Statistical Machine Learning 2016

Exercises and answers, week 7

13 October 2016

Exercise 1

We consider two distributions, with one defined conditional on the other, as

$$p(u) = \mathcal{N}(u|\mu_0, \sigma^2) \quad (1)$$
$$p(v|u) = \mathcal{N}(v|c \cdot u, s^2) \quad (2)$$

where μ_0, σ^2, c and s^2 are constant model parameters.

1. The conditional distribution $p(u|v)$ is also a Gaussian. Which equations from Bishop are relevant for computing this function?

   ANSWER: Given are two Gaussians $p(u)$ and $p(v|u)$. In order to calculate $p(u|v)$ we can use Bayes', which in terms of the equations 2.113-2.117 boils down to the expression for 2.116 given 2.113 and 2.114. For that we also need 2.117.

2. Write down an expression for the distribution $p(u|v)$ and show that the mean $\mu_{u|v}$ and variance $\sigma^2_{u|v}$ of this distribution are given by

$$\mu_{u|v} = \frac{\frac{\mu_0}{\sigma^2} + \frac{cv}{s^2}}{\frac{1}{\sigma^2} + \frac{c^2}{s^2}} \quad (3)$$

$$\frac{1}{\sigma^2_{u|v}} = \frac{1}{\sigma^2} + \frac{c^2}{s^2} \quad (4)$$

   ANSWER: Match the parameters in (1) to eq.2.113 and (2) to eq.2.114

$$\begin{aligned} (\text{standard form} &\leftrightarrow \text{exercise}) \\ \mathbf{x} &\leftrightarrow u \\ \boldsymbol{\mu} &\leftrightarrow \mu_0 \\ \boldsymbol{\Lambda}^{-1} &\leftrightarrow \sigma^2 \\ \mathbf{y} &\leftrightarrow v \\ \mathbf{A} &\leftrightarrow c \\ \mathbf{b} &\leftrightarrow 0 \\ \mathbf{L}^{-1} &\leftrightarrow s^2 \end{aligned}$$

This gives for eq.2.117

$$\begin{aligned}\boldsymbol{\Sigma} &= (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1} \\ &= (\frac{1}{\sigma^2} + c \cdot \frac{1}{s^2} \cdot c)^{-1} \\ &= (\frac{1}{\sigma^2} + \frac{c^2}{s^2})^{-1}\end{aligned}$$

which corresponds to $\sigma^2_{u|v}$, and so for the mean in eq.2.116 we find

$$\begin{aligned}\mu_{u|v} &= \boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\} \\ &= \sigma^2_{u|v}\{c\frac{1}{s^2}v + \frac{1}{\sigma^2}\mu_0\} \\ &= \sigma^2_{u|v}\{\frac{\mu_0}{\sigma^2} + \frac{cv}{s^2}\}\end{aligned}$$

Note that all parameters and variables are now just scalars.

3. Compute $p(v)$.

   ANSWER: The distribution will be a Gaussian of the form

   $$p(v) = \mathcal{N}(v|\mu_v, \sigma_v^2)$$

   We already matched the parameters in eq.2.113 and 2.114, and so filling in for eq.2.115 gives

   $$\begin{aligned}\mu_v &= \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \\ &= c \cdot \mu_0\end{aligned}$$

   and

   $$\begin{aligned}\sigma_v^2 &= \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \\ &= s^2 + c \cdot \sigma^2 \cdot c \\ &= s^2 + c^2\sigma^2\end{aligned}$$

   So the mean of v is exactly c times the mean of u, which was to be expected on comparison with (2), and the variance is a combination of the variance in v and u, which also seems reasonable.

4. Compute $p(u, v)$. Hint: using the right equations, the calculation does not get very messy.

   ANSWER: We start by using equation (2.98) from Bishop. As the product of two Gaussian distributions is again a Gaussian distribution, we can write the joint distribution in the form

   $$p(u, v) = \mathcal{N}\left(\begin{pmatrix} u \\ v \end{pmatrix} \middle| \begin{pmatrix} \mu_u \\ \mu_v \end{pmatrix}, \begin{pmatrix} \Sigma_{uu} & \Sigma_{uv} \\ \Sigma_{vu} & \Sigma_{vv} \end{pmatrix}\right).$$

   Our task is to express these new parameters in terms of the original ones.

   (a) From (2.98) we obtain that the marginal distribution $p(u) = \mathcal{N}(u|\mu_u, \Sigma_{uu})$; comparing with (1) we conclude that $\mu_u = \mu_0$ and $\Sigma_{uu} = \sigma^2$.

   (b) Again using (2.98), we obtain that the marginal distribution $p(v) = \mathcal{N}(v|\mu_v, \Sigma_{vv})$; comparing with the answer of the previous question we conclude that $\mu_v = c\mu_0$, $\Sigma_{vv} = s^2 + c^2\sigma^2$.

(c) Finally, note that because of the symmetry of the covariance matrix, $\Sigma_{uv} = \Sigma_{vu}^T = \Sigma_{vu}$ (here, the transpose is redundant as we are working with 1×1 matrices). Now the trick is to use some knowledge about one of the conditional distributions to find the remaining unknown parameter Σ_{uv}. There are several possibilities, but by far the easiest is to use equation (2.82), or in our notation:

$$\Sigma_{v|u} = \Sigma_{vv} - \Sigma_{vu}\Sigma_{uu}^{-1}\Sigma_{uv} = \Sigma_{vv} - \Sigma_{uv}^2\Sigma_{uu}^{-1},$$

where we used that all matrices are 1×1. Using $\Sigma_{v|u} = s^2$ (see (2) in the exercise) and substituting the values we already found for Σ_{uu} and Σ_{vv}, we can solve for Σ_{uv} and obtain $\Sigma_{uv} = c\sigma^2$.

Combining everything, we get:

$$p(u, v) = \mathcal{N}\left(\begin{pmatrix} u \\ v \end{pmatrix} \middle| \begin{pmatrix} \mu_0 \\ c\mu_0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & c\sigma^2 \\ c\sigma^2 & s^2 + c^2\sigma^2 \end{pmatrix}\right)$$

Exercise 2

Assume that N points x_n are independently generated according to a Gaussian $\mathcal{N}(x|\mu, \sigma^2)$.

1. Show that the vector of points $\mathbf{x} = (x_1, \ldots, x_N)^T$ is Gaussian distributed

$$p(\mathbf{x}|\mu, \sigma^2) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \sigma^2\mathbf{I}) \tag{5}$$

with $\boldsymbol{\mu} = (\mu, \ldots, \mu)^T$ and $\mathbf{I}$ the $N \times N$ identity matrix

ANSWER:

$$\begin{aligned} p(\mathbf{x}|\mu, \sigma^2) &= \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) \\ &= \frac{1}{Z}\exp(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2) \\ &= \frac{1}{Z}\exp(-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(x_n - \mu)(\delta_{nm}/\sigma^2)(x_m - \mu))) \\ &= \frac{1}{Z}\exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T(\sigma^2\mathbf{I})^{-1}(\mathbf{x} - \boldsymbol{\mu}))) \\ &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \sigma^2\mathbf{I}) \end{aligned}$$

Suppose σ is given and μ is unknown. Take as prior for μ the Gaussian distribution $\mathcal{N}(\mu|\mu_0, \sigma_0^2)$. To compute the posterior, Bayes' theorem for Gaussian variables will be used: see page 93, (2.113 to 2.117):

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) && (6) \\ p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) && (7) \\ &\Rightarrow \\ p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) && (8) \\ p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) && (9) \end{aligned}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1} \tag{10}$$

2. Use Bayes' theorem for Gaussian variables to show that the posterior is

$$p(\mu|\mathbf{x}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$$

where

$$\mu_N = \sigma_N^2 \left(\frac{N\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) \tag{11}$$

$$\sigma_N^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}} \tag{12}$$

where $\bar{x} \equiv \frac{1}{N} \sum_{n=1}^{N} x_n$

ANSWER: Using Bayes' theorem we have

$$p(\mu|\mathbf{x}) = \frac{p(\mathbf{x}|\mu)\, p(\mu)}{p(\mathbf{x})} \propto p(\mathbf{x}|\mu)\, p(\mu)$$

These are all marginal and conditional Gaussian distributions in the form

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})\, p(\mathbf{x})$$

Matching the given parameters in the exercise with the equations above results in

(variable in exercise)	$\leftrightarrow$	(variable in eqs. (6), (7) and (9))
μ	$\leftrightarrow$	$\mathbf{x}$
μ_0	$\leftrightarrow$	$\boldsymbol{\mu}$
σ_0^2	$\leftrightarrow$	$\boldsymbol{\Lambda}^{-1}$
$\mathbf{x}$	$\leftrightarrow$	$\mathbf{y}$
$(1, 1, \ldots, 1)^{\mathrm{T}}$	$\leftrightarrow$	$\mathbf{A}$
0	$\leftrightarrow$	$\mathbf{b}$
$\sigma^2\mathbf{I}$	$\leftrightarrow$	$\mathbf{L}^{-1}$

Note that $\mathbf{A}^T\mathbf{L}\mathbf{A} = (1, 1, \ldots, 1)\sigma^{-2}\mathbf{I}(1, 1, \ldots, 1)^{\mathrm{T}} = N/\sigma^2$, with N the length of vector A, and $\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{y} \equiv (1, 1, \ldots, 1)\sigma^{-2}\mathbf{I}\mathbf{x} = \sigma^{-2}(1, 1, \ldots, 1)\mathbf{x} = \sigma^{-2}(x_1 + \ldots + x_n) = N\bar{x}/\sigma^2$. Substituting in (10) gives $\boldsymbol{\Sigma} = 1/(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2})$ (so, in this specific case, $\boldsymbol{\Sigma}$ is a scalar) , which, substituted in (9), gives the result to show.

Exercise 3

A probability distribution is part of the exponential family if it can be cast in the form

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\left\{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \right\} \tag{13}$$

where the $\boldsymbol{\eta}$ are called the *natural parameters* of the distribution.

Consider the Gamma distribution over $\lambda \geq 0$ with parameters a and b, defined as

$$\mathrm{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \tag{14}$$

1. Show the Gamma distribution belongs to the exponential family by casting it in the standard representation (13). Hint: put all λ dependence in the exponential, i.e., start by rewriting $\mathrm{Gam}(\lambda|a, b) = \ldots \exp(\ldots \lambda \ldots)$.

ANSWER: Use the suggestion to write the distribution as

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \exp\{(a-1)\ln\lambda - b\lambda\}$$

Then the distribution is in standard form with $\mathbf{x} = \lambda$ and

$$\begin{aligned} h(\lambda) &= 1 \\ g(a,b) &= \frac{b^a}{\Gamma(a)} \\ \mathbf{u}(\lambda) &= \begin{pmatrix} \lambda \\ \ln\lambda \end{pmatrix} \\ \boldsymbol{\eta}(a,b) &= \begin{pmatrix} -b \\ a-1 \end{pmatrix} \end{aligned}$$

The function $g(\boldsymbol{\eta})$ ensures the distribution is normalized: $g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\} d\mathbf{x} = 1$. Taking the gradient w.r.t. $\boldsymbol{\eta}$, it is easy to show that

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})] \tag{15}$$

2. Using this result, show that the expectation value for the Gamma distribution (14) is given by $\mathbb{E}[\lambda] = \frac{a}{b}$.

   ANSWER: From the previous answer we know we are looking for

   $$\mathbb{E}[\lambda] = -\frac{\partial}{\partial \eta_1} \ln g(\boldsymbol{\eta})$$

   Since

   $$\boldsymbol{\eta} = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} -b \\ a-1 \end{pmatrix}$$

   it follows that $a = \eta_2 + 1$ and $b = -\eta_1$, and therefore

   $$g(\eta_1, \eta_2) = \frac{(-\eta_1)^{\eta_2+1}}{\Gamma(\eta_2+1)}$$

   Taking the partial derivative w.r.t. η_1 gives

   $$\begin{aligned} -\frac{\partial}{\partial \eta_1} \ln g(\eta_1, \eta_2) &= -\frac{1}{g(\eta_1,\eta_2)} \frac{\partial g(\eta_1,\eta_2)}{\partial \eta_1} \\ &= -\frac{\Gamma(\eta_2+1)}{(-\eta_1)^{\eta_2+1}} \cdot \frac{-(\eta_2+1)(-\eta_1)^{\eta_2}}{\Gamma(\eta_2+1)} = \frac{\eta_2+1}{-\eta_1} = \frac{a}{b} \end{aligned}$$

Exercise 4

Kernel density and K-nearest neighbour are two non-parametric methods to estimate an unknown probability density $p(\mathbf{x})$ in some D-dimensional space from a given set of N observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ drawn from that distribution. In essence, kernel density takes a fixed size volume and counts the number of points contained therein, whereas nearest neighbour estimates the size of the volume required to encompass the K nearest points. In the limit $N \to \infty$, both methods converge to the true probability density, provided that the kernel-volume resp. the number of neighbours scale suitably with N. However, only one of the two is a true density model whereas the other is not ...

1. Let $k(\mathbf{x})$ be a normalized probability distribution on $\mathbb{R}^d$. (So $\mathbf{x} = (x^1, \ldots, x^d)^T$).
   Show that
   $$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h^D} k\Big(\frac{\mathbf{x} - \boldsymbol{x}_n}{h}\Big)$$
   is a normalized distribution. Hint: compute $\int p(\mathbf{x})d\mathbf{x}$, and substitute $\mathbf{u} = (\mathbf{x} - \boldsymbol{x}_n)/h$, with Jacobian given by $\frac{\partial \mathbf{u}}{\partial \mathbf{x}} = \frac{1}{h}\boldsymbol{I}$.

   ANSWER: We calculate:
   $$\begin{aligned}\int p(\mathbf{x})d\mathbf{x} &= \int \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h^D} k\Big(\frac{\mathbf{x} - \boldsymbol{x}_n}{h}\Big)d\mathbf{x} = \frac{1}{N}\int \sum_{n=1}^{N} \frac{1}{h^D} k\Big(\frac{\mathbf{x} - \boldsymbol{x}_n}{h}\Big)d\mathbf{x} \\ &= \frac{1}{N}\sum_{n=1}^{N} \int \frac{1}{h^D} k\Big(\frac{\mathbf{x} - \boldsymbol{x}_n}{h}\Big)d\mathbf{x} = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h^D} \int k\Big(\frac{\mathbf{x} - \boldsymbol{x}_n}{h}\Big)d\mathbf{x}\end{aligned}$$
   To calculate the integrals, we perform variable substitution $\mathbf{u} = (\mathbf{x} - \boldsymbol{x}_n)/h$. The Jacobian is given by $\frac{\partial \mathbf{u}}{\partial \mathbf{x}} = \frac{1}{h}\boldsymbol{I}$, a diagonal matrix with values $1/h$ on the diagonal. Thus we have $d\mathbf{u} = |\det(\frac{1}{h}\boldsymbol{I})|d\mathbf{x} = \frac{1}{h^D}d\mathbf{x}$ and we find
   $$\int k\Big(\frac{\mathbf{x} - \boldsymbol{x}_n}{h}\Big)d\mathbf{x} = \int k(\mathbf{u})h^D d\mathbf{u} = h^D.$$
   Therefore,
   $$\int p(\mathbf{x})d\mathbf{x} = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h^D} h^D = 1.$$

The K-nearest neighbour density model is defined (Bishop,eq.2.246) as:
$$p(\mathbf{x}) = \frac{K}{NV(\rho)} \tag{16}$$
with ρ the distance from $\mathbf{x}$ to its K^{th} nearest neighbour, and $V(\rho)$ the volume of a D-dimensional hypersphere with radius ρ. It is, in fact, an *improper* distribution whose integral over all space is divergent. To see this, consider 1-NN in 1-dimension with *one* datapoint x_1.

2. Write down an explicit expression for $p(x)$, given the data point x_1, and show that
   $$\int p(x)dx = \infty \tag{17}$$
   What is the effect of using $K > 1$ (at least two or more neighbours)?

   ANSWER: The density $p(x)$ at a given point x is defined by making an interval that is symmetric around x, such that x_1 falls just in the interval. So, from eq. (16) the density $p(x)$ is
   $$p(x) = \frac{K}{NV(\rho)} = \frac{1}{V} \tag{18}$$
   where V is the length of the interval.
   In terms of the distance to x_1 this becomes $p(x) = \dfrac{1}{2|x - x_1|}$, so the integral is
   $$\int p(x)dx = \int_{-\infty}^{x_1} p(x)dx + \int_{x_1}^{\infty} p(x)dx$$

which is

$$\frac{1}{2}(-\ln(x_1 - x)|_{-\infty}^{x_1} + \ln(x - x_1)|_{x_1}^{\infty}) = \infty$$

Using $K > 1$ would avoid the delta-spikes centered around the data points, but it would still remain an improper distribution. (The divergence results from the fact that the density does not fall off quickly enough when distances become large, not from the delta-spike for $K = 1$).

3. Compare strengths and weaknesses of the two methods. What is the main difference between kernel density with Gaussian kernels and a Gaussian mixture model?

   ANSWER: (See Bishop, §2.3.9 and §2.5). Kernel density always produces a proper distribution and can easily alleviate the effect of artificial discontinuities by using smoother kernels. K-nearest neighbour can reveal more detail in regions where many observations are available, and can easily be extended to classification problems.
   Smooth kernel density puts a similar Gaussian over each datapoint to approximate the unknown distribution (no weighing), whereas Gaussian mixtures try to capture the essence of a distribution by approximating it by a few, well chosen Gaussians (e.g. via the EM-algorithm, chapter 9).