Statistical Machine Learning 2018

Exercises and answers, week 2

14 September 2018

TUTORIAL

Exercise 1

By repeatedly applying the product rule, show that

$$p(X, Y, Z) = p(Z|Y, X)p(Y|X)p(X). \tag{1}$$

ANSWER: The famous **product rule of probability**:

$$p(X, Y) = p(X|Y)p(Y)$$

(or: joint = conditional times marginal), so

$$p(X, Y, Z) = p(Z|Y, X)p(Y, X) = p(Z|Y, X)p(Y|X)p(X)$$

Exercise 2

Assume $p(Y) > 0$. Two equivalent criteria for independence are:

$$p(X, Y) = p(X)p(Y) \tag{2}$$
$$p(X|Y) = p(X) \tag{3}$$

Show that (2) implies (3) and vice versa. (When does the assumption $p(Y) > 0$ come into play?)

ANSWER: Use the fact that, by definition, the product rule of probability always holds. From (2) and the product rule it follows that $p(X|Y)p(Y) = p(X)p(Y)$. If also $p(Y) > 0$ then both sides can be divided by $p(Y)$ to give (3). (Here the assumption is crucial, since $a \cdot 0 = b \cdot 0$ does not imply that $a = b$.)

If (3) holds, then both sides can be multiplied by $p(Y)$; the product rule then immediately implies (2). (This is true even if $p(Y) = 0$).

Exercise 3

Suppose we have a box containing 8 apples and 4 grapefruit, and another box that contains 15 apples and 3 grapefruit. One of the boxes is selected at random ('50-50'), and then a piece of fruit is picked from the chosen box, again with equal probability for each item in the box.

1. Calculate the probability of selecting an apple.

   ANSWER: The **sum rule of probability** is defined as

   $$p(X) = \sum_Y p(X, Y)$$

   In combination with the product rule, with F representing the type of fruit (apple or grapefruit) and B the selected box (1 or 2), this can be written in the form

   $$p(F) = \sum_B p(F|B)p(B)$$

   Applying this to the case at hand we find

   $$\begin{aligned} p(F = a) &= p(B = 1)p(F = a|B = 1) + p(B = 2)p(F = a|B = 2) \\ &= \frac{1}{2}(\frac{8}{12} + \frac{15}{18}) = 0.75 \end{aligned}$$

2. The piece of fruit turns out to be an apple indeed. Use Bayes' (or Bayes's) rule to calculate the probability that it came from the first box.

   ANSWER: The posterior probability that box 1 was chosen, i.e. after observing the apple, is

   $$p(B = 1|F = a) = \frac{p(B = 1)p(F = a|B = 1)}{p(F = a)} = \frac{\frac{1}{2}\frac{8}{12}}{\frac{3}{4}} = \frac{4}{9} = 0.444...$$

3. The apple is replaced, and from the *same* box another piece of fruit is selected at random. What is the probability that this second pick is also an apple? (Note: same box, but *not* necessarily the first.)

   ANSWER:
   We found that the posterior probability on box 1, given that the first pick was an apple, is $\frac{4}{9}$ and so for box 2 it must be the remaining $\frac{5}{9}$. The probability that the next piece of fruit is again an apple is therefore

   $$\begin{aligned} p(F_2 = a|F_1 = a) &= p(F_2 = a|B = 1)p(B = 1|F_1 = a) + p(F_2 = a|B = 2)p(B = 2|F_1 = a) \\ &= \frac{4}{9}\frac{8}{12} + \frac{5}{9}\frac{15}{18} = 0.7593 \end{aligned}$$

   where we use that $p(F_2|B) = p(F_1|B) = p(F|B)$, since both pieces were selected *at random* from the same box.

Exercise 4

Consider a discrete random variable x with distribution $p(x)$. The expectation of a function $f(x)$ is

$$\mathbb{E}[f] = \sum_x p(x)f(x) \tag{4}$$

Its variance $\text{var}[f]$ is

$$\text{var}[f] = \mathbb{E}[f^2] - (\mathbb{E}[f])^2 \tag{5}$$

- Show that if c is a constant,

$$\begin{aligned} \mathbb{E}[cf] &= c\mathbb{E}[f] & (6)\\ \text{var}[cf] &= c^2\text{var}[f] & (7) \end{aligned}$$

ANSWER:

$$\begin{aligned} \mathbb{E}[cf] &= \sum_x p(x)cf(x) \\ &= c\sum_x p(x)f(x) \\ &= c\mathbb{E}[f] \\ \text{var}[cf] &= \mathbb{E}[(cf)^2] - (\mathbb{E}[cf])^2 \\ &= \mathbb{E}[c^2f^2] - (c\mathbb{E}[f])^2 \\ &= c^2\mathbb{E}[f^2] - c^2(\mathbb{E}[f])^2 \\ &= c^2\text{var}[f] \end{aligned}$$

We now consider two discrete random variables x and z with a joint probability distribution $p(x, z)$. The expectation of a function $f(x, z)$ of x and z is given by

$$\mathbb{E}[f] = \sum_{x,z} p(x, z)f(x, z) \tag{8}$$

1. Show, using (8) that the expectation of the sum of x and z satisfies

$$\mathbb{E}[x + z] = \mathbb{E}[x] + \mathbb{E}[z] \tag{9}$$

   (Hints: make use of marginal distributions $p(z) = \sum_x p(x, z)$.)

2. Show that if x and z are statistical independent, i.e.,$p(x, z) = p(x)p(z)$, the expectation of their product satisfies

$$\mathbb{E}[xz] = \mathbb{E}[x]\mathbb{E}[z] \tag{10}$$

3. Use (5) and results (9) and (10) to show that the variance of the sum of two independent variables x and z satisfies

$$\text{var}[x + z] = \text{var}[x] + \text{var}[z] \tag{11}$$

   (Hint: use that square of any sum $a + b$ satisfies $(a + b)^2 = a^2 + 2ab + b^2$)

Note: the properties of expectations and variance that are shown in this exercise hold for continuous variables as well, this can be shown in a similar way (i.e. by replacing sums by integrals.)

ANSWER:

1.

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

2.

$$\begin{aligned}
\mathbb{E}[xz] &= \sum_{x,z} p(x,z)xz \\
&= \sum_{x,z} p(x)p(z)xz \\
&= \sum_{x} p(x)x \sum_{z} p(z)z \\
&= \mathbb{E}[x]\mathbb{E}[z]
\end{aligned}$$

3.

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

Exercise 5

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

$$p_y(y) = p_x(x)\left|\frac{dx}{dy}\right| = p_x\big(g(y)\big)\big|g'(y)\big|. \tag{12}$$

Assume that this nonlinear change of variables is monotonically increasing, i.e., $g'(y) > 0$ for all y. By differentiating relationship (12), 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 $\left|\frac{dx}{dy}\right|$.

We have now shown that the maximum of a probability density is in general dependent on the choice of variable, in contrast to changing the variable in a simple function. In the particular case of a linear transformation, however, the location of the maximum transforms the same way as the variable itself. Verify that $\hat{x} = g(\hat{y})$ for a linear transformation.

ANSWER: We are often interested in finding the most probable value for some quantity. In the case of probability distributions over discrete variables this poses little problem. However, for continuous variables there is a subtlety arising from the nature of probability densities and the way they transform under non-linear changes of variable.

Consider first the way a function $f(x)$ behaves when we change to a new variable y where the two variables are related by $x = g(y)$. This defines a new function of y given by

$$\tilde{f}(y) = f\big(g(y)\big). \tag{13}$$

Suppose $f(x)$ has a mode (i.e. a maximum) at $\hat{x}$ so that $f'(\hat{x}) = 0$. The corresponding mode of $\tilde{f}(y)$ will occur for a value $\hat{y}$ obtained by differentiating both sides of (13) with respect to y

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

As $g'(\hat{y}) \neq 0$, we have $f'\big(g(\hat{y})\big) = 0$. However, we know that $f'(\hat{x}) = 0$, and so we see that the locations of the mode expressed in terms of each of the variables x and y are related by $\hat{x} = g(\hat{y})$, as one would expect. Thus, finding a mode with respect to the variable x is completely equivalent to first transforming to the variable y, then finding a mode with respect to y, and then transforming back to x.

Now consider the behaviour of a probability density $p_x(x)$ under the change of variables $x = g(y)$, where the density with respect to the new variable is $p_y(y)$ and is given by (12). As $g' > 0$, (12) can be written

$$p_y(y) = p_x\big(g(y)\big)g'(y).$$

Differentiating both sides with respect to y then gives

$$p'_y(y) = p'_x\big(g(y)\big)\big(g'(y)\big)^2 + p_x\big(g(y)\big)g''(y). \tag{14}$$

Due to the presence of the second term on the right hand side of (14) the relationship $\hat{x} = g(\hat{y})$ no longer holds. Thus the value of x obtained by maximizing $p_x(x)$ will not be the value obtained by transforming to $p_y(y)$ then maximizing with respect to y and then transforming back to x. This causes modes of densities to be dependent on the choice of variables. In the case of linear transformation, the second term on the right hand side of (14) vanishes, and so the location of the maximum transforms according to $\hat{x} = g(\hat{y})$.

Exercise 6

Properties of the univariate Gaussian distribution. The probability density of a univariate Gaussian x with mean μ and variance σ^2 is given by:

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

1. Show, using the result on page 49 of the slides, which states

$$\int_{-\infty}^{\infty} \exp(-x^2)\, dx = \sqrt{\pi}$$

   that the univariate Gaussian density is properly normalized.

2. Calculate the expected value of x (Hint: use a change of variables).

3. Calculate the variance of x (Hint: differentiate both sides of the normalization condition for $p(x)$ with respect to σ^2).

4. Calculate the mode of x (i.e., the value of x that has maximum probability density).

ANSWER:

1. We perform a change of variables $z = \frac{x-\mu}{\sqrt{2}\sigma}$, noting that the integration region $(-\infty, \infty)$ transforms into itself:

$$\begin{aligned}\int_{-\infty}^{\infty} p(x)\,dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-z^2)\,\frac{dx}{dz} dz \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-z^2)\sqrt{2}\sigma\,dz \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp(-z^2)\,dz = 1.\end{aligned}$$

2. We use the same change of variables to calculate $\mathbb{E}(x)$:

$$\begin{aligned}\mathbb{E}(x) &= \int_{-\infty}^{\infty} xp(x)\,dx \\ &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\ &= \int_{-\infty}^{\infty} (z\sqrt{2}\sigma + \mu) \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-z^2)\,\sqrt{2}\sigma dz \\ &= \sqrt{2}\sigma \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} z \exp(-z^2) dz + \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp(-z^2) dz \\ &= \sqrt{2}\sigma \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} z \exp(-z^2) dz + \mu\end{aligned}$$

   The remaining integral vanishes, as the integrand is an odd function (a function $f(x)$ is odd if $f(-x) = -f(x)$) and the integration interval is invariant under sign reversal (to show this explicitly, write the integral as the sum of two integrals, one from $-\infty$ to 0 and the other from 0 to ∞ and then show that these two integrals cancel).

3. The normalization condition reads:

$$1 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$

   Differentiating this equation with respect to σ^2 (note that we are allowed interchange the integration over x and the differentiation with respect to σ^2 as the integration domain does not depend on σ^2) yields:

$$0 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \left[-\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4}\right] dx = \mathbb{E}\left(-\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4}\right)$$

   Rewriting this (using linearity of expectation):

$$0 = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\mathbb{E}\left((x-\mu)^2\right) = \frac{1}{2\sigma^2}\left(-1 + \frac{\mathbb{V}\mathrm{ar}(x)}{\sigma^2}\right)$$

   and hence

$$\mathbb{V}\mathrm{ar}(x) = \sigma^2.$$

4. At the mode (maximum) $\hat{x}$ of $p(x)$, $p'(\hat{x}) = 0$. So

$$p'(\hat{x}) = -\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\hat{x}-\mu)^2}{2\sigma^2}\right) \frac{(\hat{x}-\mu)}{\sigma^2} = 0$$

   which only happens for $\hat{x} = \mu$. As we can check that $p''(x) < 0$, this is indeed a maximum.

BONUS PRACTICE

Exercise 7

(Exercise 1.5 from Bishop). The variance of f is defined as

$$\mathrm{var}[f] = \big\langle (f(x) - \big\langle f(x) \big\rangle)^2 \big\rangle \tag{15}$$

in which $\big\langle f(x) \big\rangle \equiv \mathbb{E}[f]$ is the expectation of a function $f(x)$ under probability distribution $p(x)$, defined as $\mathbb{E}[f] = \int f(x)p(x)\,dx$. Now show that the variance can also be written as

$$\mathrm{var}[f] = \big\langle f(x)^2 \big\rangle - \big\langle f(x) \big\rangle^2 \tag{16}$$

ANSWER: The expectation is a linear function of its operand: $\mathbb{E}[a\mathrm{X} + b] = a\mathbb{E}[\mathrm{X}] + b$, so

$$\begin{aligned}
\big\langle (f - \langle f \rangle)^2 \big\rangle &= \big\langle f^2 - 2f\langle f \rangle + \langle f \rangle^2 \big\rangle \\
&= \big\langle f^2 \big\rangle - \big\langle 2f\langle f \rangle \big\rangle + \big\langle f \big\rangle^2 \\
&= \big\langle f^2 \big\rangle - 2\big\langle f \big\rangle\big\langle f \big\rangle + \big\langle f \big\rangle^2 \\
&= \big\langle f^2 \big\rangle - \big\langle f \big\rangle^2
\end{aligned}$$

Exercise 8

Probability densities $p(x)$ should be non-negative $p(x) \geq 0$, and normalized $\int p(x)dx = 1$.

1. Consider the probability density $p(t)$ of the random variable T, defined as

$$p(t) = \begin{cases} \dfrac{1}{Z}\exp(-\lambda t) & , \quad t \geq 0 \\ 0 & , \quad t < 0 \end{cases} \tag{17}$$

   with λ a positive constant. Compute Z using the fact that p should be normalized.

   ANSWER:

$$1 = \int_{-\infty}^{\infty} p(t)dt = \int_0^{\infty} p(t)dt = \frac{1}{Z}\int_0^{\infty} \exp(-\lambda t)dt = \frac{-1}{Z\lambda}\exp(-\lambda t)\Big|_{t=0}^{\infty} = \frac{1}{\lambda Z} \implies Z = \frac{1}{\lambda}.$$

2. For the previous probability density defined in Equation (17), show how $\Pr(T > 1)$ depends on λ. Use the normalizing constant Z found in the previous part. What is the relationship between the quantity you have just computed and the cumulative distribution function $F(u) = \Pr(T \leq u)$?

   ANSWER:

$$\Pr(T > 1) = \int_1^{\infty} p(t)dt = \lambda\int_1^{\infty} \exp(-\lambda t)dt = -\exp(-\lambda t)\Big|_{t=1}^{\infty} = \exp(-\lambda). \quad \square$$

   Using the property that probabilities add up to one, we get:

$$\Pr(T > 1) = 1 - \Pr(T \leq 1) = 1 - F(1).$$

3. Let $\rho(x)$ be a normalized probability density, i.e. $\rho(x) >= 0$ and $\int_{-\infty}^{\infty} \rho(x)dx = 1$. Show that for any pair of constants μ and $\alpha > 0$, the function

$$\hat{\rho}(x) = \alpha\, \rho(\alpha(x - \mu)) \tag{18}$$

is also a normalized density.

ANSWER: You have to show that $\int_{-\infty}^{\infty} \hat{\rho}(x)dx = 1$. Change variables: $x \to y$, with $y = \alpha(x - \mu))$. That means $x = \frac{1}{\alpha}y + \mu$ and so $dx = \frac{1}{\alpha}dy$.

$$\begin{aligned}
\int_{-\infty}^{\infty} \hat{\rho}(x)dx &= \alpha \int_{-\infty}^{\infty} \rho(\alpha(x-\mu))dx \\
&= \alpha \int_{-\infty}^{\infty} \rho(y)\frac{1}{\alpha}dy \\
&= \int_{-\infty}^{\infty} \rho(y)dy = 1
\end{aligned}$$

4. Compute the normalizing constant Z of the following probability density in R^d with parameters $\lambda_i > 0$,

$$p(x_1, \ldots, x_d) = \frac{1}{Z} \exp\left\{ -\sum_{i=1}^{d} \frac{\lambda_i}{2} x_i^2 \right\} . \tag{19}$$

You may use that for $\lambda > 0$,

$$\int_{-\infty}^{\infty} \exp\left\{ -\frac{\lambda}{2} x^2 \right\} dx = \left(\frac{2\pi}{\lambda} \right)^{1/2}$$

ANSWER: Separable integrals over each dimension, so

$$\begin{aligned}
Z &= \int \ldots \int \exp\left\{ -\sum_{i=1}^{d} \frac{\lambda_i}{2} x_i^2 \right\} dx_1, \ldots dx_d = \int \ldots \int \prod_{i=1}^{d} \exp\left\{ -\frac{\lambda_i}{2} x_i^2 \right\} dx_1, \ldots dx_d \\
\Bigg(&= \int \exp\left\{ -\frac{\lambda_1}{2} x_1^2 \right\} dx_1 \ldots \int \exp\left\{ -\frac{\lambda_d}{2} x_d^2 \right\} dx_d \Bigg) \\
&= \prod_{i=1}^{d} \int \exp\left\{ -\frac{\lambda_i}{2} x_i^2 \right\} dx_i = \prod_{i=1}^{d} \left(\frac{2\pi}{\lambda_i} \right)^{1/2}
\end{aligned}$$

Exercise 9

Show that, for two (continuous) random variables X and Y, the following identities hold:

1. **The law of total expectation**: $\mathbb{E}_X[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]]$, where $\mathbb{E}_X$ and $\mathbb{E}_Y$ are the expectation values w.r.t. X and Y, respectively.

ANSWER:

$$\begin{aligned}
\mathbb{E}_X[X] &= \int xp(x)dx \\
&= \iint xp(x,y)dxdy \\
&= \iint xp(x|y)p(y)dxdy \\
&= \int \left[\int xp(x|y)dx\right] p(y)dy \\
&= \int \mathbb{E}_X[X|y]p(y)dy \\
&= \mathbb{E}_Y[\mathbb{E}_X[X|Y]]. \quad \square
\end{aligned}$$

2. **The law of total variance**: $\text{var}_X[X] = \mathbb{E}_Y[\text{var}_X[X|Y]] + \text{var}_Y[\mathbb{E}_X[X|Y]]$, where var_X and var_Y are the variances w.r.t. X and Y, respectively.

   ANSWER:

$$\begin{aligned}
\text{var}_X[X] &= \mathbb{E}_X[X^2] - (\mathbb{E}_X[X])^2 \\
&= \mathbb{E}_Y[\mathbb{E}_X[X^2|Y]] - (\mathbb{E}_Y[\mathbb{E}_X[X|Y]])^2 \\
&= \mathbb{E}_Y[\mathbb{E}_X[X^2|Y]] - \mathbb{E}_Y[(\mathbb{E}_X[X|Y])^2] + \mathbb{E}_Y[(\mathbb{E}_X[X|Y])^2] - (\mathbb{E}_Y[\mathbb{E}_X[X|Y]])^2 \\
&= \mathbb{E}_Y[\mathbb{E}_X[X^2|Y] - (\mathbb{E}_X[X|Y])^2] + \text{var}_Y[\mathbb{E}_X[X|Y]] \\
&= \mathbb{E}_Y[\text{var}_X[X|Y]] + \text{var}_Y[\mathbb{E}_X[X|Y]]. \quad \square
\end{aligned}$$

Interpretation in Bayesian inference: The process of Bayesian inference involves passing from a prior distribution over the parameter, $p(\Theta)$, to a posterior distribution $p(\Theta|\mathcal{D})$, which is dependent on the data $\mathcal{D}$. If, in the above laws, we replace X with Θ and Y with $\mathcal{D}$, we can interpret the relationship between the prior and posterior:

1. $\mathbb{E}_\Theta[\Theta] = \mathbb{E}_\mathcal{D}[\mathbb{E}_\Theta[\Theta|\mathcal{D}]]$: The prior mean of Θ is the average of all possible posterior means over the distribution of possible data.

2. $\text{var}_\Theta[\Theta] = \mathbb{E}_\mathcal{D}[\text{var}_\Theta[\Theta|\mathcal{D}]] + \text{var}_\mathcal{D}[\mathbb{E}_\Theta[\Theta|\mathcal{D}]]$: Because the posterior distribution incorporates the information from the data, the posterior variance is on average smaller than the prior variance. The amount by which they differ depends on the variation in posterior means over the distribution of possible data. The greater the latter variation, the more the potential for reducing our uncertainty with regard to Θ.