Assignemnt 2: Probabilistic ML

Marcus Hansen

November 2024

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 $y = w_1x_1 + w_2x_2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 5)$.

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

We start by defining $\mathbf{x} = (x_1\ x_2)^T$, denote $\mathbf{x}_{(i)}$ and $y_{(i)}$ as the values of $\mathbf{x}$, y for sample i. We also note that $y \sim \mathcal{N}(\mathbf{x}^T\mathbf{w}, \beta^{-1} = 5)$. The likelihood of a single sample i is therefore given by:

$$p(y_{(i)}|\mathbf{w}, \mathbf{x}_{(i)}, \beta^{-1}) = \frac{1}{\sqrt{2\pi\beta^{-1}}} \exp\left[\frac{1}{2\pi}(y_{(i)} - \mathbf{x}_{(i)}^T\mathbf{w})^2\right] \tag{1}$$

With independent samples we have the full likelihood as:

$$\begin{aligned} p(\mathbf{y}|\mathbf{w}, X, \beta^{-1}) &= \prod_{i=1}^{3} p(y_{(i)}|\mathbf{w}, \mathbf{x}_{(i)}, \beta^{-1}) \\ &= \prod_{i=1}^{3} \frac{1}{\sqrt{2\pi\beta^{-1}}} \exp\left[\frac{1}{2\pi}(y_{(i)} - \mathbf{x}_{(i)}^T\mathbf{w})^2\right] \\ &= \frac{1}{(2\pi\beta^{-1})^{3/2}} \exp\left[\frac{1}{2\pi}\sum_{i=1}^{3}(y_{(i)} - \mathbf{x}_{(i)}^T\mathbf{w})^2\right] \end{aligned}$$

With $X = [\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \mathbf{x}_{(3)}]^T$ and $\mathbf{y} = [y_{(1)}\ y_{(2)}\ y_{(3)}]^T$, we can rewrite as:

$$p(\mathbf{y}|\mathbf{w}, X, \beta^{-1}) = \frac{1}{(2\pi\beta^{-1})^{3/2}} \exp\left[\frac{1}{2\pi}\|\mathbf{y} - X\mathbf{w}\|^2\right] \tag{2}$$

In which we want to find the $\mathbf{w}$ that maximizes the likelihood of observing the outputs given the inputs. This can be done by solving:

$$\mathbf{w}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} p(\mathbf{y}|\mathbf{w}, X, \beta^{-1})$$

which is equivalent to solving:

$$\hat{\mathbf{w}}_{\mathrm{ML}} = \arg\min_{\mathbf{w}} -\log(p(\mathbf{y}|\mathbf{w}, X, \beta^{-1})) \tag{3}$$

Which we can solve by setting the derivative of the argument in expression 3 to zero:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}} \left[-\log(p(\mathbf{y}|\mathbf{w}, \mathbf{x}, \beta^{-1})) \right] = 0 \iff \\
\frac{\partial}{\partial \mathbf{w}} \left[\frac{3}{2} \log(2\pi\beta^{-1}) + \frac{1}{2\beta^{-1}} \|\mathbf{y} - X\mathbf{w}\|^2 \right] = 0 \iff \qquad (4) \\
\frac{-2}{2\beta^{-1}} X^T(\mathbf{y} - X\mathbf{w}) = 0 \iff \\
X^T X w = X^T \mathbf{y} \implies \\
\hat{\mathbf{w}}_{ML} = (X^T X)^{-1} X^T \mathbf{y} \qquad (5)
\end{aligned}$$

Inserting the values from the assignment into this equation we get:

$$\hat{\mathbf{w}}_{\mathrm{ML}} = \begin{bmatrix} 0.5200 \\ -0.4400 \end{bmatrix}$$

(ii) Now assume the prior,

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

and find w using the probabilistic approach!

We denote the covariance matrix $C_w = \alpha^{-1}\mathbf{I}_{2x2} = 0.2 \cdot \mathbf{I}_{2x2}$. we have:

$$p(\mathbf{w}) = \frac{\alpha}{2\pi} \exp\left[\frac{-\alpha}{2} \mathbf{w}^T \mathbf{w} \right] \tag{6}$$

With the new assumption of $\mathbf{w}$ being an random variable, and the additional information a prior gives we can - instead of as in (i) maximize the likelihood of observing the data - now maximize the posterior distribution of $\mathbf{w}$ given the data $X, \mathbf{y}$, called maximum a posteriori estimation (MAP). The MAP criterion is given by:

$$\hat{\mathbf{w}}_{\mathrm{MAP}} = \arg\max p(\mathbf{w} \mid \mathbf{y}, X) \tag{7}$$

Using Bayes theorem:

$$p(\mathbf{w} \mid \mathbf{y}, X, \beta^{-1}, \alpha^{-1}) = \frac{p(\mathbf{y}|\mathbf{w}, X)p(\mathbf{w})}{p(\mathbf{y})} \propto p(\mathbf{y}|\mathbf{w}, X)p(\mathbf{w})$$

and the same likelihood as in part (i), equation 2, we can rewrite equation 7 as:

$$\hat{\mathbf{w}}_{\text{MAP}} = \arg\max p(\mathbf{y}|\mathbf{w}, X)p(\mathbf{w}) \tag{8}$$

which with the same reasoning as in (i) is solved by setting the derivative of the log of the argument in equation 8 to zero and solving for $\mathbf{w}$:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}} \left[-\log\left(p(\mathbf{y}|\mathbf{w}, X) - \log\left(p(\mathbf{w})\right)\right] &= 0 \iff \\
\frac{\partial}{\partial \mathbf{w}} \left[-\frac{3}{2}\log(\frac{\beta}{2\pi}) + \frac{\beta}{2}\|(y - X\mathbf{w})\|^2 - \log(\frac{\alpha}{2}) + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \right] &= 0 \iff \qquad (9) \\
\alpha\mathbf{w} - \beta X^T(\mathbf{y} - X\mathbf{w}) &= 0 \iff \\
\left(\frac{\beta^{-1}}{\alpha^{-1}}\mathbf{I} + X^T X\right)\mathbf{w} &= X^T\mathbf{y} \implies \\
\hat{\mathbf{w}}_{\text{MAP}} = \left(\frac{\beta^{-1}}{\alpha^{-1}}\mathbf{I} + X^T X\right)^{-1} X^T\mathbf{y} &\iff \qquad (10) \\
\hat{\mathbf{w}}_{\text{MAP}} = (\lambda\mathbf{I} + X^T X)^{-1} X^T\mathbf{y} & \qquad (11)
\end{aligned}$$

Inserting the values from the assignment into this equation we get:

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

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

First, comparing the calculated values of $\hat{\mathbf{w}}_{\text{ML}}$ from (i) and $\hat{\mathbf{w}}_{\text{MAP}}$ in (ii) we clearly see that the norm of the MAP estimate, $\|\hat{\mathbf{w}}_{\text{MAP}}\|^2$ is less than that of the ML estimate, $\|\hat{\mathbf{w}}_{\text{ML}}\|^2$. This aligns with the observation from the lectures that the MAP criterion (ridge regression) adds a quadratic regularization term to the ML criterion constraining the size/norm of $\hat{\mathbf{w}}_{\text{MAP}}$, compare expression 4 to 9, which we with our machine learning googles on recognize as L2 regularization.

Focusing on expression 10 we note that as the variance of $\mathbf{w}$, α^{-1} grows large, the MAP estimator goes towards the ML estimator, which makes sense as increasing variance of $\mathbf{w}$ can be interpreted as decreasing information about $\mathbf{w}$. We see similar but opposite behavior for the variance of our output β^{-1}. We can also conclude that it for these types of tasks is the ratio of β^{-1} and α^{-1} that is of interest, not their absolute values and it can therefore can make sense for some applications to combine them into one parameter λ as in expression 11.

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 part 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.

We aim to develop a probabilistic linear regression model, $g = w^T x + m$, to predict a student's GPA based on the explanatory variable, x. Let $x = [x_{\text{reading}}, x_{\text{sports}}, x_{\text{gaming}}, x_{\text{friends}}]$, where $x_i \in [-1, 1]$ represents the time spent on each activity, $w = [w_{\text{reading}}, w_{\text{sports}}, w_{\text{gaming}}, w_{\text{friends}}]$ a random vector and a random variable. We have also received information from our colleague that it is unlikely that any of these factors contribute more than 10 points individually to the GPA, except for reading, which could explain up to around 20 points. This information will help us estimate the covariance matrix for w.

Assuming that unlikely means more than $\pm 2\sigma$ away from the mean, capturing 95% of all the samples. We set the points (10, or 20) to be equal to 4 times the standard deviation, resulting in a standard deviation of 2.5 for all random variables in w except for reading, which has a standard deviation of 5. This gives us prior knowledge about our model parameters, w. We assume that $w \sim \mathcal{N}(0, \alpha)$, where α is the covariance matrix with diagonal elements corresponding to the prior knowledge about the variance, we assume the w_i's to be independent and therefore α is 0 outside of the diagonal. Specifically,

$$\alpha_{\text{diagonal}} = [\sigma^2_{\text{reading}} = 5^2, \sigma^2_{\text{sports}} = 2.5^2, \sigma^2_{\text{gaming}} = 2.5^2, \sigma^2_{\text{friends}} = 2.5^2].$$

We have also received information that other factors not included in the study and therefore not included in our x are likely to explain at least up to 20 points. To model this we use the random variable m, mentioned above, with a standard deviation of 5. We also know that the average GPA is 200 and therefore set the expected value of m equal to 200, yielding an m that is distributed as $m \sim \mathcal{N}(\mu_m = 200, \sigma_m^2 = 5^2)$.

We have now defined all terms in our probabilistic linear regression model:

$$g = w^T x + m$$

where $w \sim \mathcal{N}(\vec{0}, \alpha)$ and $m \sim \mathcal{N}(200, 5^2)$ Where α is a diagonal matrix with $\alpha_{\text{diagonal}} = [5^2, 2.5^2, 2.5^2, 2.5^2]$. Evaluating the expected value and variance of g we find that our model is distributed as:

$$g \sim \mathcal{N}(\mu_{GPA} = 200, 5^2 + Var[w^T x])$$

Some things to note are that the grade, g, should lie within the interval $[0, 340]$. With this probabilistic model, there is a possibility that our prediction may fall outside this range, but we believe this is very unlikely. Therefore we will not make any major adjustments regarding that. Other than commenting that if it happened it would make a very bad model and that one could solve this by redefining the pdf above as the same pdf times a constant c, if $0 \leq g \leq 340$. But to be 0 if $g < 0$ and 340 if $g > 340$, the constant c is necessary to make it a valid pdf (integrate to 1). This makes the math more annoying though.

Additionally, while we have assumed fixed values for the variances based on a colleagues domain knowledge and the assumption that unlikely means more seldom than 5% of the time, a more realistic approach would be to model the variances as unknown parameters that follow a certain distribution, such as the Gamma distribution. The Gamma distribution would constrain the variances to be positive, and we would set the initial parameters of the distribution so that the expected value of the distribution matches the standard deviations we have used.

We could find $\hat{w}$ (and the variances for the case in the paragraph above) by solving the MAP criterion using the provided data similar to what we did in exercise 1.

(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?

We introduce a new random vector, $x_{\text{gender}} = [x_{\text{male}}, x_{\text{female}}]^T$, which is a one-hot encoded vector for gender. In this representation, the first position corresponds to male, and the second position corresponds to female. For example, if we are predicting for a male, we would have $x_{\text{gender}} = [1, 0]^T$.

We assume that the GPA for males and females follows two different normal distributions, both with an initial mean of 0 and variance of 2.5^2 (same reasoning as before). Using the training data, we will update the mean and variance separately for each gender to better reflect the observed data. The variable x_{gender} is then multiplied by the vector containing the random variables for

males and females:

$$\mathrm{w}_{\text{gender}} = [w_{\text{male}}, w_{\text{female}}]^T \sim [\mathcal{N}(\mu_{\text{male}} = 0, \sigma^2_{\text{male}} = 2.5^2),\ \mathcal{N}(\mu_{\text{female}} = 0, \sigma^2_{\text{female}} = 2.5^2)]^T$$

to incorporate the appropriate distribution for the gender corresponding to the data point in the model. Our new probabilistic model could therefore be written as follows:

$$g = w^T x + \mathrm{w}^T_{\text{gender}} x_{gender} + m$$

Exercise 3

Consider the Bayesian linear regression model:

$$p(y \mid w, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n; w^T x_n, \beta^{-1}), \text{ with the prior: } p(w) = \mathcal{N}(w; m_0, S_0),$$

where β, m_0, and S_0 are known.

(i)

Show that the likelihood can be expressed as a multivariate Gaussian distribution with a diagonal covariance matrix:

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

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

Define:

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

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

Solution: The likelihood can be described as

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

And the Gaussian distribution PDF is defined given($N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$)

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

Which gives that

$$\begin{aligned} p(y \mid w, \beta) &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\beta^{-1}}} \exp\left(-\frac{\beta}{2}(y_n - w^T x_n)^2\right). \\ &= \left(\frac{1}{\sqrt{2\pi\beta^{-1}}}\right)^N \exp\left(-\frac{\beta}{2}\sum_{n=1}^{N}(y_n - w^T x_n)^2\right) \\ &= \frac{1}{\sqrt{(2\pi\beta^{-1})^N}} \exp\left(-\frac{\beta}{2}\sum_{n=1}^{N}(y_n - w^T x_n)^2\right). \end{aligned}$$

Using the definition of $L_2 - norm$ where $\sum(r_i)^2 = ||r||_2^2$ and the residual can be described by $r = y - w^T X$ and where $w^T X = Xw$

$$\sum_{n=1}^{N}(y_n - w^T x_n)^2 = ||y - Xw||_2^2 = (y - Xw)^\top (y - Xw)$$

Resulting in

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

The standard definition of the multivariate Gaussian and its PDF is

$$f(\mathbf{z} : \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} \exp\left(-\frac{\boldsymbol{\Sigma}^{-1}}{2}(\mathbf{z} - \mu)^T(\mathbf{z} - \mu)\right)$$

Where $Z \sim N(\mu, \boldsymbol{\Sigma})$ and where $\boldsymbol{\Sigma} =$ is the covariance matrix. Here the $|\boldsymbol{\Sigma}|^{1/2} = det(\boldsymbol{\Sigma})^{1/2} = \sigma_1 \cdot \sigma_2 \cdot ... \cdot \sigma_n$ as the covariance matrix is diagonal.

Therefore as the variance of y is β^-1

$$\boldsymbol{\Sigma} = \beta^{-1} I_N$$

and

$$|\boldsymbol{\Sigma}| = (\beta^{-1})^N$$

As well as the vector representation of the mean is

$$\mu = Xw$$

Given the standard format, d is the number of independent univariate Gaussian distributions which is equal to the length of X. which in this case is N. Therefore

$$d = N$$

Which gives that

$$\begin{aligned} f(y; Xw, \beta^{-1} I_N) =& \frac{1}{\sqrt{(2\pi\beta^{-1})^N}} \exp\left(-\frac{\beta}{2I_N}(y - Xw)^T(y - Xw)\right) = p(y \mid w, \beta) \\ y \sim& N(y; Xw, \beta^{-1} I_N) \end{aligned}$$

And therefore the

$$p(y \mid w, \beta) = \underbrace{\prod_{n=1}^{N} \mathcal{N}(y_n; w^T x_n, \beta^{-1})}_{\text{Likelihood}} = \underbrace{\mathcal{N}(y; Xw, \beta^{-1} I_N)}_{\text{Gaussian Multivariate}}.$$

(ii)

Verify that the posterior distribution of the parameters w is:

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

where:

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

Solution: To verify that the posterior distribution of w is $\mathcal{N}(w; m_N, S_N)$ Use bayse theorem

$$p(w \mid y) = \frac{p(y \mid w)p(w)}{p(y)} \propto p(y \mid w)p(w)$$

where:

$$p(y \mid w) = \prod_{n=1}^{N} \mathcal{N}(y_n; w^T x_n, \beta^{-1})$$
$$p(w) = \mathcal{N}(w; m_0, S_0)$$

The likelihood is:

$$p(y \mid w) = \frac{1}{(\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{\beta}{2} \sum_{n=1}^{N} (y_n - w^T x_n)^2 \right) \propto \exp\left(-\frac{\beta}{2} \sum_{n=1}^{N} (y_n - w^T x_n)^2 \right),$$

which can be expressed as shown in (i) like

$$p(y \mid w) \propto \exp\left(-\frac{\beta}{2} \|y - Xw\|^2 \right),$$

As the prior is also multivariate Gaussian, the prior is :

$$p(w) \propto \exp\left(-\frac{1}{2} (w - m_0)^T S_0^{-1} (w - m_0) \right).$$

Expand the likelihood term:

$$\|y - Xw\|^2 = y^T y - 2y^T X w + w^T X^T X w.$$

Expand the prior term:

$$-\frac{1}{2}(w - m_0)^T S_0^{-1} (w - m_0) = -\frac{1}{2} w^T S_0^{-1} w + w^T S_0^{-1} m_0 - \frac{1}{2} m_0^T S_0^{-1} m_0.$$

Combining these, the log posterior becomes:

$$\log p(w \mid y) \propto -\frac{\beta}{2} y^T y - \frac{1}{2} m_0^T S_0^{-1} m_0 + w^T (\beta X^T y + S_0^{-1} m_0) - \frac{1}{2} w^T (\beta X^T X + S_0^{-1}) w.$$

As the posterior should take the Gaussian form:

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

where:

- The precision matrix S_N^{-1} is the coefficient of the quadratic term in w:

$$S_N^{-1} = S_0^{-1} + \beta X^T X.$$

- The mean m_N is determined by the linear term:

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

The derivation aligns with the provided expressions for m_N and S_N^{-1}. The posterior is:

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

with:

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

Exercise 4

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

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

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

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

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

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

then the posterior will also be a Gauss-Gamma distribution

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

where,

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

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

Solution

We want to show that the posterior will also be a Gauss-Gamma distribution:

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

If we substitute the mean m_N and variance $\beta^{-1}S_N$ into the multivariate Gaussian, along with the parameters a_N and b_N into the Gamma function, and multiply them together while applying the determinant rule $|cA| = c^n|A|$, where c is a constant, we obtain:

$$p(w, \beta \mid y) \propto \beta^{d/2} \exp\left(-\frac{\beta}{2}(w - m_N)^T s_N^{-1}(w - m_N)\right) \beta^{a_N - 1} \exp(-\beta b_N) \tag{12}$$

after some simplifications.

In Exercise 3, we demonstrated that the likelihood can be expressed as a multivariate Gaussian distribution, which can be simplified as:

$$\begin{aligned}\mathcal{N}(y; Xw, \beta^{-1} I_N) &= \frac{1}{(2\pi)^{N/2}(\beta^{-1})^{N/2}} \exp\left(-\frac{\beta}{2}(y - Xw)^T(y - Xw)\right) \\ &\propto \beta^{N/2} \exp\left(-\frac{\beta}{2}(y - Xw)^T(y - Xw)\right) \end{aligned} \tag{13}$$

From the question, we know that the prior $p(w, \beta)$ follows a Gauss-Gamma distribution, which can be simplified as:

$$\begin{aligned} p(w, \beta) &= \mathcal{N}(w; m_0, \beta^{-1} S_0) Gam(\beta; a_0, b_0) \\ &= \underbrace{\frac{1}{(2\pi)^{d/2}(\beta^{-1} S_0)^{d/2}} \exp\left(-\frac{\beta}{2}(w - m_0)^T S_0^{-1}(w - m_0)\right)}_{\mathcal{N}(w; m_0, \beta^{-1} S_0)} \underbrace{\frac{1}{\Gamma(a_0)} b_0^{a_0} \beta^{a_0 - 1} \exp(-\beta b_0)}_{Gam(\beta; a_0, b_0)} \\ &\propto \beta^{d/2} \exp\left(-\frac{\beta}{2}(w - m_0)^T S_0{}^{-}1(w - m_0)\right) \beta^{a_0 - 1} \exp(-\beta b_0) \\ &= \beta^{d/2 + a_0 - 1} \exp\left(-\frac{\beta}{2}(w - m_0)^T S_0^{-1}(w - m_0) - \beta b_0\right) \end{aligned} \tag{14}$$

By combining our likelihood, (Equation 13), and prior, (Equation 14), we obtain the posterior:

$$\begin{aligned} p(w, \beta | y) &\propto p(y \mid w, \beta) p(w, \beta) \\ &\propto \beta^{d/2 + N/2 + a_0 - 1} \exp\left(-\frac{\beta}{2}\left((y - Xw)^T(y - Xw) + (w - m_0)^T S_0^{-1}(w - m_0)\right) - \beta b_0\right) \end{aligned}$$

Let us now expand the argument of the exponent and use the information provided in the question to simplify it:

$$\begin{aligned} &-\frac{\beta}{2}\left((y - \mathbf{X}w)^T(y - \mathbf{X}w) + (w - m_0)^T S_0^{-1}(w - m_0)\right) - \beta b_0 = \\ &= -\frac{\beta}{2}\left(y^T y - 2w^T X^T y + w^T X^T X w + w^T s_0^{-1} w - 2m_0^T S_0^{-1} w + m_0^T s_0^{-1} m_0\right) - b_0 \beta \\ &= -\frac{\beta}{2}\left(y^T y + m_0^T s_0^{-1} m_0 + w^T \underbrace{\left(X^T X + S_0^{-1}\right)}_{S_N^{-1}} w - 2w^T \underbrace{\left(X^T y + m_0 \left(S_0^{-1}\right)^T\right)}_{m_N S_N^{-1}}\right) - b_0 \beta \\ &= -\frac{\beta}{2}\left(y^T y + m_0^T s_0^{-1} m_0 + w^T S_N^{-1} w - 2w^T S_N^{-1} m_N\right) - b_0 \beta \\ &= -\frac{\beta}{2}\left(y^T y + m_0^T s_0^{-1} m_0 + w^T S_N^{-1} w - 2w^T S_N^{-1} m_N + \underbrace{m_N^T S_N^{-1} m_N - m_N^T S_N^{-1} m_N}_{=0}\right) - b_0 \beta \\ &= -\beta \underbrace{\left(\frac{1}{2}\left(y^T y + m_0^T s_0^{-1} m_0 - m_N^T S_N^{-1} m_N - b_0\right)\right)}_{b_N} - \frac{\beta}{2}\left((w - m_N)^T S_N^{-1}(w - m_N)\right) \\ &= -\beta b_N - \frac{\beta}{2}\left((w - m_N)^T S_N^{-1}(w - m_N)\right) \end{aligned}$$

By inserting everything into the exponent and separating the terms, we obtain the posterior distribution.

$$
\begin{aligned}
p(w, \beta|y) &\propto \beta^{d/2} \underbrace{\exp\left(-\frac{\beta}{2}\left(w - m_N\right)^T s_N^{-1}\left(w - m_N\right)\right)}_{\mathcal{N}(w; m_0, \beta^{-1} S_0)} \underbrace{\beta^{a_N - 1} \exp\left(-\beta b_N\right)}_{\mathrm{Gam}(\beta; a, b)} \\
&= \mathcal{N}(w; m_0, \beta^{-1} S_0)\mathrm{Gam}(\beta; a, b)
\end{aligned}
$$

We have now shown that if the prior is Gauss-Gamma then the posterior is also Gauss-Gamma which means that we have shown that Gauss-Gamma prior is a conjugate prior to the Gaussian likelihood. We have achieved what we set out to show in Equation 12.