Solutions to selected exercises of Bishop (2006), Chapter 1

Jan Drugowitsch

Nov 5, 2021

Exercise 1.3

Suppose that we 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?

This exercise requires us to turn counts into probabilities, and to manipulate them to compute the desired ones. In particular, the first and second question require us compute to $p(\text{apple})$ and $p(g|\text{orange})$, respectively. To do so, let us first turn the provided fruit counts into conditional probabilities, $p(\text{fruit}|\text{box})$. We can find these for each combination of fruit and box by counting which fraction a specific fruit constitutes of the total number of fruits in this box. This yields the following table of conditional probabilities, to which we have additionally added the prior probabilities per box as the last column.

box	fruit			prior $p(\text{box})$
	apple	orange	lime	
r	0.3	0.4	0.3	0.2
b	0.5	0.5	0.0	0.2
g	0.3	0.3	0.4	0.6

Each element in this table provides $p(\text{fruit}|\text{box})$ for a specific combination of box and fruit. As required by the rules of probabilities, each row sums to one across all fruits (excluding the prior, naturally).

Equipped with these probabilities, we can now answer the exercise's questions. The first question requires us to compute $p(\text{apple})$. We can find this probability in two steps. The first is to compute the joint probability $p(\text{apple}, \text{box})$ and marginalize over box, $p(\text{apple}) = \sum_{\text{box}} p(\text{apple}, \text{box})$. The second is to find this joint probability for each box by $p(\text{apple}, \text{box}) = p(\text{apple}|\text{box})p(\text{box})$, which is the product of the above conditional probabilities and the prior for each box. Together, this results in

$$\begin{aligned} p(\text{apple}) &= p(\text{apple}|r)p(r) + p(\text{apple}|b)p(b) + p(\text{apple}|g)p(g) \\ &= 0.3 \times 0.2 + 0.5 \times 0.2 + 0.3 \times 0.6 \\ &= 0.34. \end{aligned} \tag{1}$$

Even though not required for the exercise, we can perform a similar calculation for other fruits, revealing $p(\text{orange}) = 0.36$ and $p(\text{lime}) = 0.3$. Importantly, these probabilities sum to one, that is $p(\text{apple}) + p(\text{orange}) + p(\text{lime}) = 1$, as required by the rules of probabilities. Note that we wouldn't have found these results by counting the total number of apples and dividing that by the total number of fruits, as this ignores the prior $p(\text{box})$.

The second question requires us to compute $p(g|\text{orange})$. To find this conditional probability from the ones computed above we can use Bayes rule,

$$p(g|\text{orange}) = \frac{p(\text{orange}|g)p(g)}{p(\text{orange})} = \frac{0.3 \times 0.6}{0.36} = 0.5, \tag{2}$$

where we have used the previously computed $p(\text{orange})$. Interestingly, the probability that the orange comes from the green box is higher than it coming from the red box (which turns out to be $p(r|\text{orange}) \approx 0.22$), even though there are more oranges in the red box. The reason for this is that it is a-priori more likely that fruits are drawn from the green than the red box, which biases these probabilities towards the green box.

Exercise 1.5

Using the definition (1.38, [Bishop, 2006]) show that var$[f(x)]$ satisfies (1.39, [Bishop, 2006]).

Here, (1.38, [Bishop, 2006]) refers to

$$\text{var}[f] = \mathbb{E}\left[\left(f(x) - \mathbb{E}\left[f(x)\right]\right)^2\right] \tag{3}$$

and (1.39, [Bishop, 2006]) to

$$\text{var}[f] = \mathbb{E}\left[f(x)^2\right] - \mathbb{E}\left[f(x)\right]^2. \tag{4}$$

To show that these definitions are equivalent, we expand the square in the first definition to give

$$\begin{aligned}
\text{var}[f] &= \mathbb{E}\left[\left(f(x) - \mathbb{E}\left[f(x)\right]\right)^2\right] \\
&= \mathbb{E}\left[f(x)^2 + \mathbb{E}\left[f(x)\right]^2 - 2f(x)\mathbb{E}\left[f(x)\right]\right] \\
&= \mathbb{E}\left[f(x)^2\right] + \mathbb{E}\left[f(x)\right]^2 - 2\mathbb{E}\left[f(x)\right]\mathbb{E}\left[f(x)\right] \\
&= \mathbb{E}\left[f(x)^2\right] - \mathbb{E}\left[f(x)\right]^2,
\end{aligned} \tag{5}$$

where the second equality uses that the expectation of a sum is the sum of expectations, that is $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, and that $\mathbb{E}[f(x)]$ is not a function of x such that $\mathbb{E}\left[f(x)\mathbb{E}[f(x)]\right] = \mathbb{E}[f(x)]\mathbb{E}[f(x)]$.

Exercise 1.6

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

To show this note that, by the definition of independence, $p(x, y) = p(x)p(y)$. This implies that the expectation $\mathbb{E}[xy]$ becomes

$$\mathbb{E}[xy] = \sum_{x,y} p(x, y)xy = \sum_{x,y} p(x)p(y)xy = \left(\sum_x p(x)x\right)\left(\sum_y p(y)y\right) = \mathbb{E}[x]\mathbb{E}[y], \tag{6}$$

where we have assumed that both x and y are discrete random variables. It is easy to verify that the above also holds for continuous random variables, in which case the sums get replaced by integrals. Using this property, we can re-write their covariance as

$$\text{cov}\left(x, y\right) = \mathbb{E}\left[xy\right] - \mathbb{E}[x]\mathbb{E}[y] = \mathbb{E}[x]\mathbb{E}[y] - \mathbb{E}[x]\mathbb{E}[y] = 0, \tag{7}$$

which yields the desired result.

Exercise 1.9 (univariate case only)

Show that the mode (i.e., the maximum) of the Gaussian distribution (1.46, [Bishop, 2006]) is given by μ. [... multivariate case ...]

The probability density of a Gaussian distribution is according to (1.46, [Bishop, 2006]) given by

$$\mathcal{N}\left(x|\mu, \sigma^2\right) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}\left(x - \mu\right)^2\right\}. \tag{8}$$

To find its maximum, it is easier to work with the log-probability,

$$\log \mathcal{N}\left(x|\mu,\sigma^2\right) = -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left(x-\mu\right)^2. \tag{9}$$

As the logarithm is a strictly increasing function, the location of the maximum of the log-probability is the same as for the probability. Thus, we can find this maximum by taking the first derivative of this log-probability with respect to x and setting it to zero, resulting in

$$-\frac{1}{\sigma^2}(x_{mode}-\mu) = 0. \tag{10}$$

Solving for x_{mode} result in

$$x_{mode} = \mu, \tag{11}$$

as required.

Exercise 1.11

By setting the derivatives of the log likelihood function (1.54, [Bishop, 2006]) with respect to μ and σ^2 to zero, verify the results (1.55, [Bishop, 2006]) and (1.56, [Bishop, 2006]).

The log likelihood function (1.54, [Bishop, 2006]) is given by

$$\ln p\left(\boldsymbol{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(x_n-\mu\right)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln\left(2\pi\right). \tag{12}$$

Taking the derivative with respect to μ and setting it to zero yields

$$-\frac{1}{\sigma^2}\sum_{n=1}^{N}\left(x_n-\mu_{ML}\right) = 0, \tag{13}$$

which, solved for μ_{ML} results in

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N}, \tag{14}$$

which matches (1.55, [Bishop, 2006]). Setting the derivative with respect to σ^2 to zero results in

$$\frac{1}{2\left(\sigma^2_{ML}\right)^2}\sum_{n=1}^{N}\left(x_n-\mu\right)^2 - \frac{N}{2}\frac{1}{\sigma^2_{ML}} = 0. \tag{15}$$

Solving for σ^2_{ML} gives

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}\left(x_n-\mu\right)^2, \tag{16}$$

which matches (1.56, [Bishop, 2006]) once μ is replaced by μ_{ML}.

Exercise 1.12

Using the results (1.49, [Bishop, 2006]) and (1.50, [Bishop, 2006]) show that

$$\mathbb{E}[x_n x_m] = \mu^2 + I_{nm}\sigma^2 \tag{17}$$

where x_n and y_n denote data points samples from a Gaussian distribution with mean μ and variance σ^2, and I_{nm} satisfies $I_{nm} = 1$ if $n = m$ and I_{nm} otherwise. Hence, prove the results (1.57, [Bishop, 2006]) and (1.58, [Bishop, 2006]).

The above (1.49, [Bishop, 2006]) and (1.50, [Bishop, 2006]) refer to

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x \mathrm{d}x = \mu, \tag{18}$$

and

$$\mathbb{E}\left[x^2\right] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x \mathrm{d}x = \mu^2 + \sigma^2, \tag{19}$$

respectively. For the case that $n = m$, x_n and x_m refer to the same random variable, such that $\mathbb{E}[x_n x_m] = \mathbb{E}\left[x_n^2\right] = \mu^2 + \sigma^2$. If, instead, $n \neq m$, x_n and x_m, such that $\mathbb{E}[x_n x_m] = \mathbb{E}[x_n]\mathbb{E}[x_m] = \mu^2$ (see Exercise 1.6). This confirms that $\mathbb{E}[x_n x_m] = \mu^2 + I_{nm}\sigma^2$, as required.

To use this to prove that $\mathbb{E}\left[\mu_{ML}\right] = \mu$ (1.57, [Bishop, 2006]), note that $\mu_{ML} = N^{-1}\sum_{n=1}^{N} x_n$ (Exercise 1.11 and (1.55, [Bishop, 2006])), such that

$$\mathbb{E}\left[\mu_{ML}\right] = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\left[x_n\right] = \frac{1}{N}\sum_{n=1}^{N} \mu = \mu. \tag{20}$$

Showing that $\mathbb{E}\left[\sigma_{ML}^2\right] = N^{-1}(N-1)\sigma^2$ (1.58, [Bishop, 2006]) requires more work. Noting that $\sigma_{ML}^2 = N^{-1}\sum_{n=1}(x_n - \mu_{ML})^2$ (Exercise 1.11 and (1.56, [Bishop, 2006])), we find its expectation to decompose into

$$\begin{aligned} \mathbb{E}\left[\sigma_{ML}^2\right] &= \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\left[\left(x_n - \mu_{ML}\right)^2\right] \\ &= \frac{1}{N}\sum_{n=1}^{N} \left(\mathbb{E}\left[x_n^2\right] + \mathbb{E}\left[\mu_{ML}^2\right] - 2\mathbb{E}\left[x_n \mu_{ML}\right]\right). \end{aligned} \tag{21}$$

To find the three terms in bracket, note that the first term is given by $\mathbb{E}\left[x_n^2\right] = \mu^2 + \sigma^2$. For the second term we substitute the expression for μ_{ML} and evaluate the expectation, resulting in

$$\mathbb{E}\left[\mu_{ML}^2\right] = \frac{1}{N^2}\mathbb{E}\left[\left(\sum_{n=1}^{N} x_n\right)^2\right] = \frac{1}{N^2}\sum_{mn} \mathbb{E}\left[x_n x_m\right] = \frac{1}{N^2}\sum_{nm} \left(\mu^2 + I_{nm}\sigma^2\right) = \frac{1}{N^2}\left(N^2\mu^2 + N\sigma^2\right) = \mu^2 + \frac{\sigma^2}{N}. \tag{22}$$

The same approach applied to the third term results in

$$\mathbb{E}\left[x_n \mu_{ML}\right] = \frac{1}{N}\mathbb{E}\left[x_n \sum_{m=1}^{N} x_m\right] = \frac{1}{N}\sum_{m=1}^{N} \left(\mu^2 + I_{nm}\sigma^2\right) = \frac{1}{N}\left(N\mu^2 + \sigma^2\right) = \mu^2 + \frac{\sigma^2}{N}. \tag{23}$$

Re-substituting these results into the original expression gives

$$\mathbb{E}\left[\sigma_{ML}^2\right] = \frac{1}{N}\sum_{n=1}^{N} \left(\mu^2 + \sigma^2 + \mu^2 + \frac{\sigma^2}{N} - 2\mu^2 - 2\frac{\sigma^2}{N}\right) = \frac{N-1}{N}\sigma^2, \tag{24}$$

as desired.

Exercise 1.30

Evaluate the Kullback-Leibler divergence (1.113, [Bishop, 2006]) between two Gaussians $p(x) = \mathcal{N}(x|\mu, \sigma^2)$ *and* $q(x) = \mathcal{N}(x|m, s^2)$.

By (1.113, [Bishop, 2006]), the Kullback-Leiber divergence between p and q is defined as

$$\mathrm{KL}\left(p\|q\right) = -\int p\left(\boldsymbol{x}\right) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} = \mathbb{E}_{p(\boldsymbol{x})}\left[\ln p\left(\boldsymbol{x}\right)\right] - \mathbb{E}_{p(\boldsymbol{x})}\left[\ln q\left(\boldsymbol{x}\right)\right]. \tag{25}$$

To find this divergence when p and q are Gaussians, let us first evaluate the two expectations separately. For $\mathbb{E}_{p(\boldsymbol{x})}\left[\ln q\left(\boldsymbol{x}\right)\right]$ we find

$$\begin{aligned}\mathbb{E}_{p(\boldsymbol{x})}\left[\ln q\left(\boldsymbol{x}\right)\right] &= -\frac{1}{2}\ln\left(2\pi s^2\right) - \frac{1}{2s^2}\mathbb{E}_{p(\boldsymbol{x})}\left[\left(x-m\right)^2\right] \\ &= -\frac{1}{2}\ln\left(2\pi s^2\right) - \frac{1}{2s^2}\left(\mathbb{E}_{p(\boldsymbol{x})}\left[x^2\right] + m^2 - 2m\mathbb{E}_{p(\boldsymbol{x})}\left[x\right]\right) \\ &= -\frac{1}{2}\ln\left(2\pi s^2\right) - \frac{1}{2s^2}\left(\mu^2 + \sigma^2 + m^2 - 2m\mu\right) \\ &= -\frac{1}{2}\ln\left(2\pi s^2\right) - \frac{1}{2s^2}\left(\mu - m\right)^2 - \frac{1}{2s^2}\sigma^2,\end{aligned} \tag{26}$$

where we have used $\mathbb{E}_{p(\boldsymbol{x})}\left[x^2\right] = \mu^2 + \sigma^2$. The second expectation, $\mathbb{E}_{p(\boldsymbol{x})}\left[\ln p\left(\boldsymbol{x}\right)\right]$ is the negative entropy of p, and is given by

$$\begin{aligned}\mathbb{E}_{p(\boldsymbol{x})}\left[\ln p\left(\boldsymbol{x}\right)\right] &= -\frac{1}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\mathbb{E}_{p(\boldsymbol{x})}\left[\left(x-\mu\right)^2\right] \\ &= -\frac{1}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2},\end{aligned} \tag{27}$$

where we have used $\mathbb{E}_{p(\boldsymbol{x})}\left[\left(x-\mu\right)^2\right] = \sigma^2$. Summing up the two expectations results, after some simplification, in the final expression

$$\mathrm{KL}\left(p\|q\right) = \ln\frac{s}{\sigma} + \frac{\left(\mu-m\right)^2 + \sigma^2}{2s^2} - \frac{1}{2}. \tag{28}$$

Exercise 1.37

Using the definition (1.111, [Bishop, 2006]) together with the product rule of probability, prove the result (1.112, [Bishop, 2006]).

Definition (1.111, [Bishop, 2006]) states that

$$H\left[\boldsymbol{y}|\boldsymbol{x}\right] = -\iint p\left(\boldsymbol{y},\boldsymbol{x}\right)\ln p\left(\boldsymbol{y}|\boldsymbol{x}\right)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x}. \tag{29}$$

To use this to prove $H\left[\boldsymbol{x},\boldsymbol{y}\right] = H\left[\boldsymbol{y}|\boldsymbol{x}\right] + H\left[\boldsymbol{x}\right]$ (1.112, [Bishop, 2006]), we note that, by definition, $p\left(\boldsymbol{y}|\boldsymbol{x}\right) = p(\boldsymbol{x},\boldsymbol{y})/p(\boldsymbol{y})$. Subtituting this into the conditional entropy, $H\left[\boldsymbol{y}|\boldsymbol{x}\right]$ results in

$$H\left[\boldsymbol{y}|\boldsymbol{x}\right] = -\iint p\left(\boldsymbol{y},\boldsymbol{x}\right)\ln p\left(\boldsymbol{x},\boldsymbol{y}\right)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x} + \iint p\left(\boldsymbol{y},\boldsymbol{x}\right)\ln p\left(\boldsymbol{y}\right)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x} = H\left[\boldsymbol{x},\boldsymbol{y}\right] - H\left[\boldsymbol{x}\right], \tag{30}$$

where we used $\iint p\left(\boldsymbol{y},\boldsymbol{x}\right)\ln p\left(\boldsymbol{y}\right)\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x} = \int p\left(\boldsymbol{y}\right)\ln p\left(\boldsymbol{y}\right)\mathrm{d}\boldsymbol{y}$ for the second equality. Re-expressing the above in terms of $H\left[\boldsymbol{x},\boldsymbol{y}\right]$ confirms the result. The same can be shown by substituting $\ln p\left(\boldsymbol{x},\boldsymbol{y}\right) = \ln\left(\boldsymbol{y}|\boldsymbol{x}\right) + \ln p\left(\boldsymbol{x}\right)$ into the expression for $H\left[\boldsymbol{x},\boldsymbol{y}\right]$.

Exercise 1.39

Consider two binary variables x and y having the joint distribution given in Table 1.3 [Bishop, 2006]. Evaluate the following quantities

(a)	$H[x]$	*(c)*	$H[y\|x]$	*(e)*	$H[x,y]$
(b)	$H[y]$	*(d)*	$H[x\|y]$	*(f)*	$I[x,y]$.

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

Table 1.3 of [Bishop, 2006] specifies the joint distribution of x and y to be given by

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

Summing over the above joint probabilities leads to the marginals, $p(x=0)=2/3$, $p(x=1)=1/3$, $p(y=0)=1/3$ and $p(y=1)=2/3$. This allows us to compute the single-variable entropies, (a) and (b),

$$H[x] = \sum_{x\in\{0,1\}} p(x)\ln p(x) = -\frac{2}{3}\ln\frac{2}{3} - \frac{1}{3}\ln\frac{1}{3} \approx 0.276, \qquad H[y] = -\frac{1}{3}\ln\frac{1}{3} - \frac{2}{3}\ln\frac{2}{3} \approx 0.276. \tag{31}$$

We can find the conditional entropies in two ways. The first requires computing the conditional probabilities from the above joint probabilities, using the product rule. The second relies on exploiting the relationship between conditional and joint entropy, $H[x,y] = H[y|x] + H[x]$. As we also need to compute the joint entropy, we save time and work by following the second approach. In particular, the joint entropy (e), $H[x,y]$, is given by,

$$H[x,y] = \sum_{x\in\{0,1\}}\sum_{y\in\{0,1\}} p(x,y)\ln p(x,y) = -\frac{1}{3}\ln\frac{1}{3} - \frac{1}{3}\ln\frac{1}{3} - 0\ln 0 - \frac{1}{3}\ln\frac{1}{3} = -\ln\frac{1}{3} \approx 0.477. \tag{32}$$

From this we can find the two conditional entropies, (c) and (d), by

$$H[y|x] = H[x,y] - H[x] \approx 0.477 - 0.276 = 0.201, \qquad H[x|y] = H[x,y] - H[y] \approx 0.477 - 0.276 = 0.201. \tag{33}$$

Finally, we find the mutual information, (f), by

$$I[x,y] = H[x] - H[x|y] = H[y] - H[y|x] \approx 0.076. \tag{34}$$

Note that the straight-forward difference using the above-computed values, $0.276 - 0.201 = 0.075$, results in a slightly different result, due to the accumulation of approximation errors. Drawing a diagram showing the relationship between the different quantities is left as an exercise to the reader.

Exercise 1.41

Using the sum and product rules of probability, show that the mutual information $I[x,y]$ satisfies the relation (1.121, [Bishop, 2006]).

Relationship (1.121, [Bishop, 2006]) states that

$$I[x,y] = H[x] - H[x|y] = H[y] - H[y|x]. \tag{35}$$

we will focus on the first quality, as the second can be shown in the same way by swapping x and y. This equality follows from a re-write of the definition of the mutual information,

$$\begin{aligned} I[x,y] &= -\iint p(x,y)\ln\frac{p(x)p(y)}{p(x,y)}\mathrm{d}x\mathrm{d}y \\ &= -\mathbb{E}_{p(x,y)}\left[\ln p(x)\right] - \mathbb{E}_{p(x,y)}\left[\ln p(y)\right] + \mathbb{E}_{p(x,y)}\left[\ln p(x|y) + \ln p(y)\right] \\ &= -\mathbb{E}_{p(x)}\left[\ln p(x)\right] - \mathbb{E}_{p(y)}\left[\ln p(y)\right] + \mathbb{E}_{p(x,y)}\left[\ln p(x|y)\right] + \mathbb{E}_{p(y)}\left[\ln p(y)\right] \\ &= H[x] - H[x|y], \end{aligned} \tag{36}$$

where the second equality results from using $p(x,y) = p(x|y)p(y)$ and expanding the logarithm, and the third results from marginalizing $p(x,y)$ in the expectation over either x or y.