Statistical Machine Learning 2018

Exercises and answers, week 7

19 October 2018

TUTORIAL

Exercise 1

Finally some regression! In this exercise we again use Bayes' theorem for a linear Gaussian model (p.93, eq.2.113-2.117):

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

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

Consider a Gaussian linear regression model (Bishop, §3.1) of the form

$$p(t|\mathbf{w}, x) = \mathcal{N}(t|\boldsymbol{\phi}^T(x)\mathbf{w}, \beta^{-1}) \tag{1}$$

1. Interpret this equation, i.e. what is (a) 'modelled' here, and what makes it (b) Gaussian, (c) linear and (d) a *regression* model?

   ANSWER:
   The equation in (1) models the relation between a given input variable x and the corresponding continuous target variable t, given a set of parameters $\mathbf{w}$. This type of function is typically used in regression problems, where the goal is to predict the value of t for a new value of x, given a set of N previous $\{x_n, t_n\}$ observations. The model for regression consists of a linear combination of (nonlinear) basis functions ϕ_i in x, weighted by the w_i. The model assumes that the probability of a given output t can be described by a Gaussian distribution with constant precision β around the mean given by the regression function evaluated at point x.

2. Suppose we have input data $\mathbf{x} = (x_1, \ldots, x_N)$ and output data $\mathbf{t} = (t_1, \ldots, t_N)$. Show that the likelihood can be written as

   $$p(\mathbf{t}|\mathbf{w}, \mathbf{x}) = \mathcal{N}(\mathbf{t}|\Phi\mathbf{w}, \beta^{-1}\mathbf{I}) \tag{2}$$

   where $\Phi_{nj} \equiv \phi_j(x_n)$. (Hint: give an expression for n i.i.d. observations of (1), write out the expression $\Phi\mathbf{w}$ and verify the two are equivalent.)

ANSWER: For the likelihood of n independent observations of a variable x distributed according to (1) we have

$$\prod_n p(t_n|\mathbf{w}, x_n) = \prod_n \mathcal{N}(t_n|\boldsymbol{\phi}^T(x_n)\mathbf{w}, \beta^{-1}) \tag{3}$$

This is equivalent to the likelihood $p(\mathbf{t}|\mathbf{w}, \mathbf{x})$ of one observation of an n-dimensional target variable $\mathbf{t}$ for one n-dimensional input variable $\mathbf{x}$, distributed according to a multivariate Gaussian. Each element in the n-dimensional vector of means of the Gaussian is given by $\phi^T(x_n)\mathbf{w}$ and the covariance matrix is an $n \times n$ diagonal matrix $\beta^{-1}\mathbf{I}$, as the dimensions (observations) are independent.
Since $\boldsymbol{\phi}^T(x_n) = (\phi_0(x_n), \phi_1(x_n), \ldots, \phi_M(x_n))$ and $\mathbf{w} = (w_0, w_1, \ldots, w_M)^T$, we have

$$\Phi\mathbf{w} = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \ldots & \phi_M(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \ldots & \phi_M(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_N) & \phi_1(x_N) & \ldots & \phi_M(x_N) \end{pmatrix} \cdot \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{pmatrix} = \begin{pmatrix} \phi^T(x_1)\mathbf{w} \\ \phi^T(x_2)\mathbf{w} \\ \vdots \\ \phi^T(x_N)\mathbf{w} \end{pmatrix}$$

So, $\Phi\mathbf{w}$ is exactly the required n-dimensional column vector of means in the multivariate Gaussian, which proves (2).

3. We take as prior the Gaussian $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$. Use the above stated relations for a linear Gaussian model to show that the posterior is

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \tag{4}$$

with

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\Phi^T\mathbf{t}) \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta\Phi^T\Phi \end{aligned}$$

ANSWER:

(a) Familiar steps: match prior, likelihood and posterior

$$\begin{aligned} p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) &\quad\leftrightarrow\quad p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0) \\ p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) &\quad\leftrightarrow\quad p(\mathbf{t}|\mathbf{w}, \mathbf{x}) = \mathcal{N}(\mathbf{t}|\Phi\mathbf{w}, \beta^{-1}\mathbf{I}) \\ p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) &\quad\leftrightarrow\quad p(\mathbf{w}|\mathbf{x}, \mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \end{aligned}$$

(b) Match parameters (standard form $\leftrightarrow$ exercise)

$$\begin{aligned} \mathbf{x} &\leftrightarrow \mathbf{w} \\ \boldsymbol{\mu} &\leftrightarrow \mathbf{m}_0 \\ \boldsymbol{\Lambda}^{-1} &\leftrightarrow \mathbf{S}_0 \\ \mathbf{y} &\leftrightarrow \mathbf{t} \\ \mathbf{A} &\leftrightarrow \Phi \\ \mathbf{b} &\leftrightarrow (0, 0, \ldots, 0)^T \\ \mathbf{L}^{-1} &\leftrightarrow \beta^{-1}\mathbf{I} \end{aligned}$$

(c) Calculate $\boldsymbol{\Lambda}$, $\mathbf{L}$ and $\boldsymbol{\Sigma}$

$$\begin{aligned} \boldsymbol{\Lambda} &= \mathbf{S}_0^{-1} \\ \mathbf{L} &= \beta\,\mathbf{I} \\ \boldsymbol{\Sigma} &= (\mathbf{S}_0^{-1} + \beta\Phi^T\Phi)^{-1} \qquad (\equiv \mathbf{S}_N) \end{aligned}$$

(d) ... and fill in posterior.

4. Describe what happens to m_N and S_N: (a) in the limit $S_0 \to \infty$ (broad prior), (b) when $N = 0$, and (c) in the limit $N \to \infty$? Explain.

   ANSWER:

   (a) Extremely broad prior: $\mathbf{m}_N = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{t}$ and $\mathbf{S}_N^{-1} = \beta\Phi^T\Phi$. Results independent of prior, mean even independent of β (just data).

   (b) No data implies only prior information: $\mathbf{m}_N = \mathbf{m}_0$, $\mathbf{S}_N = \mathbf{S}_0$.

   (c) Infinite data, drowns out all prior information: $\mathbf{S}_N \to 0$ (posterior gets deterministic) and $\mathbf{m}_N$ as in (a) ($\equiv$ ML solution).

5. How can the relation for the posterior (4) be used to solve the regression problem?

   ANSWER: The answer to a regression problem requires an expression for the *predictive* distribution

$$p(t|\mathbf{t}, x) = \mathcal{N}(t|\boldsymbol{\phi}^T(x)\mathbf{m}_N, \sigma_N^2(x)) \tag{5}$$

   This can be obtained by taking the posterior $p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$ as the new prior (conditioned on $\mathbf{t}$) for the regression model in (1), and apply $2.113 + 2.114 \Rightarrow 2.115$ to compute the required marginal (conditioned on $\mathbf{t}$ and x, but that does not matter), by integrating out all values of $\mathbf{w}$. Effectively we calculate

$$p(t|x, \mathbf{t}, \mathbf{x}, \alpha, \beta) = \int p(t|x, \mathbf{w}, \beta)p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \alpha, \beta)\mathrm{d}\mathbf{w} \tag{6}$$

Exercise 2

Fitting a straight line to data: Bayesian learning in a linear regression model (see Bishop, §3.3.1).

Consider a process where a target variable t is linearly dependent on the input variable x, subject to random Gaussian noise with variance β^{-1}. Suspecting such a linear relationship we use a regression model with polynomial basis functions of the form $y(x, \mathbf{w}) = w_0 + w_1 x$. For the weights we assume an isotropic, zero-mean Gaussian prior governed by precision parameter α:

$$\begin{aligned} t &= a_0 + a_1 x + \mathcal{N}(0, \beta^{-1}) && (7)\\ y(x, \mathbf{w}) &= \phi(\mathbf{x})^T\mathbf{w} = w_0 + w_1 x && (8)\\ p(\mathbf{w}|\alpha) &= \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I}) && (9)\\ p(t|x, \mathbf{w}, \beta) &= \mathcal{N}(t|\phi(\mathbf{x})^T\mathbf{w}, \beta^{-1}) && (10) \end{aligned}$$

For Bayesian linear regression, the relation between prior, likelihood and posterior can again be derived from the standard form of Bayes' theorem for linear Gaussian models (see previous exercise). In case of the prior (9), this results in (Bishop, p.153)

$$\begin{aligned} p(\mathbf{w}|\alpha) &= \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I})\\ p(\mathbf{t}|\mathbf{w}, \mathbf{x}) &= \mathcal{N}(\mathbf{t}|\Phi\mathbf{w}, \beta^{-1}\mathbf{I})\\ &\Rightarrow\\ p(\mathbf{w}|\mathbf{t}, \mathbf{x}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_N, S_N) \end{aligned}$$

with

$$\begin{aligned} \mathbf{m}_N &= \beta\mathbf{S}_N\Phi^T\mathbf{t}\\ \mathbf{S}_N^{-1} &= \alpha\mathbf{I} + \beta\Phi^T\Phi \end{aligned}$$

1. Identify the vector of basis functions $\phi(\mathbf{x})$, and write out $\Phi^T\mathbf{t}$ and $\Phi^T\Phi$ in terms of the $\{x_n, t_n\}$. (Hint: see Bishop, p.142).

   ANSWER: From (8) we see that the vector of basis functions is $\phi(\mathbf{x}) = \begin{pmatrix} 1 & x \end{pmatrix}^T$. From the definition $\Phi_{nj} \equiv \phi_j(x_n)$ (or Bishop, p.142) we then have

$$\begin{aligned}
\Phi^T\mathbf{t} &= \sum_n \boldsymbol{\phi}(x_n)t_n = N \begin{pmatrix} \bar{\mu}_t \\ \bar{\mu}_{xt} \end{pmatrix} \\
\Phi^T\Phi &= \sum_n \boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^T = N \begin{pmatrix} 1 & \bar{\mu}_x \\ \bar{\mu}_x & \bar{\mu}_{xx} \end{pmatrix} \\
\bar{\mu}_t &= \frac{1}{N}\sum_n t_n \qquad \bar{\mu}_{xt} = \frac{1}{N}\sum_n x_n t_n \\
\bar{\mu}_x &= \frac{1}{N}\sum_n x_n \qquad \bar{\mu}_{xx} = \frac{1}{N}\sum_n x_n^2
\end{aligned}$$

2. Compute the posterior $p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \alpha, \beta)$. Show that the posterior becomes independent of α, β for large N.

   ANSWER: Substituting the answers to the previous question into the expressions for $\mathbf{m}_N$ and $\mathbf{S}_N^{-1}$ in the posterior gives

$$\begin{aligned}
p(\mathbf{w}|\mathbf{t}, \mathbf{x}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \\
\mathbf{m}_N &= \beta\mathbf{S}_N\Phi^T\mathbf{t} = N\beta\mathbf{S}_N \begin{pmatrix} \bar{\mu}_t \\ \bar{\mu}_{xt} \end{pmatrix} \qquad (\equiv \mathbf{w}_{\text{MAP}}) \\
\mathbf{S}_N^{-1} &= \alpha\mathbf{I} + \beta\Phi^T\Phi = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix} + N\beta \begin{pmatrix} 1 & \bar{\mu}_x \\ \bar{\mu}_x & \bar{\mu}_{xx} \end{pmatrix}
\end{aligned}$$

   When N large, $\mathbf{S}_N^{-1}$ is dominated by the second term and becomes infinite. Thus, the width of the posterior goes to zero. Also the mean becomes independent of β, as it cancels with the β component in the $\mathbf{S}_N$ term.

The predictive distribution $p(t|x, \mathbf{t}, \mathbf{x}, \alpha, \beta)$ can be derived in the same way as before by using the previously obtained posterior $p(\mathbf{w}|\mathbf{t}, \mathbf{x})$ as the prior for a new observation and integrating out $\mathbf{w}$. In terms of Bayes' theorem:

$$\begin{aligned}
p(\mathbf{w}|\mathbf{t}, \mathbf{x}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_N, S_N) \\
p(t|\mathbf{w}, \mathbf{t}, \mathbf{x}) &= \mathcal{N}(t|\boldsymbol{\phi}(x)^T\mathbf{w}, \beta^{-1}) \\
&\Rightarrow \\
p(t|x, \mathbf{t}, \mathbf{x}) &= \mathcal{N}(t|\mathbf{m}_N^T\boldsymbol{\phi}(x), \frac{1}{\beta} + \boldsymbol{\phi}(x)^T\mathbf{S}_N\boldsymbol{\phi}(x)),
\end{aligned}$$

with $\mathbf{m}_N$ and $\mathbf{S}_N$ defined as before. To simplify notation, denote the mean of the predictive distribution by $m(x) = \mathbf{m}_N^T\boldsymbol{\phi}(x)$ and the variance by $s^2(x) = \frac{1}{\beta} + \boldsymbol{\phi}(x)^T\mathbf{S}_N\boldsymbol{\phi}(x)$.

3. Compute the predictive distribution $p(t|x, \mathbf{t}, \mathbf{x}, \alpha, \beta) = \mathcal{N}(t|m(x), s^2(x))$ in terms of known or computable quantities. Discuss the main difference with the posterior. What happens to $s(x)$ when $N \to \infty$?

ANSWER: As before, but now filling in for the marginal:

$$\begin{aligned} p(t|x,\mathbf{x},\mathbf{t}) &= \mathcal{N}(t|m(x),s^2(x)) \\ m(x) &= \boldsymbol{\phi}(x)^T\mathbf{m}_N = N\beta \begin{pmatrix} 1 & x \end{pmatrix} \mathbf{S}_N \begin{pmatrix} \bar{\mu}_t \\ \bar{\mu}_{xt} \end{pmatrix} \\ s^2(x) &= \beta^{-1} + \boldsymbol{\phi}(x)^T\mathbf{S}_N\boldsymbol{\phi}(x) = \beta^{-1} + \begin{pmatrix} 1 & x \end{pmatrix} \mathbf{S}_N \begin{pmatrix} 1 \\ x \end{pmatrix} \\ \mathbf{S}_N^{-1} &= \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix} + N\beta \begin{pmatrix} 1 & \bar{\mu}_x \\ \bar{\mu}_x & \bar{\mu}_{xx} \end{pmatrix} \end{aligned}$$

The main difference with the expression for the posterior is that the predictive distribution is now a function of x, both in the mean and in the variance.
When $N \to \infty$, $\mathbf{S}_N \to 0$ and thus $s^2(x) \to \beta^{-1}$ and

$$m(x) = \begin{pmatrix} 1 & x \end{pmatrix} \begin{pmatrix} 1 & \bar{\mu}_x \\ \bar{\mu}_x & \bar{\mu}_{xx} \end{pmatrix}^{-1} \begin{pmatrix} \bar{\mu}_t \\ \bar{\mu}_{xt} \end{pmatrix}$$

independent of α and β.

4. Consider one data point for training $x = t = 0$. Compute and sketch $m(x)$ and $s^2(x)$ around $x = 0$. Compare your result to fig. 3.8a (although different model).

   ANSWER: For the case of one data point $\{x_1, t_1\} = (0, 0)$, we have

$$\begin{aligned} \mathbf{S}_N^{-1} &= \begin{pmatrix} \alpha+\beta & 0 \\ 0 & \alpha \end{pmatrix} \\ m(x) &= 0 \\ s^2(x) &= \beta^{-1} + \begin{pmatrix} 1 & x \end{pmatrix} \begin{pmatrix} \frac{1}{\alpha+\beta} & 0 \\ 0 & \frac{1}{\alpha} \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix} \\ &= \frac{1}{\alpha}x^2 + \{\beta^{-1} + (\alpha+\beta)^{-1}\} \end{aligned}$$

   So, the variance achieves its minimum around the mean at the point $x = 0$ and increases quadratically for values further away.

Exercise 3

Bayesian model selection (Bishop, §3.4). Consider a binary experiment: x can have two outcomes a or b. Suppose there are two hypotheses: H_0 is the hypothesis that both outcomes are equally probable. Hypothesis H_1 is the hypothesis that outcomes have different probabilities: p_a and $p_b = 1 - p_a$. The prior for these probabilities is flat. Now suppose we have a dataset of N independent outcomes of the experiment, $D = \{x_1, \ldots, x_N\}$. Let N_a be the number of a's and N_b the number of b's.

1. Show that evidence for H_0 and for H_1 respectively is given by

$$P(D|H_0) = (\frac{1}{2})^N \quad \text{and} \quad P(D|H_1) = \frac{N_a!N_b!}{(N_a+N_b+1)!}$$

   Hint: use $\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ (eq.2.265), and remember that $\Gamma(x+1) = x!$.

   ANSWER: H_0: $p_a = p_b = 0.5$, so for each outcome x_i, the probability is 0.5. So $P(D|H_0) =$

$(\frac{1}{2})^N$. H_1: For a given p_a, the probability of D is $p(D|p_a) = p_a^{N_a}(1-p_a)^{N_b}$. But p_a is unknown, and should be marginalized over: With flat prior: $p(p_a) = 1$

$$P(D|H_1) = \int_0^1 p(D|p_a)p(p_a)dp_a = \int p_a^{N_a}(1-p_a)^{N_b}dp_a$$

Using 2.265 we find the integral

$$\int p_a^{N_a}(1-p_a)^{N_b}dp_a = \frac{\Gamma(N_a+1)\Gamma(N_b+1)}{\Gamma(N_a+N_b+2)}$$

and using $\Gamma(x+1) = x!$, the result follows.

2. Compute the odds-ratios $P(D|H_1)/P(D|H_0)$ for all the possible datasets of 6 outcomes. (So $(N_a = 0, N_b = 6)$, $(N_a = 1, N_b = 5)$, etc.)

   ANSWER: Table of N_a, N_b and $P(D|H_1)/P(D|H_0)$ for $D = 6$:

0	6	9.1
1	5	1.5
2	4	0.61
3	3	0.45
4	2	0.61
5	1	1.5
6	0	9.1

   We can do the same for datasets of 60 outcomes (in Matlab), where N_a and N_b are taken 10 times as large. (So $(N_a = 0, N_b = 60)$, $(N_a = 10, N_b = 50)$, etc.) Then the results are:

   Table of N_a, N_b and $P(D|H_1)/P(D|H_0)$ for $D = 60$:

0	60	$1.9 * 10^{16}$
10	50	$2.5 * 10^5$
20	40	4.5
30	30	0.16

3. Consider an arbitrary real-world binary experiment. After 6 tests you find the results (outcomes) split 3-3. Based on the answers to the second question, do you then think that the probabilities for the two outcomes are most likely to be exactly equal? (If not, explain the answer to question two; if so, is this not too much of a coincidence?)

   ANSWER: (discuss)
   Hint: In practice arbitrary binary experiments will most likely have outcomes that are not *exactly* equally probable. But the answer at two did not compare the likelihood between 'exactly equal' and 'not exactly equal', but between 'exactly equal' and 'completely random', where both were treated as a priori equally likely. A subtle difference perhaps, but with very different implications.