SSY316 Assignment 2

Edoardo Mangia
edoardom@student.chalmers.se

Nuree Kim
nuree@student.chalmers.se

November 2024

1 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 of the form:

$$y = w_1 x_1 + w_2 x_2 + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, 5)$.

(i) Find $\mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$ using the maximum likelihood approach.

(ii) Now assume the prior,

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

and find $\mathbf{w}$ using the probabilistic approach.

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

Solution

Part (i): Maximum Likelihood Estimation (MLE)

Objective: Find the weight vector $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ that maximizes the likelihood of the observed data.

Approach: For linear regression with Gaussian noise, the MLE is equivalent to the Least Squares Estimator.

Steps:

1. **Matrix Representation:**

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$$

where

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

2. **Compute $\mathbf{X}^\top \mathbf{X}$ and $\mathbf{X}^\top \mathbf{y}$:**

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} 3 & 4 & 2 \\ -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 3 & -1 \\ 4 & 2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 29 & 7 \\ 7 & 6 \end{bmatrix}$$

$$\mathbf{X}^\top \mathbf{y} = \begin{bmatrix} 3 & 4 & 2 \\ -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 12 \\ 1 \end{bmatrix}$$

3. **Compute the Inverse of $\mathbf{X}^\top \mathbf{X}$:**

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{\det(\mathbf{X}^\top \mathbf{X})} \begin{bmatrix} 6 & -7 \\ -7 & 29 \end{bmatrix} = \frac{1}{125} \begin{bmatrix} 6 & -7 \\ -7 & 29 \end{bmatrix}$$

4. **Calculate $\mathbf{w}_{\mathbf{ML}}$:**

$$\mathbf{w}_{\mathrm{ML}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \frac{1}{125} \begin{bmatrix} 6 & -7 \\ -7 & 29 \end{bmatrix} \begin{bmatrix} 12 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{65}{125} \\ \frac{-55}{125} \end{bmatrix} = \begin{bmatrix} 0.52 \\ -0.44 \end{bmatrix}$$

Result:

$$\mathbf{w}_{\mathrm{ML}} = \begin{bmatrix} 0.52 \\ -0.44 \end{bmatrix}$$

Part (ii): Bayesian Estimation with Prior

Objective: Incorporate prior information about $\mathbf{w}$ to find the posterior estimate $\mathbf{w}_{\mathrm{MAP}}$.

Given Prior:

$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w} \mid \mathbf{0}, \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}\right)$$

Approach: Use the Maximum A Posteriori (MAP) estimation, which combines the likelihood with the prior.

Steps:

1. **Define Parameters:**

- **Likelihood Variance (σ^2):** 5
- **Prior Covariance (Σ):**

$$\Sigma = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}$$

- **Prior Precision (Σ^{-1}):**

$$\Sigma^{-1} = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$$

2. **Compute $\frac{\mathbf{X}^\top \mathbf{X}}{\sigma^2} + \Sigma^{-1}$:**

$$\frac{\mathbf{X}^\top \mathbf{X}}{5} + \Sigma^{-1} = \frac{1}{5} \begin{bmatrix} 29 & 7 \\ 7 & 6 \end{bmatrix} + \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix} = \begin{bmatrix} 10.8 & 1.4 \\ 1.4 & 6.2 \end{bmatrix}$$

3. **Compute the Inverse of the Above Matrix:**

$$\left(\frac{\mathbf{X}^\top \mathbf{X}}{5} + \Sigma^{-1} \right)^{-1} = \frac{1}{\det\left(\begin{bmatrix} 10.8 & 1.4 \\ 1.4 & 6.2 \end{bmatrix} \right)} \begin{bmatrix} 6.2 & -1.4 \\ -1.4 & 10.8 \end{bmatrix} = \frac{1}{65} \begin{bmatrix} 6.2 & -1.4 \\ -1.4 & 10.8 \end{bmatrix}$$

4. **Compute $\frac{\mathbf{X}^\top \mathbf{y}}{\sigma^2}$:**

$$\frac{\mathbf{X}^\top \mathbf{y}}{5} = \frac{1}{5} \begin{bmatrix} 12 \\ 1 \end{bmatrix} = \begin{bmatrix} 2.4 \\ 0.2 \end{bmatrix}$$

5. **Calculate $\mathbf{w}_{\mathbf{MAP}}$:**

$$\mathbf{w}_{\text{MAP}} = \left(\frac{\mathbf{X}^\top \mathbf{X}}{5} + \Sigma^{-1} \right)^{-1} \frac{\mathbf{X}^\top \mathbf{y}}{5} = \frac{1}{65} \begin{bmatrix} 6.2 & -1.4 \\ -1.4 & 10.8 \end{bmatrix} \begin{bmatrix} 2.4 \\ 0.2 \end{bmatrix} = \frac{1}{65} \begin{bmatrix} 14.6 \\ -1.2 \end{bmatrix} = \begin{bmatrix} 0.2246 \\ -0.0185 \end{bmatrix}$$

Result:

$$\mathbf{w}_{\text{MAP}} \approx \begin{bmatrix} 0.2246 \\ -0.0185 \end{bmatrix}$$

Part (iii): Comparison of $\mathbf{w}_{\mathbf{ML}}$ and $\mathbf{w}_{\mathbf{MAP}}$

- **Maximum Likelihood Estimate ($\mathbf{w}_{\mathbf{ML}}$):**

$$\mathbf{w}_{\text{ML}} = \begin{bmatrix} 0.52 \\ -0.44 \end{bmatrix}$$

- **Bayesian MAP Estimate ($\mathbf{w}_{\mathbf{MAP}}$):**

$$\mathbf{w}_{\text{MAP}} \approx \begin{bmatrix} 0.2246 \\ -0.0185 \end{bmatrix}$$

$$\|W_{ML}\| = 0.691 \quad > \quad \|W_{MAP}\| = 0.225$$

MLE estimates are based solely on the data without any prior information. When the data is limited or noisy, there is a higher risk of overfitting. On the other hand, MAP incorporates a prior distribution into the estimation. In this case, the prior distribution assumes that $\mathbf{w}$ is likely to be close to 0, which leads to a tendency for $\mathbf{w}$ to converge toward values near 0.

2 Exercise 2

$$\begin{aligned}
&x_1 : \text{reading books} \quad (-1 \leq x_i \leq 1)\\
&x_2 : \text{playing computers}\\
&x_3 : \text{sports}\\
&x_4 : \text{friends}\\
&t : \text{Meta-Values} \quad (0 \leq t \leq 340)\\
&\mathbb{E}(t) = 200
\end{aligned}$$

Solution

Part (i)

(i)

1. $W_i \sim \mathcal{N}(0, \alpha^{-1})$
   - When: $W_i \leq 10 \quad (i = 2, 3, 4)$

$$\sigma = 10 \quad , \alpha = \frac{1}{\sigma^2} = 0.01$$

$$\therefore W_2, W_3, W_4 \sim \mathcal{N}(0, 100)$$

   - When: $W_i \leq 20 \quad (i = 1)$

$$\sigma = 20 \quad \alpha = \frac{1}{\sigma^2} = 0.0025$$

$$\therefore W_1 \sim \mathcal{N}(0, 400)$$

2. $\epsilon \sim \mathcal{N}(0, \beta^{-1}) \quad \Rightarrow$ other factors

$$\sigma = 20 \quad d = \frac{1}{\sigma^2} = 0.0025$$

$$\therefore \epsilon \sim \mathcal{N}(0, 400)$$

3. **When** $\mu(t, w) = x_1 w_1 + x_2 w_2 + x_3 w_3 + x_4 w_4$

$$t(t, w) = \mu(t, w) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \beta^2)$$

$$\therefore t|x \sim \mathcal{N}(\mu(t, w), 400)$$

Part (ii)

$$x_5 = \begin{cases} 0 & \text{(e.g., Male)} \\ 1 & \text{(e.g., Female)} \end{cases}$$

$$\begin{gathered} \mu(x, w) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 \\ w_5 \sim \mathcal{N}(0, 10^2) \\ t(t, w) = \mu(t, w) + \epsilon \\ \therefore t|x \sim \mathcal{N}(\mu(t, w), 400) \end{gathered}$$

3 Exercise 3

Consider the Bayesian linear regression model

$$p(\mathbf{y} \mid \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(y_n; \mathbf{w}^\top \mathbf{x}_n, \beta^{-1}),$$

with the prior

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{m}_0, \mathbf{S}_0),$$

where β, $\mathbf{m}_0$, and $\mathbf{S}_0$ are known.

Solution

Part (i)

$$\mathcal{N}(y_n | w^T x_n, \beta^{-1}) = \frac{1}{\sqrt{2\pi\beta^{-1}}} \exp\left(-\frac{\beta}{2}(y_n - w^T x_n)^2\right)$$

$$\prod_{n=1}^{N} \mathcal{N}(y_n | w^T x_n, \beta^{-1}) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\beta^{-1}}} \exp\left(-\frac{\beta}{2}(y_n - w^T x_n)^2\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\beta^{-1}}}\right)^N \exp\left(-\frac{\beta}{2} \sum_{n=1}^{N} (y_n - w^T x_n)^2\right)$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix}, \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} .X \cdot w = \begin{bmatrix} w^T x_1 \\ w^T x_2 \\ \vdots \\ w^T x_N \end{bmatrix}$$

Expanding $(\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w})$, we find:

$$(\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w}) = \sum_{n=1}^{N}(y_n - \mathbf{w}^\top\mathbf{x}_n)^2.$$

$$= \frac{1}{(2\pi)^{N/2}(\beta^{-1})^{N/2}} \exp\left(-\frac{\beta}{2}(Y - Xw)^T(Y - Xw)\right)$$

$$\beta = \beta I_N = (\beta^{-1}I_N)^{-1}, \quad \det(\beta^{-1}I_N) = \beta^{-N}$$

$$= \frac{1}{\sqrt{(2\pi)^N \det(\beta^{-1}I_N)}} \exp\left(-\frac{1}{2}(Y - Xw)^T(\beta^{-1}I_N)^{-1}(Y - Xw)\right)$$

$\because$ The general form of the probability density function for multivariate Gaussian distribution is:

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp\left(-\frac{1}{2}(x - \mu)^T\Sigma^{-1}(x - \mu)\right)$$

$\rightarrow \Sigma$ is the covariance matrix

$$\therefore \frac{1}{\sqrt{(2\pi)^N \det(\beta^{-1}I_N)}} \exp\left(-\frac{1}{2}(y - Xw)^T(\beta^{-1}I_N)^{-1}(y - Xw)\right) = \mathcal{N}(y|Xw, (\beta^{-1}I_N))$$

Part (ii)

The prior distribution of $\mathbf{w}$ is Gaussian:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{m}_0, \mathbf{S}_0).$$

From Part (i), the likelihood is Gaussian:

$$p(\mathbf{y} \mid \mathbf{w}, \beta) = \mathcal{N}(\mathbf{y}; \mathbf{X}\mathbf{w}, \beta^{-1}\mathbf{I}_N).$$

Using Bayes' theorem:

$$p(\mathbf{w} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{w})p(\mathbf{w}),$$

and since both prior and likelihood are Gaussian, the posterior $p(\mathbf{w} \mid \mathbf{y})$ is also Gaussian:

$$p(\mathbf{w} \mid \mathbf{y}) = \mathcal{N}(\mathbf{w}; \mathbf{m}_N, \mathbf{S}_N).$$

The hint can be applied to this problem by setting up w and y to correspond to x_a and x_b, respectively.

Variable Setup

- Let $x_a = w$ and $x_b = y$. Treat the **prior distribution** and **likelihood** as the distribution of x_a and the conditional distribution $x_b|x_a$, respectively.

With this setup, we have:

- $x_a = w$ with **prior distribution** $p(w) = \mathcal{N}(w; m_0, S_0)$, corresponding to the hint variables as follows:
  - $\mu_a = m_0$
  - $\Sigma_a = S_0$
- $x_b = y$ with **conditional distribution** $p(y|w) = \mathcal{N}(y; Xw, \beta^{-1}I_N)$, corresponding to the hint variables as follows:
  - $A = X$
  - $b = 0$
  - $\Sigma_{b|a} = \beta^{-1}I_N$

Using the given hints in the problem, calculate the mean (M_N) and covariance (S_N) of $p(w|y)$.

$$\mathbf{\Sigma}_{a|b} = \left(\mathbf{\Sigma}_a^{-1} + \mathbf{A}^\top \mathbf{\Sigma}_{b|a}^{-1}\mathbf{A}\right)^{-1}.$$

$$\mathbf{S}_N = (\mathbf{S}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X})^{-1},$$

Posterior mean m_N

$$m_{a|b} = \Sigma_{a|b}\left(\Sigma_a^{-1}m_a + A^T\Sigma_b^{-1}(x_b - b)\right)$$

Substitution applied: $m_N = S_N(S_0^{-1}m_0 + \beta X^T y)$

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

4 Exercise 4

1. Likelihood function $p(y|w,\beta)$: Use the likelihood from Exercise 3 as follows:

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

2. Prior distribution $p(w,\beta)$: In this problem, assume a Gaussian-Gamma prior distribution:

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

Here:

- $\mathcal{N}(w; m_0, \beta^{-1} S_0)$: A Gaussian distribution with mean m_0 and covariance matrix $\beta^{-1} S_0$.
- $\mathrm{Gam}(\beta; a_0, b_0)$: A Gamma distribution with a_0 and b_0.

$$\mathrm{Gam}(\beta; a_0, b_0) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \beta^{a_0 - 1} e^{-b_0 \beta}, \quad \beta >= 0$$

Step 1: Compute $p(y|w,\beta)$

The likelihood $p(y|w,\beta)$ is given as:

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

This can be written in the following explicit form:

$$p(y|w,\beta) = \frac{1}{(2\pi)^{N/2}(\beta^{-1})^{N/2}} \exp\left(-\frac{\beta}{2}(y - Xw)^T (y - Xw)\right)$$

Therefore,

$$p(y|w,\beta) = \frac{1}{(2\pi\beta^{-1})^{N/2}} \exp\left(-\frac{\beta}{2}\left(y^T y - 2y^T Xw + w^T X^T Xw\right)\right)$$

Step 2: Compute $p(w, \beta)$

The prior distribution $p(w, \beta)$ is given as:

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

1. Gaussian prior $w \sim \mathcal{N}(m_0, \beta^{-1} S_0)$:

$$\mathcal{N}(w; m_0, \beta^{-1} S_0) = \frac{1}{(2\pi)^{d/2} |\beta^{-1} S_0|^{1/2}} \exp\left(-\frac{\beta}{2}(w - m_0)^T S_0^{-1}(w - m_0)\right)$$

2. Gamma prior:

$$\mathrm{Gam}(\beta; a_0, b_0) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \beta^{a_0 - 1} e^{-b_0 \beta}$$

Therefore, $p(w, \beta)$ is expressed as follows:

$$p(w, \beta) = \frac{1}{(2\pi)^{d/2} |\beta^{-1} S_0|^{1/2}} \exp\left(-\frac{\beta}{2}(w - m_0)^T S_0^{-1}(w - m_0)\right) \cdot \frac{1}{\Gamma(a_0)} b_0^{a_0} \beta^{a_0 - 1} e^{-b_0 \beta}$$

Step 3: Compute $p(w, \beta | y)$

To compute the posterior distribution $p(w, \beta|y)$, we combine the likelihood $p(y|w, \beta)$ and the prior distribution $p(w, \beta)$:

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

Substituting the expressions for $p(y|w, \beta)$ and $p(w, \beta)$, we get:

$$p(y|w, \beta) p(w, \beta) \propto \exp\left(-\frac{\beta}{2}\left(y^T y - 2y^T X w + w^T X^T X w + (w - m_0)^T S_0^{-1}(w - m_0)\right)\right) \beta^{a_0 - 1} e^{-b_0 \beta}$$

Now, by rearranging this equation, we will separate the posterior distributions of w and β into their respective forms.

Step 4: w Related Terms

From the equation above, we will isolate the terms related to w. By combining the likelihood and prior distributions, we construct a quadratic equation for w. Using this, we can derive the mean and covariance of the posterior distribution of w.

The exponential terms for w can be rearranged as follows:

$$-\frac{\beta}{2}\left(w^T \left(X^T X + S_0^{-1}\right) w - 2\left(X^T y + S_0^{-1} m_0\right)^T w\right)$$

From this expression, we can identify that the conditional posterior distribution of w follows a Gaussian distribution. That is:

$$w|\beta, y \sim \mathcal{N}(m_N, \beta^{-1}S_N)$$

where:

- Posterior mean:
$$m_N = S_N \left(S_0^{-1}m_0 + X^T y\right)$$
- Posterior covariance:
$$S_N^{-1} = S_0^{-1} + X^T X$$

Step 5: Terms Related to β

Having derived the posterior distribution for w, we now proceed to isolate the terms related to β and express them in the form of a Gamma distribution.

From the resulting expression, the exponential terms related to β can be rearranged as follows:

$$\exp\left(-\beta\left(b_0 + \frac{1}{2}\left(y^T y + m_0^T S_0^{-1} m_0 - m_N^T S_N^{-1} m_N\right)\right)\right)$$

Thus, the posterior distribution of β can be expressed in the form of a Gamma distribution as follows:

$$\beta|y \sim \mathrm{Gam}(a_N, b_N)$$

where:

- Updated shape parameter:
$$a_N = a_0 + \frac{N}{2}$$
- Updated scale parameter:
$$b_N = b_0 + \frac{1}{2}\left(y^T y + m_0^T S_0^{-1} m_0 - m_N^T S_N^{-1} m_N\right)$$

Final Result: Posterior $p(w, \beta|y)$

Through this process, the joint posterior distribution of w and β can be expressed as the Gaussian-Gamma distribution (Gauss-Gamma distribution):

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

where:

$$m_N = S_N \left(S_0^{-1}m_0 + X^T y\right),$$

$$S_N^{-1} = S_0^{-1} + X^T X,$$
$$a_N = a_0 + \frac{N}{2},$$
$$b_N = b_0 + \frac{1}{2}\left(y^T y + m_0^T S_0^{-1} m_0 - m_N^T S_N^{-1} m_N\right).$$