Chapter 1: Introduction

Notes for Pattern Recognition and Machine Learning by Christopher Bishop

Contents

1 Exercises

Exercise 1.1

Proof. According to Eq. (1.2), we can get the error function:

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

Calculating the derivative of this function with respect to w_i we can get:

$$\frac{\partial E(\mathbf{w})}{\partial w_i} = \sum_{n=1}^{N} x_n\{y\left(x_n, \mathbf{w}\right) - t_n\}. \tag{1}$$

Let Eq. (1) equals to 0, then

$$\begin{aligned}
&\sum_{n=1}^{N} x_n\{y\left(x_n, \mathbf{w}\right) - t_n\} = 0 \\
&\qquad \Rightarrow \sum_{n=1}^{N} y(x_n, \mathbf{w})x_n^i = \sum_{n=1}^{N} t_n x_n^i \\
&\Rightarrow \sum_{n=1}^{N}\{x_n^i(\sum_{j=0}^{M} w_j x_n^j)\} = \sum_{n=1}^{N} t_n x_n^i \\
&\Rightarrow \sum_{n=1}^{N}\sum_{j=0}^{M} w_j (x_n)^{i+j} = \sum_{n=1}^{N} t_n x_n^i \\
&\Rightarrow \sum_{j=0}^{M}\sum_{n=1}^{N} w_j (x_n)^{i+j} = \sum_{n=1}^{N} t_n x_n^i.
\end{aligned}$$

According to Eq. (1.122) and (1.123), we can get

$$\begin{aligned}
\sum_{j=0}^{M}\sum_{n=1}^{N} w_j (x_n)^{i+j} &= \sum_{n=1}^{N} t_n x_n^i \\
\Rightarrow \sum_{j=0}^{M}\{\sum_{n=1}^{N}(x_n)^{i+j}\}w_j &= \sum_{n=1}^{N} t_n x_n^i \\
\Rightarrow \sum_{j=0}^{M} A_{ij} w_j &= T_i,
\end{aligned}$$

as desired. □

Exercise 1.2

Proof. According to Eq. (1.4), we can get the error function:

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y\left(x_n, \mathbf{w}\right) - t_n\} + \frac{\lambda}{2}\|\mathbf{w}\|_2^2.$$

Calculating the derivative of this function with respect to w_i we can get:

$$\frac{\partial \widetilde{E}(\mathbf{w})}{\partial w_i} = \sum_{n=1}^{N} x_n\{y\left(x_n, \mathbf{w}\right) - t_n\} + \lambda w_i. \tag{2}$$

Let Eq. (2) equals to 0, then

$$\begin{aligned}
&\lambda w_i + \sum_{n=1}^{N} x_n\{y\left(x_n, \mathbf{w}\right) - t_n\} = 0 \\
&\Rightarrow \lambda w_i + \sum_{n=1}^{N} y(x_n, \mathbf{w})x_n^i = \sum_{n=1}^{N} t_n x_n^i \\
&\Rightarrow \lambda w_i + \sum_{j=0}^{M}\{\sum_{n=1}^{N}(x_n)^{i+j}\}w_j = \sum_{n=1}^{N} t_n x_n^i \\
&\Rightarrow \sum_{j=0}^{M}\{\sum_{n=1}^{N}(x_n)^{i+j} + \lambda\mathbb{I}_{(i=j)}\}w_j = \sum_{n=1}^{N} t_n x_n^i,
\end{aligned} \tag{3}$$

where $\mathbb{I}_{(i=j)} = 1$ if and only if $i = j$, otherwise 0. According to Eq. (3), (1.122) and (1.123), we can rewrite similar set of equations

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

where

$$A_{ij} = \sum_{n=1}^{N}\{(x_n)^{i+j} + \lambda\mathbb{I}_{(i=j)}\},\ T_i = \sum_{n=1}^{N}(x_n)^i t_n.$$

□

Exercise 1.3

Proof. We suppose that the probability of selecting an apple from boxes is $p(a)$, then

$$p(a) = p(a|r)p(r) + p(a|b)p(b) + p(a|g)p(g) = 0.34.$$

Suppose that the probability of selecting an orange from boxes is $p(o)$. There is an equation

$$p(og) = p(o|g)p(g) = p(g|o)p(o).$$

Hence we can get

$$p(g|o) = \frac{p(o|g)p(g)}{p(o)} = \frac{p(o|g)p(g)}{p(o|r)p(r) + p(o|b)p(b) + p(o|g)p(g)} = 0.5.$$

□

Exercise 1.4

Proof. According to Eq. (1.27), we can get

$$p_y(y) = p_x(g(y))|g'(y)|.$$

By differentiating Eq. (1.27), we can get

$$\begin{aligned} \frac{dp_y(y)}{dy} &= \frac{dp_x(g(y))|g'(y)|}{dy} \\ &= |g'(y)|\frac{dp_x(g(y))}{dy} + p_x(g(y))\frac{d|g'(y)|}{dy}. \end{aligned} \tag{4}$$

Calculating the derivative of $p_x(x)$, it is obvious that it equals to 0 when $x = \hat{x}$, so

$$\left.\frac{dp_x(x)}{dx}\right|_{\hat{x}} = 0.$$

When $g(y) = \hat{x}$, Eq. (4) is equivalent to

$$\begin{aligned} \frac{dp_y(y)}{dy} &= |g'(y)|\frac{dp_x(g(y))}{dy} + p_x(g(y))\frac{d|g'(y)|}{dy} \\ &= |g'(y)|\frac{dp_x(g(y))}{dg(y)}\frac{dg(y)}{dy} + p_x(g(y))\frac{d|g'(y)|}{dy} \\ &= |g'(y)| \cdot 0 \cdot \frac{dg(y)}{dy} + p_x(g(y))\frac{d|g'(y)|}{dy} \\ &= p_x(g(y))\frac{d|g'(y)|}{dy}, \end{aligned}$$

Equation $p_x(g(y))\frac{d|g'(y)|}{dy}$ is always 0 means that if $\hat{x}$ is the location of the maximum of $p_x(x)$, $\hat{y}$, where $g(\hat{y}) = \hat{x}$, is also the location of the maximum of $p_y(y)$. By observing the last term $p_x(g(y))\frac{d|g'(y)|}{dy}$, $p_x(g(y))$ is a density probability hence it is not always 0. $\frac{d|g'(y)|}{dy} = 0$, if and only if $g(y)$ is a linear function, which means in the case of a linear transformation, the location of the maximum transforms in the same way as the variable itself, as desired. □

Exercise 1.5

Proof.

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

$$
\begin{aligned}
&= \mathbb{E}\left[f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2\right] \\
&= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2 \\
&= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]^2 + \mathbb{E}[f(x)]^2 \\
&= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2,
\end{aligned}
$$

as desired. □

Exercise 1.6

Proof. If x, y are discrete independent variables, they follow

$$
\begin{aligned}
\mathbb{E}_{x,y}[xy] &= \sum_x \sum_y p(xy)xy \\
&= \sum_x \sum_y xyp_x(x)p_y(y) \\
&= \sum_x xp_x(x) \sum_y yp_y(y) \\
&= \mathbb{E}[x]\mathbb{E}[y]. \qquad (5)
\end{aligned}
$$

If x, y are continuous independent variables, they follow

$$
\begin{aligned}
\mathbb{E}_{x,y}[xy] &= \int_x \int_y p(xy)xydxdy \\
&= \int_x \int_y xyp_x(x)p_y(y)dxdy \\
&= \int_x xp_x(x)dx \int_y yp_y(y)dy \\
&= \mathbb{E}[x]\mathbb{E}[y]. \qquad (6)
\end{aligned}
$$

With the combination of Eq. (5) and (6), we can get that if x, y are independent variables, they satisfy

$$
\mathbb{E}_{x,y}[xy] = \mathbb{E}[x]\mathbb{E}[y].
$$

Hence $cov[x, y]$ follows

$$
\begin{aligned}
cov[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\
&= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \\
&= \mathbb{E}[x]\mathbb{E}[y] - \mathbb{E}[x]\mathbb{E}[y] \\
&= 0,
\end{aligned}
$$

as desired. □

Exercise 1.7

Proof. According to Eq. (1.125), we have

$$\begin{aligned}
I^2 &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}(x^2+y^2)\right) dxdy \\
&= \int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) rdrd\theta \\
&= \frac{1}{2}\int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) dr^2d\theta \\
&= -\sigma^2\int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) d\left(-\frac{1}{2\sigma^2}\right) r^2d\theta \\
&= \sigma^2\int_0^{2\pi} d\theta \\
&= 2\pi\sigma^2,
\end{aligned}$$

where we apply $x = rcos\theta, y = rsin\theta$. Thus $I = (2\pi\sigma^2)^{\frac{1}{2}}$. Then

$$\begin{aligned}
\int_{-\infty}^{+\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) dx &= \int_{-\infty}^{+\infty} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right\} dx \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right\} d(x-\mu) \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}(2\pi\sigma^2)^{\frac{1}{2}} = 1,
\end{aligned}$$

as desired. □

Exercise 1.8

Proof.

$$\begin{aligned}
\mathbb{E}[x] &= \int_{-\infty}^{+\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) xdx = \int_{-\infty}^{+\infty} \frac{x}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right\} dx \\
&= \int_{-\infty}^{+\infty} \frac{(x-\mu+\mu)}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right\} dx \\
&= \int_{-\infty}^{+\infty} \frac{(x-\mu)}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right\} d(x-\mu) + \int_{-\infty}^{+\infty} \frac{\mu}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right\} d(x-\mu) \\
&= 0 + \frac{\mu}{(2\pi\sigma^2)^{\frac{1}{2}}}(2\pi\sigma^2)^{\frac{1}{2}} \\
&= \mu,
\end{aligned}$$

as desired.

$$
\begin{aligned}
\mathbb{E}[x^2] &= \int_{-\infty}^{+\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x^2 dx = \int_{-\infty}^{+\infty} \frac{x^2}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx \\
&= \int_{-\infty}^{+\infty} \frac{(x-\mu)^2 + 2\mu x - \mu^2}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx \\
&= \int_{-\infty}^{+\infty} \frac{(x-\mu)^2}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx \\
&+ 2\mu \int_{-\infty}^{+\infty} \frac{x}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx \\
&- \int_{-\infty}^{+\infty} \frac{\mu^2}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx \\
&= \int_{-\infty}^{+\infty} \frac{(x-\mu)^2}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx \\
&+ 2\mu \int_{-\infty}^{+\infty} \frac{x}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx - \mu^2 \\
&= \int_{-\infty}^{+\infty} \frac{(x-\mu)^2}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx + 2\mu \int_{-\infty}^{+\infty} \frac{x}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx - \mu^2 \\
&= \int_{-\infty}^{+\infty} \frac{(x-\mu)^2}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} d(x-\mu) + \mu^2 \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \int_{-\infty}^{+\infty} (x-\mu)^2 \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} d(x-\mu) + \mu^2 \\
&= \frac{\sqrt{2}\sigma \cdot 2\sigma^2}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2 \exp\left\{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2\right\} d\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) + \mu^2 \\
&= \frac{\sqrt{2}\sigma \cdot 2\sigma^2}{\sqrt{2\pi}\sigma} \cdot \frac{\sqrt{\pi}}{2} + \mu^2 \\
&= \mu^2 + \sigma^2,
\end{aligned}
$$

where we utilize $\int_{-\infty}^{+\infty} x^2 \exp(-x^2) dx = \frac{\sqrt{\pi}}{2}$ [Proof]. Hence we have $var[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$. □

Exercise 1.9

Proof. Let us get the derivative of $\mathcal{N}(x|\mu,\sigma^2)$

$$
\frac{\partial \mathcal{N}(x|\mu,\sigma^2)}{\partial x} = -\frac{2(x-\mu)}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}.
$$

Obviously, $\mathcal{N}(x|\mu,\sigma^2)$ is monotonically increasing on the interval $(-\infty,\mu)$ and decreasing on the interval $(\mu,+\infty)$. Hence mode (i.e. the maximum) of the univariate Gaussian distribution

is given by μ.
For multivariate Gaussian distribution, we can get its derivative [Matrix Calculus]

$$\frac{\partial \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \mathbf{x}} = -\frac{1}{2}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu})(\boldsymbol{\Sigma}^{-\mathbf{1}} + (\boldsymbol{\Sigma}^{-\mathbf{1}})^{\mathbf{T}})(\mathbf{x} - \boldsymbol{\mu}).$$

Hence we have $\frac{\partial \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \mathbf{x}} = 0$ when $\mathbf{x} = \boldsymbol{\mu}$. Thus the mode of the multivariate Gaussian is given by $\boldsymbol{\mu}$. □

Exercise 1.10

Proof. If x, z are continuous independent variables, we have

$$\begin{aligned}
\mathbb{E}[x + z] &= \int_z \int_x (x + z)p(x, z)dxdz \\
&= \int_z \int_x (x + z)p(x)p(z)dxdz \\
&= \int_z \int_x xp(x)p(z) + zp(x)p(z)dxdz \\
&= \int_z \int_x xp(x)p(z)dxdz + \int_z \int_x zp(x)p(z)dxdz \\
&= \int_z p(z)dz \int_x xp(x)dx + \int_x p(x)dx \int_z zp(z)dz \\
&= \mathbb{E}[x] + \mathbb{E}[z],
\end{aligned}$$

as desired.
If x, z are independent discrete variables, we have

$$\begin{aligned}
\mathbb{E}[x + z] &= \sum_z \sum_x (x + z)p(x, z) \\
&= \sum_z \sum_x (x + z)p(x)p(z) \\
&= \sum_z \sum_x xp(x)p(z) + zp(x)p(z) \\
&= \sum_z \sum_x xp(x)p(z) + \sum_z \sum_x zp(x)p(z) \\
&= \sum_z p(z) \sum_x xp(x) + \sum_x p(x) \sum_z zp(z) \\
&= \mathbb{E}[x] + \mathbb{E}[z],
\end{aligned}$$

as desired.
For variance, we have

$$\begin{aligned}
var[x + z] &= \mathbb{E}[(x + z)^2] - \mathbb{E}[x + z]^2 \\
&= \mathbb{E}[x^2 + 2xz + z^2] - \mathbb{E}[x + z]^2
\end{aligned}$$

$$= \mathbb{E}[x^2] + 2\mathbb{E}[xz] + \mathbb{E}[z^2] - (\mathbb{E}[x]^2 + \mathbb{E}[z]^2 - 2\mathbb{E}[x][z])$$
$$= (\mathbb{E}[x^2] - \mathbb{E}[x]^2) + (\mathbb{E}[z^2] - \mathbb{E}[z]^2)$$
$$= var[x] + var[z],$$

as desired. □

Exercise 1.11

Proof. By setting the derivatives of the log likelihood function with respect to μ, we have

$$\frac{\partial lnp(\mathbf{x}|\mu, \sigma^{\mathbf{2}})}{\partial \mu} = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu) = 0.$$

Thus $\mu_{ML} = \frac{\sum_{n=1}^{N} x_n}{N}$. By setting the derivatives of the log likelihood function with respect to σ^2, we have

$$\frac{\partial lnp(\mathbf{x}|\mu, \sigma^{\mathbf{2}})}{\partial \sigma^2} = \frac{1}{2\sigma^4}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2\sigma^2} = 0.$$

Hence we have $\sigma^2_{ML} = \frac{\sum_{n=1}^{N}(x_n - \mu_{ML})^2}{N}$. □

Exercise 1.12

Proof. If $n = m$, $\mathbb{E}[x_n x_m] = \mathbb{E}[x_n^2] = \mu^2 + \sigma^2$, based on solution of Exercise 1.8. If $n \neq m$, $\mathbb{E}[x_n x_m] = \mathbb{E}[x_n]\mathbb{E}[x_m] = \mu^2$ for x_n and x_m are i.i.d. Hence $\mathbb{E}[x_n x_m] = \mu^2 + I_{nm}\sigma^2$, as desired. For $\mathbb{E}[\mu_{ML}]$, we have

$$\begin{aligned}\mathbb{E}[\mu_{ML}] &= \mathbb{E}[\frac{\sum_{n=1}^{N} x_n}{N}] \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n] \\ &= \mu,\end{aligned}$$

as desired. For $\mathbb{E}[\sigma^2_{ML}]$, we have

$$\begin{aligned}\mathbb{E}[\sigma^2_{ML}] &= \mathbb{E}[\frac{\sum_{n=1}^{N}(x_n - \mu_{ML})^2}{N}] \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[(x_n - \mu_{ML})^2] \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n^2 + \mu^2_{ML} - 2x_n\mu_{ML}]\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n^2] + \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}\mu_{ML}^2] - \frac{2}{N}\mathbb{E}[\sum_{n=1}^{N}x_n\mu_{ML}] \\
&= \mu^2 + \sigma^2 + \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}(\frac{\sum_{n=1}^{N}x_n}{N})^2] - \frac{2}{N}\mathbb{E}[\sum_{n=1}^{N}x_n(\frac{\sum_{i=1}^{N}x_i}{N})] \\
&= \mu^2 + \sigma^2 + \frac{1}{N^2}\mathbb{E}[(\sum_{n=1}^{N}x_n)^2] - \frac{2}{N^2}\mathbb{E}[\sum_{n=1}^{N}x_n(\sum_{i=1}^{N}x_i)] \\
&= \mu^2 + \sigma^2 + \frac{1}{N^2}\mathbb{E}[(\sum_{n=1}^{N}x_n)^2] - \frac{2}{N^2}\mathbb{E}[(\sum_{n=1}^{N}x_n)^2] \\
&= \mu^2 + \sigma^2 - \frac{1}{N^2}\mathbb{E}[(\sum_{n=1}^{N}x_n)^2] \\
&= \mu^2 + \sigma^2 - \frac{1}{N^2}\{N(\mu^2 + 1\cdot\sigma^2) + (N^2 - N)(\mu^2 + 0\cdot\sigma^2)\} \\
&= \left(\frac{N-1}{N}\right)\sigma^2,
\end{aligned}
$$

as desired. □

Exercise 1.13

Proof. We let

$$\widetilde{\sigma}_{ML}^2 = \frac{\sum_{n=1}^{N}(x_n - \mu)^2}{N}.$$

Then

$$
\begin{aligned}
\mathbb{E}[\widetilde{\sigma}_{ML}^2] &= \mathbb{E}[\frac{\sum_{n=1}^{N}(x_n - \mu)^2}{N}] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[(x_n - \mu)^2] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n^2 + \mu^2 - 2\mu x_n] \\
&= \frac{1}{N}\sum_{n=1}^{N}(\mathbb{E}[x_n^2] + \mathbb{E}[\mu^2] - 2\mu\mathbb{E}[x_n]) \\
&= \frac{1}{N}\sum_{n=1}^{N}(\mu^2 + \sigma^2 - \mu^2) \\
&= \sigma^2,
\end{aligned}
$$

as desired. □

Exercise 1.14

Proof. w_{ij} and w_{ji} satisfies

$$\begin{aligned} w_{ij} &= w_{ij}^S + w_{ij}^A; \\ w_{ji} &= w_{ji}^S + w_{ji}^A = w_{ij}^S - w_{ij}^A, \end{aligned}$$

where we can see that w_{ij}^S and w_{ij}^A provide two degrees of freedom to w_{ij} and w_{ji}, so w_{ij} and w_{ji} can be in terms of w_{ij}^S and w_{ij}^A. In other words, w_{ij} and w_{ji} can be written as follow

$$\begin{aligned} w_{ij}^S &= \frac{w_{ij} + w_{ji}}{2}; \\ w_{ij}^A &= \frac{w_{ij} - w_{ji}}{2}, \end{aligned}$$

as desired. Then

$$\begin{aligned} \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_i x_j &= \sum_{i=1}^{D}\sum_{j=1}^{D}\{(w_{ij}^S + w_{ij}^A)x_i x_j\} \\ &= \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}\{(w_{ij}^S + w_{ij}^A)x_i x_j + (w_{ji}^S + w_{ji}^A)x_j x_i\} \\ &= \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}\{(w_{ij}^S + w_{ij}^A)x_i x_j + (w_{ij}^S - w_{ij}^A)x_j x_i\} \\ &= \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^S x_i x_j, \end{aligned}$$

which shows that the contribution of w_{ij}^A vanishes. The number of independent elements in symmetric matrix is given by $D + \frac{D^2-D}{2} = \frac{(D+1)D}{2}$ because the elements on the diagonal of a matrix are independent, while half of the remaining elements are independent. □

Exercise 1.15

Proof. According to Eq. (1.134), we let us replace D with $D+1$

$$\begin{aligned} &\sum_{i_1=1}^{D+1}\sum_{i_2=1}^{i_1}\cdots\sum_{i_M=1}^{i_{M-1}} \widetilde{w}_{i_1 i_2 \ldots i_M} x_{i_1} x_{i_2} \ldots x_{i_M} \\ =&\sum_{i_1=1}^{D}\sum_{i_2=1}^{i_1}\cdots\sum_{i_M=1}^{i_{M-1}} \widetilde{w}_{i_1 i_2 \ldots i_M} x_{i_1} x_{i_2} \ldots x_{i_M} + \sum_{i_2=1}^{D+1}\sum_{i_3=1}^{i_2}\cdots\sum_{i_M=1}^{i_{M-1}} \widetilde{w}_{i_1 i_2 \ldots i_M} x_{(D+1)} x_{i_2} \ldots x_{i_M}. \end{aligned}$$

Then we can get

$$\begin{aligned} n(D+1,M) &= n(D,M) + n(D+1,M-1) \\ &= \sum_{i=1}^{D} n(i,M-1) + n(D+1,M-1) \\ &= \sum_{i=1}^{D+1} n(i,M-1), \end{aligned}$$

which means

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

as desired. If $D = 1$, we have

$$\begin{aligned} \sum_{i=1}^{D} \frac{(i+M-2)!}{(i-1)!(M-1)!} &= \sum_{i=1}^{1} \frac{(i+M-2)!}{(i-1)!(M-1)!} \\ &= \frac{(M-1)!}{0!(M-1)!} = 1, \end{aligned}$$

while

$$\frac{(D+M-1)!}{(D-1)!M!} = \frac{(M)!}{0!M!} = 1,$$

as desired. Thus we have

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

for all M when $D = 1$. Assuming that it is correct for dimension D, for $D+1$ dimension, we have

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

as desired. For $M = 2$, we have

$$n(D,2) = \sum_{i=1}^{D} n(i,1)$$

$$
\begin{aligned}
&= n(1,1) + n(2,1) + ... + n(D,1) \\
&= 1 + 2 + ... + D \\
&= \frac{D(D+1)}{2} \\
&= \frac{(D+1)!}{(D-1)!2!},
\end{aligned}
$$

as desired. Assuming that result holds at order $M-1$, for order M, we have

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

as desired. □

Exercise 1.16

Proof. For $M = 0$, we have

$$
N(D,0) = \sum_{m=0}^{0} n(D,m) = n(D,0) = 1 = \frac{(D+0)!}{D!0!},
$$

as desired. Assuming Eq. (1.138) is correct for M, for $M+1$, we have

$$
\begin{aligned}
N(D,M+1) &= \sum_{m=0}^{M+1} n(D,m) \\
&= \sum_{m=0}^{M} n(D,m) + n(D,M+1) \\
&= N(D,M) + n(D,M+1) \\
&= \frac{(D+M)!}{D!M!} + \frac{(D+M)!}{(D-1)!(M+1)!} \\
&= \frac{(D+M+1)!}{D!(M+1)!},
\end{aligned}
$$

as desired. For $D \gg M$, we have

$$
\begin{aligned}
N(D,M) &= \frac{(D+M)!}{D!M!} \\
&\simeq \frac{(D+M)^{(D+M)}}{D^D M^M}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{M^M}\left(\frac{D+M}{D}\right)^D (D+M)^M \\
&= \frac{1}{M^M}\left[\left(1+\frac{M}{D}\right)^{\frac{D}{M}}\right]^M (D+M)^M \\
&\simeq \left(\frac{e}{M}\right)^M (D+M)^M \\
&= \left(\frac{e}{M}\right)^M \left(1+\frac{M}{D}\right)^M D^M \\
&= \left(\frac{e}{M}\right)^M \left[\left(1+\frac{M}{D}\right)^{\frac{D}{M}}\right]^{M^2} D^M \\
&\simeq \frac{e^{M^2+M}}{M^M} D^M \\
&\simeq D^M,
\end{aligned}
$$

as desired. For $M \gg D$, we can simply get $N(D,M) \simeq M^D$ by symmetry. Finally, we can calculate $N(10,3) = 286$ and $N(100,3) = 176851$ directly. □

Exercise 1.17

Proof. According to Eq. (1.141), we have

$$
\begin{aligned}
\Gamma(x+1) &= \int_0^{+\infty} u^x e^{-u} du \\
&= -\int_0^{+\infty} u^x de^{-u} \\
&= -u^x e^{-u}\big|_0^{+\infty} + \int_0^{+\infty} e^{-u} du^x \\
&= \int_0^{+\infty} x e^{-u} u^{x-1} du \\
&= x\Gamma(x),
\end{aligned}
$$

as desired. $\Gamma(1) = \int_0^{+\infty} e^{-u} du = 1$. Hence we have

$$
\Gamma(x+1) = x\Gamma(x) = x(x-1)\Gamma(x-1) = ... = x!,
$$

when x is an integer. □

Exercise 1.18

Proof. Let

$$
I = \int_{-\infty}^{+\infty} e^{-x^2} dx.
$$

Then

$$\begin{aligned} I^2 &= \int_{-\infty}^{+\infty} e^{-x^2} dx \int_{-\infty}^{+\infty} e^{-y^2} dy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-(x^2+y^2)} dx dy \\ &= \int_0^{2\pi} \int_0^{+\infty} e^{-r^2} r dr d\theta \\ &= \frac{1}{2} \int_0^{2\pi} \int_0^{+\infty} e^{-r^2} dr^2 d\theta \\ &= \pi. \end{aligned}$$

Thus $I = \sqrt{\pi}$. Considering the left side of Eq. (1.142). we have

$$\prod_{i=1}^{D} \int_{-\infty}^{+\infty} e^{-x_i^2} dx_i = \pi^{\frac{D}{2}}.$$

Considering the right side of Eq. (1.142), we can get

$$\begin{aligned} S_D \int_0^{+\infty} e^{-r^2} r^{D-1} dr &= S_D \int_0^{+\infty} e^{-u} u^{\frac{D-1}{2}} \cdot \frac{1}{2\sqrt{u}} du \\ &= S_D \int_0^{+\infty} e^{-u} u^{\frac{D}{2}-1} du \\ &= \frac{S_D}{2} \Gamma(\frac{D}{2}), \end{aligned}$$

where we utilize $r = \sqrt{u}$. Hence we can get

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

The surface of D-dimensional sphere is proportional to the r^{D-1}, so we can get the surface of D-dimensional sphere is

$$S_D(r) = \frac{2\pi^{\frac{D}{2}}}{\Gamma(\frac{D}{2})} r^{D-1}.$$

For $dV = S_D(r)dr$, we have

$$\begin{aligned} V_D(r) &= \int_0^1 \frac{2\pi^{\frac{D}{2}}}{\Gamma(\frac{D}{2})} r^{D-1} dr \\ &= S_D \int_0^1 r^{D-1} dr \\ &= \frac{S_D}{D}, \end{aligned}$$

as desired. Thus

$$
\begin{aligned}
S_2 &= \frac{2\pi}{\Gamma(1)} = 2\pi; \\
V_2 &= \pi; \\
S_3 &= \frac{2\pi^{\frac{3}{2}}}{\Gamma(3/2)} = 4\pi; \\
V_3 &= \frac{4}{3}\pi.
\end{aligned}
$$

□

Exercise 1.19

Proof. By tuition, we can get the volume of D-dimensional cube with side $2a$ is $(2a)^D$. Then

$$
\frac{\text{Volume of sphere}}{\text{Volume of cube}} = \frac{2\pi^{\frac{D}{2}} a^D}{\Gamma\left(D/2\right) D(2a)^D} = \frac{\pi^{\frac{D}{2}}}{D2^{D-1}\Gamma(D/2)}. \tag{7}
$$

When $D \to \infty$, we have

$$
\begin{aligned}
\frac{\text{Volume of sphere}}{\text{Volume of cube}} &= \frac{\pi^{\frac{D}{2}}}{D2^{D-1}\Gamma(D/2)} \\
&\simeq \frac{\pi^{\frac{D}{2}}}{D2^{D-1}(2\pi)^{\frac{1}{2}} e^{\left(-\frac{D}{2}+1\right)} \left(\frac{D}{2}-1\right)^{\frac{D}{2}-\frac{1}{2}}} \\
&\simeq \frac{\pi^{\frac{D}{2}} e^{\frac{D}{2}-1}}{D2^{D-1} \left(\frac{D}{2}-1\right)^{\frac{D}{2}-\frac{1}{2}}} \\
&= \frac{1}{D} \cdot \frac{\pi^{\frac{D}{2}}}{2^{D-1}} \cdot \frac{e^{\frac{D}{2}-1}}{\left(\frac{D}{2}-1\right)^{\frac{D}{2}-\frac{1}{2}}} \\
&\simeq 0 \cdot C \cdot 0 \\
&= 0,
\end{aligned}
$$

where C is a constant. Thus Eq. (7) goes to zero when $D \to \infty$. Then

$$
\frac{\text{Distance from center to a corner}}{\text{Distance from center to a side}} = \frac{\sqrt{D \cdot a^2}}{a} = \sqrt{D}, \tag{8}
$$

which means Eq. (8) goes zero as $D \to \infty$. □

Exercise 1.20

Proof. A D-dimensional Gaussian distribution is given by

$$
p(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{\|\mathbf{x}\|_2^2}{2\sigma^2}\right).
$$

The volume of thin shell of radius r is given by

$$V_{shell} = S_D(r) r^{D-1} \epsilon.$$

Thus the probability of density over a thin shell of radius r and thickness ε is

$$p(r) = \int_{shell} p(\mathbf{x}) dV = p(\mathbf{x}) \int_{shell} dV = \frac{S_D r^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right),$$

as desired. Calculating the derivative of $p(r)$, we have

$$\frac{dp(r)}{dr} = \frac{S_D(D-1)r^{D-2}}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) + \frac{S_D r^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \cdot \left(-\frac{r}{\sigma^2}\right). \tag{9}$$

Let Eq. (9) equals to 0. Then

$$\begin{aligned}
\frac{S_D(D-1)r^{D-2}}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) &= \frac{S_D r^D}{(2\pi\sigma^2)^{D/2}\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) \\
\Rightarrow r^2 &= (D-1)\sigma^2 \ (r > 0) \\
\Rightarrow r &= \sigma\sqrt{D-1}.
\end{aligned}$$

Thus we can conclude Eq. (9) has a single stationary point located, for large D, at $\hat{r} \simeq \sqrt{D}\sigma$. Considering $p(\hat{r} + \epsilon)$ divide by $p(\hat{r})$, we have

$$\begin{aligned}
\frac{p(\hat{r}+\epsilon)}{p(\hat{r})} &= \left(1 + \frac{\epsilon}{\hat{r}}\right)^{D-1} \exp\left(-\frac{\epsilon^2 + 2\hat{r}\epsilon}{2\sigma^2}\right) \\
&= \exp\left\{-\frac{\epsilon^2 + 2\hat{r}\epsilon}{2\sigma^2} + (D-1) ln\left(1 + \frac{\epsilon}{r}\right)\right\} \\
&\simeq \exp\left\{-\frac{\epsilon^2 + 2\hat{r}\epsilon}{2\sigma^2} + (D-1)\left(\frac{\epsilon}{\hat{r}} - \frac{\epsilon^2}{2\hat{r}^2}\right)\right\} \\
&\simeq \exp\left(-\frac{\epsilon^2 + 2\hat{r}\epsilon}{2\sigma^2} + \frac{-\epsilon^2 + 2\hat{r}\epsilon}{2\sigma^2}\right) \\
&= \exp\left(-\frac{\epsilon^2}{\sigma^2}\right),
\end{aligned}$$

where we utilize Taylor Theorem and $\hat{r} \simeq \sqrt{D}\sigma$. Thus

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

as desired. Let $\mathbf{x} = 0$, and we have

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

Let $\|\mathbf{x}\|_2^2 = r^2$ and we have

$$p(\mathbf{x})\Big|_{\|x\|_2^2 = r^2} = \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}}} \exp\left(-\frac{\hat{r}^2}{2\sigma^2}\right) \simeq \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}}} \exp\left(-\frac{D}{2}\right),$$

where we utilize $\hat{r} \simeq \sqrt{D}\sigma$. Thus the probability density $p(\mathbf{x})$ is larger at the origin than at the radius $\hat{r}$ by a factor of $\exp(D/2)$. □

Exercise 1.21

Proof. Obviously $a \leq \sqrt{ab}$ for $a^2 \leq ab$ and $b \geq a \geq 0$. For $p(mistake)$, we have

$$\begin{aligned} p(mistake) &= p(x \in \mathcal{R}_1, \mathcal{C}_2) + p(x \in R_2, \mathcal{R}_1) \\ &= \int_{\mathcal{R}_1} p(x, \mathcal{C}_2) dx + \int_{\mathcal{R}_2} p(x, \mathcal{C}_1) dx. \end{aligned}$$

In decision region $\mathcal{R}_1$, we have

$$p(x, \mathcal{C}_2) \leq p(x, \mathcal{C}_1).$$

The same can be applied to decision region $\mathcal{R}_2$ and we can get $p(x, \mathcal{C}_1) \leq p(x, \mathcal{C}_2)$. Thus we have

$$p(x, \mathcal{C}_2) \leq \sqrt{p(x, \mathcal{C}_1) p(x, \mathcal{C}_2)}$$

for region $\mathcal{R}_1$ and

$$p(x, \mathcal{C}_1) \leq \sqrt{p(x, \mathcal{C}_1) p(x, \mathcal{C}_2)}$$

for region $\mathcal{R}_2$. Thus we have

$$\begin{aligned} p(mistake) &= \int_{\mathcal{R}_1} p(x, \mathcal{C}_2) dx + \int_{\mathcal{R}_2} p(x, \mathcal{C}_1) dx \\ &\leq \int_{\mathcal{R}_1} \sqrt{p(x, \mathcal{C}_1) p(x, \mathcal{C}_2)} dx + \int_{\mathcal{R}_2} \sqrt{p(x, \mathcal{C}_1) p(x, \mathcal{C}_2)} dx \\ &= \int_{\mathcal{R}_1 + \mathcal{R}_2} \sqrt{p(x, \mathcal{C}_1) p(x, \mathcal{C}_2)} dx \\ &= \int \sqrt{p(x, \mathcal{C}_1) p(x, \mathcal{C}_2)} dx, \end{aligned}$$

as desired. □

Exercise 1.22

Proof. Because $L_{kj} = 1 - I_{kj}$, we have

$$\begin{aligned} \sum_k L_{kj} p(\mathcal{C}_k | x) &= \sum_k (1 - I_{kj}) p(\mathcal{C}_k | x) \\ &= \sum_k p(\mathcal{C}_k | x) - \sum_k I_{kj} p(\mathcal{C}_k | x) \\ &= 1 - p(\mathcal{C}_j | x), \end{aligned} \tag{10}$$

where C_j is the true class of x. If we want to minimize Eq. (10), we should maximize $p(\mathcal{C}_j | x)$. Thus minimizing this loss function means to choose the class having the largest posterior probability. The meaning of this loss function is that it will increase the loss by one when there is a misclassification, while it will not increase or decrease the loss for a correct classification. This is because the elements on the diagonal of the matrix L_{kj} are 0, while the remaining elements are 1. □

Exercise 1.23

Proof.

$$\begin{aligned}\mathbb{E}[L] &= \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} \\ &= \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}|\mathcal{C}_k) p(\mathcal{C}_k) d\mathbf{x} \\ &= \sum_k \sum_j \int_{\mathcal{R}_j} L^*_{kj} p(\mathbf{x}|\mathcal{C}_k) d\mathbf{x},\end{aligned}$$

where $L^*_{kj} = L_{kj} p(\mathcal{C}_k)$. □

Exercise 1.24

Proof. The decision criterion is based on the following

$$\begin{cases} \text{classify as } C_l, \text{ if } \min \sum_k L_{kj} p(\mathcal{C}_k|x) \leq \lambda \\ \text{reject to classify, otherwise} \end{cases},$$

where $l = \arg\min_j \sum_k L_{kj} p(\mathcal{C}_k|x)$. Obviously this decision criterion will give minimum loss. If the loss matrix is given by $L_{kj} = 1 - I_{kj}$, we have

$$\sum_k L_{kj} p(\mathcal{C}_k|x) = 1 - p(\mathcal{C}_j|x).$$

Then

$$\begin{aligned}&\min \sum_k L_{kj} p(\mathcal{C}_k|x) \leq \lambda \\ &\quad \Rightarrow \max_j p(\mathcal{C}_j|x) \geq 1 - \lambda.\end{aligned}$$

Let $\theta = 1 - \lambda$, and we can conclude that $\mathbf{x}$ will be assigned to C_l, where $l = \arg\max_j p(\mathcal{C}_j|x)$, if the largest of posterior probability is more than θ, otherwise will be rejected to assign, as desired. □

Exercise 1.25

Proof. Calculating the partial derivative of $\mathbb{E}[L(\mathbf{t}, \mathbf{y}(\mathbf{x}))]$ with respect to $\mathbf{y}(\mathbf{x})$ and we can get

$$\frac{\partial \mathbb{E}[L(\mathbf{t}, \mathbf{y}(\mathbf{x}))]}{\partial \mathbf{y}(\mathbf{x})} = 2 \int \{\mathbf{y}(\mathbf{x}) - \mathbf{t}\} p(\mathbf{x}, \mathbf{t}) d\mathbf{t}. \tag{11}$$

Let Eq. (11) equals to 0, we have

$$\begin{aligned}
\int \mathbf{y}(\mathbf{x})p(\mathbf{x},\mathbf{t})d\mathbf{t} &= \int \mathbf{t}p(\mathbf{x},\mathbf{t})d\mathbf{t} \\
\Rightarrow \mathbf{y}(\mathbf{x})\int p(\mathbf{x},\mathbf{t})d\mathbf{t} &= \int \Rightarrow \mathbf{t}p(\mathbf{x},\mathbf{t})d\mathbf{t} \\
\Rightarrow \mathbf{y}(\mathbf{x})p(\mathbf{x}) &= \int \mathbf{t}p(\mathbf{x},\mathbf{t})d\mathbf{t} \\
\Rightarrow \mathbf{y}(\mathbf{x}) &= \int \Rightarrow \mathbf{t}p(\mathbf{x}|\mathbf{t})d\mathbf{t} \\
\Rightarrow \mathbf{y}(\mathbf{x}) &= \mathbb{E}[\mathbf{x}|\mathbf{t}],
\end{aligned}$$

as desired. □

Exercise 1.26

Proof. There some mistakes and errata in Eq. (1.90). Let us derive it again. It follows

$$\begin{aligned}
\mathbb{E}[L] &= \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x},t)d\mathbf{x}dt \\
&= \iint \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x},t)d\mathbf{x}dt \\
&= \iint \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x},t)d\mathbf{x}dt + \iint \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x},t)d\mathbf{x}dt \\
&+ 2\iint \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}\}\{\mathbb{E}[t|\mathbf{x}] - t\}d\mathbf{x}dt \\
&= \iint \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x},t)d\mathbf{x}dt + \iint \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x},t)d\mathbf{x}dt \\
&= \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x})d\mathbf{x} + \iint \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x},t)d\mathbf{x}dt \\
&= \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x})d\mathbf{x} + \iint \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(t|\mathbf{x})p(\mathbf{x})d\mathbf{x}dt \\
&= \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x})d\mathbf{x} + \int \left[\int \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(t|\mathbf{x})dt\right] p(\mathbf{x})d\mathbf{x} \\
&= \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x})d\mathbf{x} + \int var[t|\mathbf{x}]p(\mathbf{x})d\mathbf{x}.
\end{aligned}$$

You can read [this] for more errata. We should make it clear that $E[t|\mathbf{x}]$, or $E_t[t|\mathbf{x}]$, is a function with respect to $\mathbf{x}$. It may be $E[t|\mathbf{x}] = \mathbf{x}^T\mathbf{x} + 2$, but it cannot be $E[t|\mathbf{x}] = \mathbf{x}^T\mathbf{x} + t^2 - t$ or other functions that contain variable $\mathbf{t}$. Hence for vector $\mathbf{t}$ we have

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}]\}^2 p(\mathbf{x})d\mathbf{x} + \int var[\mathbf{t}|\mathbf{x}]p(\mathbf{x})d\mathbf{x}.$$

□

Exercise 1.27

Proof. Calculating the derivative of $\mathbb{E}[L_q]$ we have

$$\begin{aligned}\frac{\partial \mathbb{E}[L_q]}{\partial y(\mathbf{x})} &= \int q|y(\mathbf{x}) - t|^{q-1} sign(y(\mathbf{x}) - t) p(\mathbf{x}, t) dt \\ &= \int_{-\infty}^{y(\mathbf{x})} q|y(\mathbf{x}) - t|^{q-1} p(\mathbf{x}, t) dt - \int_{y(\mathbf{x})}^{+\infty} q|y(\mathbf{x}) - t|^{q-1} p(\mathbf{x}, t) dt. \end{aligned} \tag{12}$$

Let Eq. (12) equals to 0 and we can get that $y(\mathbf{x})$ should satisfy

$$\int_{-\infty}^{y(\mathbf{x})} q|y(\mathbf{x}) - t|^{q-1} p(\mathbf{x}, t) dt = \int_{y(\mathbf{x})}^{+\infty} q|y(\mathbf{x}) - t|^{q-1} p(\mathbf{x}, t) dt.$$

If $q = 1$, we have

$$\int_{-\infty}^{y(\mathbf{x})} qp(\mathbf{x}, t) dt = \int_{y(\mathbf{x})}^{+\infty} qp(\mathbf{x}, t) dt,$$

which means that $y(\mathbf{x})$ must be the conditional median of t. For $q \to 0$, let us consider $\|y(\mathbf{x}) - t\|_2^q$. $\|y(\mathbf{x}) - t\|_2^q = 1$ if and only if $y(\mathbf{x}) \neq t$, which means misclassification. $\|y(\mathbf{x}) - t\|_2^q = 0$ if and only if $y(\mathbf{x}) = t$, which means correct classification. Hence for $q \to 0$ we can rewrite loss function, similar to Eq. (1.81), where $L_{kj} = 1 - I_{kj}$. By exercise 1.22, we can draw a same conclusion that minimizing the loss function, when $q \to 0$, is given by the conditional mode. □

Exercise 1.28

Proof. Let us assume that f is the relation between h and p in the form of a function $h(p)$, i.e. $h(x) = f(p(x))$. We have

$$h(x, y) = f(p(x, y)) = f(p(x)p(y)),$$

while

$$h(x, y) = h(x) + h(y) = f(p(x)) + f(p(y)).$$

Thus we have $f(p(x)p(y)) = f(p(x)) + f(p(y))$, or $f(ab) = f(a) + f(b)$, which means $f(p^2) = 2f(p)$. Proving by induction, we should suppose that $f(p^n) = nf(p)$ is correct. By doing so, we have

$$f(p^{n+1}) = f(p^n p) = nf(p) + f(p) = (n + 1)f(p),$$

which means $f(p^n) = nf(p)$ is correct for all positive integer n. We also can get

$$f(p^n) = f\left(\left(p^{\frac{n}{m}}\right)^m\right) = mf\left(p^{\frac{n}{m}}\right) = nf(p),$$

which implies $f\left(p^{\frac{n}{m}}\right) = \frac{n}{m}f(p)$ for m is a positive integer. Next let we derive the relationship between $f(x)$ and $\ln x$. Calculating the both sides' derivative of $f(xy) = f(x) + f(y)$, we have

$$f'(xy)(xdy + ydx) = f'(x)dx + f'(y)dy,$$

which says

$$\begin{aligned} xf'(xy) &= f'(y) \\ yf'(xy) &= f'(x). \end{aligned}$$

Multiply both sides of the two equations by y and x respectively, we have

$$xyf'(xy) = yf'(y) = xf'(x) = C,$$

where C is a constant. C is not zero for $f(x)$ is monotonic. Thus we have

$$f'(x) = \frac{C}{x},$$

which means $f(x) = C\ln(x) + D$, D is a constant. Thus $f(x) \propto \ln(x)$. □

Exercise 1.29

Proof. The entropy of M-state discrete variable X is

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

We choose $f(x) = \ln(x)$ as the **concave** function. It follows

$$\begin{aligned} H[x] &= \sum_{i=1}^{M} p(x_i)\frac{1}{\ln p(x_i)} \\ &\geq \ln\left\{\sum_{i=1}^{M} p(x_i)\frac{1}{p(x_i)}\right\} \\ &= \ln M, \end{aligned}$$

as desired. The equality holds if and only if X follows the uniform distribution. □

Exercise 1.30

Proof. Because $p(x) = \mathcal{N}(x|\mu, \sigma^2)$ and $q(x) = \mathcal{N}(x|m, s^2)$, we have

$$KL(p\|q) = -\int p(x)ln\left\{\frac{q(x)}{p(x)}\right\}dx$$

$$
\begin{aligned}
&= \int \mathcal{N}(x|\mu,\sigma^2) ln \left\{ \frac{p(x)}{q(x)} \right\} dx \\
&= \int \mathcal{N}(x|\mu,\sigma^2) ln(p(x)) dx - \int \mathcal{N}(x|\mu,\sigma^2) ln(q(x)) dx \\
&= \int \mathcal{N}(x|\mu,\sigma^2) ln \left\{ \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{\sigma^2} \right) \right\} dx \\
&- \int \mathcal{N}(x|\mu,\sigma^2) ln \left\{ \frac{1}{\sqrt{2\pi}s} \exp\left(-\frac{(x-m)^2}{s^2} \right) \right\} dx \\
&= \ln\left(\frac{1}{\sqrt{2\pi}\sigma} \right) \int \mathcal{N}(x|\mu,\sigma^2) dx - \ln\left(\frac{1}{\sqrt{2\pi}s} \right) \int \mathcal{N}(x|\mu,\sigma^2) dx \\
&+ \int \mathcal{N}(x|\mu,\sigma^2) \left(-\frac{(x-\mu^2)}{2\sigma^2} \right) dx - \int \mathcal{N}(x|\mu,\sigma^2) \left(-\frac{(x-m^2)}{2s^2} \right) dx \\
&= \ln\left(\frac{s}{\sigma} \right) - \frac{1}{2\sigma^2} \int (x-\mu)^2 \mathcal{N}(x|\mu,\sigma^2) dx + \int \mathcal{N}(x|\mu,\sigma^2) \left(\frac{x^2+m^2-2mx}{2s^2} \right) dx \\
&= \ln\left(\frac{s}{\sigma} \right) - \frac{1}{2\sigma^2} \mathbb{E}[(x-\mu)^2] + \frac{1}{2s^2} \int x^2 \mathcal{N}(x|\mu,\sigma^2) dx + \frac{m^2}{2s^2} \int \mathcal{N}(x|\mu,\sigma^2) dx \\
&- \frac{m}{s^2} \int x \mathcal{N}(x|\mu,\sigma^2) dx \\
&= ln\left(\frac{s}{\sigma} \right) - \frac{1}{2} + \frac{1}{2s^2} \mathbb{E}[x^2] + \frac{m^2}{2s^2} - \frac{m}{s^2} \mathbb{E}[x] \\
&= ln\left(\frac{s}{\sigma} \right) + \frac{\sigma^2 + \mu^2 + m^2 - 2\mu m}{2s^2} - \frac{1}{2}.
\end{aligned}
$$

□

Exercise 1.31

Proof. According to Eq. (1.112) and (1.121), we have

$$
\begin{aligned}
H[\mathbf{x}] + H[\mathbf{y}] - H[\mathbf{x},\mathbf{y}] &= H[\mathbf{x}] + H[\mathbf{y}] - (H[\mathbf{x}] + H[\mathbf{y}|\mathbf{x}]) \\
&= H[\mathbf{x}] - H[\mathbf{y}|\mathbf{x}]) \\
&= I[\mathbf{x},\mathbf{y}] \geq 0
\end{aligned}
$$

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

Exercise 1.32

Proof. By Eq. (1.27) we have

$$
p_{\mathbf{y}}(\mathbf{y}) d\mathbf{y} = p_{\mathbf{x}}(\mathbf{x}) |\mathbf{A}|^{-1} d\mathbf{x}.
$$

It follows

$$
H[\mathbf{y}] = -\int p_{\mathbf{y}}(\mathbf{y}) \ln p_{\mathbf{y}}(\mathbf{y}) d\mathbf{y}
$$

$$
\begin{aligned}
&= -\int p_{\mathbf{y}}(\mathbf{y}) \ln \left\{ p_{\mathbf{x}}(\mathbf{x}) |\mathbf{A}|^{-1} \right\} d\mathbf{y} \\
&= -\int |\mathbf{A}| p_{\mathbf{y}}(\mathbf{y}) \ln \left\{ p_{\mathbf{x}}(\mathbf{x}) |\mathbf{A}|^{-1} \right\} d\mathbf{x} \\
&= -\int p_{\mathbf{x}}(\mathbf{x}) \ln \left\{ p_{\mathbf{x}}(\mathbf{x}) |\mathbf{A}|^{-1} \right\} d\mathbf{x} \\
&= -\int p_{\mathbf{x}}(\mathbf{x}) \ln p_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} + \ln |\mathbf{A}| \int p_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\
&= H[\mathbf{x}] + \ln |\mathbf{A},
\end{aligned}
$$

as desired. □

Exercise 1.33

Proof. Let $H[y|x] = 0$. Then we have

$$
\begin{aligned}
\sum_i \sum_j p(x_i, y_j) &= 0 \\
\Rightarrow p(x_i, y_j) ln p(y_j|x_i) &= 0 \\
\Rightarrow p(y_j|x_i) p(x_i) ln p(y_j|x_i) &= 0 \\
\Rightarrow p(y_j|x_i) ln p(y_j|x_i) &= 0
\end{aligned}
$$

$p(y|x) \neq 0$ because both variables x and y are possible. Thus $p(y_j|x_i) = 1$, which means that for each x there is only one variable y such that $p(y|x) \neq 0$. □

Exercise 1.34

Proof. Let us assume two functional F and G

$$
\begin{aligned}
F[p(x)] &= \int_{-\infty}^{+\infty} p(x) \ln p(x) dx; \\
G[p(x)] &= \int_{-\infty}^{+\infty} f(x) p(x) dx.
\end{aligned}
$$

We have

$$
F[p(x) + \epsilon\eta(x)] = \int p(x) \ln(p(x) + \epsilon\eta(x)) dx + \int \epsilon\eta(x) \ln(p(x) + \epsilon\eta(x)) dx. \tag{13}
$$

By Taylor Theorem, we have

$$
\ln(p(x) + \epsilon\eta(x)) = \ln p(x) + \frac{\epsilon\eta(x)}{p(x)} + O(\epsilon^2).
$$

Thus Eq. (13) can be written as

$$\begin{aligned} F[p(x)+\epsilon\eta(x)] &= \int p(x)\left(\ln p(x)+\frac{\epsilon\eta(x)}{p(x)}\right)dx+\epsilon\int\left(\ln p(x)+\frac{\epsilon\eta(x)}{p(x)}\right)\eta(x)dx+O(\epsilon^2) \\ &= \int p(x)\ln p(x)dx+\epsilon\int(\ln p(x)+1)\eta(x)dx+O(\epsilon^2) \end{aligned} \tag{14}$$

By Eq. (D.3), $F[p(x)+\epsilon\eta(x)]$ is equivalent to

$$F[p(x)+\epsilon\eta(x)] = \int p(x)\ln p(x)dx+\epsilon\int\frac{\partial F}{\partial p(x)}\eta(x)dx+O(\epsilon^2). \tag{15}$$

Comparing Eq. (14) with (15), we can get

$$\frac{\partial F}{\partial p(x)} = \ln p(x)+1. \tag{16}$$

Similarly, for G we have

$$\frac{\partial G}{\partial p(x)} = f(x). \tag{17}$$

Suppose

$$\begin{aligned} T = &-\int_{-\infty}^{+\infty}p(x)\ln p(x)dx+\lambda_1\left(\int_{-\infty}^{+\infty}p(x)dx-1\right)+\lambda_2\left(\int_{-\infty}^{+\infty}xp(x)dx-\mu\right) \\ &+\lambda_3\left(\int_{-\infty}^{+\infty}(x-\mu)^2p(x)dx-\sigma^2\right). \end{aligned}$$

By Eq. (16) and (17) we can get

$$\frac{\partial T}{\partial p(x)} = -(\ln p(x)+1)+\lambda_1+\lambda_2x+\lambda_3(x-\mu)^2,$$

which means

$$p(x) = \exp\{-1+\lambda_1+\lambda_2x+\lambda_3(x-\mu)^2\},$$

when $\frac{\partial T}{\partial p(x)}=0$, as desired. Let us assume

$$p(x) = \exp(-a(x-b)^2+c).$$

By Eq. (1.105) we have

$$\begin{aligned} \int_{-\infty}^{+\infty}p(x)dx &= \int_{-\infty}^{+\infty}\exp\{-a(x-b)^2+c\}dx \\ &= \frac{e^c}{\sqrt{a}}\int_{-\infty}^{+\infty}\exp\{-a(x-b)^2\}d\{\sqrt{a}(x-b)\}dx \\ &= \sqrt{\frac{\pi}{a}}e^c = 1, \end{aligned} \tag{18}$$

where we utilize the Gaussian integral. By Eq. (1.106) we have

$$\begin{aligned}
\int_{-\infty}^{+\infty} xp(x)dx &= \int_{-\infty}^{+\infty} x\exp\{-a(x-b)^2+c\}dx \\
&= \int_{-\infty}^{+\infty} (x-b+b)\exp\{-a(x-b)^2+c\}dx \\
&= \int_{-\infty}^{+\infty} (x-b)\exp\{-a(x-b)^2+c\}dx + b\int_{-\infty}^{+\infty} \exp\{-a(x-b)^2+c\}dx \\
&= \frac{e^c}{a}\int_{-\infty}^{+\infty} \sqrt{a}(x-b)\exp\{-a(x-b)^2+c\}d\sqrt{a}(x-b) + \sqrt{\frac{\pi}{a}}e^c b \\
&= \int_{-\infty}^{+\infty} t\exp(-t^2)dt + b \\
&= b = \mu, \qquad (19)
\end{aligned}$$

where we observe that $t\exp(-t^2)$ is an odd function and utilize its properties. By Eq. (1.107) we have

$$\begin{aligned}
\int_{-\infty}^{+\infty} (x-\mu)^2 p(x)dx &= \int_{-\infty}^{+\infty} x^2 p(x)dx - 2\mu\int_{-\infty}^{+\infty} xp(x)dx + \mu^2\int_{-\infty}^{+\infty} p(x)dx \\
&= \int_{-\infty}^{+\infty} (x-b+b)^2 p(x)dx - 2\mu^2 + \mu^2 \\
&= \int_{-\infty}^{+\infty} (x-b)^2 p(x)dx + b^2\int_{-\infty}^{+\infty} p(x)dx + 2b\int_{-\infty}^{+\infty} (x-b)p(x)dx - \mu^2 \\
&= \int_{-\infty}^{+\infty} (x-b)^2\exp\{-a(x-b)^2+c\}dx \\
&= \frac{e^c}{a\sqrt{a}}\int_{-\infty}^{+\infty} t^2 exp(-t^2)dt \\
&= \frac{e^c\sqrt{\pi}}{2a\sqrt{a}} = \sigma^2. \qquad (20)
\end{aligned}$$

With the combination of Eq. (18-20), we can get

$$\begin{aligned}
a &= \frac{1}{2\sigma^2}; \\
b &= \mu; \\
c &= -\ln(2\pi\sigma^2)^{\frac{1}{2}},
\end{aligned}$$

which means

$$\begin{aligned}
\lambda_1 &= 1 - \frac{1}{2}\ln(2\pi\sigma^2); \\
\lambda_2 &= 0; \\
\lambda_3 &= -\frac{1}{2\sigma^2},
\end{aligned}$$

as desired. □

Exercise 1.35

Proof. Similar to Exercise 1.30, we have

$$\begin{aligned}
H[x] &= -\int p(x)\ln p(x)dx \\
&= -\int \mathcal{N}(x|\mu,\sigma^2)\ln\left\{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right\}dx \\
&= \ln\sqrt{2\pi}\sigma\int \mathcal{N}(x|\mu,\sigma^2)dx + \int \mathcal{N}(x|\mu,\sigma^2)\frac{x^2+\mu^2-x\mu}{2\sigma^2}dx \\
&= \ln\sqrt{2\pi}\sigma + \frac{\mu^2+\sigma^2+\mu^2-\mu^2}{2\sigma^2} \\
&= \frac{1}{2}\{1+\ln(2\pi\sigma^2)\},
\end{aligned}$$

as desired. □

Exercise 1.36

Proof. (1) Suppose that $f''(x) > 0$ and

$$H(x) = \lambda f(x) + (1-\lambda)f(x_2) - f(\lambda x + (1-\lambda)x_2).$$

It follows

$$\begin{aligned}
H'(x) &= \lambda f'(x) - \lambda f'(\lambda x + (1-\lambda)x_2) \\
&= \lambda(f'(x) - f'(\lambda x + (1-\lambda)x_2)).
\end{aligned}$$

Suppose $x_1 < x_2$. $f'(x_1) < f'(x_2)$ for $f''(x) > 0$. Thus

$$H'(x) = \lambda(f'(x_1) - f'(\lambda x_1 + (1-\lambda)x_2)) < 0,$$

which means $H(x_1) < H(x_2) = 0$, i.e.,

$$\lambda f(x_1) + (1-\lambda)f(x_2) < f(\lambda x_1 + (1-\lambda)x_2),$$

as desired.
(2) Suppose $f(x)$ is a convex function and $x_1 < x_3 < x_2$. Let $\lambda = \frac{x_2-x_3}{x_2-x_1}$ and hence $0 < \lambda < 1$ and $x_3 = \lambda x_1 + (1-\lambda)x_2$. It follows

$$\begin{aligned}
&f(x_3) < \lambda f(x_1) + (1-\lambda)f(x_2) \\
\Rightarrow &\frac{f(x_2)-f(x_3)}{x_2-x_3} = \frac{f(x_2)-f(x_3)}{\lambda(x_2-x_1)} > \frac{f(x_2)-f(x_1)}{x_2-x_1} > \frac{f(x_3)-f(x_1)}{(1-\lambda)(x_2-x_1)} = \frac{f(x_3)-f(x_1)}{\lambda(x_3-x_1)}.
\end{aligned}$$

Thus we can get

$$\begin{aligned}
f'(x_2) &= \lim_{x_3\to x_2}\frac{f(x_2)-f(x_3)}{x_2-x_3} \geq \frac{f(x_2)-f(x_4)}{x_2-x_4} > \frac{f(x_2)-f(x_1)}{x_2-x_1} \\
&> \frac{f(x_5)-f(x_1)}{x_5-x_1} \geq \lim_{x_3\to x_2}\frac{f(x_6)-f(x_1)}{x_6-x_1} = f'(x_1),
\end{aligned}$$

where $x_2 > x_3 > x_4 > x_5 > x_6 > x_1$. Hence we can get $f''(x) > 0$. □

Exercise 1.37

Proof.

$$\begin{aligned}
H[\mathbf{x},\mathbf{y}] &= -\iint p(\mathbf{x},\mathbf{y})\ln p(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y} \\
&= -\iint p(\mathbf{x},\mathbf{y})\ln\{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\}d\mathbf{x}d\mathbf{y} \\
&= -\iint p(\mathbf{x},\mathbf{y})\ln p(\mathbf{y}|\mathbf{x})d\mathbf{x}d\mathbf{y} - \iint p(\mathbf{x},\mathbf{y})\ln p(\mathbf{x})d\mathbf{x}d\mathbf{y} \\
&= H[\mathbf{y}|\mathbf{x}] - \int \ln p(\mathbf{x})\left(\int p(\mathbf{x},\mathbf{y})d\mathbf{y}\right)d\mathbf{x} \\
&= H[\mathbf{y}|\mathbf{x}] - \int p(\mathbf{x})lnp(\mathbf{x})d\mathbf{x} \\
&= H[\mathbf{y}|\mathbf{x}] + H[\mathbf{x}],
\end{aligned}$$

as desired. □

Exercise 1.38

Proof. Eq. (1.115) reduces to Eq.(1.114) when $M = 2$, which means Eq. (1.115) is correct when $M = 2$. Let us assume that Eq. (115) is correct when $M = M'$. For $M = M' + 1$ we have

$$\lambda_{M'+1}f(x_{M'+1}) + (1-\lambda_{M'+1})f(b) \geq f(\lambda_{M'+1}x_{M'+1} + (1-\lambda_{M'+1})f(b)),$$

when $a = b = x_{M'+1}$ in Eq. (1.114). Thus

$$\begin{aligned}
f\left(\sum_{i=1}^{M'+1}\lambda_i x_i\right) &= f\left(\sum_{i=1}^{M'}\lambda_i x_i + \lambda_{M'+1}x_{M'+1}\right) \\
&= f\left((1-\lambda_{M'+1})\sum_{i=1}^{M'}\eta_i x_i + \lambda_{M'+1}x_{M'+1}\right) \\
&\leq \lambda_{M'+1}f(x_{M'+1}) + (1-\lambda_{M'+1})f\left(\sum_{i=1}^{M'}\eta_i x_i\right) \\
&\leq \lambda_{M'+1}f(x_{M'+1}) + \sum_{i=1}^{M'}\left(\frac{\eta_i}{1-\lambda_{M'+1}}\right)f(x_i) \\
&= \sum_{i=1}^{M'+1}\lambda_i f(x_i),
\end{aligned}$$

where we utilize $\eta_i = \frac{\lambda_i}{1-\lambda_{M'+1}}$ and thus $\sum_{i=1}^{M}\eta_i = 1$, as desired. □

Exercise 1.39

Proof.

$$
\begin{aligned}
&H[x] = -\sum_x p(x)\ln p(x) = -\frac{2}{3}\ln\frac{2}{3} - \frac{1}{3}\ln\frac{1}{3} \simeq 0.64;\\
&H[y] = H[x] \simeq 0.64;\\
&H[y|x] = -\sum_x\sum_y p(x,y)\ln p(y|x) \simeq 0.46;\\
&H[x|y] = -\sum_x\sum_y p(x,y)\ln p(x|y) \simeq 0.46;\\
&H[x,y] = H[y|x] + H[x] \simeq 1.10;\\
&I[x,y] = H[x] - H[x|y] \simeq 0.17.
\end{aligned}
$$

Diagram is shown in Fig. 1. □

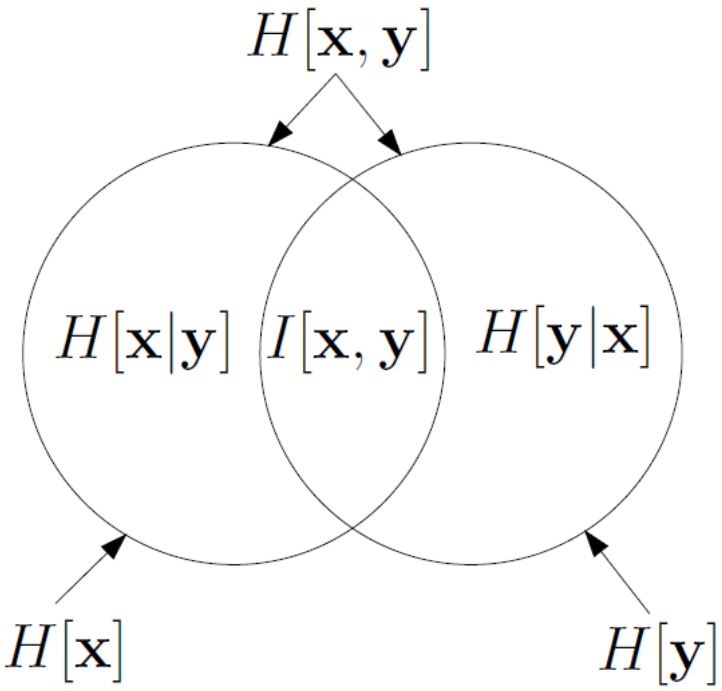

Figure 1: A diagram showing the arithmetic relationship between entropy, conditional entropy, and mutual information.

Exercise 1.40

Proof.

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

which means

$$\frac{\sum_{i=1}^{M} x_i}{M} > \left(\prod_{i=1}^{M} x_i\right)^{\frac{1}{M}},$$

as desired. □

Exercise 1.41

Proof.

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

By symmetry, we can get

$$I[\mathbf{x},\mathbf{y}] = H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}],$$

as desired. □