Institute for Data Science in Mechanical Engineering
RWTH Aachen University

Fundamentals of Machine Learning

Prof. Dr. Sebastian Trimpe

Winter Semester 2022 – 2023

Problem Set 2: Linear Models for Regression and Classification

Last updated: November 25, 2022

Preamble

- The problem sets help you train and deepen your understanding of the topics taught in class. We recommend that you try solving the problems without looking at the solutions. Only consult the solution to validate your result, or when you have problems.
- This problem set will be completed and updated as we progress in the course.
- Difficulty is indicated by means of stars: ★ is the lowest difficulty level, and ★★★ the highest. Experienced difficulty may vary depending on your background.
- Please report any errors found in this problem set to the teaching assistants (fml@dsme.rwth-aachen.de).
- Finally, do not get scared if you find these problems hard: the goal is to train you!

1 Regression

1.1 Flavors of linear regression

1. **Standard linear regression** ★

Assume we have a data set $(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_N, \mathbf{y}_N)$, where $\mathbf{x} = (\mathbf{x}_{n,1} \ \dots \ \mathbf{x}_{n,D})^\top \in \mathbb{R}^D$ are input values and $\mathbf{y}_n \in \mathbb{R}$ are output values, for $n = 1, \dots, N$. Define the extended data matrix as

$$X = \begin{pmatrix} 1 & \mathbf{x}_{1,1} & \dots & \mathbf{x}_{1,D} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathbf{x}_{N,1} & \dots & \mathbf{x}_{N,D} \end{pmatrix} \in \mathbb{R}^{N \times (D+1)}. \tag{1}$$

We assume that X has full column rank. For $x \in \mathbb{R}^D$, we consider the linear model

$$h(x, w) = w_0 + \sum_{i=1}^{D} w_i \cdot x_i = w^\top \cdot \tilde{x}, \tag{2}$$

where $w = (w_0 \ \dots \ w_D)^\top \in \mathbb{R}^{D+1}$ is the vector of weights and $\tilde{x} = (1 \ x^\top)^\top \in \mathbb{R}^{D+1}$ is the extended input.

(a) Compute the optimal weights $w^\star$ according to the least squares criterion.

(b) Let $D = 1$, and the data be as given in Table 1. Compute the optimal weights $w^\star$ according to the least-squares criterion for this data.

x	-1	-0.5	0	1.5
y	0	-1	2	1

Table 1: Data for standard linear regression

(c) Compute the sum of squared errors for the model you found in the previous question.

2. **Regularized linear regression** ★

Assume we have a data set $(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_N, \mathbf{y}_N)$, where $\mathbf{x} = (\mathbf{x}_{n,1} \ \dots \ \mathbf{x}_{n,D})^\top \in \mathbb{R}^D$ are input values and $\mathbf{y}_n \in \mathbb{R}$ are output values, for $n = 1, \dots, N$. Define the extended data matrix as

$$X = \begin{pmatrix} 1 & \mathbf{x}_{1,1} & \dots & \mathbf{x}_{1,D} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathbf{x}_{N,1} & \dots & \mathbf{x}_{N,D} \end{pmatrix} \in \mathbb{R}^{N \times (D+1)}. \tag{3}$$

We assume that X has full column rank. For $x \in \mathbb{R}^D$, we consider the linear model

$$h(x, w) = w_0 + \sum_{i=1}^{D} w_i \cdot x_i = w^\top \cdot \tilde{x}, \tag{4}$$

where $w = (w_0 \ \dots \ w_D)^\top \in \mathbb{R}^{D+1}$ is the vector of weights and $\tilde{x} = (1 \ x^\top)^\top \in \mathbb{R}^{D+1}$ is the extended input.

(a) Compute the optimal weights $w^\star$ according to the least squares criterion with L_2-regularization of parameter λ.

(b) Let $D = 1$, and the data be as given in Table 2. Compute the optimal weights $w^\star$ according to the regularized least-squares criterion for this data, with $\lambda = 10$.

x	-1.5	-1	0.5	1.5
y	2	1.5	1.5	1

Table 2: Data for regularized linear regression.

(c) i. Compute the sum of squared errors for the model you found in the previous question. *Hint: this error does not have the regularization term.*

ii. Compute the optimal weights $w_0^\star$ according to the *unregularized* least squares criterion for the data set in Table 2.

iii. Compute the sum of squared errors for the unregularized model $w_0^\star$. Is it higher or lower than the sum of squared errors for the weights $w^\star$? Why?

3. **Generalized linear regression** ★

Assume we have a data set $(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_N, \mathbf{y}_N)$, where $\mathbf{x} = (\mathbf{x}_{n,1} \ \dots \ \mathbf{x}_{n,D})^\top \in \mathbb{R}^D$ are input values and $\mathbf{y}_n \in \mathbb{R}$ are output values, for $n = 1, \dots, N$. Define $M+1$ basis functions $\phi_0, \dots, \phi_M$ from $\mathbb{R}^D$ to $\mathbb{R}$, with $M \in \mathbb{N}$, and the feature matrix Φ as

$$\Phi = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \dots & \phi_M(\mathbf{x}_1) \\ \vdots & \vdots & \vdots \\ \phi_0(\mathbf{x}_N) & \dots & \phi_M(\mathbf{x}_N) \end{pmatrix} \in \mathbb{R}^{N \times (M+1)}. \tag{5}$$

We assume that Φ has full column rank. For $x \in \mathbb{R}^D$, we consider the generalized linear model

$$h(x, w) = \sum_{i=0}^{D} w_i \cdot \phi_i(x) = w^\top \cdot \phi(x), \tag{6}$$

where $w = (w_0 \ \dots \ w_D)^\top \in \mathbb{R}^{D+1}$ is the vector of weights and $\phi(x) = (\phi_0(x) \ \dots \ \phi_M(x))^\top \in \mathbb{R}^{M+1}$ is the generalized input.

(a) Like in Problem 1, define the extended data matrix X as

$$X = \begin{pmatrix} 1 & \mathbf{x}_{1,1} & \dots & \mathbf{x}_{1,D} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathbf{x}_{N,1} & \dots & \mathbf{x}_{N,D} \end{pmatrix} \in \mathbb{R}^{N \times (D+1)}. \tag{7}$$

Propose M and $\phi_0, \dots, \phi_M$ such that $\Phi = X$.

In what follows, we do *not* assume that this condition is satisfied.

(b) Explain in your own words why h is called a *linear* model.

(c) Compute the optimal weights $w^\star$ according to the least squares criterion.

(d) Let $D = 1$, and the data be as given in Table 3. Take $M = 2$, and $\phi_j(x) = x^j$. Compute the optimal weights $w^\star$ according to the least-squares criterion for this data.

x	-3	1	9
y	2.8	0.4	0.2

Table 3: Data for generalized linear regression.

(e) Re-do the same question with $M = 1$, $\phi_0(x) = 1$, and

$$\phi_1(x) = \operatorname{sign}(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ -1, & \text{otherwise.} \end{cases} \tag{8}$$

4. **A regression pipeline** ★

The goal of this problem is to illustrate the influence of the choice of basis functions on the learned model. We will demonstrate the importance of *feature engineering*, and show how one can use noticeable patterns in the data to come up with relevant features.

We will illustrate this by learning multiple models for a synthetic dataset that describes solid friction. The model is taken from [Per+11].

The questions are in the accompanying notebook `FML-WS22-PS2-04_Regression_pipeline.ipynb`.

5. **Multidimensional linear regression** ★★

Let $f : \mathbb{R}^D \to \mathbb{R}^K$ be a function that we want to approximate with generalized linear regression. Instead of the setting of the lecture where the function f takes values in $\mathbb{R}$, it is now vector-valued.

We propose to run a *multidimensional linear regression* on f. We take $M+1$ basis functions $\phi_0, \dots, \phi_M : \mathbb{R}^D \to \mathbb{R}$, and model f as

$$h(x, W) = \sum_{j=0}^{M} \phi_j(x) \cdot w_j, \tag{9}$$

where, for all $j \in \{1, \dots, M\}$, $w_j \in \mathbb{R}^K$ is a *weights vector.* The weights are grouped in the matrix $W = (w_0 \ \dots \ w_M)^\top \in \mathbb{R}^{(M+1)\times K}$.

Assume we have a data set $(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_N, \mathbf{y}_N)$, where $\mathbf{x} \in \mathbb{R}^D$ are input values and $\mathbf{y}_n \in \mathbb{R}^K$ are output values, for $n = 1, \dots, N$. We define the feature matrix Φ as

$$\Phi = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \dots & \phi_M(\mathbf{x}_1) \\ \vdots & \vdots & \vdots \\ \phi_0(\mathbf{x}_N) & \dots & \phi_M(\mathbf{x}_N) \end{pmatrix} \in \mathbb{R}^{N\times(M+1)}. \tag{10}$$

We assume that Φ has full column rank. Finally, we define the output matrix as

$$Y = (\mathbf{y}_1 \ \dots \ \mathbf{y}_N)^\top \in \mathbb{R}^{N\times K}. \tag{11}$$

We propose to use the least squares criterion L to fit the model:

$$L(W) = \frac{1}{2} \sum_{i=1}^{N} \|\mathbf{y}_i - h(\mathbf{x}_i, W)\|^2. \tag{12}$$

(a) Express $h(x, W)$ as a matrix product.

(b) Show that $L(W) = \frac{1}{2} \mathrm{Tr}\left[(Y - \Phi W)^\top \cdot (Y - \Phi W)\right]$, where $\mathrm{Tr}[A]$ is the trace of the matrix A.

(c) Find the optimal weights matrix for the least squares criterion L.
Hint: you can use the fact that $\nabla_A \mathrm{Tr}[A^\top A] = A$.

(d) How does this setting differ from running K independent linear regressions on each dimension of f? Do you find the same results?

1.2 Properties of the least-squares estimate

6. **The least-squares estimator** ★

In this exercise, we assume that the data is a random variable. The least-squares solution is, therefore, a random variable itself, which we call the *least squares estimator.* We postulate a form for the true function f, and find how this estimator relates to the ground truth, that is, to the true parameters of f.

Assume we have $N \in \mathbb{N}_>$ data points $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^D$ and $M+1$ basis functions $\phi_0, \dots, \phi_M$. Define the feature matrix as

$$\Phi = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \dots & \phi_M(\mathbf{x}_1) \\ \vdots & \vdots & \vdots \\ \phi_0(\mathbf{x}_N) & \dots & \phi_M(\mathbf{x}_N) \end{pmatrix} \in \mathbb{R}^{N\times(M+1)} \tag{13}$$

We assume that Φ has full column rank. We assume that the *true* f has the following form, that is,

$$f(x) = \sum_{i=0} w_i \cdot \phi_i(x) = w^\top \cdot \phi(x), \tag{14}$$

where $w = (w_0 \ \dots \ w_M)^\top \in \mathbb{R}^{M+1}$ is a fixed vector of weights and $\phi(x) = (\phi_0(x) \ \dots \ \phi_N(x))$. Finally, we consider the case where we do not measure the value of f directly, but a noisy value, that is,

$$Y = \Phi^\top \cdot w + \epsilon, \tag{15}$$

where ϵ is an N-dimensional random variable. At this point, we make no other assumptions on ϵ.

We model f with a linear model with basis functions $\phi_0, \dots, \phi_M$, and denote by w_{LS} the least-squares estimator of the weights of the model.

(a) Give the expression of w_{LS}, and explain where the randomness comes from.

(b) Assume that $\mathbb{E}[\epsilon] = 0$. Show that w_{LS} is an unbiased estimator of w, i.e., $\mathbb{E}[w_{\text{LS}}] = w$.

(c) Find the variance of w_{LS} (sometimes also called the covariance matrix).

(d) Assume that $\epsilon \sim \mathcal{N}(0, \Sigma)$ is Gaussian, where Σ is symmetric, positive definite. Show that w is also Gaussian and specify its parameters.

7. A geometric perspective ★★★

Assume we have a data set $(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_N, \mathbf{y}_N)$, where $\mathbf{x} = (\mathbf{x}_{n,1} \ \dots \ \mathbf{x}_{n,D})^\top \in \mathbb{R}^D$ are input values and $\mathbf{y}_n \in \mathbb{R}$ are output values, for $n = 1, \dots, N$. Define $M + 1$ basis functions $\phi = (\phi_0, \dots, \phi_M)^\top$ from $\mathbb{R}^D$ to $\mathbb{R}$, with $M \in \mathbb{N}$, and the feature matrix $\Phi \in \mathbb{R}^{N \times (M+1)}$. We assume that Φ has full column rank.

We consider the linear regression problem with basis functions $\phi_0, \dots, \phi_M$:

$$h(x, w) = \phi(x) \cdot w, \tag{16}$$

where $w \in \mathbb{R}^{M+1}$ is the vector of weights. We denote by w_{LS}. the optimal weigths of the model according to the least squares criterion. Finally, we define $\hat{y}_n = h(\mathbf{x}_n, w_{\text{LS}})$ the output of the model at the point $\mathbf{x}_n$ for all $n \in \{1, \dots, N\}$.

(a) Find a symmetric matrix $A \in \mathbb{R}^{N \times N}$ such that

$$\hat{y} = A \cdot \mathbf{y}, \tag{17}$$

where $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)$.

(b) Consider the mapping

$$p : y \in \mathbb{R}^N \mapsto A \cdot y. \tag{18}$$

Show that p is the orthogonal projection.
Hint: recall that a linear function p is a projector if, and only if, $p(p(y)) = p(y)$. Additionally, it is an orthogonal projection if, and only if, $A^\top = A$, where A is the matrix that represents p.

(c) Show that the image set of p is the space spanned by the column vectors of Φ.

(d) Give a geometric interpretation of the last two results: how do you obtain $\hat{y}$ from y?

8. Interpreting the bias term ★

Assume we have a data set $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$, where $\mathbf{x} = (\mathbf{x}_{n,1} \ \dots \ \mathbf{x}_{n,D})^\top \in \mathbb{R}^D$ are input values and $y_n \in \mathbb{R}$ are output values, for $n = 1, \dots, N$. Define $M+1$ basis functions $\phi_0, \dots, \phi_M$ from $\mathbb{R}^D$ to $\mathbb{R}$, with $M \in \mathbb{N}$, and the feature matrix $\Phi \in \mathbb{R}^{N \times (M+1)}$ as usual. We assume that Φ has full column rank.

We consider the case where ϕ_0 is constant equal to 1. We model the relationship between x and y as the linear model

$$h(x, w) = \sum_{i=0}^{D} w_i \cdot \phi_i(x) = w_0 + \sum_{i=1}^{D} \phi_i(x). \tag{19}$$

We consider the optimal weights $w^\star$ according to the least squares criterion. The goal of this exercise is to interpret the term $w_0^\star$, and to justify the denomination "bias term".

(a) Give the expression of the optimal weights $w^\star$ as a function of the data.

(b) Express the term $w_0^\star$ as a sum over the data points.

(c) Interpret your result: what does $w_0^\star$ represent?

9. Interpreting the precision of a Gaussian likelihood model ★★

Consider a data set $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, \mathbf{y}_N)$, where $\mathbf{x}_n = (\mathbf{x}_{n,1} \ \dots \ \mathbf{x}_{n,D})^\top \in \mathbb{R}^D$ are input values and $\mathbf{y}_n \in \mathbb{R}$ are output values, for $n = 1, \dots, N$. Let X be the data matrix. Assume that we we have fit a model $h(x, w)$ to this regression problem; for instance, w can be the outcome of a generalized linear regression with the least-squares criterion.

Unfortunately, our model is imperfect, and $h(\mathbf{x}_n, w)$ is not equal to $\mathbf{y}_n$ for all n. We model the remaining unexplained variance by Gaussian noise. More precisely, we assume that the output vector $\mathbf{y} = (\mathbf{y}_1 \ \dots \ \mathbf{y}_N) \in \mathbb{R}^N$ is a realization of a Gaussian random variable $Y \in \mathbb{R}^N$ with known mean $(h(\mathbf{x}_1, w) \dots h(\mathbf{x}_N, w))^\top$ and unknown variance $\beta^{-1}I$, where I is the identity matrix of size N:

$$p_Y(\mathbf{y} \mid X, w, \beta) = \mathcal{N}\left(\mathbf{y} \,\middle|\, \begin{pmatrix} h(\mathbf{x}_1, w) \\ \vdots \\ h(\mathbf{x}_N, w) \end{pmatrix}, \ \beta^{-1} I\right). \tag{20}$$

The goal of this exercise is to find and interpret the *maximum likelihood estimate* of β, which we denote by β_{ML}.

(a) Show that the likelihood p_Y factorizes as

$$p_Y(\mathbf{y}|X, w, \beta) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{y}_i \mid h(\mathbf{x}_i, w), \beta^{-1}). \tag{21}$$

(b) Let $L(\beta)$ be the negative log-likelihood of Y, i.e.,

$$L(\beta) = -\ln\left[p_Y(\mathbf{y} \mid X, w, \beta)\right]. \tag{22}$$

Compute L as a function of X, w, and β.

(c) Give the definition of β_{ML} as an extremum of L, and compute β_{ML}.

(d) Interpret your result: what does β_{ML}^{-1} represent?

10. **Gauss-Markov theorem** ★★★

The goal of this exercise is to prove the Gauss-Markov theorem, which we state after question d.

Assume we have $N \in \mathbb{N}_>$ data points $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^D$ and $M+1$ basis functions $\phi_0, \dots, \phi_M$. Define the feature matrix $\Phi \in \mathbb{R}^{N \times (M+1)}$, and assume that Φ has full column rank. We consider the case where we measure a noisy value of a linear function, that is,

$$Y = \Phi \cdot \bar{w} + \epsilon. \tag{23}$$

Here, $\bar{w} = (\bar{w}_0 \ \dots \ \bar{w}_M)^\top \in \mathbb{R}^{M+1}$ is a fixed vector of weights, $\phi(x) = (\phi_0(x) \ \dots \ \phi_N(x))$ is the feature vector, and $\epsilon = (\epsilon_1, \dots, \epsilon_N)$ is a CRV representing the noise. We make the following assumptions on the noise:

1. ϵ has zero mean: $\mathbb{E}[\epsilon] = 0$;
2. the variance of ϵ is $\text{Var}[\epsilon] = \sigma^2 \cdot I$, where I is the identity matrix of size n.

Consider w_{LS}, the least-squares estimator of the weights $\bar{w}$. Like in Problem 6, we consider w_{LS} as a random variable. It is an example of the general class of *linear estimators*, that is, a random variable w of the form

$$w = C \cdot Y, \tag{24}$$

where $C \in \mathbb{R}^{(M+1)\times N}$.

(a) Show that w_{LS} is indeed a linear estimator, and specify $C_{\text{LS}} \in \mathbb{R}^{(M+1)\times N}$ such that $w_{\text{LS}} = C_{\text{LS}} Y$.

The bias of w is its expected difference with the quantity it estimates, that is, $\mathbb{E}[w - \bar{w}]$.

(b) Show that w_{LS} is an unbiased estimator of $\bar{w}$.
Hint: you can use the results of Problem 6 if you have solved it.

A common way to evaluate the quality of an estimator is through its mean squared error $\text{MSE}(w)$, defined as

$$\text{MSE}(w) = \mathbb{E}[\|w - \bar{w}\|^2]. \tag{25}$$

(c) For an arbitrary estimator w, verify that the following bias-variance decomposition holds:

$$\mathrm{MSE}(w) = \|\mathbb{E}[w - \bar{w}]\|^2 + \mathrm{Tr}\left[\mathrm{Var}[w]\right], \tag{26}$$

where $\mathrm{Tr}[A]$ denotes the trace of the square matrix A, i.e., the sum of its diagonal elements. *Hint: Show that both sides of the equation are equal to*

$$\mathbb{E}[\|w\|^2] + \|\bar{w}\|^2 - 2 \cdot \mathbb{E}[w] \cdot \bar{w}. \tag{27}$$

For this, use the identity

$$\mathrm{Tr}[v^\top v] = \|v\|^2, \quad \forall v \in \mathbb{R}^n. \tag{28}$$

(d) Show that, for a linear estimator $w = C \cdot Y$,

$$\begin{aligned} \mathbb{E}[w - \bar{w}] &= (C\Phi - I) \cdot \bar{w} \\ \mathrm{Var}[w] &= \sigma^2 C \cdot C^\top \end{aligned} \tag{29}$$

The Gauss-Markov theorem states that if w is a linear unbiased estimator (LUE), then

$$\mathrm{MSE}(w) \geq \mathrm{MSE}(w_{\mathrm{LS}}). \tag{30}$$

In other words, w_{LS} is the *best* linear unbiased estimator (BLUE) as measured by the MSE. To prove this theorem, we consider an arbitrary LUE $w = C \cdot Y$.

(e) Deduce from the previous questions that

$$\mathrm{MSE}[w] = \mathrm{Tr}\left[\mathrm{Var}[w]\right]. \tag{31}$$

For LUEs, minimizing the MSE thus boils down to minimizing the (trace of the) variance.

(f) Let $D = C - C_{\mathrm{LS}} \in \mathbb{R}^{(M+1)\times N}$. Show that $D \cdot \Phi \cdot \bar{w} = 0$.
Hint: start by computing $\mathbb{E}[w - w_{\mathrm{LS}}]$.

Since this is valid for an arbitrary $\bar{w}$ and since C, C_{LS} and Φ don't depend on $\bar{w}$, we deduce that $D \cdot \Phi = 0$.

(g) Using the previous results, show

$$\mathrm{Var}[w] = \mathrm{Var}[w_{\mathrm{LS}}] + \sigma^2 D \cdot D^\top. \tag{32}$$

Hint: what is the value of $D \cdot C_{\mathrm{LS}}$*? You will need the expression of* C_{LS}.

(h) Conclude: show that $\mathrm{MSE}[w] \geq \mathrm{MSE}[w_{\mathrm{LS}}]$, and deduce the Gauss-Markov theorem.

Cultural note Importantly, there exist linear estimators that achieve a lower MSE than the least-squares estimator. However, according to the Gauss-Markov theorem, they must be biased. A relevant example is the regularized least-squares estimator introduced in the lecture. This estimator accepts a little bias to reduce variance and, in total, achieve a lower MSE. This is an interpretation of why regularized weights that are more stable, i.e., less sensitive to the individual data set.

1.3 Regularization

11. **Maximum a posteriori estimate with Laplace prior** ★★

In this exercise, we generalize the result of the lecture on the connection between the maximum a posteriori (MAP) estimate and regularization.

We assume that we have a data set $D = ((\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N))$, where $\mathbf{x}_n \in \mathbb{R}^{D+1}, \mathbf{y}_n \in \mathbb{R}$, and $M+1$ basis functions $\phi = (\phi_0, \ldots, \phi_M)^\top$. We denote by $\phi \in \mathbb{R}^{N\times(M+1)}$ the feature matrix. We also assume the following observation model: the output vector $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N) \in \mathbb{R}^N$ is the realization of a Gaussian random variables with independent components, mean $\Phi \cdot w$, and variance $\beta^{-1} I$, where $w \in \mathbb{R}^{M+1}$ is a weights vector and $\beta > 0$. In other words,

$$p_{Y|W}(\mathbf{y}|w) = \mathcal{N}(\mathbf{y}|\Phi \cdot w, \beta^{-1} I). \tag{33}$$

We put a prior $p_W(w)$ on w, and combine that with the above likelihood model to get the posterior distribution given the data:

$$p_{W|Y}(w|\mathbf{y}) = \frac{p_{Y|W}(\mathbf{y}|w) \cdot p_W(w)}{p_Y(\mathbf{y})}. \tag{34}$$

(a) Justify that the likelihood function factorizes as

$$p_{Y|W}(\mathbf{y}|w) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{y}_i|\phi(\mathbf{x}_i)^\top w, \beta^{-1}). \tag{35}$$

Consider the maximum a posteriori estimate

$$w_{\text{MAP}} = \operatorname*{argmax}_{w} p_{W|Y}(w|\mathbf{y}). \tag{36}$$

(b) Justify that

$$w_{\text{MAP}} = \operatorname*{argmin}_{w} L_{\text{NLL}}(w) - \ln\left[p_W(w)\right], \tag{37}$$

where the negative log-likelihood L_{NLL} is defined as

$$L_{\text{NLL}}(w) = -\ln\left[p_{Y|W}(\mathbf{y}|w)\right]. \tag{38}$$

(c) Verify that

$$L_{\text{NLL}}(w) = K + \beta \sum_{i=1}^{N} (\mathbf{y}_i - \phi(\mathbf{x}_i)^\top \cdot w)^2, \tag{39}$$

where K is a constant that does not depend on w.

(d) We now put a Laplace prior on W, that is, we assume that

$$p_W(w) = \frac{1}{2\alpha} \exp\left[-\frac{\sum_{i=0}^{M} |w_i|}{\alpha}\right] = \frac{1}{2\alpha} \exp\left[-\frac{\|w\|_1}{\alpha}\right], \tag{40}$$

where $\alpha > 0$ is a constant.

i. Only for this question, assume $M = 0$. Verify that p_W is indeed a PDF.

ii. Compute $-\ln\left[p_W(w)\right]$.

iii. Interpret w_{MAP} as the solution to the following regularized least-squares problem

$$w_{\text{MAP}} = \operatorname*{argmin}_{w} \frac{1}{2} \sum_{i=1}^{N} (\mathbf{y}_i - \phi(\mathbf{x}_i)^\top \cdot w)^2 + \lambda \|w\|_1, \tag{41}$$

where you will specify the value of λ.

Cultural note This regularized least-squares problem is called Lasso regularization. Compared to the quadratic/Ridge regularization seen in the lecture, the regularization term involves the sum of absolute values of the weights.

(e) We generalize the result of the previous question by assuming a prior of the general form

$$p_W(w) = C \cdot \exp\left[-\frac{\|w\|_p^p}{p \cdot \alpha^p}\right], \tag{42}$$

where $p \in \mathbb{N}_>$, $\alpha > 0$ is a constant, C is a normalization constant, and

$$\|w\|_p^p = \sum_{i=0}^{M} |w_i|^p \tag{43}$$

is the p-norm of w to the power of p. For instance, $p = 1$ yields the previous case, and $p = 2$ yields the case of the Gaussian prior seen in the lecture.

i. Compute $-\ln[p_W(w)]$.

ii. Interpret w_{MAP} as the solution to the following regularized least-squares problem

$$w_{\mathrm{MAP}} = \operatorname*{argmin}_w \frac{1}{2}\sum_{i=1}^{N}(\mathbf{y}_i - \phi(\mathbf{x}_i)^\top \cdot w)^2 + \lambda \|w\|_p^p, \tag{44}$$

where you will specify the value of λ as a function of p, α, and C.

12. **Gaussian input noise and ridge regression** ★★

Assume we have a data set $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)$, where $\mathbf{x} = (\mathbf{x}_{n,1} \ \ldots \ \mathbf{x}_{n,D})^\top \in \mathbb{R}^D$ are input values and $\mathbf{y}_n \in \mathbb{R}$ are output values, for $n = 1, \ldots, N$. Consider a linear model of the form

$$h(x, w) = (1 \ x^\top) \cdot w, \tag{45}$$

together with the usual least-squared criterion

$$L_D(w) = \frac{1}{2}\sum_{i=1}^{N}(\mathbf{y}_n - h(\mathbf{x}_n, w))^2. \tag{46}$$

Assume now that we have additive noise on the inputs, that is, $X_n = \mathbf{x}_n + \epsilon_n$ for all n, where $\epsilon_1, \ldots, \epsilon_N$ are i.i.d. according to $\mathcal{N}(0, \sigma^2 I)$. Consider the corresponding adapted least-squares criterion

$$L_D^\epsilon(w) = \frac{1}{2}\sum_{i=1}^{N}(\mathbf{y}_n - h(X_n, w))^2. \tag{47}$$

This is now a random variable, since the $(X_n)_n$'s themselves are random.

Show that minimizing the expected least-squares criterion $\mathbb{E}[L_D^\epsilon(w)]$ corresponds to minimizing $L_D(w)$ with the addition of a regularization term, which you will precise.

1.4 Bayesian linear regression

13. **Working with Gaussian random variables: theory** ★★

In Bayesian linear regresison, we often assume Gaussian prior and likelihood. One reason is that this enables *closed-form solutions*, meaning that we can calculate exactly the distributions of the variables involved (posteriors, marginals, ...). In this exercise, we prove the calculation rules for Gaussian random variables mentioned in the lecture.

Let X be an n-dimensional Gaussian vector that we partition in two disjoing subsets X_1 and X_2. We assume without loss of generality that X_1 corresponds to the m first dimensions of X, and X_2 to the other $p = n - m$ ones:

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma), \tag{48}$$

where $\Sigma \in \mathbb{R}^{n \times n}$ is positive semi-definite[1]. We partition the mean and covariance similarly:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}. \tag{49}$$

[1] This means that Σ is symmetric and that $\forall v \in \mathbb{R}^n, v^\top \Sigma v \geq 0$.

The symmetry of Σ imposes that Σ_{11} and Σ_{22} are also symmetric and that $\Sigma_{12}^\top = \Sigma_{21}$. Finally, for convenience, we introduce the precision matrix Λ as the inverse of the covariance matrix:

$$\Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}. \tag{50}$$

Here again, Λ_{11} and Λ_{22} are symmetric and $\Lambda_{12}^\top = \Lambda_{21}$. We emphasize that, in general, we have $\Lambda_{ij} \neq \Sigma_{ij}^{-1}$.

(a) Let $A \in \mathbb{R}^{q \times n}$ be a matrix and $b \in \mathbb{R}^q$ be a vector, where $q \in \mathbb{N}$. Show that the affine transformation of X, $AX + b$, is Gaussian as well, i.e.

$$AX \sim \mathcal{N}(A\mu + b, A\Sigma A^\top). \tag{51}$$

Remark: if $b = 0$, the above is summarized as "a linear transformation of a GRV is a GRV".

(b) Show that

$$X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}). \tag{52}$$

Hint: use the previous question with a good choice of q, A, b.

(c) We want to show that the variable X_1 conditioned on X_2 is also Gaussian, i.e., has PDF $p_{1|2}$ defined as

$$p_{1|2}(x_1|x_2) = \mathcal{N}(x_1|\mu_{1|2}, \Sigma_{1|2}), \tag{53}$$

where

$$\begin{aligned} \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned} \tag{54}$$

For this, we introduce

$$e(x) = -\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu). \tag{55}$$

i. Show that

$$e(x) = -\frac{1}{2}(x_1 - \mu_1)^\top \Lambda_{11}(x_1 - \mu_1) - \frac{1}{2}(x_2 - \mu_2)^\top \Lambda_{22}(x_2 - \mu_2) - (x_1 - \mu_1)^\top \Lambda_{12}(x_2 - \mu_2). \tag{56}$$

ii. Deduce

$$e(x) = -\frac{1}{2}x_1^\top \Lambda_{11} x_1 + x_1^\top \Lambda_{11}(\mu_1 - m) + K, \tag{57}$$

where K is a constant that does not depend on x_1 and

$$m = \Lambda_{11}^{-1}\Lambda_{12}(x_2 - \mu_2). \tag{58}$$

iii. Deduce that the conditional PDF of X_1 conditioned on X_2 can be written as

$$p_{1|2}(x_1|x_2) = C(x_2) \cdot \exp\left[-\frac{1}{2}x_1^\top \Lambda_{11} x_1 + x_1^\top \Lambda_{11}(\mu_1 - m)\right], \tag{59}$$

where $C(x_2)$ is a coefficient that does not depend on x_1.

We deduce from the above expression that X_1 conditioned on X_2 is a Gaussian random variable:

$$X_1|X_2 = x_2 \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1||2}) \tag{60}$$

We skip the proof of this statement, since it is purely calculatory. In short, it consists in completing the square in the exponential to recognize the usual quadratic form exponent of the Gaussian PDF.

iv. Show that

$$\begin{aligned} \mu_{1|2} &= \mu_1 - m, \\ \Sigma_{1|2} &= \Lambda_{11}^{-1}. \end{aligned} \tag{61}$$

Hint: you know that $X_1 \mid X_2 = x_2$ is Gaussian. Express its PDF w.r.t. $\mu_{1|2}$ and $\Sigma_{1|2}$.

Known results from linear algebra [Wikb, Section 3.7, Equation (2)] tell us that

$$\begin{aligned}\Lambda_{11} &= \left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1},\\ \Lambda_{12} &= -\left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1}\Sigma_{12}\Sigma_{22}^{-1}.\end{aligned} \tag{62}$$

v. Deduce the expression of $\mu_{1|2}$ and $\Sigma_{1|2}$ announced in (54).

14. **Working with Gaussian random variables: practice** ★

In this problem, we apply the formulas of the previous problem to compute PDFs of transformations of GRVs.

We first consider a 2-dimensional Gaussian vector X such that

$$X \sim \mathcal{N}(\mu, \Sigma), \quad \text{with } \mu = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}. \tag{63}$$

(a) Find the marginal PDFs of X_1 and X_2.

(b) Find the PDF of X_2 conditioned on $X_1 = 1$.

(c) Find the PDF of $X_1 + X_2$.

We now consider a 3-dimensional GRV Y:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} \sim \mathcal{N}(\mu', \Sigma'), \quad \text{with } \mu = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \text{ and } \Sigma' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.8 \\ 0 & 0.8 & 1 \end{pmatrix}. \tag{64}$$

We also introduce the random variable $Z = (Y_1\ Y_2)^\top$.

(d) Find the marginal PDFs of Y_1, Y_2, Y_3, and Z.

(e) Find the PDF of Z conditioned on $Y_3 = 1$.

15. **The framework of Bayesian linear regression** ★★

The basic framework of Bayesian regression consists of 5 ingredients:

- a CRV $\hat{Y}$ taking values in $\mathbb{R}$ that depends on an input $\hat{x} \in \mathbb{R}^D$, which we denote $\hat{Y}(\hat{x})$. It represents the prediction of our model, and its density $p(\hat{y}|\hat{x})$ is called the prior predictive distribution;
- a CRV Y taking values in $\mathbb{R}^N$ that represents our observations. We make its connection with $\hat{Y}$ explicit below;
- a random variable W with its prior distribution $p(w)$. In general, W can be discrete or continuous; we assume that W is a CRV in what follows;
- a predictive model, that describes how $\hat{Y}$ varies with W and $\hat{x}$. We denote it by $\hat{Y} = f(\hat{x}, W)$ or, equivalently, by the conditional PDF $p(\hat{y}|\hat{x}, w)$;
- an obervation model, also called *likelihood*, that describes how Y varies with W. We denote it by $p(y|w)$.

A common assumption is that the observation results from applying the same function f, and corrupting the output with noise, i.e.,

$$Y = f(X, W) + \epsilon, \tag{65}$$

where $\epsilon \in \mathbb{R}^N$ is residual noise independent of W and such that $\mathbb{E}[\epsilon] = 0$, and where $X = (\mathbf{x}_1\ \dots\ \mathbf{x}_N)^\top \in \mathbb{R}^{N\times D}$ corresponds to the fixed locations of the observations in $\mathbb{R}^D$. In that case, we write $p(y|w, X)$ for the likelihood model to emphasize its dependency in X.[2]

We are now equipped to state the two main steps of Baysian regression.

[2] We emphasize that the notations $p(y|w, X)$, $p(\hat{y}|\hat{x})$ and $p(\hat{y}|\hat{x}, w)$ are abusive since X and $\hat{x}$ are not random; it does not make sense to "condition" on them. Instead, they should be understood as a functional dependency: $\hat{x} \mapsto p(\hat{y}|\hat{x})$ is a function that depends on $\hat{x}$. It is always clear from context whether a quantity appearing after the conditioning operator | is random or fixed. You can have a look at this stackexchange post for a longer discussion.

Prediction The usual assumptions are that the prior $p(w)$ and the predictive model $p(\hat{y}|\hat{x}, w)$ are known, but the prior predictive distribution $p(\hat{y}|\hat{x})$ is not. Instead, it is *computed* from the prior and prediction model according to the rules of probability theory.

(a) Express the predictive distribution $p(\hat{y}|\hat{x})$ as an integral involving the prior and the predictive model $p(\hat{y}|\hat{x}, w)$.

We say that we are doing Bayesian *linear* regression when f is linear in its second argument, that is, we have a (known) vector of basis functions $\phi(x) = (\phi_0(x), \dots, \phi_M(x)) \in \mathbb{R}^{M+1}$ such that

$$\hat{Y} = \phi(\hat{x})^\top W, \tag{66}$$

This assumption is already enough to compute the expectation and variance of $\hat{Y}$.

(b) Compute the expectation and variance of $\hat{Y}$ as functions of the expectation and variance of W.

In general, we cannot compute the integral of question a analytically. An important exception is when the prior is a Gaussian random variable.

(c) Assume that $W \sim \mathcal{N}(\bar{w}, \Sigma_W)$. Show that Y and W are jointly Gaussian, with

$$\begin{pmatrix} W \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \bar{w} \\ \phi(x)^\top \bar{w} \end{pmatrix}, \begin{pmatrix} \Sigma_W & \Sigma_W \phi(x) \\ \phi(x)^\top \Sigma_W & \phi(x)^\top \Sigma_W \phi(x) \end{pmatrix}\right) \tag{67}$$

Hint: Use the results of Problem 13.a and pick A to obtain the variable $(W^\top \ \phi(x)^\top W)^\top$.

(d) Under the same assumptions, deduce the prior predictive distribution $p(\hat{y}|\hat{x})$.

Posterior update, or Bayesian inference A core feature of Bayesian regression is its ability to adapt the distribution of the weights as new measurements come. We now assume that we have a realization $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)^\top$ of Y. Importantly, according to the definition of Y in (65), all of these observations are generated with the *same* realization of W. The noise, however, is different for every realization. We are interested in how to update our prior over the weights to reflect this noisy, incomplete observation of W. This is done straightforwardly through Bayes' theorem.

(e) Express the posterior of the weights $p(w|\mathbf{y})$ as a function of the prior $p(w)$ and the likelihood $p(\mathbf{y}|w)$.

This is as far as we can go without further assumptions on W and ϵ. Now assume that they are Gaussian and independent:

$$W \sim \mathcal{N}(w, \Sigma_W), \quad \epsilon \sim \mathcal{N}(0, \sigma_{\mathrm{n}}^2 I). \tag{68}$$

We can then compute the posterior distribution analytically.

(f) Show that W and Y are jointly Gaussian, i.e.,

$$\begin{pmatrix} W \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \bar{w} \\ \Phi \bar{w} \end{pmatrix}, \begin{pmatrix} \Sigma_W & \Sigma_W \Phi^\top \\ \Phi \Sigma_W & \Phi \Sigma_W \Phi^\top + \sigma_n^2 I \end{pmatrix}\right) \tag{69}$$

(g) Show that the posterior distribution of W is Gaussian, i.e.,

$$W|\mathbf{y} \ \sim \mathcal{N}(\mu_{\text{post}}, \ \Sigma_{\text{post}}), \tag{70}$$

with

$$\begin{aligned} \mu_{\text{post}} &= \bar{w} + \Sigma_W \Phi^\top (\Phi \Sigma_W \Phi^\top + \sigma_n^2 I)^{-1} (\mathbf{y} - \Phi \bar{w}), \\ \sigma_{\text{post}} &= \Sigma_W - \Sigma_W \Phi^\top (\Phi \Sigma_W \Phi^\top + \sigma_n^2 I)^{-1} \Phi \Sigma_W. \end{aligned} \tag{71}$$

Hint: Use the results of Problem 13.c.

We observe that the prior and the posterior are both Gaussian distributions. Therefore, we can perform both prediction and inference again if new data becomes available by using the posterior as a new prior.

This property that the prior and posterior are of the same family of distributions is specific to Gaussian distributions; you can have a look here [Wika] for more information.

References

[Per+11] Tasneem Pervez et al. "Stick-slip Friction Modeling in Tube Expansion". In: *TMT*. 2011.

[Wika] Wikipedia. *Conjugate prior*. https://en.wikipedia.org/wiki/Conjugate_prior. Last visited on Nov. 25th, 2022.

[Wikb] Wikipedia. *Invertible Matrix*. https://en.wikipedia.org/wiki/Invertible_matrix. Last visited on Nov. 24th, 2022.