Assignment 2 SSY316

Newson Matthew, Erik Norlin

November 2023

Exercise 1

Given observations

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

and the linear regression model $y = w_1x_1 + w_2x_2 + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \beta^{-1})$, $\beta^{-1} = 5$

i)

We want to find $w = (w_1, w_2)^T$ using the maximum likelihood approach. Since ϵ is assumed to be Gaussian distributed we find w_{ML} by minimising least squares $L_D(w)$ and solve w_{ML} in closed form.

$$\begin{aligned} L_D(w) &= \sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{M} w_{ij}x_{ij}\right)^2 \\ &= (y - Xw)^T(y - Xw) \end{aligned}$$

Where N is the number of points and M is the number of dimensions. Taking the derivative and setting it to zero to solve for w_{ML}

$$\begin{aligned} &\frac{\partial L_D(w)}{\partial w_{ML}} = -2X^T(y - Xw_{ML}) = 0 \\ &\Rightarrow w_{ML} = (X^TX)^{-1}X^Ty \end{aligned}$$

Where

$$X = \begin{pmatrix} 3 & -1 \\ 4 & 2 \\ 2 & 1 \end{pmatrix}, y = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$$

We get $w_{ML} = \frac{1}{25}\begin{pmatrix} 13 \\ -11 \end{pmatrix} \approx \begin{pmatrix} 0.52 \\ -0.44 \end{pmatrix}$

ii)

Now we use a probabilistic approach instead and assume the prior

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

We choose a Gaussian distributed likelihood

$$p(y|X, w, \beta) = \prod_{i=1}^{N} \mathcal{N}(y_i|w^T x_i, \beta^{-1})$$

Since both the likelihood and the prior are Gaussian the posterior will also be Gaussian distributed

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

where

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

and we can see from the prior that

$$\begin{aligned} m_0 &= (0, 0)^T \\ S_0 &= \alpha^{-1} I \\ \alpha &= 5 \end{aligned}$$

Putting it all together we get that the posterior is

$$p(w|y, X, \beta) \approx \mathcal{N}\left(w| \begin{pmatrix} 0.22 \\ -0.02 \end{pmatrix}, \begin{pmatrix} 0.1 & -0.02 \\ -0.02 & 0.17 \end{pmatrix}\right)$$

iii)

Comparing the results from (i) and (ii) we can see that in the probabilistic approach we get a family of models where as computing the maximum likelihood estimate gives us one model. In the probabilistic approach, computing the square root of the eigenvalues of the covariance matrix of the posterior, we get the standard deviation magnitude in the direction of the largest spread of the data.

$$m_N \pm \sqrt{\lambda} \approx \begin{pmatrix} 0.22 \\ -0.02 \end{pmatrix} \pm \begin{pmatrix} 0.32 \\ 0.41 \end{pmatrix}$$

From this we can see that w_{ML} roughly falls within the deviation of the distribution of the posterior.

Exercise 2

i)

The goal is to predict the probability distribution for a student's grade which corresponds to the merit value we shall call y. The data she has obtain during her research is the following :

- x_1 : reading books and comics (normalized) with a standard deviation of 20.
- x_2 : playing computer games (normalized) with a standard deviation of 10.
- x_3 : playing sports (normalized) with a standard deviation of 10.
- x_4 : spending time with friends(normalized)with a standard deviation of 10.
- w_0 : merit value which varies from 0 up to 340 with an average of 200 with a standard deviation of 10.

A probabilistic linear regression model for this case would be the following :

$$y = w_0 + w_1 * x_1 + w_2 * x_2 + w_3 * x_3 + w_4 * x_4 + \epsilon$$

with w_0, w_1, w_2, w_3, w_4 and ϵ corresponding to the following :

- w_0 : $\mathcal{N}(200, 10^2)$
- w_1 : $\mathcal{N}(0, 20^2)$
- w_2 : $\mathcal{N}(0, 10^2)$
- w_3 : $\mathcal{N}(0, 10^2)$
- w_4 : $\mathcal{N}(0, 10^2)$
- ϵ : $\mathcal{N}(0, 10^2)$

ii)

If we were to include gender then it would change the following :

$$y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + w_6x_6 + \epsilon$$

with x_5 corresponding to males and x_6 to females for examples.

- w_5 : $\mathcal{N}(0, 10^2)$
- w_6 : $\mathcal{N}(0, 10^2)$

Exercise 3

Consider the Bayesian linear regression model

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

with the prior

$$p(w) = \mathcal{N}(w; m_0, S_0)$$

where β, m_0 and S_0 are known

i)

We want to show that

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

$$
\begin{aligned}
\mathcal{N}(y; Xw, \beta^{-1} I_N) &= \frac{\exp\left(-\frac{1}{2}(y - Xw)^T(\beta^{-1} I_N)^{-1}(y - Xw)\right)}{\sqrt{(2\Pi)^N |\beta^{-1} I_N|}} \\
&= \frac{\exp\left(-\frac{1}{2\beta^{-1}}(y - Xw)^T(y - Xw)\right)}{\sqrt{(2\Pi)^N \beta^{-N}}} \\
&= \frac{\exp\left(-\frac{1}{2\beta^{-1}}\sum_{n=1}^{N}(y_n - x_n^T w)^2\right)}{\sqrt{2\Pi\beta^{-1}}^N} \\
&= \prod_{n=1}^{N} \frac{\exp\left(-\frac{1}{2\beta^{-1}}(y_n - x_n^T w)^2\right)}{\sqrt{2\Pi\beta^{-1}}^n} \\
&= \prod_{n=1}^{N} \mathcal{N}(y_n; w^T x_n, \beta^{-1}) \\
&= p(y|w, \beta)
\end{aligned}
$$

ii)

Now we want to verify that the posterior distribution of the parameters w is

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

where

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

We will utilize the important following result for conditional Gaussians. Assume that x_a, as well as x_b conditioned on x_a, are Gaussian distributed according to

$$
\begin{aligned}
p(x_a) &= \mathcal{N}(x_a|\mu_a, \Sigma_a) \\
p(x_b|x_a) &= \mathcal{N}(x_b|Ax_a + c, \Sigma_{b|a})
\end{aligned}
$$

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

$$
\begin{aligned}
\mu_{a|b} &= \Sigma_{a|b}\left(\Sigma_a^{-1}\mu_a + A^T\Sigma_{b|a}^{-1}(x_b - c)\right) \\
\Sigma_{a|b} &= \left(\Sigma_a^{-1} + A^T\Sigma_{b|a}^{-1}A\right)^{-1}
\end{aligned}
$$

Now let

$$
\begin{cases}
x_a = w \\
\mu_a = m_0 \\
\Sigma_a = S_0 \\
x_b = y \\
A = X \\
\Sigma_{b|a} = \beta^{-1} I \\
c = 0
\end{cases}
$$

Repeating the result for conditional Gaussians we have the following. Assume that w, as well as y conditioned on w, are Gaussian distributed according to

$$
\begin{aligned}
p(w) &= \mathcal{N}(w|m_0, S_0) \\
p(y|w, \beta) &= \mathcal{N}(y|Xw, \beta^{-1} I)
\end{aligned}
$$

Then the conditional distribution

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

is given by

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

with

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

This holds and verifies the claim.

Exercise 4

Now assume that both β and w are unknown random variables which means that we need to have a prior for both β and w and solve

$$
p(w, \beta|y) \propto p(y|\beta, w)p(w, \beta)
$$

We want to show that if we consider the same Gaussian likelihood distribution as in Exercise 3

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

and the following Gaussian-Gamma prior

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

where

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

then the posterior will also be Gaussian-Gamma distributed as

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

where

$$
\begin{aligned}
m_N &= S_N(S_0^{-1}m_0 + X^T y)\\
S_N^{-1} &= S_0^{-1} + X^T X\\
a_N &= a_0 + \frac{N}{2}\\
b_N &= b_0 + \frac{1}{2}\left(m_0^T S_0^{-1} m_0 - m_N^T S_N^{-1} m_N + \sum_{n=1}^{N} y_n^2\right)
\end{aligned}
$$

We show this by expanding and manipulating the term $p(w,\beta|y) \propto p(y|\beta,w)p(w,\beta)$

$$
p(y|\beta,w)p(w,\beta) = \mathcal{N}(y|Xw,\beta^{-1}I_N)\mathcal{N}(w|m_0,\beta^{-1}S_0)\mathrm{Gam}(\beta|a_0,b_0)
$$

$$
\begin{aligned}
&= \frac{\exp\left[-\frac{1}{2}(y-Xw)(\beta^{-1}I)^{-1}(y-Xw)\right]}{\sqrt{(2\pi)^N|\beta^{-1}I|}} \cdot \frac{\exp\left[-\frac{1}{2}(w-m_0)(\beta^{-1}S_0)^{-1}(w-m_0)\right]}{\sqrt{(2\pi)^M|\beta^{-1}S_0|}} \cdot \frac{b_0^{a_0}\beta^{a_0-1}\exp\left[-b_0\beta\right]}{\Gamma(a_0)}\\
&= \frac{\beta^{N/2}\exp\left[-\frac{\beta}{2}(y-Xw)(y-Xw)\right]}{\sqrt{(2\pi)^N}} \cdot \frac{\beta^{M/2}\exp\left[-\frac{\beta}{2}(w-m_0)S_0^{-1}(w-m_0)\right]}{\sqrt{(2\pi)^M|S_0|}} \cdot \frac{b_0^{a_0}\beta^{a_0-1}\exp\left[-b_0\beta\right]}{\Gamma(a_0)}
\end{aligned}
$$

Now we simplify and disregard normalising constants that do not depend on β

$$
\begin{aligned}
&\Rightarrow \beta^{a_0+\frac{N+M}{2}-1}\exp\left[-\frac{\beta}{2}(y-Xw)(y-Xw)\right] \cdot \exp\left[-\frac{\beta}{2}(w-m_0)S_0^{-1}(w-m_0)\right] \cdot \exp\left[-b_0\beta\right]\\
&= \beta^{a_0+\frac{N+M}{2}-1}\exp\left[-\frac{\beta}{2}\left[(y-Xw)(y-Xw)+(w-m_0)S_0^{-1}(w-m_0)\right]\right] \cdot \exp\left[-b_0\beta\right]
\end{aligned}
$$

We handle the squares in the second exponent in isolation

$$
\begin{aligned}
&\Rightarrow (y-Xw)(y-Xw)+(w-m_0)S_0^{-1}(w-m_0)\\
&= y^T y - 2w^T X^T y + w^T X^T X w + w^T S_0^{-1} w - 2w^T S_0^{-1} m_0 + m_o^T S_o^{-1} m_0\\
&= y^T y - w^T(X^T X + S_0^{-1})w - 2w^T(X^T y + S_0^{-1} m_0) + m_o^T S_o^{-1} m_0
\end{aligned}
$$

Let

$$
\begin{aligned}
B &= S_0^{-1} m_0 + X^T y\\
S_N^{-1} &= S_0^{-1} + X^T X
\end{aligned}
$$

then we have

$$
\begin{aligned}
&\Rightarrow y^T y + w^T S_N^{-1} w - 2w^T B + m_0^T S_0^{-1} m_0\\
&= y^T y + (w - S_N B)^T S_N^{-1}(w - S_N B) - B S_N B + m_0^T S_0^{-1} m_0
\end{aligned}
$$

Let $m_N = S_N B$, then we get that

$$
\Rightarrow y^T y + (w-m_N)^T S_N^{-1}(w-m_N) - m_N^T S_N^{-1} m_N + m_0^T S_0^{-1} m_0
$$

Putting this back into the second exponent

$$
\begin{aligned}
&\Rightarrow \beta^{a_0+\frac{N+M}{2}-1}\exp\left[-\frac{\beta}{2}\left[y^Ty+(w-m_N)^TS_N^{-1}(w-m_N)-m_N^TS_N^{-1}m_N+m_0^TS_0^{-1}m_0\right]\right]\cdot\exp\left[-b_0\beta\right]\\
&=\beta^{\frac{M}{2}}\exp\left[-\frac{\beta}{2}\left[(w-m_N)^TS_N^{-1}(w-m_N)\right]\right]\cdot\beta^{a_0+\frac{N}{2}-1}\exp\left[-\beta\left[b_0+(y^Ty-m_N^TS_N^{-1}m_N+m_0^TS_0^{-1}m_0)\right]\right]\\
&=\frac{\exp\left[-\frac{1}{2}\left[(w-m_N)^T(\beta^{-1}S_N)^{-1}(w-m_N)\right]\right]}{\sqrt{|\beta^{-1}I|}}\cdot\beta^{a_0+\frac{N}{2}-1}\exp\left[-\beta\left[b_0+\frac{1}{2}\left(m_0^TS_0^{-1}m_0-m_N^TS_N^{-1}m_N+\sum_{n=1}^{N}y_n^2\right)\right]\right]\\
&\propto\mathcal{N}(w|m_N,\beta^{-1}S_N)\mathrm{Gam}(\beta|a_N,b_N)\\
&=p(w,\beta|y)
\end{aligned}
$$

where

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

which proves the claim.