Statistical Machine Learning 2018

Exercises, week 6

12 October 2018

TUTORIAL

Exercise 1

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

$$p(\mathbf{x}|\mu, \sigma^2) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \sigma^2\mathbf{I}) \tag{1}$$

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

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):

$$\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) && (2)\\
p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) && (3)\\
&\Rightarrow \\
p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) && (4)\\
p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) && (5)
\end{aligned}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1} \tag{6}$$

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

$$\begin{aligned}
\mu_N &= \sigma_N^2\left(\frac{N\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) && (7)\\
\sigma_N^2 &= \frac{1}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}} && (8)
\end{aligned}$$

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

Exercise 2

Consider a mixture of K Gaussian densities of the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{9}$$

Show that if the mixing coefficients satisfy

$$\pi_k \geq 0 \quad \text{and} \quad \sum_k \pi_k = 1$$

then the mixture of Gaussians in (9) is positive and normalized. (You may assume that the components of the mixture are normalized)

Exercise 3

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:

$$p(\mathbf{x}) = \frac{K}{NV(\rho)} \tag{10}$$

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
   $$\int p(x) dx = \infty \tag{11}$$
   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?

BONUS PRACTICE

Exercise 4

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

$$p(u) = \mathcal{N}(u|\mu_0, \sigma^2) \tag{12}$$
$$p(v|u) = \mathcal{N}(v|c{\cdot}u, s^2) \tag{13}$$

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?
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

$$\mu_{u|v} = \frac{\dfrac{\mu_0}{\sigma^2} + \dfrac{cv}{s^2}}{\dfrac{1}{\sigma^2} + \dfrac{c^2}{s^2}} \tag{14}$$

$$\frac{1}{\sigma^2_{u|v}} = \frac{1}{\sigma^2} + \frac{c^2}{s^2} \tag{15}$$

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

Exercise 5

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

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\right\} \tag{16}$$

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

$$\mathrm{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \tag{17}$$

1. Show the Gamma distribution belongs to the exponential family by casting it in the standard representation (16). Hint: put all λ 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

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})] \tag{18}$$

2. Using this result, show that the expectation value for the Gamma distribution (17) is given by $\mathbb{E}[\lambda] = \dfrac{a}{b}$.

Exercise 6

Matrix identities (Exercises 2.24 and 2.26 in Bishop).

- Prove the *partitioned matrix inversion formula*:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix},$$

  where $\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}$. This identity is used, for example, to simplify the expression of the inverse of the precision in a linear Gaussian model (see Bishop (2.104) and (2.105)).

- The *Woodbury matrix inversion formula* (see below) is useful when we have a large diagonal matrix $\mathbf{A}$, which is easy to invert, while $\mathbf{B}$ has many rows, but few columns (and conversely for $\mathbf{D}$), so that the right-hand side is much cheaper to evaluate than the left-hand side. A common application is finding the inverse of a low-rank update $\mathbf{A} + \mathbf{B}\mathbf{C}\mathbf{D}$ of $\mathbf{A}$, for example in the Kalman filter algorithm. Prove the correctness of the identity, which is given by:

$$(\mathbf{A} + \mathbf{B}\mathbf{C}\mathbf{D})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{D}\mathbf{A}^{-1}.$$