Statistical Machine Learning 2016

Exercises, 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.

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.

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 $\lambda > 0$,

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

Exercise 2

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

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

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

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

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,

$$\begin{aligned} \mathbb{E}[cf] &= c\mathbb{E}[f] && (8) \\ \mathrm{var}[cf] &= c^2 \mathrm{var}[f] && (9) \end{aligned}$$

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

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

   (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.)

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

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

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

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

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.
   (b) Show that the algorithm converges in one step if $\eta = 1/\lambda$.
   (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.
   (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 $\eta > 2/\lambda$.
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.
   (a) Write down the gradient descent iteration rule. Verify that the minimum of E is a fixed point.
   (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?
   (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 $\lambda_2 \gg \lambda_1$)?

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