Homework 2

Deadline: November 21, 15:00

Exercise 1

We have made the following observations

sample	input x_1	input x_2	output y
(1)	3	-1	2
(2)	4	2	1
(3)	2	1	1

and want to learn a linear regression model on the form $y = w_1x_1 + w_2x_2 + \epsilon$, where $e \sim \mathcal{N}(0, 5)$.

(i) Find $w = (w_1 \;\; w_2)^T$ using the maximum likelihood approach.

(ii) Now assume the prior,

$$p(w) = \mathcal{N}\left(w| \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.2 & 0 \\ 0 & 0.2 \end{pmatrix}\right)$$

and find w using the probabilistic approach!

(iii) Compare the results from (i) and (ii).

Exercise 2

A colleague of yours is doing a study on the Swedish lower secondary school. She has collected some data about grades, and asked you for help in assembling a model. Her ultimate goal is to predict the probability distribution for a student's grade, based on some data about how he/she spends his/her spare time. The data contains the merit-value (the Swedish equivalent to GPA) for a number of students, which is on the scale $0 - 340$ points with an average somewhere around 200 points. Her data also concerns how much time each student spends on reading books and comics, playing computer games, taking parts in sports activities, and hanging out with friends. Each of these are normalized on a scale $[-1, 1]$ (where 0 is the average student), and she can see no reason (based on the outset of the study itself) to favor any activity in the explanation. In fact, your colleague tells you, it would be rather unlikely if either of these factors explained more than about 10 points each (apart from the reading, which she thinks could be likely to explain up to around 20 points). She also tells you that she does not expect these factors to explain the merit-value perfectly, but she thinks other factors not included in the study are quite likely to explain at least up to 20 points.

(i) Write down a probabilistic linear regression model (with all distributions specified!) for the problem.

(ii) If you were to include gender (likely to explain not much more than 10 points, according to your colleague) in the model as well, how would you do that?

Exercise 3

Consider the Bayesian linear regression model

$$p(y|w,\beta) = \prod_1^N \mathcal{N}(y_n; w^T x_n, \beta^{-1}) \quad \text{with the prior} \quad p(w) = \mathcal{N}(w; m_0, S_0)$$

where β, m_0 and S_0 are known.

(i) Show that the likelihood can be expressed as a multivariate Guassian distribution with a diagonal covariance matrix, i.e.,

$$p(y|w,\beta) = \prod_1^N \mathcal{N}(y_n; w^T x_n, \beta^{-1}) = \mathcal{N}(y; Xw, \beta^{-1} I_N)$$

where I_N is the identity matrix of size $N \times N$ and,

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix} \quad \text{and,} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Hint: The determinant of a diagonal matrix is equal to the product of the diagonal elements.

(ii) verify that the posterior distribution of the parameters w is

$$p(w|y) = \mathcal{N}(w; m_N, S_N)$$

where

$$\begin{aligned} m_N &= S_N(S_0^{-1} m_0 + \beta X^T y) \\ S_N^{-1} &= S_0^{-1} + \beta X^T X \end{aligned} \tag{1}$$

Hint: Assume that x_a, as well as x_b conditioned on x_a, are Gaussian distributed according to

$$p(x_a) = \mathcal{N}(x_a; \mu_a, \Sigma_a)$$

$$p(x_b|x_a) = \mathcal{N}(x_b; Ax_a + b, \Sigma_{b|a})$$

Then the conditional distribution

$$p(x_a|x_b) = \frac{p(x_a, x_b)}{p(x_b)} = \frac{p(x_b|x_a)p(x_a)}{p(x_b)}$$

is given by

$$p(x_a|x_b) = \mathcal{N}(x_a; \mu_{a|b}, \Sigma_{a|b})$$

with

$$\mu_{a|b} = \Sigma_{a|b}\Big(\Sigma_a^{-1}\mu_a + A^T\Sigma_{b|a}^{-1}(x_b - b)\Big)$$
$$\Sigma_{a|b} = \Big(\Sigma_a^{-1} + A^T\Sigma_{b|a}^{-1}A\Big)^{-1}$$

Exercise 4

In Exercise 3, we assumed that the precision β is known. Now assume that β is unknown and treat is as a random variable. That means we need to have a prior for both w and β and solve

$$p(w, \beta|y) = \frac{p(y|w, \beta)p(w, \beta)}{p(y)} \propto p(y|w, \beta)p(w, \beta)$$

Show that if we consider the likelihood $p(y|w, \beta)$ in Exercise 3 and the following Gauss-Gamma prior

$$p(w, \beta) = \mathcal{N}(w; m_0, \beta^{-1}S_0)Gam(\beta; a_0, b_0)$$

where $Gam(\beta; a, b)$ is the Gamma distribution

$$Gam(\beta; a, b) = \frac{1}{\Gamma(a)}b^a\beta^{a-1}e^{-b\beta}, \qquad \beta \in [0, \infty)$$

then the posterior will also be a Gauss-Gamma distribution

$$p(w, \beta|y) = \mathcal{N}(w; m_N, \beta^{-1}S_N)Gam(\beta; a_N, b_N),$$

where,

$$\begin{aligned}
m_N &= S_N(S_0^{-1}m_0 + X^Ty) \\
S_N^{-1} &= S_0^{-1} + X^TX \\
a_N &= a_0 + \frac{N}{2} \\
b_N &= b_0 + \frac{1}{2}\Big(m_0^TS_0^{-1}m_0 - m_N^TS_N^{-1}m_N + \sum_{n=1}^{N} y_n^2\Big)
\end{aligned}$$

This means that the Gauss-Gamma prior is a conjugate prior to the Gaussian likelihood with unknown w and β.