Statistical Machine Learning 2016

Exercises and answers, week 3

15 September 2016

Exercise 1

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

1. Consider the probability density $p(t)$ defined as

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

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

ANSWER:

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

So $Z = \frac{1}{\lambda}$.

2. Let $\rho(x)$ be a normalised 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

$$\hat{\rho}(x) = \alpha\, \rho(\alpha(x - \mu)) \tag{2}$$

is also a normalised 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 $dx = \frac{1}{\alpha}dy$.

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

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

$$p(x_1, \ldots, x_d) = \frac{1}{Z}\exp\left\{-\sum_{i=1}^{d}\frac{\lambda_i}{2}x_i^2\right\}. \tag{3}$$

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

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

Exercise 2

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

$$\mathrm{var}[f] = \left\langle (f(x) - \left\langle f(x) \right\rangle)^2 \right\rangle \tag{4}$$

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

$$\mathrm{var}[f] = \left\langle f(x)^2 \right\rangle - \left\langle f(x) \right\rangle^2 \tag{5}$$

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

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

Exercise 3

More about expectation values and variances.

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

$$\mathbb{E}[f] = \sum_x p(x)f(x) \tag{6}$$

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

$$\mathrm{var}[f] = \mathbb{E}[f^2] - (\mathbb{E}[f])^2 \tag{7}$$

- Show that if c is a constant,

$$\mathbb{E}[cf] = c\mathbb{E}[f] \tag{8}$$
$$\text{var}[cf] = c^2\text{var}[f] \tag{9}$$

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

$$\mathbb{E}[f] = \sum_{x,z} p(x, z)f(x, z) \tag{10}$$

1. Show, using (10) that the expectation of the sum of x and z satisfies

$$\mathbb{E}[x + z] = \mathbb{E}[x] + \mathbb{E}[z] \tag{11}$$

   (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

$$\mathbb{E}[xz] = \mathbb{E}[x]\mathbb{E}[z] \tag{12}$$

3. Use (7) and results (11) and (12) to show that the variance of the sum of two independent variables x and z satisfies

$$\text{var}[x + z] = \text{var}[x] + \text{var}[z] \tag{13}$$

   (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 4

We consider the Gaussian distribution in one dimension (see Bishop, p. 27-28)

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{14}$$

with parameters μ and $\sigma^2 > 0$. Now suppose we have a data set of observations χ

$$\chi = \{x_1, \ldots, x_N\}$$

The observations are drawn independently from a Gaussian distribution whose mean μ and variance σ^2 are unknown. The probability of the data set χ, given these unknown parameters is

$$p(\chi|\mu,\sigma^2) = \prod_{i=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2)$$

1. Show that the log likelihood function can be written in the form

$$\ln p(\chi|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln(\sigma^2) - \frac{N}{2}\ln(2\pi) \tag{15}$$

ANSWER: Make use of the fact that $\ln \prod_i \ldots = \sum_i \ln \ldots$, and in particular $\ln(ab) = \ln(a) + \ln(b)$, furthermore, $\ln(1/a) = -\ln(a)$, and $\ln(\sqrt{a}) = 1/2 \ln(a)$. In addition $\ln \exp(a) = a$, and finally $\sum_{i=1}^{N} c = Nc$ if c does not depend on i. Then it should more or less easily follow:

$$\begin{aligned}
\ln p(\chi|\mu, \sigma^2) &= \ln \prod_{i=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) \\
&= \sum_{i=1}^{N} \ln \mathcal{N}(x_n|\mu, \sigma^2) \\
&= \sum_{i=1}^{N} \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} \right) \\
&= N \ln(\frac{1}{\sqrt{2\pi\sigma^2}}) - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \\
&= -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \ln(\sigma^2) - \frac{N}{2} \ln(2\pi)
\end{aligned}$$

2. By maximizing (15) with respect to μ (i.e., take the partial derivative with respect to μ and set to zero), we obtain the maximum likelihood solution μ_{ML}. Verify that it is given by

$$\mu_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n \equiv \bar{x} \tag{16}$$

3. In the previous item, you may have noticed that the maximum likelihood solution μ_{ML} does not depend on σ^2. We can now substitute the solution $\mu = \mu_{\mathrm{ML}} = \bar{x}$ in (15) and maximize the result with respect to σ^2_{ML} (i.e., take the partial derivative with respect to σ^2 and set to zero), we then obtain the maximum likelihood solution σ^2_{ML}. Verify that it is given by

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})^2 \tag{17}$$

ANSWER: Taking the derivative of (15) with respect to μ gives $\frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu)$. Set equal to zero and rearranging terms, this gives the solution

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

Note that this solution does not depend on σ^2.

Maximizing with respect to σ^2: substitute $\mu(= \mu_{\mathrm{ML}}) = \bar{x}$ and $v = \sigma^2$ in (15). Take the derivative w.r.t. v and set equal to zero.

$$\frac{1}{2v^2} \sum_{n=1}^{N} (x_n - \bar{x})^2 - \frac{N}{2v} = 0$$

So

$$\frac{N}{2v^2} \left(\frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})^2 - v \right) = 0$$

This means that either $v = \sigma^2 = \infty$ (but that is actually a minimum since $\ln(\sigma^2) \to \infty$), or $v = \sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2$ which is indeed a maximum and therefore the ML solution.

Exercise 5

In this exercise, we will have a closer look at the gradient descent algorithm for function minimization. When the function to be minimized is $E(\mathbf{x})$, the gradient descent iteration is

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \eta \nabla E(\mathbf{x}_n) \tag{18}$$

where $\eta > 0$ is the so-called learning-rate.

1. Consider the function $E(x) = \frac{\lambda}{2}(x-a)^2$ with parameters $\lambda > 0$, and a arbitrary.

   (a) Write down the gradient descent iteration rule. Verify that the minimum of E is a and that a is a fixed point[1] of the gradient descent iteration rule.

   ANSWER:

   $$x_{n+1} = x_n - \eta\lambda(x_n - a) = (1 - \eta\lambda)x_n + \eta\lambda a$$

   The minimum of E is $x^* = a$ (for this value E is zero, for any other value, it is larger). Fixed point: fill in $a = (1 - \eta\lambda)a + \eta\lambda a = a$.

   (b) Show that the algorithm converges in one step if $\eta = 1/\lambda$.

   ANSWER: With $\eta = 1/\lambda$,

   $$x_{n+1} = x_n - (x_n - a) = a$$

   So $x_1 = a$ for any x_0.

   (c) Define $d_n = x_n - a$. Show that if $0 < \eta < 1/\lambda$, subsequent d_n's have the same signs. Also show that if $\eta > 1/\lambda$, subsequent d_n's have opposite signs.

   ANSWER: In terms of d_n the iteration rule is

   $$d_{n+1} = d_n - \eta\lambda d_n = (1 - \eta\lambda)d_n$$

   If $0 < \eta < 1/\lambda$ then $(1 - \eta\lambda) > 0$ and if $\eta > 1/\lambda$ then $(1 - \eta\lambda) < 0$

   (d) The distance to the fixed point is $|d_n|$. Show that $|d_{n+1}| = |(1 - \eta\lambda)||d_n|$. Show that this implies that the algorithm converges to the fixed point if $0 < \eta < 2/\lambda$, and that it diverges if $\eta > 2/\lambda$.

   ANSWER: In terms of d_n the iteration rule is

   $$d_{n+1} = d_n - \eta\lambda d_n = (1 - \eta\lambda)d_n$$

   so

   $$|d_n| = |(1 - \eta\lambda)|^n |d_0|$$

   If $0 < \eta < 2/\lambda$, then $|(1 - \eta\lambda)| < 1$ and $|(1 - \eta\lambda)|^n \to 0$. If $\eta > 2/\lambda$ then $|(1 - \eta\lambda)| > 1$ and $|(1 - \eta\lambda)|^n \to \infty$

2. Consider now the function $E(x, y) = \frac{\lambda_1}{2}(x-a_1)^2 + \frac{\lambda_2}{2}(y-a_2)^2$ with parameters $0 < \lambda_1 < \lambda_2$, and a_i arbitrary.

[1] A fixed point x^* of an iteration $x_{n+1} = F(x_n)$ satisfies $x^* = F(x^*)$.

(a) Write down the gradient descent iteration rule. Verify that the minimum of E is a fixed point.

ANSWER:

$$x_{n+1} = (1-\eta\lambda_1)x_n + \eta\lambda_1 a_1 \quad (19)$$
$$y_{n+1} = (1-\eta\lambda_2)y_n + \eta\lambda_2 a_2 \quad (20)$$

The minimum of E is (a_1, a_2). Two equations are decoupled. Same as previous.

(b) We want to find the learning rate η that leads to the fasted convergence in both x and y direction. This optimal learning rate is the one for which both $|1-\eta\lambda_1|$ and $|1-\eta\lambda_2|$ are as small as possible. For the optimal learning rate, the equation $|1-\eta\lambda_1| = |1-\eta\lambda_2|$ must therefore hold. Since $\lambda_1 < \lambda_2$, this can only hold if $\eta\lambda_1 < 1$ and $\eta\lambda_2 > 1$.

- Show that solving the equation leads to $\eta^* = 2/(\lambda_2 + \lambda_1)$ (which is the optimal learning rate). What happens if η is smaller than the optimal value? What happens if it is larger?

ANSWER: The solution is to set $|1-\eta\lambda_1| = |1-\eta\lambda_2|$, where $\eta\lambda_1 < 1$ and $\eta\lambda_2 > 1$. So $1-\eta\lambda_1 = \eta\lambda_2 - 1$. So $\eta = 2/(\lambda_2+\lambda_1)$. When η smaller: slows down in the flat direction. η larger: more overshoot in the steep direction, causing slowing down.

(c) What is the value of $|1-\eta^*\lambda_i|$ in both directions? What does this say about the applicability of gradient descent to functions with steep hills and flat valleys (i.e., if $\lambda_2 \gg \lambda_1$)?

ANSWER:

$$\left|1 - 2\frac{\lambda_i}{\lambda_2+\lambda_1}\right| = \left|\frac{\lambda_2-\lambda_1}{\lambda_2+\lambda_1}\right|$$

If $\lambda_2 \gg \lambda_1$, then this value is approximately $1-2\lambda_1/\lambda_2$, which is only a little bit smaller than 1, i.e. gradient descent will converge only very slowly.