# Statistical Machine Learning 2016

Exercises, 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 $\mu_0$, $\sigma^2$, $c$ 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?
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

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

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

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

## 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

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

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

Suppose $\sigma$ is given and $\mu$ is unknown. Take as prior for $\mu$ 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

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

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

$$\begin{aligned} \mu_N &= \sigma_N^2\left(\frac{N\bar{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right) && (11)\\ \sigma_N^2 &= \frac{1}{\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}} && (12) \end{aligned}$$

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

# Exercise 3

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

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

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

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

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

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

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

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

# 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 $\frac{\partial \mathbf{u}}{\partial \mathbf{x}} = \frac{1}{h}\boldsymbol{I}$.

The $K$-nearest neighbour density model is defined (Bishop,eq.2.246) as:

$$p(\mathbf{x}) = \frac{K}{NV(\rho)} \tag{16}$$

with $\rho$ the distance from $\mathbf{x}$ to its $K^{\text{th}}$ nearest neighbour, and $V(\rho)$ the volume of a $D$-dimensional hypersphere with radius $\rho$. 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)?

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?