Exercises 2

Bayesian linear regression

Exercise 2.1 Continuous random variables
Consider a continuous random variable X with mean $\mathbb{E}[X] = \mu$ and variance $\mathrm{Var}[X] = \sigma^2$. Consider also the transformation $Y = aX + b$. Use the definition of the mean and the variance to compute $\mathbb{E}[Y]$ and $\mathrm{Var}[Y]$.

Hint: For a scalar continuous random variable X the following relation holds

$$\mathbb{E}[f(X)] = \int_{-\infty}^{\infty} f(x)p(x)dx$$

where $f(X)$ is a function of X and $p(x)$ is the density function for X. Moreover, the variance is defined as

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sigma^2.$$

Exercise 2.2 Gaussian random variables
Consider a Gaussian random variable X with

$$X \sim \mathcal{N}\left(x;\ \mu,\ \sigma^2\right), \qquad \text{where} \qquad \mathcal{N}\left(x;\ \mu,\ \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

(a) Show that the mean of X is μ.

(b) Show that the variance of X is σ^2. *Hint:* $\int_{z=0}^{\infty} z^2 e^{-z^2} dz = \frac{\sqrt{\pi}}{4}$

(c) Assume that another random variable Y is related to X as $Y = aX + b$ where $a,\ b \in \mathbb{R}$. What is the density function of Y? *Hint: Use the cumulative distribution function of Y.*

Exercise 2.3 The Gaussian distribution
Consider two independent random variables $X \sim p_X(x)$ and $Y \sim p_Y(y)$ and their sum $Z = X + Y$. Then, the probability density function of $Z \sim p_Z(z)$ is

$$p_Z(z) = \int_{-\infty}^{\infty} p_Y(z - x)p_X(x)dx. \tag{4}$$

If X and Y are Gaussian random variables with $p_X(x) = \mathcal{N}\left(x;\ \mu_X,\ \sigma_X^2\right)$ and $p_Y(y) = \mathcal{N}\left(y;\ \mu_Y,\ \sigma_Y^2\right)$ we get that

$$p_Z(z) = \int_{-\infty}^{\infty} \mathcal{N}\left(z - x;\ \mu_Y,\ \sigma_Y^2\right) \mathcal{N}\left(x;\ \mu_X,\ \sigma_X^2\right) dx = \mathcal{N}\left(z;\ \mu_X + \mu_Y,\ \sigma_X^2 + \sigma_Y^2\right) \tag{5}$$

(a) Assume that $p_X(x) = \mathcal{N}\left(x;\ 0,\ 1^2\right)$ and $p_Y(y) = \mathcal{N}\left(y;\ 2,\ 2^2\right)$. What is $p_Z(z)$?

(b) Assume the same expressions for $p_X(x)$ and $p_Y(y)$ as in (a). In addition, we receive the observation $X = 1$. What is the conditional probability density function of Z under this observation, i.e. what is $p_{Z|X}(z|X = 1)$?

(c) Assume the same expressions for $p_X(x)$ and $p_Y(y)$ as in (a). In addition, we receive the observation $Z = 1$. What is the conditional probability density function of X under this observation, i.e. what is $p_{X|Z}(x|Z = 1)$? *Hint: Use Bayes' theorem* $p(x|Z = z) = \frac{p(z|x)p(x)}{p(z)}$.

(d) Extra: Prove (4).

(e) Extra: Prove (5).

Exercise 2.4 Conditional Gauss (scalar)
Consider the joint probability density function

$$p\left(x_a,\ x_b\right) = \mathcal{N}\left(\mathbf{x};\ \boldsymbol{\mu},\ \boldsymbol{\Sigma}\right)$$

where

$$\mathbf{x} = \begin{bmatrix} x_a \\ x_b \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{aa} & \sigma_{ab} \\ \sigma_{ab} & \sigma_{bb} \end{bmatrix}$$

where x_a and x_b are scalars.

(a) Show that

$$p(x_a) = \mathcal{N}\left(x_a;\ \mu_a,\ \sigma_{aa}\right). \tag{7}$$

(b) Use result in (7) and show that

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

where

$$\mu_{a|b} = \mu_a + \frac{\sigma_{ab}}{\sigma_{bb}}(x_b - \mu_b), \qquad \sigma_{a|b} = \sigma_{aa} - \frac{\sigma_{ab}^2}{\sigma_{bb}}$$

Exercise 2.5 Conditional Gauss (multivariate)
Consider the joint probability density function

$$p\left(\mathbf{x}_a,\ \mathbf{x}_b\right) = \mathcal{N}\left(\mathbf{x};\ \boldsymbol{\mu},\ \boldsymbol{\Sigma}\right)$$

where

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix}, \qquad \mu = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ab}^{\mathsf{T}} & \boldsymbol{\Sigma}_{bb} \end{bmatrix}$$

where $\mathbf{x}_a$ and $\mathbf{x}_b$ are vectors.

(a) Show that

$$p(\mathbf{x}_a) = \mathcal{N}\left(\mathbf{x};\ \boldsymbol{\mu}_a,\ \boldsymbol{\Sigma}_{aa}\right). \tag{10}$$

(b) Use result in (10) and show that

$$p\left(\mathbf{x}_a|\ \mathbf{x}_b\right) = \mathcal{N}\left(\mathbf{x}_a;\ \boldsymbol{\mu}_{a|b},\ \boldsymbol{\Sigma}_{a|b}\right)$$

where

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b), \qquad \boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ab}^{\mathsf{T}}$$

Exercise 2.6 Marginal and Conditional Gauss (multivariate)
Consider a marginal Gaussian distribution for $\mathbf{x}_b$ and a conditional Gaussian distribution for $\mathbf{x}_a$ given $\mathbf{x}_b$

$$p\left(\mathbf{x}_a\right) = \mathcal{N}\left(\mathbf{x}_a;\ \boldsymbol{\mu}_a,\ \boldsymbol{\Sigma}_{aa}\right)$$
$$p\left(\mathbf{x}_b|\ \mathbf{x}_a\right) = \mathcal{N}\left(\mathbf{x}_b;\ \mathbf{A}\mathbf{x}_a + \mathbf{b},\ \boldsymbol{\Sigma}_{b|a}\right)$$

where $\mathbf{x}_a$ and $\mathbf{x}_b$ are vectors.

(a) Show that

$$p(\mathbf{x}_b) = \mathcal{N}\left(\mathbf{x};\ \boldsymbol{\mu}_b,\ \boldsymbol{\Sigma}_{bb}\right).$$

where

$$\begin{aligned}\boldsymbol{\mu}_b &= \mathbf{A}\boldsymbol{\mu}_a + \mathbf{b}\\ \boldsymbol{\Sigma}_{bb} &= \boldsymbol{\Sigma}_{b|a} + \mathbf{A}\boldsymbol{\Sigma}_{aa}\mathbf{A}^\mathsf{T}\end{aligned}$$

(b) Show that

$$p\left(\mathbf{x}_a|\ \mathbf{x}_b\right) = \mathcal{N}\left(\mathbf{x}_a;\ \boldsymbol{\mu}_{a|b},\ \boldsymbol{\Sigma}_{a|b}\right)$$

where

$$\begin{aligned}\boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\Big(\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{b|a}^{-1}(\mathbf{x}_b - \mathbf{b}) + \boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\mu}_a\Big) = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{aa}\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \mathbf{b} - \mathbf{A}\boldsymbol{\mu}_a)\\ \boldsymbol{\Sigma}_{a|b} &= \left(\boldsymbol{\Sigma}_{aa}^{-1} + \mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{b|a}^{-1}\mathbf{A}\right)^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{aa}\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}_{bb}^{-1}\mathbf{A}\boldsymbol{\Sigma}_{aa}\end{aligned}$$

Exercise 2.7 (adapted from [2])
Consider the Bayesian linear regression model

$$p(\mathbf{y}|\mathbf{w},\beta) = \prod_{n=1}^{N}\mathcal{N}\left(y_n;\ \mathbf{w}^\mathsf{T}\mathbf{x}_n,\ \beta^{-1}\right) \quad \text{with the prior} \quad p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w};\ \mathbf{m}_0,\ \mathbf{S}_0\right)$$

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

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

$$p(\mathbf{y}|\mathbf{w},\beta) = \prod_{n=1}^{N}\mathcal{N}\left(y_n;\ \mathbf{w}^\mathsf{T}\mathbf{x}_n,\ \beta^{-1}\right) = \mathcal{N}\left(\mathbf{y};\ \mathbf{X}\mathbf{w},\ \beta^{-1}\mathbf{I}_N\right)$$

where $\mathbf{I}_N$ is the identity matrix of size $N \times N$ and where

$$\mathbf{X} = \begin{bmatrix}\mathbf{x}_1^\mathsf{T}\\ \vdots\\ \mathbf{x}_N^\mathsf{T}\end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix}y_1\\ \vdots\\ y_N\end{bmatrix}.$$

b) Use the result in Exercise 2.6(b) to verify that the posterior distribution of the parameters $\mathbf{w}$ is

$$p(\mathbf{w}|\mathbf{y}) = \mathcal{N}\left(\mathbf{w};\ \mathbf{m}_N,\ \mathbf{S}_N\right)$$

where

$$\begin{aligned}\mathbf{m}_N &= \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{X}^\mathsf{T}\mathbf{y}),\\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta\mathbf{X}^\mathsf{T}\mathbf{X},\end{aligned} \tag{15}$$

Exercise 2.8 Bayesian linear regression
We have made the 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 $y = w_1x_1 + w_2x_2 + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0,5)$.

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

(b) Now assume the prior

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

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

(c) Compare the results from (a) and (b).

Exercise 2.9 Bayesian linear regression example
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.

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

(b) 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 2.10 Regularization and priors
Consider the following Gaussian data distribution

$$p(y_i|\mathbf{w}) = \mathcal{N}\left(y_i;\ \mathbf{x}_i^\mathsf{T}\mathbf{w},\ \sigma^2\right)$$

We are interested in a so-called *maximum a posteriori estimate* of $\mathbf{w}$,

$$\widehat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{y}).$$

Hint: Remember the definition $p(\mathbf{w}|\mathbf{y}) = \frac{\prod_{i=1}^N p(y_i|\mathbf{w})}{\prod_{i=1}^N p(y_i)} p(\mathbf{w})$*, which implies that* $\log p(\mathbf{w}|\mathbf{y}) = \sum_{i=1}^N \log p(y_i|\mathbf{w}) + \log p(\mathbf{w})$

Show that $\widehat{\mathbf{w}}$ is ...

(a) ... the same as the solution to the least square problem

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{ \sum_{i=1}^N (y_i - \mathbf{x}_i^\mathsf{T}\mathbf{w})^2 \right\}$$

if we choose an uninformative prior

$$p(\mathbf{w}) \propto 1$$

(b) ... the solution to linear regression with ridge regression

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{ \sum_{i=1}^N (y_i - \mathbf{x}_i^\mathsf{T}\mathbf{w})^2 + \lambda \sum_{j=1}^p |w_j|^2 \right\} \tag{18a}$$

if we choose a Gaussian prior

$$p(\mathbf{w}) = \prod_{j=1}^p p(w_j) = \prod_{j=1}^p \mathcal{N}\left(w_j;\ 0,\ \alpha^2\right)$$

What is is the value of α?

(c) ... the solution to linear regression with LASSO

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{ \sum_{i=1}^{N} (y_i - x_i^{\mathsf{T}}\mathbf{w})^2 + \lambda \sum_{j=1}^{p} |w_j| \right\} \tag{19a}$$

if we choose a Laplacian prior

$$p(\mathbf{w}) = \prod_{j=1}^{p} p(w_j) = \prod_{j=1}^{p} \mathcal{L}(w_j \mid 0,\ \alpha)$$

(Hint: the Laplace distribution has the density function $\mathcal{L}(x \mid \mu,\ \alpha) = \frac{1}{2\alpha} \exp\left(-\frac{|x-\mu|}{\alpha}\right)$*).*

What is the value of α?

Exercise 2.11 (adapted from [2])
In Exercise (2.7) 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 $\mathbf{w}$ and β and solve

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

Show that if we consider the likelihood $p(\mathbf{y}|\mathbf{w}, \beta)$ in Exercise (2.7) and the following *Gauss-Gamma prior*

$$p(\mathbf{w}, \beta) = \mathcal{N}\left(\mathbf{w};\ \mathbf{m}_0,\ \beta^{-1}\mathbf{S}_0\right) \text{Gam}\left(\beta; a_0,\ b_0\right),$$

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

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

then the posterior will also be a Gauss-Gamma distribution

$$p(\mathbf{w}, \beta|\mathbf{y}) = \mathcal{N}\left(\mathbf{w};\ \mathbf{m}_N,\ \beta^{-1}\mathbf{S}_N\right) \text{Gam}\left(\beta; a_N,\ b_N\right),$$

where

$$\begin{aligned}
\mathbf{m}_N &= \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{X}^{\mathsf{T}}\mathbf{y}), \\
\mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \mathbf{X}^{\mathsf{T}}\mathbf{X}, \\
a_N &= a_0 + \frac{N}{2}, \\
b_N &= b_0 + \frac{1}{2}\left(\mathbf{m}_0^{\mathsf{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 - \mathbf{m}_N^{\mathsf{T}}\mathbf{S}_N^{-1}\mathbf{m}_N + \sum_{n=1}^{N} y_n^2\right).
\end{aligned}$$

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

Solutions 2

Bayesian linear regression

Solution to Exercise 2.1
Denote the probability distribution of X with $p(x)$. Then for the mean we have

$$\mathbb{E}[Y] = \mathbb{E}[aX + b] = \int (ax + b)p(x)dx = a \underbrace{\int xp(x)dx}_{=\mathbb{E}[X]=\mu} + b \underbrace{\int p(x)dx}_{1} = a\mu + b.$$

Consequently, the mean operator is a linear operator. (The derivation is similar for the multivariate case.)
For the variance we have

$$\mathrm{Var}[Y] = \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \mathbb{E}[(aX + b - a\mu - b)^2] = \mathbb{E}[a^2(X - \mu)^2] = a^2 \underbrace{\int_{x=-\infty}^{\infty} (x - \mu)^2 p(x)dx}_{\mathbb{E}[(X-\mathbb{E}[X])^2]} = a^2\mathrm{Var}[X] = a^2\sigma^2.$$

Solution to Exercise 2.2 (a) The definition of the mean and the assumption that $X \sim \mathcal{N}\left(x;\ \mu,\ \sigma^2\right)$ gives

$$\begin{aligned}
\mathbb{E}[X] &= \int_{x=-\infty}^{\infty} x\mathcal{N}\left(x;\ \mu,\ \sigma^2\right) dx \\
&= \int_{x=-\infty}^{\infty} x \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\
&= /z = x - \mu/ = \int_{z=-\infty}^{\infty} (z + \mu) \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{z^2}{2\sigma^2}} dz \\
&= \underbrace{\int_{z=-\infty}^{\infty} z \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{z^2}{2\sigma^2}} dz}_{=0 \text{ (odd function)}} + \mu \underbrace{\int_{z=-\infty}^{\infty} \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{z^2}{2\sigma^2}} dz}_{=1 \text{ (integral of a pdf)}} = \mu
\end{aligned}$$

(b) The variance of X is

$$\begin{aligned}
\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] &= \int_{x=-\infty}^{\infty} (x - \mu)^2 \mathcal{N}\left(x;\ \mu,\ \sigma^2\right) dx \\
&= \int_{x=-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\
&= /z = \frac{x - \mu}{\sqrt{2\sigma^2}}/ = \int_{z=-\infty}^{\infty} 2z^2\sigma^2 \frac{1}{\sqrt{2\sigma^2\pi}} e^{-z^2} \sqrt{2\sigma^2} dz \\
&= \frac{2\sigma^2}{\sqrt{\pi}} \int_{z=-\infty}^{\infty} z^2 e^{-z^2} dz = \frac{4\sigma^2}{\sqrt{\pi}} \underbrace{\int_{z=0}^{\infty} z^2 e^{-z^2} dz}_{=\sqrt{\pi}/4} = \sigma^2
\end{aligned}$$

(c) Let $F_Y(t)$ be the cumulative distribution function of Y. Then

$$F_Y(t) = \int_{y=-\infty}^{t} p(y)dy = \Pr[Y \leq t] = \Pr[aX + b \leq t] = \Pr\left[X \leq \frac{t-b}{a}\right] = \int_{x=-\infty}^{\frac{t-b}{a}} \mathcal{N}\left(x;\ \mu,\ \sigma^2\right) dx$$
$$= /y = ax + b/ = \int_{y=-\infty}^{t} \frac{1}{a}\mathcal{N}\left(\frac{y-b}{a};\ \mu,\ \sigma^2\right) dy$$

Further, we see that

$$\frac{1}{a}\mathcal{N}\left(\frac{y-b}{a};\ \mu,\ \sigma^2\right) = \frac{1}{a}\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{1}{2\sigma^2}\left(\frac{y-b}{a}-\mu\right)^2} = \frac{1}{\sqrt{2a^2\sigma^2\pi}} e^{-\frac{1}{2a^2\sigma^2}(y-b-a\mu)^2} = \mathcal{N}\left(y;\ a\mu+b,\ a^2\sigma^2\right)$$

From this follows that

$$\int_{y=-\infty}^{t} p(y)dy = \int_{y=-\infty}^{t} \mathcal{N}\left(y;\ a\mu+b,\ a^2\sigma^2\right) dy \quad \Rightarrow \quad p(y) = \mathcal{N}\left(y;\ a\mu+b,\ a^2\sigma^2\right).$$

Consequently, Y is also a Gaussian random variable with the mean $a\mu + b$ and the variance $a^2\sigma^2$.

Solution to Exercise 2.3 (a) According to (5) we have $p_Z(z) = \mathcal{N}\left(z;\ \mu_X + \mu_Y,\ \sigma_X^2 + \sigma_Y^2\right) = \mathcal{N}\left(z;\ 0+2,\ 1^2+2^2\right) = \mathcal{N}\left(z;\ 2,\ \sqrt{5}^2\right)$

(b) Now we have $Z = 1 + Y$ which gives $p_Z(z|X=1) = \mathcal{N}\left(z;\ 1+\mu_Y,\ \sigma_Y^2\right) = \mathcal{N}\left(z;\ 1+2,\ 2^2\right) = \mathcal{N}\left(z;\ 3,\ 2^2\right)$. You can see this by (i) either using $\mu_X = 1$ and $\sigma_X^2 = 0$ in (5), or (ii) using the relation derived in Exercise 2.1.

(c) Use Bayes' theorem

$$p(x|Z=z) = \frac{p(z|x)p(x)}{p(z)} = \frac{\mathcal{N}\left(z;\ x+\mu_Y,\ \sigma_Y^2\right)\mathcal{N}\left(x;\ \mu_X,\ \sigma_X^2\right)}{\mathcal{N}\left(z;\ \mu_X+\mu_Y,\ \sigma_X^2+\sigma_Y^2\right)} = \frac{\mathcal{N}\left(z;\ x+2,\ 2^2\right)\mathcal{N}\left(x;\ 0,\ 1^2\right)}{\mathcal{N}\left(z;\ 2,\ \sqrt{5}^2\right)}$$
$$= \frac{\frac{1}{\sqrt{2\pi 2^2}} e^{-\frac{1}{2}\frac{(z-(x+2))^2}{2^2}} \frac{1}{\sqrt{2\pi 1^2}} e^{-\frac{1}{2}\frac{x^2}{1^2}}}{\frac{1}{\sqrt{2\pi\sqrt{5}^2}} e^{-\frac{1}{2}\frac{(z-2)^2}{\sqrt{5}^2}}}$$
$$= \frac{1}{\sqrt{2\pi(4/5)}} e^{-\frac{1}{2}\left(\frac{(z-(x+2))^2}{2^2}+\frac{x^2}{1^2}-\frac{(z-2)^2}{\sqrt{5}^2}\right)}$$

Using $z = 1$ gives

$$p(x|Z=1) = \frac{1}{\sqrt{2\pi(4/5)}} e^{-\frac{1}{2}\left(\frac{x^2+2x+1}{2^2}+\frac{x^2}{1^2}-\frac{1}{\sqrt{5}^2}\right)}$$

Now, complete the squares of the exponent

$$\frac{x^2+2x+1}{4} + \frac{x^2}{1} - \frac{1}{5} = \frac{x^2+2x+1+4x^2-4/5}{4}$$
$$= \frac{5x^2+2x+1/5}{4}$$
$$= \frac{5(x^2+2/5x+1/25)}{4}$$
$$= \frac{5(x+1/5)^2}{4}$$
$$= \frac{(x-(-1/5))^2}{4/5}.$$

Consequently, we have that

$$p(x|Z=1) = \frac{1}{\sqrt{2\pi(4/5)}} e^{-\frac{(x-(-1/5))^2}{(4/5)}} = \mathcal{N}\left(x;\ -\frac{1}{5},\ \left(\frac{2}{\sqrt{5}}\right)^2\right)$$

(d) Starting with the joint probability density function of X and Z $p_{X,Z}(x,z)$ and using the definition of conditioning $p_{Z|X}(z|X=x) = \frac{p_{X,Z}(x,z)}{p_X(x)}$ we get

$$\begin{aligned} p_Z(z) &= \int_{-\infty}^{\infty} p_{X,Z}(x,z)dx \\ &= \int_{-\infty}^{\infty} p_{Z|X}(z|X=x)p_X(x)dx \\ &= \int_{-\infty}^{\infty} p_Y(z-x)p_X(x)dx \end{aligned}$$

(e) -

Solution to Exercise 2.4 a) -

b) We start with the definition of the conditioning

$$p\left(x_a|\ x_b\right) = \frac{p\left(x_a,\ x_b\right)}{p\left(x_b\right)} \Rightarrow \log p\left(x_a|\ x_b\right) = \log p\left(x_a,\ x_b\right) - \log p\left(x_b\right)$$

From the exercise we also have that

$$\log p\left(x_a|\ x_b\right) = -\frac{1}{2\sigma_{a|b}}(x_a - \mu_{a|b})^2 + const. \tag{29a}$$

$$\log p\left(x_b\right) = -\frac{1}{2\sigma_{bb}}(x_b - \mu_b)^2 + const. \tag{29b}$$

$$\begin{aligned} \log p\left(x_a,\ x_b\right) &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) + const. = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}) + const. \\ &= -\frac{\lambda_a}{2}(x_a-\mu_a)^2 - \lambda_{ab}(x_a-\mu_a)(x_b-\mu_b) - \frac{\lambda_{bb}}{2}(x_b-\mu_b)^2 + const. \end{aligned} \tag{29c}$$

where we have used the precision matrix instead of the covariance matrix

$$\begin{aligned} \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} \Rightarrow \begin{bmatrix} \lambda_{aa} & \lambda_{ab} \\ \lambda_{ab} & \lambda_{bb} \end{bmatrix} &= \begin{bmatrix} \sigma_{aa} & \sigma_{ab} \\ \sigma_{ab} & \sigma_{bb} \end{bmatrix}^{-1} = \frac{1}{\sigma_{aa}\sigma_{bb} - \sigma_{ab}^2} \begin{bmatrix} \sigma_{bb} & -\sigma_{ab} \\ -\sigma_{ab} & \sigma_{aa} \end{bmatrix}^{-1} \\ \Rightarrow \frac{1}{\lambda_{aa}} = \sigma_{aa} - \frac{\sigma_{ab}^2}{\sigma_{bb}}, \qquad \frac{1}{\lambda_{bb}} &= \sigma_{bb} - \frac{\sigma_{ab}^2}{\sigma_{aa}} \qquad \frac{1}{\lambda_{ab}} = \sigma_{ab} - \frac{\sigma_{aa}\sigma_{bb}}{\sigma_{ab}} \end{aligned} \tag{30a}$$

Combining (29b) and (29c) gives

$$\begin{aligned} E \triangleq \log p\left(x_a,\ x_b\right) - \log p\left(x_b\right) = &-\frac{\lambda_a}{2}(x_a-\mu_a)^2 - \lambda_{ab}(x_a-\mu_a)(x_b-\mu_b) \\ &- \frac{\lambda_{bb}}{2}(x_b-\mu_b)^2 + \frac{1}{2\sigma_{bb}}(x_b-\mu_b)^2 + const. \end{aligned}$$

To get the form (29a) we first expand the parenthesis and collect the quadratic and the linear terms of x_a, and then complete the squares with respect to x_a.

$$\begin{aligned} E &= -\frac{\lambda_a}{2}x_a^2 + \Big(\lambda_a\mu_a - \lambda_{ab}(x_b-\mu_b)\Big)x_a + const. \\ &= -\frac{\lambda_a}{2}\left(x_a - \left(\mu_a - \frac{\lambda_{ab}}{\lambda_a}(x_b-\mu_b)\right)\right)^2 + const. \end{aligned}$$

which by comparing with (29a) gives that

$$\begin{aligned} \mu_{a|b} &= \mu_a - \frac{\lambda_{ab}}{\lambda_a}(x_b-\mu_b) = \mu_a - \frac{\sigma_{aa} - \frac{\sigma_{ab}^2}{\sigma_{bb}}}{\sigma_{ab} - \frac{\sigma_{aa}\sigma_{bb}}{\sigma_{ab}}}(x_b-\mu_b) = \mu_a + \frac{\sigma_{ab}}{\sigma_{bb}}(x_b-\mu_b) \\ \sigma_{a|b} &= \frac{1}{\lambda_{aa}} = \sigma_{aa} - \frac{\sigma_{ab}^2}{\sigma_{bb}} \end{aligned}$$

where the identities in (30a) have been used.

Solution to Exercise 2.5 a) -

b) We start with the definition of the conditioning

$$p\left(\mathbf{x}_a|\ \mathbf{x}_b\right) = \frac{p\left(\mathbf{x}_a,\ \mathbf{x}_b\right)}{p\left(\mathbf{x}_b\right)} \Rightarrow \log p\left(\mathbf{x}_a|\ \mathbf{x}_b\right) = \log p\left(\mathbf{x}_a,\ \mathbf{x}_b\right) - \log p\left(\mathbf{x}_b\right)$$

From the exercise we also have that

$$\log p\left(\mathbf{x}_a|\ \mathbf{x}_b\right) = -\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_{a|b})^\mathsf{T}\boldsymbol{\Sigma}_{a|b}^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_{a|b}) + const. \tag{32a}$$

$$\log p\left(\mathbf{x}_b\right) = -\frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^\mathsf{T}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) + const. \tag{32b}$$

$$\begin{aligned}\log p\left(\mathbf{x}_a,\ \mathbf{x}_b\right) &= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) + const. \\ &= -\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\mathsf{T}\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a - \boldsymbol{\mu}_a) - (\mathbf{x}_a - \boldsymbol{\mu}_a)^\mathsf{T}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\ &\quad -\frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^\mathsf{T}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b - \boldsymbol{\mu}_b) + const.\end{aligned} \tag{32c}$$

where we have used the precision matrix instead of the covariance matrix

$$\begin{aligned}\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} \Rightarrow \begin{bmatrix}\boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ab} & \boldsymbol{\Lambda}_{bb}\end{bmatrix} &= \begin{bmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ab} & \boldsymbol{\Sigma}_{bb}\end{bmatrix}^{-1} \\ \Rightarrow \quad \boldsymbol{\Lambda}_{aa} &= (\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ab}^\mathsf{T})^{-1}, \qquad \boldsymbol{\Lambda}_{bb} = (\boldsymbol{\Sigma}_{bb} - \boldsymbol{\Sigma}_{ab}^\mathsf{T}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab})^{-1}, \qquad (33a) \\ \boldsymbol{\Lambda}_{ab} &= -(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ab}^\mathsf{T})^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1} \qquad (33b)\end{aligned}$$

Combining (32b) and (32c) gives

$$\begin{aligned}E \triangleq \log p\left(\mathbf{x}_a,\ \mathbf{x}_b\right) - \log p\left(\mathbf{x}_b\right) &= -\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\mathsf{T}\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a - \boldsymbol{\mu}_a) - (\mathbf{x}_a - \boldsymbol{\mu}_a)^\mathsf{T}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\ &\quad -\frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^\mathsf{T}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b - \boldsymbol{\mu}_b) + \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^\mathsf{T}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) + const.\end{aligned}$$

To get the form (32a) we first expand the parenthesis and collect the quadratic and the linear terms of $\mathbf{x}_a$, and then complete the squares with respect to $\mathbf{x}_a$.

$$\begin{aligned}E &= -\frac{1}{2}\mathbf{x}_a^\mathsf{T}\boldsymbol{\Lambda}_{aa}\mathbf{x}_a + \mathbf{x}_a^\mathsf{T}\Big(\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\Big) + const. \\ &= -\frac{1}{2}\left(\mathbf{x}_a - \Big(\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\Big)\right)^\mathsf{T}\boldsymbol{\Lambda}_{aa}\left(\mathbf{x}_a - \Big(\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\Big)\right) + const.\end{aligned}$$

which by comparing with (32a) and using the identities (33a) and (33b) gives that

$$\begin{aligned}\boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ab}^\mathsf{T}\end{aligned}$$

Solution to Exercise 2.7 a) The likelihood can be written as

$$\begin{aligned}\mathcal{N}\left(\mathbf{y};\ \mathbf{X}\mathbf{w},\ \mathbf{I}_N\beta^{-1}\right) &= \frac{1}{(2\pi)^{N/2}\sqrt{\det \mathbf{I}_N\beta^{-1}}}\exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{I}_N\beta^{-1})^{-1}(\mathbf{y} - \mathbf{X}\mathbf{w})\right) \\ &= \frac{1}{(2\pi)^{N/2}\sqrt{\beta^{-N}}}\exp\left(-\frac{1}{2\beta^{-1}}(\mathbf{y} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{w})\right) \\ &= \frac{1}{(2\pi)^{N/2}\sqrt{\beta^{-N}}}\exp\left(-\frac{1}{2\beta^{-1}}\sum_{n=1}^{N}(y - \mathbf{x}_n^\mathsf{T}\mathbf{w})^2\right) \\ &= \prod_{n=1}^{N}\frac{1}{\sqrt{2\pi\beta^{-1}}}\exp\left(-\frac{1}{2\beta^{-1}}(y - \mathbf{x}_n^\mathsf{T}\mathbf{w})^2\right) \\ &= \prod_{n=1}^{N}\mathcal{N}\left(y_n;\ \mathbf{w}^\mathsf{T}\mathbf{x}_n,\ \beta^{-1}\right)\end{aligned}$$

b) Together with the prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w};\ \mathbf{m}_0,\ \mathbf{S}_0)$ we can by comparing with Exercise 2.6 identify

$$\begin{aligned} \mathbf{x}_a &= \mathbf{w}, \qquad \boldsymbol{\mu}_a = \mathbf{m}_0, \qquad \boldsymbol{\Sigma}_{aa} = \mathbf{S}_0 \\ \mathbf{x}_b &= \mathbf{y}, \qquad \mathbf{A} = \mathbf{X}, \qquad \boldsymbol{\Sigma}_{b|a} = \beta^{-1}\mathbf{I}_N, \\ \boldsymbol{\mu}_{a|b} &= \mathbf{m}_N, \qquad \boldsymbol{\Sigma}_{a|b} = \mathbf{S}_N \end{aligned}$$

This inserted in equations in Exercise 2.6(b) gives

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N(\beta\mathbf{X}^\mathsf{T}\mathbf{y} + \mathbf{S}_0^{-1}\mathbf{m}_0), \\ \mathbf{S}_N &= (\mathbf{S}_0^{-1} + \beta\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}. \end{aligned}$$

Solution to Exercise 2.8 a) Since the assumption is that ε is Gaussian distributed, $\widehat{\mathbf{w}}$ is found using least squares

$$\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = \left(\begin{pmatrix} 3 & 4 & 2 \\ -1 & 2 & 1 \end{pmatrix}\begin{pmatrix} 3 & -1 \\ 4 & 2 \\ 2 & 1 \end{pmatrix}\right)^{-1}\begin{pmatrix} 3 & 4 & 2 \\ -1 & 2 & 1 \end{pmatrix}\begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} = \cdots = \frac{1}{25}\begin{pmatrix} 13 \\ -11 \end{pmatrix} = \begin{pmatrix} 0.52 \\ -0.44 \end{pmatrix}$$

b) By using equation (15), we get

$$\begin{aligned} \mathbf{S}_N = (\mathbf{S}_0^{-1} + \beta\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} &= \left(\begin{pmatrix} 1/5 & 0 \\ 0 & 1/5 \end{pmatrix}^{-1} + \frac{1}{5}\begin{pmatrix} 3 & 4 & 2 \\ -1 & 2 & 1 \end{pmatrix}\begin{pmatrix} 3 & -1 \\ 4 & 2 \\ 2 & 1 \end{pmatrix}\right)^{-1} = \dots \\ &= \frac{1}{325}\begin{pmatrix} 31 & -7 \\ -7 & 54 \end{pmatrix} \approx \begin{pmatrix} 0.10 & -0.02 \\ -0.02 & 0.17 \end{pmatrix} \end{aligned}$$

and

$$\begin{aligned} \mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{X}^\mathsf{T}\mathbf{y}) &= \mathbf{S}_N\left(\begin{pmatrix} 1/5 & 0 \\ 0 & 1/5 \end{pmatrix}^{-1}\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \frac{1}{5}\begin{pmatrix} 3 & 4 & 2 \\ -1 & 2 & 1 \end{pmatrix}\begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}\right) = \dots \\ &= \frac{1}{325}\begin{pmatrix} 73 \\ -6 \end{pmatrix} \approx \begin{pmatrix} 0.22 \\ -0.02 \end{pmatrix} \end{aligned}$$

Thus, the answer is

$$p(\mathbf{w}\,|\,\mathbf{y}) = \mathcal{N}\left(\mathbf{w}\,|\,\begin{pmatrix} 0.22 \\ -0.02 \end{pmatrix}, \begin{pmatrix} 0.10 & -0.02 \\ -0.02 & 0.17 \end{pmatrix}\right).$$

c) The most apparent difference is that the maximum likelihood solution is a number, whereas the probabilistic solution is a distribution. Further, the mean of the probabilistic solution is smaller than the maximum likelihood estimate, which is due to the fact that the posterior is also influenced by the (rather narrow) prior (akin to regularization).

Solution to Exercise 2.9 (a) We list all variables involved:

- y: merit-value
- x_1: (normalized) time spent on reading books and comics
- x_2: (normalized) time spent on gaming
- x_3: (normalized) time spent on sports activities
- x_4: (normalized) time spent with friends

Putting this together in a probabilistic linear regression model yields

$$y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + \varepsilon$$

where our colleague also has provided the following information

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

(other interpretations of the text are also possible)

(b) Including also the gender (which is a binary variable) can be done as

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + w_6 x_6 + \varepsilon$$

where $x_5 = 1$ and $x_6 = 0$ or vice versa, depending on the student's gender, and $w_5 \sim \mathcal{N}(0, 10^2)$ and $w_6 \sim \mathcal{N}(0, 10^2)$.

Solution to Exercise 2.10 (a) Using the uninformative prior $p(\mathbf{w}) \propto 1$ together with the likelihood, we get the posterior

$$\begin{aligned} p(\mathbf{w}|\mathbf{y}) &\propto \prod_{i=1}^{N} p(y_i|\mathbf{w}) \\ &= \prod_{i=1}^{N} \mathcal{N}\left(y_i;\ \mathbf{x}_i^\mathsf{T}\mathbf{w},\ \sigma^2\right) \\ &\propto \prod_{i=1}^{N} e^{-\frac{1}{2\sigma^2}\left(y_i - \mathbf{x}_i^\mathsf{T}\mathbf{w}\right)^2} \end{aligned}$$

Further, since $\log$ is a monotonically increasing function, we can write

$$\begin{aligned} \widehat{\mathbf{w}} &= \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{y}) \\ &= \arg\max_{\mathbf{w}} \log p(\mathbf{w}|\mathbf{y}) \\ &= \arg\max_{\mathbf{w}} \left\{ \sum_{i=1}^{N} -\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^\mathsf{T}\mathbf{w})^2 \right\} \\ &= \arg\min_{\mathbf{w}} \left\{ \sum_{i=1}^{N} (y_i - \mathbf{x}_i^\mathsf{T}\mathbf{w})^2 \right\} \end{aligned}$$

(b) Using the Gaussian prior together with the likelihood, we get the posterior

$$\begin{aligned} p(\mathbf{w}|\mathbf{y}) &\propto \prod_{i=1}^{N} p(y_i|\mathbf{w}) p(\mathbf{w}) \\ &= \prod_{i=1}^{N} \mathcal{N}\left(y_i;\ \mathbf{x}_i^\mathsf{T}\mathbf{w},\ \sigma^2\right) \prod_{j=1}^{p} \mathcal{N}\left(w_j;\ 0,\ \alpha^2\right) \\ &\propto \prod_{i=1}^{N} e^{-\frac{1}{2\sigma^2}\left(y_i - \mathbf{x}_i^\mathsf{T}\mathbf{w}\right)^2} \prod_{j=1}^{p} e^{-\frac{1}{2\alpha^2} w_j^2} \end{aligned}$$

As in the previous exercise, we make use of the fact that $\log$ is a monotonically increasing function

$$\begin{aligned}\widehat{\mathbf{w}} &= \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{y}) \\ &= \arg\max_{\mathbf{w}} \log p(\mathbf{w}|\mathbf{y}) \\ &= \arg\max_{\mathbf{w}} \left\{ \sum_{i=1}^{N} -\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^\mathsf{T}\mathbf{w})^2 + \sum_{j=1}^{p} -\frac{1}{2\alpha^2} w_j^2 \right\} \\ &= \arg\min_{\mathbf{w}} \left\{ \sum_{i=1}^{N} (y_i - \mathbf{x}_i^\mathsf{T}\mathbf{w})^2 + \frac{\sigma^2}{\alpha^2} \sum_{j=1}^{p} w_j^2 \right\}\end{aligned}$$

which is equivalent to (18a) if $\alpha = \frac{\sigma}{\sqrt{\lambda}}$

(c) In a similar fashion as in the previous exercise we get

$$\begin{aligned}p(\mathbf{w}|\mathbf{y}) &= \prod_{i=1}^{N} \mathcal{N}\left(y_i;\ \mathbf{x}_i^\mathsf{T}\mathbf{w},\ \sigma^2\right) \prod_{j=1}^{p} \mathcal{L}(w_j;\ 0,\ \alpha) \\ &\propto \prod_{i=1}^{N} e^{-\frac{1}{2\sigma^2}\left(y_i - \mathbf{x}_i^\mathsf{T}\mathbf{w}\right)^2} \prod_{j=1}^{p} e^{-\frac{|w_j|}{\alpha}}\end{aligned}$$

Further, since $\log$ is a monotonically increasing function, we can write

$$\begin{aligned}\widehat{\mathbf{w}} &= \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{y}) \\ &= \arg\max_{\mathbf{w}} \log p(\mathbf{w}|\mathbf{y}) \\ &= \arg\max_{\mathbf{w}} \left\{ \sum_{i=1}^{N} -\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^\mathsf{T}\mathbf{w})^2 + \sum_{j=1}^{p} -\frac{|w_j|}{\alpha} \right\} \\ &= \arg\min_{\mathbf{w}} \left\{ \sum_{i=1}^{N} (y_i - \mathbf{x}_i^\mathsf{T}\mathbf{w})^2 + \frac{2\sigma^2}{\alpha} \sum_{j=1}^{p} |w_j| \right\}\end{aligned}$$

which is equivalent to (19a) if $\alpha = \frac{2\sigma^2}{\lambda}$

Solution to Exercise 2.11
The solution can be simplified by considering the logarithm of Bayes' theorem

$$\ln p(\mathbf{w}, \beta|\mathbf{y}) = \ln p(\mathbf{y}|\mathbf{w}, \beta) + \ln p(\mathbf{w}, \beta) + \textit{const.},$$

where *const.* is a constant that depends neither on $\mathbf{w}$, nor β. This gives

$$\begin{aligned}\ln p(\mathbf{w}, \beta|\mathbf{y}) &= \ln \mathcal{N}\left(\mathbf{y};\ \mathbf{X}\mathbf{w},\ \beta^{-1}\mathbf{I}_N\right) + \ln \mathcal{N}\left(\mathbf{w};\ \mathbf{m}_0,\ \beta^{-1}\mathbf{S}_0\right) + \ln \text{Gam}\left(\beta; a_0,\ b_0\right) + \textit{const.} \\ &= \frac{N}{2}\ln\beta - \frac{1}{2}\beta(\mathbf{y} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \frac{p}{2}\ln\beta - \frac{1}{2}\beta(\mathbf{w} - \mathbf{m}_0)^\mathsf{T}\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0) + (a_0 - 1)\ln\beta - b_0\beta + \textit{const.},\end{aligned} \tag{35}$$

where p is the dimension of $\mathbf{w}$. We have used the fact that $|\gamma\mathbf{A}| = \gamma^p|\mathbf{A}|$ for a matrix $\mathbf{A}$ with dimension $p \times p$ and we have also again disregarded all terms that do not depend on $\mathbf{w}$ or β.

Similarly, we also compute the logarithm of the suggested Gauss-Gamma form of the posterior

$$\begin{aligned}\ln p(\mathbf{w}, \beta|\mathbf{y}) &= \ln \mathcal{N}\left(\mathbf{w};\ \mathbf{m}_N,\ \beta^{-1}\mathbf{S}_N\right) + \ln \text{Gam}\left(\beta; a_N,\ b_N\right) + \textit{const.} \\ &= \frac{p}{2}\ln\beta - \frac{1}{2}\beta(\mathbf{w} - \mathbf{m}_N)^\mathsf{T}\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N) + (a_N - 1)\ln\beta - b_N\beta + \textit{const.}\end{aligned} \tag{36}$$

To bring (35) into the form (36) we first need to complete the squares in (35) with respect to $\mathbf{w}$

$$\begin{aligned}-\frac{1}{2}\beta(\mathbf{y}-\mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{y}-\mathbf{X}\mathbf{w})-\frac{1}{2}\beta(\mathbf{w}-\mathbf{m}_0)^\mathsf{T}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)& \\ = -\frac{1}{2}\beta\left(\mathbf{y}^\mathsf{T}\mathbf{y}-2\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{y}+\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}+\mathbf{w}^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{w}-2\mathbf{w}^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0+\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0\right)& \\ = -\frac{1}{2}\beta\left(\mathbf{w}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X}+\mathbf{S}_0^{-1})\mathbf{w}-2\mathbf{w}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{y}+\mathbf{S}_0^{-1}\mathbf{m}_0)+\mathbf{y}^\mathsf{T}\mathbf{y}+\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0\right).&\end{aligned}$$

To simplify the notation, denote $\mathbf{A}=\mathbf{X}^\mathsf{T}\mathbf{X}+\mathbf{S}_0^{-1}$ and $\mathbf{b}=\mathbf{X}^\mathsf{T}\mathbf{y}+\mathbf{S}_0^{-1}\mathbf{m}_0$ and get

$$-\frac{1}{2}\beta\left(\mathbf{w}^\mathsf{T}\mathbf{A}\mathbf{w}-2\mathbf{w}^\mathsf{T}\mathbf{b}+\mathbf{y}^\mathsf{T}\mathbf{y}+\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0\right)=-\frac{1}{2}\beta\left((\mathbf{w}-\mathbf{A}^{-1}\mathbf{b})^\mathsf{T}\mathbf{A}(\mathbf{w}-\mathbf{A}^{-1}\mathbf{b})-\mathbf{b}^\mathsf{T}\mathbf{A}^{-1}\mathbf{b}+\mathbf{y}^\mathsf{T}\mathbf{y}+\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0\right). \tag{37}$$

By comparing (37) with the quadratic form of $\mathbf{w}$ in (36), we can identify

$$\begin{aligned}\mathbf{S}_N^{-1}&=\mathbf{A}=\mathbf{X}^\mathsf{T}\mathbf{X}+\mathbf{S}_0^{-1},\\ \mathbf{m}_N&=\mathbf{A}^{-1}\mathbf{b}=\mathbf{S}_N(\mathbf{X}^\mathsf{T}\mathbf{y}+\mathbf{S}_0^{-1}\mathbf{m}_0),\end{aligned}$$

where (37) expressed in $\mathbf{S}_N^{-1}$ and $\mathbf{m}_N$ will be

$$-\frac{1}{2}\beta\left((\mathbf{w}-\mathbf{m}_N)^\mathsf{T}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)-\mathbf{m}_N^\mathsf{T}\mathbf{S}_N^{-1}\mathbf{m}_N+\mathbf{y}^\mathsf{T}\mathbf{y}+\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0\right).$$

By plugging this expression back into (35) we get

$$\begin{aligned}\ln p(\mathbf{w},\beta|\mathbf{y})&=-\frac{1}{2}\beta(\mathbf{w}-\mathbf{m}_N)^\mathsf{T}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)+\frac{1}{2}\beta\mathbf{m}_N^\mathsf{T}\mathbf{S}_N^{-1}\mathbf{m}_N-\frac{1}{2}\beta\mathbf{y}^\mathsf{T}\mathbf{y}-\frac{1}{2}\beta\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0\\ &\qquad+\frac{N}{2}\ln\beta+\frac{p}{2}\ln\beta+(a_0-1)\ln\beta-b_0\beta+\textit{const.}\\ &=\frac{p}{2}\ln\beta-\frac{1}{2}\beta(\mathbf{w}-\mathbf{m}_N)^\mathsf{T}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)\\ &\qquad+\left(a_0+\frac{N}{2}-1\right)\ln\beta-\left(b_0-\frac{1}{2}\mathbf{m}_N^\mathsf{T}\mathbf{S}_N^{-1}\mathbf{m}_N+\frac{1}{2}\mathbf{y}^\mathsf{T}\mathbf{y}+\frac{1}{2}\beta\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0\right)\beta+\textit{const.}\end{aligned}$$

and by comparing with (36) we get

$$\begin{aligned}a_N&=a_0+\frac{N}{2},\\ b_N&=b_0+\frac{1}{2}\left(\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0-\mathbf{m}_N^\mathsf{T}\mathbf{S}_N^{-1}\mathbf{m}_N+\sum_{n=1}^N y_n^2\right).\end{aligned}$$

Bibliography

[1] David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[2] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[3] Kevin B Korb and Ann E Nicholson. *Bayesian artificial intelligence*. CRC press, 2010.

[4] Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988.

[5] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.