Statistical Machine Learning 2018

Exercises, 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?

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.)

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

   $$p(\mathbf{w}|\mathbf{x}, \mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \tag{3}$$

   with

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

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.

5. How can the relation for the posterior (3) be used to solve the regression problem?

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:

$$\begin{aligned} t &= a_0 + a_1 x + \mathcal{N}(0, \beta^{-1}) && (4)\\ y(x, \mathbf{w}) &= \phi(\mathbf{x})^T \mathbf{w} = w_0 + w_1 x && (5)\\ p(\mathbf{w}|\alpha) &= \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I}) && (6)\\ p(t|x, \mathbf{w}, \beta) &= \mathcal{N}(t|\phi(\mathbf{x})^T \mathbf{w}, \beta^{-1}) && (7) \end{aligned}$$

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 (6), 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).

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.

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 $N \to \infty$?

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).

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

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.)

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?)