by making use of the result $0! = 1$, then assuming it is correct for dimension D and verifying that it is correct for dimension $D + 1$. Finally, use the two previous results, together with proof by induction, to show

$$n(D, M) = \frac{(D+M-1)!}{(D-1)!\,M!}. \tag{1.141}$$

To do this, first show that the result is true for $M = 2$, and any value of $D \geq 1$, by comparison with the result of Exercise 1.14. Then make use of (1.139), together with (1.140), to show that, if the result holds at order $M - 1$, then it will also hold at order M.

64

Proof. [3] The redundancy in (1.137) arises from the interchange symmetries between the indices i_k. Enforcing the appearance order of the indices i_k can remove this redundancy as in (1.138). To derive (1.139), the number of independent parameters which appear at order M can be written as

$$n(D, M) = \sum_{i_1=1}^{D}\sum_{i_2=1}^{i_1}\cdots\sum_{i_M=1}^{i_{M-1}} 1 \tag{1.142}$$

[3] I referenced the official manual for the first part of the exercise.

which can be rewritten as

$$n(D,M) = \sum_{i_1=1}^{D} \left(\sum_{i_2=1}^{i_1} \cdots \sum_{i_M=1}^{i_{M-1}} 1 \right) = \sum_{i_1=1}^{D} n(i_1, M-1) \tag{1.143}$$

65 which is equivalent to (1.139).

Now we use proof by induction to show (1.140). For $D = 1$ and arbitrary M, using the result $0! = 1$, we have

$$\sum_{i=1}^{1} \frac{(1+M-2)!}{(1-1)!\,(M-1)!} = \frac{(M-1)!}{0!\,(M-1)!} = 1 = \frac{1+M-1}{(1-1)!\,M!}. \tag{1.144}$$

which shows (1.140) is correct when $D = 1$. Assuming (1.140) is correct for dimension D, then for dimension $D+1$, we have

$$\begin{aligned}
\sum_{i=1}^{D+1} \frac{(i+M-2)!}{(i-1)!\,(M-1)!} &= \frac{(D+M-1)!}{(D-1)!\,M!} + \frac{(D+1+M-2)!}{(D+1-1)!\,(M-1)!} && (1.145)\\
&= \frac{(D+M-1)!}{(D-1)!\,M!} + \frac{(D+M-1)!}{D!\,(M-1)!} && (1.146)\\
&= \frac{(D+M-1)!\,(D+M)}{D!\,M!} && (1.147)\\
&= \frac{(D+M)!}{D!\,M!} && (1.148)
\end{aligned}$$

as desired. Once again, we show (1.141) with proof by induction. When $M = 2$ and $D \geq 1$, it is exactly the case of Exercise 1.14. Then we have

$$n(D,2) = \frac{(D+2-1)!}{(D-1)!\,2!} = \frac{(D+1)!}{(D-1)!\,2} = \frac{D(D+1)}{2} \tag{1.149}$$

which shows that (1.141) holds when $M = 2$ and any $D \geq 1$. Assuming (1.141) is correct for $M-1$, then we have

$$n(D, M-1) = \frac{(D+M-2)!}{(D-1)!\,(M-1)!} \tag{1.150}$$

Combining this with (1.139) and (1.140), we get

$$n(D,M) = \sum_{i=1}^{D} n(i, M-1) = \sum_{i=1}^{D} \frac{(i+M-2)!}{(i-1)!\,(M-1)!} = \frac{(D+M-1)!}{(D-1)!\,M!}. \tag{1.151}$$

66 This completes the proof. □

Exercise 1.16

In Exercise 1.15, we proved the result (1.139) for the number of independent parameters in the M^{th} order term of a D-dimensional polynomial. We now find an expression for the total number $N(D, M)$ of independent paramaters in all of the terms up to and including the Mth order. First show that $N(D, M)$ satisfies

$$N(D, M) = \sum_{m=0}^{M} n(D, m) \tag{1.152}$$

where $n(D, m)$ is the number of independent parameters in the term of order m. Now make use of the result (1.141), together with proof by induction, to show that

$$N(D, M) = \frac{(D + M)!}{D!\, M!}. \tag{1.153}$$

This can be done by first proving that the result holds for $M = 0$ and arbitrary $D \geq 1$, then assuming that it holds at order M, and hence showing that it holds at order $M + 1$. Finally, make use of Stirling's approximation in the form

$$n! \simeq n^n e^{-n} \tag{1.154}$$

for large n to show that, for $D \gg M$, the quantity $N(D, M)$ grows like D^M, and for $M \gg D$ it grows like M^D. Consider a cubic ($M = 3$) polynomial in D dimensions, and evaluate numerically the total number of independent parameters for (i) $D = 10$ and (ii) $D = 100$, which correspond to typical small-scale and medium-scale machine learning applications.

67

Proof. We have proved the number of independent parameters in the Mth order term of a D-dimensional polynomial can be written as (1.141). Since $N(D, M)$ represents the total number of independent parameters includeing the 0th order to the Mth order. Therefore, summing over all the orders gives the total number of independent parameters as follows.

$$N(D, M) = \sum_{m=0}^{M} n(D, m) \tag{1.155}$$

We use (1.141), together with proof by induction to show (1.153). For $M = 0$ and arbitrary $D \geq 0$,

$$N(D, 0) = n(D, 0) = \frac{(D + 0 - 1)!}{(D - 1)!\, 0!} = \frac{(D - 1)!}{(D - 1)!\, 0!} = 1 = \frac{(D + 0)!}{D!\, 0!}. \tag{1.156}$$

The case when $M = 0$ obviously holds for arbitrary $D \geq 1$. Assume (1.155) holds at order M, then

$$N(D, M + 1) = \frac{(D + M)!}{D!\, M!} + n(D, M + 1) \tag{1.157}$$

$$= \frac{(D + M)!}{D!\, M!} + \frac{(D + M)!}{(D - 1)!\, (M + 1)!} \tag{1.158}$$

$$= \frac{(D + M)!\, (M + 1)}{D!\, (M + 1)!} + \frac{(D + M)!\, D}{D!\, (M + 1)!} \tag{1.159}$$

$$= \frac{(D + M)!\, (M + 1)}{D!\, (M + 1)!} + \frac{(D + M)!\, D}{D!\, (M + 1)!} \tag{1.160}$$

$$= \frac{(D + M)!\, (D + M + 1)}{D!\, (M + 1)!} = \frac{(D + M + 1)!}{D!\, (M + 1)!} \tag{1.161}$$

68 as desired.

Finally, when $D \gg M$, substituting Stirling's approximation into (1.153) yields

$$
\begin{aligned}
N(D,M) =& \frac{(D+M)!}{D!\,M!} && (1.162)\\
\approx& \frac{(D+M)^{(D+M)}e^{-(D+M)}}{D^D e^{-D} M!} && (1.163)\\
=& \frac{(D+M)^{(D+M)}e^{-M}}{D^D M!} && (1.164)\\
=& \frac{D^M (D+M)^{(D+M)}e^{-M}}{D^{(D+M)} M!} && (1.165)\\
=& \frac{D^M e^{-M}}{M!}\left(1+\frac{M}{D}\right)^{(D+M)} && (1.166)\\
\approx& \frac{D^M e^{-M}}{M!}\left(1+\frac{M(D+M)}{D}\right) && (1.167)\\
\approx& \frac{D^M e^{-M}}{M!}(1+M) && (1.168)
\end{aligned}
$$

Thus, for $D \gg M$, the quantity $N(D,M)$ grows like D^M. Likewise, for $M \gg D$, we have

$$N(D,M) \approx \frac{M^D e^{-D}}{D!}(1+D). \tag{1.169}$$

69 Hence, for $M \gg D$ it grows like M^D. When $M = 3$, we employ (1.153) to get $N(10,3) = 286$ and
70 $N(100,3) = 176851$. □

Exercise 1.17

The gamma function is defined by

$$\Gamma(x) \equiv \int_0^\infty u^{x-1} e^{-u} \mathrm{d}u. \tag{1.170}$$

Using integration by parts, prove the relation $\Gamma(x+1) = x\Gamma(x)$. Show also that $\Gamma(1) = 1$ and hence that $\Gamma(x+1) = x!$ when x is an integer.

71

Proof.

$$
\begin{aligned}
\Gamma(x+1) =& \int_0^\infty u^x e^{-u} \mathrm{d}u && (1.171)\\
=& -\left[u^x e^{-u}\Big|_0^\infty - x\int_0^\infty u^{x-1}e^{-u}\mathrm{d}u \right] && (1.172)\\
=& x\int_0^\infty u^{x-1}e^{-u}\mathrm{d}u = x\Gamma(x). && (1.173)
\end{aligned}
$$

To calculate $\Gamma(1)$, we have

$$\Gamma(1) = \int_0^\infty e^{-u}\mathrm{d}u = -e^{-u}\Big|_0^\infty = 1. \tag{1.174}$$

72 Thus, $\Gamma(x+1) = x!$ when x is an integer. □

Exercise 1.18

We can use the result (1.46) to derive an expression for the surface area S_D, and the volume V_D, of a sphere of a unit radius in D dimensions. To do this, consider the following result, which is obtained by transforming from Cartesian to polar coordinates

$$\prod_{i=1}^{D} \int_{-\infty}^{\infty} e^{-x_i^2} \mathrm{d}x_i = S_D \int_0^{\infty} e^{-r^2} r^{D-1} \mathrm{d}r. \tag{1.175}$$

Using the definition (1.170) of the Gamma function, together with (1.46), evaluate both sides of this equation, and hence show that

$$S_D = \frac{2\pi^{D/2}}{\Gamma(D/2)}. \tag{1.176}$$

Next, by integrating with respect to radius from 0 to 1, show that the volume of the unit sphere in D dimensions is given by

$$V_D = \frac{S_D}{D}. \tag{1.177}$$

Finally, use the results $\Gamma(1) = 1$ and $\Gamma(3/2) = \sqrt{\pi}/2$ to show that (1.176) and (1.177) reduce to the usual expressions for $D = 2$ and $D = 3$.

Note that it is necessary to clarify that S_D is not the surface area and $S_D r^{D-1}$ represents the true surface area. In this exercise, in numerical sense they are equal because they only consider a sphere of a unit radius in D dimensions. If you are interested in this topic, please check out https://zhangyk8.github.io/teaching/file/Exercise_4_insight.pdf.

Proof. Given $\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right) \mathrm{d}x = \sqrt{2\pi\sigma^2}$, we have $\int_{-\infty}^{\infty} \exp\left(-x_i^2\right) \mathrm{d}x = \sqrt{\pi}$.

$$\prod_{i=1}^{D} \int_{-\infty}^{\infty} e^{-x_i^2} \mathrm{d}x_i = \pi^{D/2}. \tag{1.178}$$

Let $u = r^2$, then $\mathrm{d}r = \frac{1}{2\sqrt{u}}\mathrm{d}u$. Then the right hand side can be evaluated as follows.

$$S_D \int_0^{\infty} e^{-r^2} r^{D-1} \mathrm{d}r = S_D \cdot \frac{1}{2} \int_0^{\infty} u^{D/2-1} e^{-u} \mathrm{d}u = \frac{S_D}{2}\Gamma(D/2) = \pi^{D/2} \tag{1.179}$$

which shows

$$S_D = \frac{2\pi^{D/2}}{\Gamma(D/2)}. \tag{1.180}$$

To compute the volume, after observing (1.175), we drop the terms $e^{-x_i^2}$ and e^{-r^2}, and integrate w.r.t radius from 0 to 1 as follows.

$$V_D = \int \cdots \int_{x_1^2+x_2^2+\cdots+x_n^2 \le 1} \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_n = S_D \int_0^1 r^{D-1} \mathrm{d}r = \frac{S_D}{D} \tag{1.181}$$

where an excellent derivation for the integrand r^{D-1} can be found at https://zhangyk8.github.io/teaching/file/Exercise_4_insight.pdf. When $D = 2$, we have

$$S_2 = \frac{2\pi}{\Gamma(1)} = 2\pi, \quad V_2 = \pi. \tag{1.182}$$

And when $D = 3$, we get

$$S_3 = \frac{2\pi^{3/2}}{\Gamma(3/2)} = \frac{2\pi^{3/2}}{\sqrt{\pi}/2} = 4\pi, \quad V_2 = 4\pi/3. \tag{1.183}$$

$\square$

Exercise 1.19

Consider a sphere of radius a in D-dimensions together with the concentric hypercube of side $2a$, so that the sphere touches the hypercube at the centers of each of its sides By using the results of Exercise 1.18, show that the ratio of the volume of the sphere to the volume of the cube is given by

$$\frac{\text{volume of sphere}}{\text{volume of cube}} = \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)}. \tag{1.184}$$

Now make use of Stirling's formula in the form

$$\Gamma(x+1) \approx (2\pi)^{1/2} e^{-x} x^{x+1/2} \tag{1.185}$$

which is valid for $x \gg 1$, to show that, as $D \to \infty$, the ratio (1.184) goes to zero. Show also that the ratio of the distance from the center of the hypercube to one of the corners, divided by the perpendicular distance to one of the sides, is $\sqrt{D}$, which therefore goes to ∞ as $D \to \infty$. From these results we see that, in as space of high dimensionality, most of the volume of a cube concentrated in the large number of corners, which themselves become very long 'spikes'!

Proof. Without loss of generality, we consider the case when $a = 1$. Then we can use the results of Exercise 1.18 directly as follows.

$$\frac{\text{volume of sphere}}{\text{volume of cube}} = \frac{2\pi^{D/2}}{D\Gamma(D/2) \times 2^D} = \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)}. \tag{1.186}$$

By Stirling's formula, namely, (1.185), we get

$$\begin{aligned}
\frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)} &= \frac{\pi^{D/2}}{D2^{D-1}(2\pi)^{1/2}e^{-D/2}(D/2)^{D/2+1/2}} && (1.187)\\
&= \frac{2\sqrt{2}\pi^{D/2}e^{D/2}}{D^{3/2}4^{D/2}(2\pi)^{1/2}(D/2)^{D/2}} && (1.188)\\
&= \frac{2\sqrt{2}}{D^{3/2}(2\pi)^{1/2}} \cdot \left(\frac{\pi e}{2D}\right)^{D/2} && (1.189)
\end{aligned}$$

which goes to 0 as $D \to \infty$. The distance from the center of the hypercube to one of the corners is $\sqrt{1^2 + 1^2 + \cdots + 1^2} = \sqrt{D}$, but the perpendicular distance to one of the sides is 1. Hence, the ratio between them is $\sqrt{D}$. Thus, it goes to ∞ as $D \to \infty$. $\square$

Exercise 1.20

In this exercise, we explore the behavior of the Gaussian distribution in high-dimensional spaces. Consider a Gaussian distribution in D dimensions given by

$$p(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right). \tag{1.190}$$

We wish to find the density with respect to radius in polar coordinates in which the direction variables have been integrated out. To do this, show that the integral of the probability density over a thin shell of radius r and thickness ϵ, where $\epsilon \ll 1$, is given by $p(r)\epsilon$ where

$$p(r) = \frac{S_D r^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \tag{1.191}$$

where S_D is the surface area of a unit sphere in D dimensions. Show that the function $p(r)$ has a single stationary point located, for large D, at $\hat{r} \approx \sqrt{D}\sigma$. By considering $p(\hat{r}+\epsilon)$ where $\epsilon \ll \hat{r}$, show that for large D,

$$p(\hat{r}+\epsilon) = p(\hat{r}) \exp\left(-\frac{\epsilon^2}{\sigma^2}\right) \tag{1.192}$$

which shows that $\hat{r}$ is a maximum of the radial probability density and also that $p(r)$ decays exponentially away from its maximum at $\hat{r}$ with length scale σ. We have already seen that $\sigma \ll \hat{r}$ for large D, and so we see that most of the probability mass is concentrated in a thin shell at large radius. Finally, show that the probability density $p(\mathbf{x})$ is larger at the origin than at the radius $\hat{r}$ by a factor of $\exp(D/2)$. We therefore see that most of the probability mass in a high-dimensional Gaussian distribution is located at a different radius from the region of high probability density. This property of distributions in spaces of high dimensionality will have important consequences when we consider Bayesian inference of model parameters in later chapters.

83

Proof. From Exercise 1.18, the surface area of a sphere of radius r in D dimensions can be represented as $S_D r^{D-1}$. Given this result and $\epsilon \ll 1$, we have

$$\int_{\text{shell}} p(\mathbf{x}) \mathrm{d}\mathbf{x} = p(\mathbf{x}) V_{\text{shell}} = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{\|r\|^2}{2\sigma^2}\right) S_D r^{D-1} \epsilon = p(r)\epsilon. \tag{1.193}$$

which implies that $p(r)$ can be written as

$$p(r) = \frac{S_D r^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right). \tag{1.194}$$

By setting the gradient $\nabla_r p(r)$ to 0, we get

$$\nabla_r p(r) = \frac{S_D}{(2\pi\sigma^2)^{D/2}} \left[(D-1) r^{D-2} \exp\left(-\frac{r^2}{2\sigma^2}\right) + \frac{r^D}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right)\right] = 0 \Longleftrightarrow \hat{r} = \sqrt{D-1}\sigma. \tag{1.195}$$

Therefore, for large D, $\hat{r} \approx \sqrt{D}\sigma$. Provided $\epsilon \ll \hat{r}$, we have

$$\frac{p(\hat{r}+\epsilon)}{p(\hat{r})} = \frac{(r+\epsilon)^{D-1} \exp\left(-\frac{(r+\epsilon)^2}{2\sigma^2}\right)}{r^{D-1} \exp\left(-\frac{r^2}{2\sigma^2}\right)} \tag{1.196}$$

$$= \left(1 + \frac{\epsilon}{\hat{r}}\right)^{D-1} \exp\left(-\frac{2\epsilon\hat{r} + \epsilon^2}{2\sigma^2}\right) \tag{1.197}$$

$$= \exp\left((D-1)\ln(1 + \frac{\epsilon}{\hat{r}}) - \frac{2\epsilon\hat{r} + \epsilon^2}{2\sigma^2}\right) \tag{1.198}$$

$$\approx \exp\left((D-1)(\frac{\epsilon}{\hat{r}} - \frac{\epsilon^2}{2\hat{r}^2}) - \frac{2\epsilon\hat{r} + \epsilon^2}{2\sigma^2}\right) \quad (\ln(1+x) = x - \frac{x^2}{2} + O(x^3)) \tag{1.199}$$

$$= \exp\left(\frac{2\epsilon\hat{r} - \epsilon^2}{2\sigma^2} - \frac{2\epsilon\hat{r} + \epsilon^2}{2\sigma^2}\right) \quad (D - 1 = \frac{\hat{r}^2}{\sigma^2}) \tag{1.200}$$

$$= \exp\left(-\frac{\epsilon^2}{\sigma^2}\right) \tag{1.201}$$

which shows that for large D,

$$p(\hat{r} + \epsilon) = p(\hat{r})\exp\left(-\frac{\epsilon^2}{\sigma^2}\right). \tag{1.202}$$

This indicates that $\hat{r}$ is a maximum of radial probability density and also that $p(r)$ decays exponentially away from its maximum at $\hat{r}$ with length scale σ. At the origin, we have

$$p(\mathbf{0}) = \frac{1}{(2\pi\sigma^2)^{D/2}}. \tag{1.203}$$

At the radius $\hat{r}$, we have

$$p(\|\mathbf{x}\| = \hat{r}) = \frac{1}{(2\pi\sigma^2)^{D/2}}\exp\left(-\frac{\hat{r}^2}{2\sigma^2}\right) \approx \frac{1}{(2\pi\sigma^2)^{D/2}}\exp\left(-\frac{D}{2}\right). \tag{1.204}$$

Thus,

$$\frac{p(\mathbf{0})}{p(\|\mathbf{x}\| = \hat{r})} = \exp\left(\frac{D}{2}\right). \tag{1.205}$$

84 This completes the proof. □

Exercise 1.21

Consider two nonnegative numbers a and b, show that, if $a \leq b$, then $a \leq (ab)^{1/2}$. Use this result to show that, if the decision regions of a two-class classification problem are chosen to minimize the probability of misclassification, this probability will satisfy

$$p(\text{mistake}) \leq \int \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2}\mathrm{d}\mathbf{x}. \tag{1.206}$$

85

Proof. Given $a, b \geq 0$, we have

$$a \leq b \iff a^2 \leq ab \iff a \leq (ab)^{1/2} \tag{1.207}$$

where the last 'iff' follows from the square root function is increasing on $\mathbb{R}_+$. Combining this result and (1.78) in PRML book,

$$p(\text{mistake}) = p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \tag{1.208}$$

$$= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\mathrm{d}\mathbf{x} \tag{1.209}$$

$$\leq \int_{\mathcal{R}_1} \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2}\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2}\mathrm{d}\mathbf{x} \tag{1.210}$$

$$= \int \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2} \mathrm{d}\mathbf{x} \tag{1.211}$$

where the inequality follows from $p(\mathbf{x}, \mathcal{C}_1) > p(\mathbf{x}, \mathcal{C}_2)$ for $\mathbf{x} \in \mathcal{R}_1$ and $p(\mathbf{x}, \mathcal{C}_2) > p(\mathbf{x}, \mathcal{C}_1)$ for $\mathbf{x} \in \mathcal{R}_2$, since the decision regions are chosen to minimize the probability of misclassification. This completes the proof. □

Exercise 1.22

Given a loss matrix with elements L_{kj}, the expected risk is minimized if, for each $\mathbf{x}$, we choose the class that minimizes (1.81) in PRML book

$$\sum_k L_{kj} p(\mathcal{C}_k \mid \mathbf{x}). \tag{1.212}$$

Verify that, when the loss matrix is given by $L_{kj} = 1 - I_{kj}$, where I_{kj} are the elements of the identity matrix, this reduces to the criterion of choosing the class having the largest posterior probability. What is the interpretation of this form of loss matrix?

Proof. Plugging $L_{kj} = 1 - I_{kj}$ into (1.212) gives

$$\begin{aligned}
\sum_k L_{kj} p(\mathcal{C}_k \mid \mathbf{x}) &= \sum_k (1 - I_{kj}) p(\mathcal{C}_k \mid \mathbf{x}) && (1.213)\\
&= \sum_k p(\mathcal{C}_k \mid \mathbf{x}) - \sum_k I_{kj} p(\mathcal{C}_k \mid \mathbf{x}) && (1.214)\\
&= 1 - p(\mathcal{C}_k \mid \mathbf{x}) && (1.215)
\end{aligned}$$

which implies that minimizing $\sum_k L_{kj} p(\mathcal{C}_k \mid \mathbf{x})$ is equivalent to maximizing $p(\mathcal{C}_k \mid \mathbf{x})$.

This kind of loss matrix gives a loss of one if the sample is misclassified and a loss of zero if it is classified correctly. The above proof shows that minimizing the expected risk in the sense of this loss matrix is equivalent to minimizing the misclassification rate. □

Exercise 1.23

Derive the criterion for minimizing the expected loss when there is a general loss matrix and general prior probabilities for the classes.

Solution: From (1.212), minimizing the expected loss is equivalent to minimizing

$$\sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x}) = \frac{1}{p(\mathbf{x})} \sum_k L_{kj} p(\mathbf{x} | \mathcal{C}_k) p(\mathcal{C}_k). \tag{1.216}$$

□

Exercise 1.24

Consider a classification problem in which the loss incurred when an input vector from class $\mathcal{C}_k$ is classified as belonging to class $\mathcal{C}_j$ is given by loss matrix L_{kj}, and for which the loss incurred in selecting the reject option is λ. Find the decision criterion that will give the minimum expected loss. Verify that this reduces to the reject criterion discussed in Section 1.5.3 when the loss matrix is given by $L_{kj} = 1 - I_{kj}$. What is the relationship between λ and the rejection threshold θ?

Proof. From Section 1.5.2, the decision rule that minimizes the expected loss is the one that assigns each new $\mathbf{x}$ to the class j for which the quantity

$$\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}) \tag{1.217}$$

is a minimum. Equivalently, we should assign a new $\mathbf{x}$ to class j for which $j = \operatorname{argmin}_l L_{kl} p(\mathcal{C}_k|\mathbf{x})$ and $L_{kj} p(\mathcal{C}_k|\mathbf{x}) < \lambda$, otherwise we reject $\mathbf{x}$. Given a loss matrix $L_{kj} = 1 - I_{kj}$, according to Exercise 1.22, $\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}) = 1 - p(\mathcal{C}_j)$. If the smallest $1 - p(\mathcal{C}_j) < \lambda$, or equivalently the largest $p(\mathcal{C}_j) > 1 - \lambda$, we assign $\mathbf{x}$ to j, otherwise we reject $\mathbf{x}$. In other words, $1 - \lambda = \theta$. □

Exercise 1.25

Consider the generalization of the squared loss function (1.87) in PRML book for a single target variable t to the case of multiple target variables described by the vector $\mathbf{t}$ given by

$$\mathbb{E}[L(t, y(\mathbf{x}))] = \iint \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{t}. \tag{1.218}$$

Using the calculus of variations, show that the function $\mathbf{y}(\mathbf{x})$ for which this expected loss is minimized is given by $\mathbf{y}(\mathbf{x}) = \mathbb{E}_{\mathbf{t}}[\mathbf{t}|\mathbf{x}]$. Show that this result reduces to $y(\mathbf{x}) = \mathbb{E}_t[t|\mathbf{x}]$ for the case of a single variable target t.

Proof. Our goal is to choose $\mathbf{y}(\mathbf{x})$ so as to minimize $\mathbb{E}[L(t, y(\mathbf{x}))]$. If we assume a completely flexible function $y(\mathbf{x})$, we can do this formally using the calculus of variations to give

$$\frac{\delta \mathbb{E}[L(t, y(\mathbf{x}))]}{\delta \mathbf{y}(\mathbf{x})} = 2 \int (\mathbf{y}(\mathbf{x}) - \mathbf{t}) p(\mathbf{x}, \mathbf{t}) \mathrm{d}\mathbf{t} = 0. \tag{1.219}$$

Solving for $\mathbf{y}(\mathbf{x})$, and using the sum and product rules of probability, we obtain

$$\mathbf{y}(\mathbf{x}) = \frac{\int \mathbf{t} p(\mathbf{x}, \mathbf{t}) \mathrm{d}\mathbf{t}}{p(\mathbf{x})} = \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) \mathrm{d}\mathbf{t} = \mathbb{E}_{\mathbf{t}}[\mathbf{t}|\mathbf{x}] \tag{1.220}$$

which is the conditional average of $\mathbf{t}$ conditioned on $\mathbf{x}$. When t is a scalar,

$$y(\mathbf{x}) = \frac{\int t p(\mathbf{x}, t) \mathrm{d}t}{p(\mathbf{x})} = \int t p(t|\mathbf{x}) \mathrm{d}t = \mathbb{E}_t[t|\mathbf{x}] \tag{1.221}$$

which is equivalent to (1.87) in the PRML book. □

Exercise 1.26

By expansion of the square in (1.218), derive a result analogous to (1.90) in the PRML book, which is

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}_t[t|\mathbf{x}]\} p(\mathbf{x}) \mathrm{d}\mathbf{x} + \int \operatorname{var}[t|\mathbf{x}] p(\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{1.222}$$

Using the calculus of variations, show that the function $\mathbf{y}(\mathbf{x})$ that minimizes the expected squared loss for the case of a vector $\mathbf{t}$ of target variables is again given by the conditional expectation of $\mathbf{t}$.

Proof. Following Section 1.5.5, we have

$$\|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 = \|\mathbf{y}(\mathbf{x}) - \mathbb{E}(\mathbf{t}|\mathbf{x}) + \mathbb{E}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 \tag{1.223}$$

$$=\|\mathbf{y}(\mathbf{x}) - \mathbb{E}(\mathbf{t}|\mathbf{x})\|^2 + (\mathbf{y}(\mathbf{x}) - \mathbb{E}(\mathbf{t}|\mathbf{x}))^T(\mathbb{E}(\mathbf{t}|\mathbf{x}) - \mathbf{t}) \tag{1.224}$$

$$+ (\mathbb{E}(\mathbf{t}|\mathbf{x}) - \mathbf{t})^T(\mathbf{y}(\mathbf{x}) - \mathbb{E}(\mathbf{t}|\mathbf{x})) + \|\mathbb{E}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2. \tag{1.225}$$

Substituting this into (1.218) and performing the integral over $\mathbf{t}$, we see that the cross term vanishes and the last term is exactly the definition of variance $\mathbb{E}[\mathbf{t}|\mathbf{x}]$. Thus,

$$\mathbb{E}[L] = \int \{\mathbf{y}(\mathbf{x}) - \mathbb{E}_{\mathbf{t}}[\mathbf{t}|\mathbf{x}]\}p(\mathbf{x})\mathrm{d}\mathbf{x} + \int \mathrm{var}[\mathbf{t}|\mathbf{x}]p(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{1.226}$$

$\square$

Exercise 1.27

Consider the expected loss for regression problems under the L_q loss function given by (1.91) in the PRML book, i.e.,

$$\mathbb{E}\left[L_q\right] = \iint |y(\mathbf{x}) - t|^q p(\mathbf{x}, t)\mathrm{d}\mathbf{x}\mathrm{d}t. \tag{1.227}$$

Write down the condition that $y(\mathbf{x})$ must satisfy in order to minimize $\mathbb{E}\left[L_q\right]$. Show that, for $q = 1$, this solution represents the conditional median, i.e., the function $y(\mathbf{x})$ such that the probability mass for $t < y(\mathbf{x})$ is the same as for $t \geq y(\mathbf{x})$. Also show that the minimum expected L_q loss for $q \to 0$ is given by the conditional mode, i.e., by the function $y(\mathbf{x})$ equal to the value of t that maximizes $p(t|\mathbf{x})$ for each $\mathbf{x}$.

Proof. We follow the logic from the official solution manual. (1.227) can be rewritten as

$$\mathbb{E}\left[L_q\right] = \int \left(\int |y(\mathbf{x}) - t|^q p(t|\mathbf{x})\mathrm{d}t\right) p(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{1.228}$$

Since $y(\mathbf{x})$ can be chosen independently for each $\mathbf{x}$, the minimum of $\mathbb{E}[L_q]$ can be obtained via minimizing the following integrand

$$\int |y(\mathbf{x}) - t|^q p(t|\mathbf{x})\mathrm{d}t \tag{1.229}$$

for each $\mathbf{x}$. Setting the derivative to 0 yields

$$q \int |y(\mathbf{x}) - t|^{q-1}\mathrm{sign}(y(\mathbf{x}) - t)p(t|\mathbf{x})\mathrm{d}t = 0. \tag{1.230}$$

Then,

$$\int_{y(\mathbf{x})}^{\infty} |y(\mathbf{x}) - t|^{q-1}p(t|\mathbf{x})\mathrm{d}t = \int_{-\infty}^{y(\mathbf{x})} |y(\mathbf{x}) - t|^{q-1}p(t|\mathbf{x})\mathrm{d}t. \tag{1.231}$$

When $q = 1$, we have

$$\int_{y(\mathbf{x})}^{\infty} p(t|\mathbf{x})\mathrm{d}t = \int_{-\infty}^{y(\mathbf{x})} p(t|\mathbf{x})\mathrm{d}t \tag{1.232}$$

which implies that the conditional median is the solution for the case of $q = 1$. As $q \to 0$, $|y(\mathbf{x}) - t|^q$ is close to 1 except in a small neighborhood around $y(\mathbf{x})$ where it falls to 0. Therefore, the value of (1.229) is close to 1. Normally, $p(t|\mathbf{x})$ is normalized, but in this case there is a notch at $t = y(\mathbf{x})$. Therefore, the value of (1.229) is slightly reduced from 1. To minimize (1.229), we can choose $y(\mathbf{x})$ to coincide with the maximum $p(t|\mathbf{x})$, namely, the conditional mode. $\square$

Exercise 1.28

In Section 1.6, we introduced the idea of entropy $h(x)$ as the information gained on observing the value of a random variable x having distribution $p(x)$. We saw that, for independent variables x and y for which $p(x, y) = p(x)p(y)$, the entropy functions are additive, so that $h(x, y) = h(x) + h(y)$. In this exercise, we derive the relation between h and p in the form of a function $h(p)$. First show that $h(p^2) = 2h(p)$, and hence by induction that $h(p^n) = nh(p)$ where n is a positive integer. Hence show that $h(p^{n/m}) = (n/m)h(p)$ where m is also a positive integer. This implies that $h(p^x) = xh(p)$ where x is a positive rational number, and hence by continuity when it is a positive real number. Finally, show that this implies $h(p)$ must take the form $h(p) \propto \ln p$.

111

Proof. Since $h(x, y) = h(x) + h(y)$, we have $h(p^2) = h(p, p) = h(p) + h(p) = 2h(p)$. Then suppose that $h(p^{n-1}) = (n-1)h(p)$, and we get $h(p^n) = h(p^{n-1}) + h(p) = nh(p)$. With this, we have

$$h(p^{n/m}) = \frac{mh(p^{n/m})}{m} = \frac{h(p^n)}{m} = \frac{nh(p)}{m}. \tag{1.233}$$

Hence, by continuity, $h(p^x) = xh(p)$ holds for any positive real number. Suppose $p = q^x$, then we have

$$\frac{h(p)}{\ln p} = \frac{h(q^x)}{\ln q^x} = \frac{xh(q)}{x \ln q} = \frac{h(q)}{\ln q} \tag{1.234}$$

112 which implies $h(p) \propto \ln p$. □

Exercise 1.29

Consider an M-state discrete random variable x, and use Jensen's inequality in the form of

$$f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \le \sum_{i=1}^{M} \lambda_i f(x_i) \tag{1.235}$$

where $\lambda_i \ge 0$ and $\sum_i \lambda_i = 1$, for any set of points $\{x_i\}$, to show that the entropy of its distribution $p(x)$ satisfies $H[x] \le \ln M$.

113

Proof. By the definition of entropy, we have

$$H[x] = -\sum_{i=1}^{M} p(x_i) \ln p(x_i) = \sum_{i=1}^{M} p(x_i) \ln \frac{1}{p(x_i)}. \tag{1.236}$$

Since $\ln \frac{1}{x}$ is concave, by Jensen's inequality, we get

$$H[x] = \sum_{i=1}^{M} p(x_i) \ln \frac{1}{p(x_i)} \le \ln \sum_{i=1}^{M} p(x_i) \frac{1}{p(x_i)} = \ln M. \tag{1.237}$$

114 which completes the proof. □

Exercise 1.30

Evaluate the Kullback-Leibler divergence

$$\mathrm{KL}(p||q) = \int p(x) \ln \frac{p(x)}{q(x)} \mathrm{d}x \tag{1.238}$$

between two Gaussians, $p(x) = \mathcal{N}(x|\mu, \sigma^2)$ and $q(x) = \mathcal{N}(x|m^2, s^2)$.

115

Solution:

$$\begin{aligned}
\mathrm{KL}(p||q) &= \int \mathcal{N}(x|\mu,\sigma^2) \ln \frac{\mathcal{N}(x|\mu,\sigma^2)}{\mathcal{N}(x|m^2,s^2)} \mathrm{d}x && (1.239)\\
&= \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \ln \frac{\sqrt{s^2/\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{e^{-\frac{(x-m)^2}{2s^2}}} \mathrm{d}x && (1.240)\\
&= \frac{1}{2} \ln \frac{s^2}{\sigma^2} + \int \left(\frac{(x-m)^2}{2s^2} - \frac{(x-\mu)^2}{2\sigma^2} \right) \mathcal{N}(x|\mu,\sigma^2) \mathrm{d}x && (1.241)\\
&= \frac{1}{2} \ln \frac{s^2}{\sigma^2} + \int \frac{(x-m)^2}{2s^2} \mathcal{N}(x|\mu,\sigma^2) \mathrm{d}x - \frac{\sigma^2}{2\sigma^2} && (1.242)\\
&= \frac{1}{2} \ln \frac{s^2}{\sigma^2} + \int \frac{x^2 - 2mx + m^2}{2s^2} \mathcal{N}(x|\mu,\sigma^2) \mathrm{d}x - \frac{1}{2} && (1.243)\\
&= \frac{1}{2} \ln \frac{s^2}{\sigma^2} + \frac{\mu^2 + \sigma^2 - 2m\mu + m^2}{2s^2} - \frac{1}{2}. && (1.244)
\end{aligned}$$

116 □

Exercise 1.31

Consider two variables x and y having joint distribution $p(\mathbf{x}, \mathbf{y})$. Show that the differential entropy of this pair of variables satisfies

$$\mathrm{H}[\mathbf{x}, \mathbf{y}] \le \mathrm{H}[\mathbf{x}] + \mathrm{H}[\mathbf{y}] \tag{1.245}$$

with equality if, and only if, $\mathbf{x}$ and $\mathbf{y}$ are statistically independent.

117

Proof. From Exercise 1.41, we have $\mathrm{I}(\mathbf{x}, \mathbf{y}) = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]$. Since the mutual information is a form of KL divergence, $\mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}] \ge 0$ which implies $\mathrm{H}[\mathbf{y}] \ge \mathrm{H}[\mathbf{y}|\mathbf{x}]$. With the relation $\mathrm{H}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}]$ which will be shown in Exercise 1.37, we have

$$\mathrm{H}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}] \le \mathrm{H}[\mathbf{y}] + \mathrm{H}[\mathbf{x}]. \tag{1.246}$$

If $\mathbf{x}$ and $\mathbf{y}$ are statistically independent, we have $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$. Thus,

$$\begin{aligned}
\mathrm{H}(\mathbf{x}, \mathbf{y}) &= -\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} && (1.247)\\
&= -\iint p(\mathbf{x})p(\mathbf{y}) \ln p(\mathbf{x})p(\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} && (1.248)\\
&= -\iint p(\mathbf{x})p(\mathbf{y}) \ln p(\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} - \iint p(\mathbf{x})p(\mathbf{y}) \ln p(\mathbf{y}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} && (1.249)\\
&= \mathrm{H}[\mathbf{x}] + \mathrm{H}[\mathbf{y}] && (1.250)
\end{aligned}$$

which shows the sufficiency. For the necessity, if the equality in (1.246) holds, then $\mathrm{H}[\mathbf{y}|\mathbf{x}] = \mathrm{H}[\mathbf{y}]$. According to Exercise 1.41, $\mathrm{I}(\mathbf{x}, \mathbf{y}) = 0$. By the definition of $\mathrm{I}(\mathbf{x}, \mathbf{y})$, the KL divergence of two distributions is 0 if and only if the two distributions are identical, i.e. $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$. In other words, $\mathbf{x}$ and $\mathbf{y}$ are statistically independent. This completes the proof. □

Exercise 1.32

Consider a vector $\mathbf{x}$ of continuous variables with distribution $p(\mathbf{x})$ and corresponding entropy $\mathrm{H}[\mathbf{x}]$. Suppose that we make a nonsingular linear transformation of $\mathbf{x}$ to obtain a new variable $\mathbf{y} = \mathbf{A}\mathbf{x}$. Show that the corresponding entropy is given by $\mathrm{H}[\mathbf{y}] = \mathrm{H}[\mathbf{x}] + \ln|\mathbf{A}|$ where $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$.

Proof. By the change of variables, we have

$$p(\mathbf{x}) = p(\mathbf{y})|\mathbf{A}|. \tag{1.251}$$

Then,

$$\begin{aligned}
\mathrm{H}[\mathbf{y}] &= -\int p(\mathbf{y}) \ln p(\mathbf{y}) \mathrm{d}\mathbf{y} && (1.252)\\
&= -\int p(\mathbf{x}) \ln p(\mathbf{x}) |\mathbf{A}|^{-1} \mathrm{d}\mathbf{x} && (1.253)\\
&= -\int p(\mathbf{x})(\ln p(\mathbf{x}) - \ln|\mathbf{A}| \mathrm{d}\mathbf{x} && (1.254)\\
&= -\int p(\mathbf{x}) \ln p(\mathbf{x}) \mathrm{d}\mathbf{x} + \int p(\mathbf{x}) \ln|\mathbf{A}| \mathrm{d}\mathbf{x} && (1.255)\\
&= \mathrm{H}[\mathbf{x}] + \ln|\mathbf{A}| && (1.256)
\end{aligned}$$

where the second equality we used $p(\mathbf{y})\mathrm{d}\mathbf{y} = p(\mathbf{x})\mathrm{d}\mathbf{x}$. This completes the proof. □

Exercise 1.33

Suppose that the conditional entropy $\mathrm{H}[y|x]$ between two discrete random variables x and y is zero. Show that, for all values of x such that $p(x) > 0$, the variable y must be a function of x, in other words for each x there is only one value of y such that $p(y|x) = 0$.

Proof. By definition, the conditional entropy $\mathrm{H}[y|x]$ can be written as

$$\mathrm{H}[y|x] = \sum_{i=1}^{n} \sum_{j=1}^{n} -p(y_j|x_i)p(x_i) \ln p(y_j|x_i). \tag{1.257}$$

It is easy to see that each summand is nonnegative. Since $\mathrm{H}[y|x] = 0$, each summand should be 0. For each $p(x_i) \neq 0$, $-p(y_j|x_i)p(x_i) \ln p(y_j|x_i) = 0$ if and only if $p(y_j|x_i)$ is equal to either 0 or 1. By the constraint of $\sum_{j=1}^{n} p(y_j|x_i) = 1$, there exists only one y_j such that $p(y_j|y_i) = 1$ for each $p(x_i) \neq 0$, as desired. □

Exercise 1.34

Use the calculus of variations to show that the stationary point of the functional

$$- \int_{-\infty}^{\infty} p(x) \ln p(x) \mathrm{d}x + \lambda_1 \left(\int_{-\infty}^{\infty} p(x) \mathrm{d}x - 1 \right) \tag{1.258}$$

$$+ \lambda_2 \left(\int_{-\infty}^{\infty} x p(x) \mathrm{d}x - \mu \right) + \lambda_3 \left(\int_{-\infty}^{\infty} (x-\mu)^2 p(x) \mathrm{d}x - \sigma^2 \right) \tag{1.259}$$

is given by

$$p(x) = \exp\{-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2\}. \tag{1.260}$$

Then use the constraints

$$\int_{-\infty}^{\infty} p(x) \mathrm{d}x = 1 \tag{1.261}$$

$$\int_{-\infty}^{\infty} x p(x) \mathrm{d}x = \mu \tag{1.262}$$

$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x) \mathrm{d}x = \sigma^2 \tag{1.263}$$

to eliminate the Lagrange multipliers and hence show that the maximum entropy solution is given by the Gaussian

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}. \tag{1.264}$$

129

Proof. In essence, this problem is to optimize a functional with respect to a probability distribution. As provided in the Appendix D of the PRML book, we denote the functional derivative of $F[f]$ with respect to $f(x)$ by $\delta F / \delta f(x)$, and define it by the following relation:

$$F[y(x) + \epsilon \eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)} \eta(x) \mathrm{d}x + O(\epsilon^2). \tag{1.265}$$

Consider the functional

$$I(p(x)) = \int p(x) f(x) \mathrm{d}x. \tag{1.266}$$

Under a small variation $p(x) \to p(x) + \epsilon \eta(x)$,

$$I(p(x) + \epsilon \eta(x)) = \int p(x) f(x) \mathrm{d}x + \epsilon \int f(x) p(x) \mathrm{d}x. \tag{1.267}$$

Comparing with (1.265), we deduce that

$$\frac{\delta I}{\delta p(x)} = f(x). \tag{1.268}$$

Now we consider

$$J[p(x)] = \int p(x) \ln p(x) \mathrm{d}x. \tag{1.269}$$

Under a small variation $p(x) \to p(x) + \epsilon \eta(x)$,

$$J[p(x) + \epsilon \eta(x)] = \int (p(x) + \epsilon \eta(x)) \ln (p(x) + \epsilon \eta(x)) \, \mathrm{d}x \tag{1.270}$$

$$= \int (p(x) + \epsilon\eta(x)) \left(\ln p(x) + \frac{\epsilon\eta(x)}{p(x)}) + O(\epsilon^2) \right) \mathrm{d}x \tag{1.271}$$

$$= \int (p(x) \ln p(x) + \epsilon\eta(x)) (\ln p(x) + \epsilon\eta(x)) \,\mathrm{d}x + O(\epsilon^2) \tag{1.272}$$

$$= \int (p(x) \ln p(x) \mathrm{d}x + \epsilon \int (\ln p(x) + 1)\eta(x) \mathrm{d}x + O(\epsilon^2) \tag{1.273}$$

where the second equality follows from the Taylor expansion of $\ln(p(x)+\epsilon\eta(x))$ at $p(x)$. From (1.265), we deduce that

$$\frac{\delta J}{\delta p(x)} = 1 + \ln p(x). \tag{1.274}$$

Therefore, the derivative of the functional with respect to $p(x)$ is

$$-(1 + \ln p(x)) + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 = 0. \tag{1.275}$$

Setting it to 0 gives

$$p(x) = \exp\{-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2\}. \tag{1.276}$$

By comparing it to a Gaussian with mean of μ and variance of σ^2, we have

$$e^{-1+\lambda_1} = \frac{1}{\sqrt{2\pi\sigma^2}}, \quad \lambda_2 = 0, \quad \lambda_3 = -\frac{1}{2\sigma^2} \tag{1.277}$$

which results in

$$\lambda_1 = 1 - \frac{1}{2}\ln 2\pi\sigma^2, \quad \lambda_2 = 0, \quad \lambda_3 = -\frac{1}{2\sigma^2}. \tag{1.278}$$

Just now we cheated since we supposed $p(x)$ is a Gaussian. Now we derive $p(x)$ analytically. The exponent of (1.276) can be written as

$$-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 = \lambda_3 x^2 - (2\mu\lambda_3 - \lambda_2)x + \lambda_3\mu^2 + \lambda_1 - 1 \tag{1.279}$$

$$= ax^2 - bx + c \tag{1.280}$$

$$= a(x-d)^2 + f \tag{1.281}$$

where $a = \lambda_3$, $b = 2\mu\lambda_3 - \lambda_2$, $c = \lambda_3\mu^2 + \lambda_1 - 1$, $d = \frac{b}{2a}$ and $f = c - \frac{b^2}{4a}$. Now we plug $p(x) = e^{a(x-d)^2+f}$ into the three constraints one by one.

$$\int_{-\infty}^{\infty} p(x)\mathrm{d}x = \int_{-\infty}^{\infty} e^{a(x-d)^2+f}\mathrm{d}x = e^f \int_{-\infty}^{\infty} e^{a(x-d)^2}\mathrm{d}x = 1 \tag{1.282}$$

Let $u = x - d$, then $\mathrm{d}x = \mathrm{d}u$.

$$e^f \int_{-\infty}^{\infty} e^{a(x-d)^2}\mathrm{d}x = e^f \int_{-\infty}^{\infty} e^{au^2}\mathrm{d}u = 2e^f \int_{0}^{\infty} e^{au^2}\mathrm{d}u = 1 \tag{1.283}$$

where the second last equality follows from that the integrand is an even function. Let $au^2 = -t$, then $u = \sqrt{t/(-a)}$ and $\mathrm{d}u = \mathrm{d}t/(2\sqrt{-at})$.

$$2e^f \int_0^{\infty} e^{au^2}\mathrm{d}u = 2e^f \int_0^{\infty} \frac{t^{-1/2}}{2\sqrt{-a}} e^{-t}\mathrm{d}t = \frac{e^f}{\sqrt{-a}} \int_0^{\infty} t^{\frac{1}{2}-1} e^{-t}\mathrm{d}t = \frac{e^f}{\sqrt{-a}}\Gamma(\frac{1}{2}) = 1 \tag{1.284}$$

where $\Gamma(z) = \int_0^{\infty} t^{z-1}e^{-z}\mathrm{d}z$. For non-negative integers of z, we have

$$\Gamma(n) = \frac{(2n)!}{4^n n!}\sqrt{\pi}. \tag{1.285}$$

Now we consider the second constraint.

$$\int_{-\infty}^{\infty} xp(x)\mathrm{d}x = \int_{-\infty}^{\infty} xe^{a(x-d)^2+f}\mathrm{d}x = e^f \int_{-\infty}^{\infty} xe^{a(x-d)^2}\mathrm{d}x = \mu. \tag{1.286}$$

Let $u = x - d$, then $x = u + d$ and $\mathrm{d}x = \mathrm{d}u$.

$$e^f \int_{-\infty}^{\infty} xe^{a(x-d)^2}\mathrm{d}x = e^f \int_{-\infty}^{\infty} (u+d)e^{au^2}\mathrm{d}u \tag{1.287}$$

$$= e^f \int_{-\infty}^{\infty} ue^{au^2}\mathrm{d}u + e^f \int_{-\infty}^{\infty} de^{au^2}\mathrm{d}u \tag{1.288}$$

$$= 2de^f \int_{0}^{\infty} e^{au^2}\mathrm{d}u \tag{1.289}$$

$$= d = \mu \tag{1.290}$$

where the second last line follows from that ue^{au^2} is an odd function and de^{au^2} is an even function and the last line follows from (1.284). Thus,

$$d = \frac{b}{2a} = \frac{2\mu\lambda_3 - \lambda_2}{2\lambda_3} = \mu \tag{1.291}$$

which implies $\lambda_2 = 0$. With $d = \mu$ and $\lambda_2 = 0$, we move on to the third constraint.

$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x)\mathrm{d}x = e^f \int_{-\infty}^{\infty} (x-\mu)^2 e^{a(x-\mu)^2}\mathrm{d}x = \sigma^2. \tag{1.292}$$

Let $u = x - \mu$, then $\mathrm{d}x = \mathrm{d}u$.

$$e^f \int_{-\infty}^{\infty} (x-\mu)^2 e^{a(x-\mu)^2}\mathrm{d}x = e^f \int_{-\infty}^{\infty} u^2 e^{au^2}\mathrm{d}u \tag{1.293}$$

$$= 2e^f \int_{0}^{\infty} u^2 e^{au^2}\mathrm{d}u \tag{1.294}$$

$$= \sigma^2. \tag{1.295}$$

Let $-t = au^2$, then $u = \sqrt{-t/a}$ and $\mathrm{d}u = \mathrm{d}t/(2\sqrt{-at})$. With this,

$$2e^f \int_{0}^{\infty} u^2 e^{au^2}\mathrm{d}u = 2e^f \int_{0}^{\infty} -\frac{t}{2a\sqrt{-at}}e^{-t}\mathrm{d}t \tag{1.296}$$

$$= -\frac{e^f}{a\sqrt{-a}} \int_{0}^{\infty} \sqrt{t}e^{-t}\mathrm{d}t \tag{1.297}$$

$$= -\frac{e^f}{a\sqrt{-a}} \int_{0}^{\infty} t^{\frac{3}{2}-1}e^{-t}\mathrm{d}t \tag{1.298}$$

$$= -\frac{e^f}{a\sqrt{-a}}\Gamma(\frac{3}{2}) = -\frac{e^f}{a\sqrt{-a}}\Gamma(1+\frac{1}{2}) \tag{1.299}$$

$$= -\frac{e^f}{2a\sqrt{-a}}\Gamma(\frac{1}{2}) \tag{1.300}$$

$$= -\frac{1}{2a} = \sigma^2 \tag{1.301}$$

where the second last line follows from (1.284). This indicates $a = \lambda_3 = -\frac{1}{2\sigma^2}$. From (1.284), we have

$$\frac{e^f}{\sqrt{-a}}\Gamma(\frac{1}{2}) = 1 \Rightarrow e^f = \frac{\sqrt{-\lambda_3}}{\sqrt{\pi}} = \frac{1}{\sqrt{2\pi\sigma^2}} \Rightarrow f = -\frac{1}{2}\ln 2\pi\sigma^2 \tag{1.302}$$

Since $f = c - \frac{b^2}{4a} = \lambda_1 - 1$, we have

$$\lambda_1 = 1 - \frac{1}{2}\ln 2\pi\sigma^2. \tag{1.303}$$

Now we can plug λ_1, λ_2, and λ_3 into (1.276). Then we get

$$p(x) = \exp\{-1 + 1 - \frac{1}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}(x-\mu)^2\} = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{1.304}$$

130 □

Exercise 1.35

Use the results (1.261) and (1.263) to show that the entropy of the univariate Gaussian (1.264) is given by

$$\mathrm{H}[x] = \frac{1}{2}\left\{1 + \ln(2\pi\sigma^2)\right\}. \tag{1.305}$$

131

Proof. By definition, the entropy of the univariate Gaussian can be computed as follows.

$$\begin{aligned}
\mathrm{H}(x) &= -\int_{-\infty}^{\infty} p(x)\ln p(x)\mathrm{d}x && (1.306)\\
&= -\int_{-\infty}^{\infty} p(x)\ln\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x && (1.307)\\
&= -\int_{-\infty}^{\infty} p(x)\left(-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x && (1.308)\\
&= \frac{1}{2}\ln(2\pi\sigma^2)\int_{-\infty}^{\infty} p(x)\mathrm{d}x + \int_{-\infty}^{\infty}\frac{(x-\mu)^2}{2\sigma^2}p(x)\mathrm{d}x && (1.309)\\
&= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{\sigma^2}{2\sigma^2} && (1.310)\\
&= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2} = \frac{1}{2}\left\{1 + \ln(2\pi\sigma^2)\right\} && (1.311)
\end{aligned}$$

132 where the two terms in the second last line follow from (1.261) and (1.263), respectively. □

Exercise 1.36

A strictly convex function is defined as one for which every chord lies above the function. Show that this is equivalent to the condition that the second derivative of the function be positive.

133

Proof. [4] Let us first formalize the definition of a strictly convex function. Given a function $f(x)$, it is strictly convex if and only if for any two points x_1 and x_2 in the domain of f,

$$f((1-\lambda)x_1 + \lambda x_2) < (1-\lambda)f(x_1) + \lambda f(x_2) \tag{1.312}$$

where $\lambda \in (0, 1)$. Let $x = (1-\lambda)x_1 + \lambda x_2$, then

$$\begin{aligned}
(1-\lambda)f(x_1) + \lambda f(x_2) - f(x) &> 0 && (1.313)\\
\lambda(f(x_2) - f(x)) - (1-\lambda)(f(x) - f(x_1)) &> 0 && (1.314)\\
\lambda f'(\beta)(x_2 - x) - (1-\lambda)f'(\alpha)(x - x_1) &> 0 && (1.315)
\end{aligned}$$

[4] https://math.stackexchange.com/questions/513887/second-derivative-positive-implies-convex

$$\lambda(1-\lambda)(x_2-x_1)(f'(\beta)-f'(\alpha)) > 0 \tag{1.316}$$
$$f'(\beta)-f'(\alpha) > 0 \tag{1.317}$$

where $\alpha \in (x_1, x), \beta \in (x, x_2)$ and the second last line follows from the mean value theorem. This indicates f' is an increasing function. Equivalently, $f'' > 0$. Conversely, the argument still holds. This completes the proof. □

Exercise 1.37

Using the definition of the conditional entropy of $\mathbf{y}$ given $\mathbf{x}$

$$\mathrm{H}(\mathbf{y}|\mathbf{x}) = -\iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \tag{1.318}$$

together with the product rule of probability, prove the result

$$\mathrm{H}(\mathbf{x},\mathbf{y}) = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}]. \tag{1.319}$$

Proof. By the definition of the conditional entropy, we have

$$\begin{aligned}
\mathrm{H}(\mathbf{y}|\mathbf{x}) &= -\iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} && (1.320)\\
&= -\iint p(\mathbf{x},\mathbf{y}) \ln \frac{p(\mathbf{y},\mathbf{x})}{p(\mathbf{x})} \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} && (1.321)\\
&= -\iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y},\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} + \iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} && (1.322)\\
&= \mathrm{H}[\mathbf{x},\mathbf{y}] - \mathrm{H}[\mathbf{x}]. && (1.323)
\end{aligned}$$

After rearrangement, we get

$$\mathrm{H}[\mathbf{x},\mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}]. \tag{1.324}$$

□

Exercise 1.38

Using proof by induction, show that the inequality

$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda) f(b) \tag{1.325}$$

for convex functions implies the result

$$f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leq \sum_{i=1}^{M} \lambda_i f(x_i) \tag{1.326}$$

where $\lambda_i \leq 0$ and $\sum_i \lambda_i = 1$, for any set of points $\{x_i\}$.

Proof. We prove the Jensen's inequality by induction. The base case when $M = 2$ is true. Assume that (1.326) holds for $M = m$, then we consider the case when $M = m + 1$. Let $s = 1 - \lambda_{m+1}$, then we have

$$f\left(\sum_{i=1}^{m+1} \lambda_i x_i\right) = f\left(s \sum_{i=1}^{m} \frac{\lambda_i}{s} x_i + \lambda_{i+1} x_{i+1}\right) \leq s f\left(\sum_{i=1}^{m} \frac{\lambda_i}{s} x_i\right) + \lambda_{i+1} f(x_{i+1}) \tag{1.327}$$

		y	
		0	1
x	0	1/3	1/3
	1	0	1/3

Table 1: The joint distribution $p(x, y)$ for two binary variables x and y used in Exercise 1.39.

$$\leq s \sum_{i=1}^{m} \frac{\lambda_i}{s} f(x_i) + \lambda_{i+1} f(x_{i+1}) \tag{1.328}$$

$$= \sum_{i=1}^{m+1} \lambda_i f(x_i) \tag{1.329}$$

where the first inequality follows from the base case and the second inequality follows from the inductive hypothesis. This completes the proof. □

Exercise 1.39

Consider two binary variables x and y having the joint distribution given in Table 1. Evaluate the following quantities

$$\textbf{(a)}\ \mathrm{H}[x] \quad \textbf{(c)}\ \mathrm{H}[y|x] \quad \textbf{(e)}\ \mathrm{H}[x,y] \tag{1.330}$$

$$\textbf{(b)}\ \mathrm{H}[y] \quad \textbf{(d)}\ \mathrm{H}[x|y] \quad \textbf{(f)}\ \mathrm{I}[x,y]. \tag{1.331}$$

Draw a diagram to show the relationship between these various quantities.

Solution: From Table 1, we have $p(x=0) = 2/3$, $p(x=1) = 1/3$, $p(y=0) = 1/3$, $p(y=1) = 2/3$, $p(x=0|y=1) = 1/2$, $p(x=1|y=1) = 1/2$, $p(y=0|x=1) = 0$, $p(y=1|x=1) = 1$, $p(x=0|y=0) = 1$, $p(x=1|y=0) = 0$, $p(x=0|y=1) = 1/2$, and $p(x=1|y=1) = 1/2$.

(a) $\mathrm{H}[x] = -p(x=0)\ln p(x=0) - p(x=1)\ln p(x=1) = -\frac{2}{3}\ln\frac{2}{3} - \frac{1}{3}\ln\frac{1}{3} = -\frac{1}{3}\ln\frac{4}{27}$.

(b) $\mathrm{H}[y] = -p(y=0)\ln p(y=0) - p(y=1)\ln p(y=1) = -\frac{1}{3}\ln\frac{1}{3} - \frac{2}{3}\ln\frac{2}{3} = -\frac{1}{3}\ln\frac{4}{27}$.

(c) $\mathrm{H}[y|x] = -p(y=0,x=0)\ln p(y=0|x=0) - p(y=1,x=0)\ln p(y=1|x=0) - p(y=0,x=1)\ln p(y=0|x=1) - p(y=1,x=1)\ln p(y=1|x=1) = -\frac{1}{3}\ln\frac{1}{2} - \frac{1}{3}\ln\frac{1}{2} - 0 - 0 = \frac{1}{3}\ln 4$.

(d) $\mathrm{H}[x|y] = -p(y=0,x=0)\ln p(x=0|y=0) - p(y=0,x=1)\ln p(x=1|y=0) - p(x=0,y=1)\ln p(x=0|y=1) - p(y=1,x=1)\ln p(x=1|y=1) = -\frac{1}{3}\ln 1 - 0\ln 0 - \frac{1}{3}\ln\frac{1}{2} - \frac{1}{3}\ln\frac{1}{2} = \frac{1}{3}\ln 4$.

(e) $\mathrm{H}[x,y] = -p(y=0,x=0)\ln p(x=0,y=0) - p(y=0,x=1)\ln p(x=1,y=0) - p(x=0,y=1)\ln p(x=0,y=1) - p(y=1,x=1)\ln p(x=1,y=1) = -\frac{1}{3}\ln\frac{1}{3} - 0\ln 0 - \frac{1}{3}\ln\frac{1}{3} - \frac{1}{3}\ln\frac{1}{3} = \ln 3$.

(f) $\mathrm{I}[x,y] = -p(y=0,x=0)\ln p(x=0)p(y=0)/p(x=0,y=0) - p(y=0,x=1)\ln p(x=1)p(y=0)/p(x=1,y=0) - p(x=0,y=1)\ln p(x=0)p(y=1)/p(x=0,y=1) - p(y=1,x=1)\ln p(x=1)p(y=1)/p(x=1,y=1) = -\frac{1}{3}\ln\frac{2}{3}\cdot\frac{1}{3}/\frac{1}{3} - 0 - \frac{1}{3}\ln\frac{2}{3}\cdot\frac{1}{3}/\frac{1}{3} - \frac{1}{3}\ln\frac{2}{3}\cdot\frac{2}{3}/\frac{1}{3} = -\frac{1}{3}\ln\frac{16}{27}$.

The relationship between these quantities:

$$\mathrm{H}(\mathbf{x},\mathbf{y}) = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}] = \mathrm{H}[\mathbf{x}|\mathbf{y}] + \mathrm{H}[\mathbf{y}] \tag{1.332}$$

$$\mathrm{I}(\mathbf{x},\mathbf{y}) = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]. \tag{1.333}$$

□

Exercise 1.40

By applying Jensen's inequality (1.326) with $f(x) = \ln x$, show that the arithmetic mean of a set of real numbers is never less than their geometrical mean.

158

Proof. Since $f(x) = \ln x$ is a concave function, by Jensen's inequality, i.e. (1.326), we have

$$\ln \sum_{i=1}^{M} \frac{x_i}{M} \geq \sum_{i=1}^{M} \frac{1}{M} \ln x_i = \sum_{i=1}^{M} \ln x_i^{\frac{1}{M}} = \ln \left(\prod_{i=1}^{M} x_i \right)^{\frac{1}{M}}. \tag{1.334}$$

Since $f(x) = \ln x$ is increasing, then

$$\sum_{i=1}^{M} \frac{x_i}{M} \geq \sqrt[M]{\prod_{i=1}^{M} x_i} \tag{1.335}$$

159 as desired. $\square$

Exercise 1.41

Using the sum and product rules of probability, show that the mutual information $\mathrm{I}(\mathbf{x}, \mathbf{y})$ satisfies the relation

$$\mathrm{I}(\mathbf{x}, \mathbf{y}) = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]. \tag{1.336}$$

160

Proof. By the definition of mutual information,

$$\begin{aligned}
\mathrm{I}(\mathbf{x}, \mathbf{y}) &\equiv \mathrm{KL}(p(\mathbf{x}, \mathbf{y}) || p(\mathbf{x})p(\mathbf{y})) && (1.337)\\
&= \iint p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} && (1.338)\\
&= \iint p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x}|\mathbf{y})}{p(\mathbf{x})} \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} && (1.339)\\
&= \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}|\mathbf{y}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} - \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} && (1.340)\\
&= -\int p(\mathbf{x}) \ln p(\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d} - \left(-\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}|\mathbf{y}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \right) && (1.341)\\
&= \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}]. && (1.342)
\end{aligned}$$

Similarly, we can get that

$$\mathrm{I}(\mathbf{x}, \mathbf{y}) = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]. \tag{1.343}$$

161 $\square$

162 # 2 Chapter 2 Probability Distributions

Exercise 2.1

Verify that the Bernoulli distribution

$$\mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x} \tag{2.1}$$

satisfies the following properties

$$\sum_{x=0}^{1} p(x|\mu) = 1 \tag{2.2}$$

$$\mathbb{E}[x] = \mu \tag{2.3}$$

$$\mathrm{var}[x] = \mu(1-\mu). \tag{2.4}$$

Show that the entropy $\mathrm{H}[x]$ of a Bernoulli distributed random binary variable x is given by

$$\mathrm{H}[x] = -\mu \ln \mu - (1-\mu)\ln(1-\mu). \tag{2.5}$$

163 *Proof.*

$$\sum_{x=0}^{1} p(x|\mu) = p(x=0|\mu) + p(x=1|\mu) = 1 - \mu + \mu = 1 \tag{2.6}$$

$$\mathbb{E}[x] = 0 \cdot p(x=0|\mu) + 1 \cdot p(x=1|\mu) = \mu \tag{2.7}$$

$$\begin{aligned} \mathrm{var}[x] &= (0-\mu)^2 \cdot p(x=0|\mu) + (1-\mu)^2 \cdot p(x=1|\mu) && (2.8)\\ &= \mu^2(1-\mu) + (1-\mu)^2\mu && (2.9)\\ &= \mu(1-\mu) && (2.10) \end{aligned}$$

$$\begin{aligned} \mathrm{H}[x] &= -p(x=1|\mu)\ln p(x=1|\mu) - p(x=0|\mu)\ln p(x=0|\mu) && (2.11)\\ &= -\mu\ln\mu - (1-\mu)\ln(1-\mu). && (2.12) \end{aligned}$$

164 □

Exercise 2.2

The form of the Bernoulli distribution given by (2.1) is not symmetric between the values of x. In some situations, it will be more convenient to use an equivalent formulation for which $x \in \{-1, 1\}$, in which case the distribution can be

$$p(x|\mu) = \left(\frac{1-\mu}{2}\right)^{(1-x)/2} \left(\frac{1+\mu}{2}\right)^{(1+x)/2} \tag{2.13}$$

where $\mu \in [-1, 1]$. Show that the distribution (2.13) is normalized, and evaluate its mean, variance, and entropy.

165 *Proof.*

$$p(x=1|\mu) + p(x=-1|\mu) = \frac{1+\mu}{2} + \frac{1-\mu}{2} = 1 \tag{2.14}$$

$$\mathbb{E}[x] = \frac{1+\mu}{2} - \frac{1-\mu}{2} = \mu \tag{2.15}$$

$$\begin{aligned}
\mathrm{var}[x] &= (1-\mu)^2 \cdot p(x=1|\mu) + (-1-\mu)^2 \cdot p(x=-1|\mu) && (2.16)\\
&= (1-\mu)^2 \frac{1+\mu}{2} + (-1-\mu)^2 \frac{1-\mu}{2} && (2.17)\\
&= \frac{(1-\mu^2)(1-\mu)}{2} + \frac{(1-\mu^2)(1+\mu)}{2} && (2.18)\\
&= 1-\mu^2 && (2.19)
\end{aligned}$$

$$\begin{aligned}
\mathrm{H}[x] &= -p(x=1|\mu)\ln p(x=1|\mu) - p(x=-1|\mu)\ln p(x=-1|\mu) && (2.20)\\
&= -\frac{1+\mu}{2}\ln\frac{1+\mu}{2} - \frac{1-\mu}{2}\ln\frac{1-\mu}{2} && (2.21)
\end{aligned}$$

166 □

Exercise 2.3

In this exercise, we prove that the binomial distribution

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m} \tag{2.22}$$

is normalized. First use the definition

$$\binom{N}{m} \equiv \frac{N!}{(N-m)!\,m!} \tag{2.23}$$

of the number of combinations of m identical objects chosen from a total of N to show that

$$\binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m}. \tag{2.24}$$

Use this result to prove by induction the following result

$$(1+x)^N = \sum_{m=0}^{N}\binom{N}{m}x^m \tag{2.25}$$

which is known as the *binomial theorem*, and which is valid for all real values of x. Finally, show that the binomial distribution is normalized, so that

$$\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m} = 1 \tag{2.26}$$

which can be done by first pulling out a factor $(1-\mu)^N$ out of the summation and then making use of the binomial theorem.

167

Proof.

$$\binom{N}{m} + \binom{N}{m-1} = \frac{N!}{(N-m)!\,m!} + \frac{N!}{(N-m+1)!\,(m-1)!} \tag{2.27}$$

$$
\begin{aligned}
&= \frac{N!}{(N-m)!\,(m-1)!}\left(\frac{1}{m}+\frac{1}{N-m+1}\right) && (2.28)\\
&= \frac{N!}{(N-m)!\,(m-1)!}\frac{N+1}{m(N-m+1)} && (2.29)\\
&= \frac{(N+1)!}{(N+1-m)!\,(m)!} = \binom{N+1}{m}. && (2.30)
\end{aligned}
$$

When $N = 1$, (2.25) holds. Assume that (2.25) holds for $N = M$, then for $N = M + 1$ we have

$$
\begin{aligned}
(1+x)^{M+1} &= (1+x)^M(1+x) && (2.31)\\
&= \sum_{m=0}^{M}\binom{M}{m}x^m(1+x) && (2.32)\\
&= \sum_{m=0}^{M}\binom{M}{m}x^m + \sum_{m=0}^{M}\binom{M}{m}x^{m+1} && (2.33)\\
&= 1 + \sum_{m=1}^{N}\binom{M}{m}x^m + \sum_{m=1}^{N}\binom{M}{m-1}x^m && (2.34)\\
&= 1 + \sum_{m=1}^{M}\left[\binom{M}{m}+\binom{M}{m-1}\right]x^m && (2.35)\\
&= 1 + \sum_{m=1}^{M}\binom{M+1}{m}x^m = \sum_{m=0}^{M}\binom{M+1}{m}x^m && (2.36)
\end{aligned}
$$

where the second line follows from the induction assumption. Hence, the binomial theorem follows. Then,

$$
\begin{aligned}
\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m} &= (1-\mu)^N\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{-m} && (2.37)\\
&= (1-\mu)^N\sum_{m=0}^{N}\binom{N}{m}\left(\frac{\mu}{1-\mu}\right)^m && (2.38)\\
&= (1-\mu)^N\left(1+\frac{\mu}{1-\mu}\right)^N && (2.39)\\
&= (1-\mu)^N\frac{1}{(1-\mu)^N} = 1 && (2.40)
\end{aligned}
$$

which completes the proof. □

Exercise 2.4

Show that the mean of the binomial distribution is given by

$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\mathrm{Bin}(m|N,\mu) = N\mu. \tag{2.41}$$

To do this, differentiate both sides of the normalization condition (2.26) with respect to μ and then rearrange to obtain an expression for the mean of n. Similarly, by differentiating twice with respect to μ and making use of the result (2) for the mean of the binomial distribution prove the result

$$\mathrm{var}[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \mathrm{Bin}(m|N,\mu) = N\mu(1-\mu). \tag{2.42}$$

for the variance of the binomial.

Proof. Differentiating both sides of the normalization condition (2.26) with respect to μ gives

$$\sum_{m=0}^{N} \binom{N}{m} \left[m\mu^{m-1}(1-\mu)^{N-m} - (N-m)\mu^m(1-\mu)^{N-m-1}\right] = 0 \tag{2.43}$$

$$\sum_{m=0}^{N} \binom{N}{m} \mu^{m-1}(1-\mu)^{N-m-1} \left[m(1-\mu) - (N-m)\mu\right] = 0 \tag{2.44}$$

$$\sum_{m=0}^{N} \binom{N}{m} \mu^{m-1}(1-\mu)^{N-m-1} \left[m - N\mu\right] = 0. \tag{2.45}$$

After rearranging, we get

$$\sum_{m=0}^{N} \binom{N}{m} m\mu^{m-1}(1-\mu)^{N-m-1} = N\mu \sum_{m=1}^{N-1} \binom{N}{m} \mu^{m-1}(1-\mu)^{N-m-1}. \tag{2.46}$$

Multiplying both sides by $\mu(1-\mu)$ yields

$$\sum_{m=0}^{N} \binom{N}{m} m\mu^{m}(1-\mu)^{N-m} = N\mu \sum_{m=0}^{N} \binom{N}{m} \mu^{m}(1-\mu)^{N-m} = N\mu \tag{2.47}$$

where the last equality follows from (2.26). Differentiating both sides of (2) with respect to μ gives

$$\sum_{m=0}^{N} \binom{N}{m} \left[m^2\mu^{m-1}(1-\mu)^{N-m} - (N-m)m\mu^m(1-\mu)^{N-m-1}\right] = N \tag{2.48}$$

$$\sum_{m=0}^{N} \binom{N}{m} \left[m(m-N\mu)\mu^{m-1}(1-\mu)^{N-m-1}\right] = N \tag{2.49}$$

$$\sum_{m=0}^{N} \binom{N}{m} \left[(m - \mathbb{E}[m] + \mathbb{E}[m])(m - \mathbb{E}[m])\mu^m(1-\mu)^{N-m}\right] = N\mu(1-\mu) \tag{2.50}$$

$$\sum_{m=0}^{N} \binom{N}{m} \left[(m - \mathbb{E}[m])(m - \mathbb{E}[m])\mu^m (1-\mu)^{N-m} + \mathbb{E}[m](m - \mathbb{E}[m])\mu^m (1-\mu)^{N-m}\right] = N\mu(1-\mu) \tag{2.51}$$

$$\mathrm{var}[m] + \mathbb{E}[m] \sum_{m=0}^{N} \binom{N}{m} (m - \mathbb{E}[m])\mu^m (1-\mu)^{N-m} = N\mu(1-\mu) \tag{2.52}$$

$$\mathrm{var}[m] = N\mu(1-\mu) \tag{2.53}$$

170 which completes the proof. □

Exercise 2.5

In this exercise, we prove that the beta distribution, given by

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} \tag{2.54}$$

is correctly normalized, so that

$$\int_0^1 \mathrm{Beta}(\mu|a,b)\mathrm{d}\mu = 1 \tag{2.55}$$

holds. This is equivalent to showing that

$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1}\mathrm{d}\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{2.56}$$

From the definition (1.170) of the gamma function, we have

$$\Gamma(a)\Gamma(b) = \int_0^\infty x^{a-1}e^{-x}\mathrm{d}x \int_0^\infty y^{b-1}e^{-y}\mathrm{d}y. \tag{2.57}$$

Use this expression to prove (2.56) as follows. First bring the integral over y inside the integrand of the integral over x, next make the change of variable $t = y + x$ where x is fixed, then interchange the order of the x and t integrations, and finally make the change of variable $x = t\mu$ where t is fixed.

171

Proof. First, we bring the integral over y inside the integrand of the integral over x

$$\Gamma(a)\Gamma(b) = \int_0^\infty \int_0^\infty x^{a-1}y^{b-1}e^{-(x+y)}\mathrm{d}y\mathrm{d}x. \tag{2.58}$$

Let $t = y + x$ where x is fixed, then

$$\Gamma(a)\Gamma(b) = \int_0^\infty \int_x^\infty x^{a-1}(t-x)^{b-1}e^{-t}\mathrm{d}t\mathrm{d}x \tag{2.59}$$

$$= \int_0^\infty \int_0^t x^{a-1}(t-x)^{b-1}e^{-t}\mathrm{d}x\mathrm{d}t \tag{2.60}$$

Let $x = t\mu$, then

$$\Gamma(a)\Gamma(b) = \int_0^\infty \int_0^1 (t\mu)^{a-1}(t-t\mu)^{b-1}e^{-t}t\mathrm{d}\mu\mathrm{d}t \tag{2.61}$$

$$
\begin{aligned}
&= \int_0^\infty t^{a+b-1}e^{-t}\int_0^1 \mu^{a-1}(1-\mu)^{b-1}\mathrm{d}\mu\mathrm{d}t && (2.62)\\
&= \Gamma(a+b)\int_0^1 \mu^{a-1}(1-\mu)^{b-1}\mathrm{d}\mu. && (2.63)
\end{aligned}
$$

After rearrangement, we get the desired (2.56). □

Exercise 2.6

Make use of the result (2.56) to show that the mean, variance, and mode of the beta distribution (2.54) are given respectively by

$$
\begin{aligned}
\mathbb{E}[\mu] &= \frac{a}{a+b} && (2.64)\\
\mathrm{var}[\mu] &= \frac{ab}{(a+b)^2(a+b+1)} && (2.65)\\
\mathrm{mode}[\mu] &= \frac{a-1}{a+b-2}. && (2.66)
\end{aligned}
$$

Proof.

$$
\begin{aligned}
\mathbb{E}[\mu] &= \int_0^1 \mu\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}\mathrm{d}\mu && (2.67)\\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 \mu^{a}(1-\mu)^{b-1}\mathrm{d}\mu && (2.68)\\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} && (2.69)\\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)} && (2.70)\\
&= \frac{a}{a+b}. && (2.71)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[\mu^2] &= \int_0^1 \mu^2\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}\mathrm{d}\mu && (2.72)\\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 \mu^{a+1}(1-\mu)^{b-1}\mathrm{d}\mu && (2.73)\\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} && (2.74)\\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{a(a+1)\Gamma(a)\Gamma(b)}{(a+b)(a+b+1)\Gamma(a+b)} && (2.75)\\
&= \frac{a(a+1)}{(a+b)(a+b+1)}. && (2.76)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{var}[\mu] &= \mathbb{E}[\mu^2]-\mathbb{E}[\mu]^2 && (2.77)\\
&= \frac{a(a+1)}{(a+b)(a+b+1)} - \frac{a^2}{(a+b)^2} && (2.78)\\
&= \frac{ab}{(a+b)^2(a+b+1)}. && (2.79)
\end{aligned}
$$

Taking derivatives of (2.54) with respect to μ and setting it to 0 gives

$$\nabla_\mu \text{Beta}(\mu) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}((a-1)\mu^{a-2}(1-\mu)^{b-1} + (b-1)\mu^{a-1}(1-\mu)^{b-2}) \tag{2.80}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}(a-1-(a+b-2)\mu) = 0. \tag{2.81}$$

Thus,

$$\text{mode}[\mu] = \frac{a-1}{a+b-2}. \tag{2.82}$$

174 □

Exercise 2.7

Consider a binomial random variable x given by (2.22), with prior distribution for μ given by the beta distribution (2.54), and suppose we have observed m occurrences of $x = 1$ and l occurrences of $x = 0$. Show that the posterior mean value of μ lies between the prior mean and the maximum likelihood estimate for μ. To do this, show that the posterior mean can be written as λ times the prior mean plus $1 - \lambda$ times the maximum likelihood estimate, where $0 \leq \lambda \leq 1$. This illustrates the concept of the posterior distribution being a compromise between the prior distribution and the maximum likelihood solution.

175

Proof. From (2.18) in the textbook, the posterior distribution is given by

$$p(\mu|m, l, a, b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1}. \tag{2.83}$$

According to (2.64), the posterior mean value of x is given by

$$\frac{m+a}{m+a+l+b} \tag{2.84}$$

which can be expressed as

$$\frac{a+b}{m+a+l+b} \cdot \frac{a}{a+b} + (1 - \frac{a+b}{m+a+l+b}) \cdot \frac{m}{m+l} \tag{2.85}$$

176 which indicates $\lambda = (a+b)/(m+a+l+b)$. □

Exercise 2.8

Consider two variables x and y with joint distribution $p(x, y)$. Prove the following two results

$$\mathbb{E}[x] = \mathbb{E}_y[\mathbb{E}_x[x|y]] \tag{2.86}$$

$$\text{var}[x] = \mathbb{E}_y[\text{var}_x[x|y]] + \text{var}_y[\mathbb{E}_x[x|y]]. \tag{2.87}$$

Here $\mathbb{E}_x[x|y]$ denotes the expectation of x under the conditional distribution $p(x|y)$, with a similar notation for the conditional variance.

177

Proof.

$$\mathbb{E}_y[\mathbb{E}_x[x|y]] = \int \left(\int xp(x|y)\mathrm{d}x \right) p(y)\mathrm{d}y \tag{2.88}$$

$$
\begin{aligned}
&= \int \int x p(x|y) p(y) \mathrm{d}x \mathrm{d}y && (2.89)\\
&= \int \int x p(x, y) \mathrm{d}x \mathrm{d}y && (2.90)\\
&= \int x \int p(x, y) \mathrm{d}y \mathrm{d}x && (2.91)\\
&= \int x p(x) \mathrm{d}x = \mathbb{E}[x]. && (2.92)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_y[\mathrm{var}_x[x|y]] + \mathrm{var}_y[\mathbb{E}_x[x|y]] &= \mathbb{E}_y[\mathbb{E}_x[x^2|y] - (\mathbb{E}_x[x|y])^2] + \mathbb{E}_y[(\mathbb{E}_x[x|y])^2] - (\mathbb{E}_y[\mathbb{E}_x[x|y]])^2 && (2.93)\\
&= \mathbb{E}_y[\mathbb{E}_x[x^2|y]] - \mathbb{E}_y[(\mathbb{E}_x[x|y])^2] + \mathbb{E}_y[(\mathbb{E}_x[x|y])^2] - (\mathbb{E}[x])^2 && (2.94)\\
&= \mathbb{E}[x^2] - (\mathbb{E}[x])^2 = \mathrm{var}[x] && (2.95)
\end{aligned}
$$

□

Exercise 2.9

In this exercise, we prove the normalization of the Dirichlet distribution (2.38)

$$
\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \tag{2.96}
$$

using induction. We have already shown in Exercise 2.5 that the beta distribution, which is a special case of the Dirichlet for $M = 2$, is normalized. We now assume that the Dirichlet distribution is normalized for $M - 1$ variables and prove that it is normalized for M variables. To do this, consider the Dirichlet distribution over M variables, and take account of the constraint $\sum_{k=1}^{M} \mu_k = 1$ by eliminating μ_M, so that the Dirichlet is written

$$
p_M(\mu_1, \ldots, \mu_{M-1}) = C_M \prod_{k=1}^{M-1} \mu_k^{\alpha_k - 1} \left(1 - \sum_{j=1}^{M-1} \mu_j\right)^{\alpha_M - 1} \tag{2.97}
$$

and our goal is to find an expression for C_M. To do this, integrate over μ_{M-1}, taking care over the limits of integration, and then make a change of variable so that this integral has limits 0 and 1. By assuming the correct result for C_{M-1} and making use of (2.54), derive the expansion for C_M.

Proof. We integrate (2.97) over μ_{M-1} as follows.

$$
\begin{aligned}
&p_{M-1}(\mu_1, \ldots, \mu_{M-2}) && (2.98)\\
=&\int_0^{1-\sum_{j=1}^{M-2}\mu_j} p_M(\mu_1, \ldots, \mu_{M-1}) \mathrm{d}\mu_{M-1} && (2.99)\\
=&C_M \prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1} \int_0^{1-\sum_{j=1}^{M-2}\mu_j} \mu_{M-1}^{\alpha_{M-1}-1} \left(1 - \sum_{j=1}^{M-2} \mu_j - \mu_{M-1}\right)^{\alpha_M - 1} \mathrm{d}\mu_{M-1} && (2.100)
\end{aligned}
$$

where the upper limit follows from the constraints $\mu_M = 1 - \sum_{j=1}^{M-2} \mu_j - \mu_{M-1} \geq 0$ and $\mu_{M-1} \geq 0$. Let $\mu_{M-1} = (1 - \sum_{j=1}^{M-2} \mu_j)t$, then

$$
p_{M-1}(\mu_1, \ldots, \mu_{M-2}) \tag{2.101}
$$

$$
\begin{aligned}
&= C_M \prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1} \left(1 - \sum_{j=1}^{M-2} \mu_j\right)^{\alpha_{M-1}+\alpha_M - 1} \int_0^1 t^{\alpha_{M-1}-1} (1-t)^{\alpha_M - 1} \, \mathrm{d}t && (2.102) \\
&= C_M \prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1} \left(1 - \sum_{j=1}^{M-2} \mu_j\right)^{\alpha_{M-1}+\alpha_M - 1} \frac{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}{\Gamma(\alpha_{M-1}+\alpha_M)} && (2.103)
\end{aligned}
$$

where the last line follows from (2.56). The right hand side of (2.103) can be considered as a normalized Dirichlet distribution over $M-1$ variables, i.e. $\alpha_1, \ldots, \alpha_{M-2}, \alpha_{M-1} + \alpha_M$. Note that the last two categories have been combined. By the inductive assumption and comparison with (2.96), we get

$$
\begin{aligned}
C_M &= \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_M)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_{M-1} + \alpha_M)} \cdot \frac{\Gamma(\alpha_{M-1} + \alpha_M)}{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)} && (2.104) \\
&= \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_M)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_{M-1})\Gamma(\alpha_M)} && (2.105)
\end{aligned}
$$

as desired. □

3 Chapter 4 Linear Models for Classification

3.1 Discriminant Functions

3.1.1 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.

$$
\mathbf{w}^T\mathbf{x} = \|\mathbf{w}\|\|\mathbf{x}\|\cos\theta = \|\mathbf{w}\| \cdot d \Longrightarrow \frac{\mathbf{w}^T\mathbf{x}}{\|\mathbf{w}\|} = d. \tag{3.1}
$$

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

$$
\mathbf{w}^T\mathbf{x} = -w_0 \Longrightarrow \frac{\mathbf{w}^T\mathbf{x}}{\|\mathbf{w}\|} = d = \frac{-w_0}{\|\mathbf{w}\|}. \tag{3.2}
$$

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

$$
r = \frac{y(\mathbf{0})}{\|\mathbf{w}\|} = \frac{w_0}{\|\mathbf{w}\|}. \tag{3.3}
$$

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* $\mathbf{x}$.

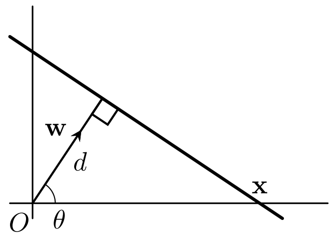

Figure 1: Illustration of the geometry of a linear discriminant function in two dimensions. The direction of $\mathbf{w}$ depends on the form of the decision surface, shown in thick, $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$. This makes θ in range $[0, \pi]$.

188
3.2 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.

189

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,

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

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

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

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

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.

$\square$