# Statistical Machine Learning 2018

Exercises, week 2

14 September 2018

## TUTORIAL

## Exercise 1

By repeatedly applying the product rule, show that

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

## 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?)

## 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.
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.
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.)

## 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 $\mathrm{var}[f]$ is

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

- Show that if $c$ is a constant,

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

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

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

   (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.)

# 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

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

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.

# Exercise 6

Properties of the univariate Gaussian distribution. The probability density of a univariate Gaussian $x$ with mean $\mu$ and variance $\sigma^2$ is given by:

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

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 $\sigma^2$).

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

# BONUS PRACTICE

## Exercise 7

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

$$\mathrm{var}[f] = \left\langle (f(x) - \langle f(x) \rangle)^2 \right\rangle \tag{13}$$

in which $\langle f(x) \rangle \equiv \mathbb{E}[f]$ 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

$$\mathrm{var}[f] = \left\langle f(x)^2 \right\rangle - \left\langle f(x) \right\rangle^2 \tag{14}$$

## 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

$$p(t) = \begin{cases} \dfrac{1}{Z}\exp(-\lambda t) & , \quad t \geq 0 \\ 0 & , \quad t < 0 \end{cases} \tag{15}$$

   with $\lambda$ a positive constant. Compute $Z$ using the fact that $p$ should be normalized.

2. For the previous probability density defined in Equation (15), show how $\Pr(T > 1)$ depends on $\lambda$. 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 $F(u) = \Pr(T \leq u)$?

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 $\mu$ and $\alpha > 0$, the function

$$\hat{\rho}(x) = \alpha\,\rho(\alpha(x - \mu)) \tag{16}$$

   is also a normalized density.

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

$$p(x_1, \ldots, x_d) = \frac{1}{Z}\exp\left\{ -\sum_{i=1}^{d} \frac{\lambda_i}{2} x_i^2 \right\} . \tag{17}$$

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

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

## 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.

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

*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 $\Theta$ 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 $\Theta$ is the average of all possible posterior means over the distribution of possible data.

2. $\mathrm{var}_\Theta[\Theta] = \mathbb{E}_\mathcal{D}[\mathrm{var}_\Theta[\Theta|\mathcal{D}]] + \mathrm{var}_\mathcal{D}[\mathbb{E}_\Theta[\Theta|\mathcal{D}]]$: 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 $\Theta$.