Exercise 8-1  Linear Regression with Gaussian Noise

Let $D, d_i = (x_{i,1}, \ldots, x_{i,M}, y_i)^T$, be a dataset of size $N$ with $M$ features and an output $y$ which depends linearly on $X$. Due to erroneous measurements the inputs are noisy, i.e.:

$$y_i = x_i^T w + \epsilon_i,$$

where $\epsilon_i$ is the noise of data point $i$. Furthermore, assume $\epsilon$ to be gaussian distributed:

$$P(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\epsilon_i^2}.$$

Given the variables $X$ and the model $w$, we can then model the distribution of $y$ as follows:

$$P(y_i|x_i, w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - x_i^T w)^2}.$$

(a) Determine the parameter $\hat{w}$ which maximizes the probability of the training data $P(D|w)$, using the maximum-likelihood estimator: $\hat{w}_{\text{ML}} = \arg\max_w P(D|w)$.

You may assume that the $w$ are distributed independently of the input data $X$.

(b) A common assumption for the a priori distribution of random variables is:

$$P(w) = \frac{1}{(2\pi\alpha^2)^{\frac{M}{2}}} e^{\left(-\frac{1}{2\sigma^2}\sum_{j=0}^{M-1} w_j^2\right)}$$

Compute the parameter $\hat{w}$ which maximizes $P(w)P(D|w)$. Does this give an alternative interpretation to the $\lambda$-term of the penalized least squares function (PLS)?
Possible Solution:

a) Observation: \( L(w) = P(D|w) = P(y, X|w) \). \( P(y, X, w) \) is given. We can use this by \( P(y, X|w) = P(y|X, w) \cdot P(X|w) \). We know that \( X \) is independent of \( w \), hence, \( P(X|w) = P(X) \). Thus, we have the following likelihood:

\[
L(w) = P(y|X, w) \cdot P(X).
\]

However, we do not know \( P(X) \). We will see later on, that this is not important, as \( P(X) \) is independent of \( w \).

Also, we do not have \( P(y|X, w) \), but “only” \( P(y_i|x_i, w) \). Assuming that our samples have been drawn independently from the same distribution (i.i.d. = “independent, identically distributed”), we may write:

\[
L(w) = \prod_{i=1}^{N} P(y_i, x_i|w) = \prod_{i=1}^{N} P(y_i|x_i, w) \cdot P(x_i)
\]

\[
= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - x_i^T w)^2} \cdot P(x_i).
\]

which we have to derive now. Instead of deriving the product over all \((x_i, y_i) \in D\), we derive the log-likelihood, applying \( \ln(a \cdot b) = \ln a + \ln b \) (which is not the same as \( e^{a+b} = e^a \cdot e^b \)).

\[
\mathcal{L}(w) = \ln L(w) = \ln \left( \prod_{i=1}^{N} P(y_i|x_i, w) \cdot P(x_i) \right) = \sum_{i=1}^{N} \ln \left( P(y_i|x_i, w) \cdot P(x_i) \right) = \sum_{i=1}^{N} \ln P(x_i) = \sum_{i=1}^{N} \ln \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - x_i^T w)^2} \right) \cdot P(x_i) \]

\[
= \sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi\sigma^2}} + \sum_{i=1}^{N} \ln \left( e^{-\frac{1}{2\sigma^2}(y_i - x_i^T w)^2} \right) + \sum_{i=1}^{N} \ln P(x_i) = \sum_{i=1}^{N} \ln \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) - \sum_{i=1}^{N} \ln \left( e^{-\frac{1}{2\sigma^2}(y_i - x_i^T w)^2} \right) + \sum_{i=1}^{N} \ln P(x_i) \]

\[
= -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - x_i^T w)^2 + \sum_{i=1}^{N} \ln P(x_i).
\]

\[
\frac{\partial \mathcal{L}(w)}{\partial w} = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (-x_i) \cdot 2 \cdot (y_i - x_i^T w) = \frac{1}{\sigma^2} \sum_{i=1}^{N} x_i \cdot (y_i - x_i^T w)
\]

b.w.
Possible Solution:

We set this term equal to 0.

\[
\frac{\partial L(\hat{\mathbf{w}}_{ML})}{\partial \hat{\mathbf{w}}_{ML}} = 0 = \left( \begin{array}{c} 0 \\ 0 \\ \vdots \\ 0 \end{array} \right) = \frac{1}{\sigma^2} \underbrace{\mathbf{X}^T}_{M \times N} \left( \mathbf{y} - \underbrace{\mathbf{X}\hat{\mathbf{w}}_{ML}}_{N \times 1} \right) = \mathbf{0} \]

\[\Leftrightarrow 0 = \mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{x}_{ML} \]

\[\Leftrightarrow \mathbf{X}^T \mathbf{x}_{ML} = \mathbf{X}^T \mathbf{y} \]

\[\Leftrightarrow \mathbf{x}_{ML} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \]

This is exactly the solution of the Least Squares (LS) method.

Alternatively directly by matrix solution:

\[
L(\mathbf{w}) = P(\mathbf{X}) \cdot \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{x} \mathbf{w})^2} = 
\]

\[
= P(\mathbf{X}) \cdot \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{x} \mathbf{w})^T (\mathbf{y} - \mathbf{x} \mathbf{w})} = 
\]

\[
= P(\mathbf{X}) \cdot \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{1}{2\sigma^2} (\mathbf{y}^T \mathbf{y} - \mathbf{w}^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \mathbf{w} + \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w})} = 
\]

\[
= P(\mathbf{X}) \cdot \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{1}{2\sigma^2} (\mathbf{y}^T \mathbf{y} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{y} + \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w})} . 
\]

Derivative:

\[
\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial \ln L(\mathbf{w})}{\partial \mathbf{w}} = -\frac{1}{2\sigma^2} (0 - 2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \mathbf{w}) = 
\]

\[= \frac{1}{\sigma^2} (\mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{X} \mathbf{w}) 
\]

Rest is as before b.w.
Possible Solution:

b) We are looking for \( \hat{w}^{\text{ML}} \) für \( L(w) = P(w)P(D|w) = P(w)P(y|X, w)P(X) = \hat{w}^{\text{MAP}} \), the maximum-a-posteriori estimator.

Log-Likelihood:
\[
L(w) = \ln L(w) = \ln P(w) + \ln P(y|X, w) + \ln P(X) = \\
= \ln \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}w^T w} \right) + \ln \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y^T y - 2w^T X^T y + w^T X^T X w)} \right) + \ln P(X) = \\
= \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}w^T w + \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(y^T y - 2w^T X^T y + w^T X^T X w) + \ln P(X).
\]

Derivative:
\[
\frac{\partial L(w)}{\partial w} = -\frac{1}{2\sigma^2}2w - \frac{1}{2\sigma^2}(-2X^T y + 2X^T X w) = \\
= -\frac{1}{\sigma^2}w + \frac{1}{\sigma^2}(X^T y - X^T X w)
\]

Set equal to 0:
\[
\frac{\partial L(\hat{w}^{\text{MAP}})}{\partial \hat{w}^{\text{MAP}}} = 0
\]
\[
0 = \frac{1}{\sigma^2}X^T y - \frac{1}{\sigma^2}X^T X \hat{w}^{\text{MAP}} - \frac{1}{\sigma^2}w^{\text{MAP}}
\]
\[
\frac{1}{\sigma^2}X^T X \hat{w}^{\text{MAP}} + \frac{1}{\sigma^2} \hat{w}^{\text{MAP}} = \frac{1}{\sigma^2}X^T y
\]
\[
\left( \frac{1}{\sigma^2}X^T X + \frac{1}{\sigma^2}I \right) \hat{w}^{\text{MAP}} = \frac{1}{\sigma^2}X^T y
\]
\[
\left( \frac{1}{\sigma^2}X^T X + \frac{1}{\sigma^2}I \right)^{-1} \left( \frac{1}{\sigma^2}X^T X + \frac{1}{\sigma^2}I \right) \hat{w}^{\text{MAP}} = \left( \frac{1}{\sigma^2}X^T X + \frac{1}{\sigma^2}I \right)^{-1} \frac{1}{\sigma^2}X^T y
\]
\[
\hat{w}^{\text{MAP}} = \left( \frac{1}{\sigma^2}X^T X + \frac{1}{\sigma^2}I \right)^{-1} \frac{1}{\sigma^2}X^T y
\]
\[
\hat{w}^{\text{MAP}} = \frac{1}{\sigma^2} \left( X^T X + \frac{\sigma^2}{\alpha^2} I \right)^{-1} \frac{1}{\sigma^2}X^T y
\]
\[
\hat{w}^{\text{MAP}} = \left( X^T X + \frac{\sigma^2}{\alpha^2} I \right)^{-1} X^T y
\]

The MAP estimator corresponds to the model of the regularized cost function where \( \lambda = \frac{\sigma^2}{\alpha^2} \). The noisy model is thereby a special case of the regularized cost function.

Recall:
\[
\hat{w}_{\text{pen}} = \left( X^T X + \lambda I \right)^{-1} X^T y, \text{ where } \text{cost}_{\text{pen}}(w) = \sum_{i=1}^{N} (y_i - f(x_i, w))^2 + \lambda \sum_{i=0}^{M-1} w_i^2.
\]

Exercise 8-2  Optimal Separating Hyperplane 1

Consider the following dataset consisting of points \( \left( x_i \right) \) in \( \mathbb{R}^2 \). Using a hyperplane, points marked by \( \times \) are to be mapped onto \( \geq 1 \), points marked by \( \circ \) are to be mapped onto \( \leq -1 \).
(a) Find the support vectors.
(b) Determine the equation of one separating hyperplane \( h = x^T w \), optimize it and draw it within the figure.
(c) Compute the margin \( C \).

Possible Solution:

Find support vectors that maximize the margin:

\[ x_1 = \left( \frac{1}{3} \right) \text{ with } y = 1 \text{ and } x_2 = \left( \frac{3}{1} \right), \text{ and } x_3 = \left( \frac{3}{3} \right) \text{ with } y = -1. \]
Possible Solution:

b) 

\[ s_1 = \left( \frac{2}{2} \right), \quad s_2 = \left( \frac{2}{3} \right) \]

linear equation: 
\[ h = 0 = w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 \]

\[ \Rightarrow I): \quad w_0 + w_1 \cdot 2 + w_2 \cdot 2 = 0 \]

\[ II): \quad w_0 + w_1 \cdot 2 + w_2 \cdot 3 = 0 \]

\[ II) - I): \quad w_2 = 0 \implies 2 \cdot w_1 = -w_0 \]

define \( w_1 = 1 \implies w_0 = -2 \]

\[ w = \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} \]

Test if the hyperplane \( w = (-2, 1, 0)^T \) classifies correctly:

\[ x_1 : y_1 (x_1^T w) = 1 \cdot (-2 + 1 \cdot 1 + 0 \cdot 3) = -1 \implies \text{not correct.} \]

Change the signs:

\[ w = \begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix} \implies y_1 (x_1^T w) = 1 \cdot (2 \cdot 1 - 1 \cdot 1) = 1 \quad \text{The solution is correct.} \]

Possible Solution:

c) 

\[ C = \frac{1}{\|\tilde{w}_{opt}\|} \text{, with } \tilde{w}_{opt} \text{ being the optimal } w. \]

Since \( x_1 : y_1 (x_1^T w) = 1 \) we already have the optimal \( w_{opt}. \)

\( \tilde{w}_{opt} \) is \( w_{opt} \) without the bias dimension, therefore \( \tilde{w}_{opt} = \begin{pmatrix} w_{opt1} \\ w_{opt2} \end{pmatrix}. \) Thus the margin is:

\[ C = \frac{1}{\|w_{opt}\|} = \frac{1}{\sqrt{(-1)^2 + (0)^2}} = 1, \quad \text{with } \tilde{w}_{opt} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} \]

Exercise 8-3 Optimal Separating Hyperplane 2

Determine the optimal separating hyperplane of the following dataset, partitioned into two classes \( A \) and \( B \):

\[
A = \left\{ p_1 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}, p_2 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}, p_3 = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}, p_4 = \begin{pmatrix} 2.5 \\ 3 \end{pmatrix}, p_5 = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right\},
\]

\[
B = \left\{ p_6 = \begin{pmatrix} 0.5 \\ 1.5 \end{pmatrix}, p_7 = \begin{pmatrix} -1 \\ 3 \end{pmatrix}, p_8 = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix} \right\}
\]

Instances of class \( A \) shall be labeled with 1, instances of class \( B \) with \(-1\).

Name the support vectors, compute the optimal separating hyperplane and visualize the result. How wide is the margin?
Possible Solution:

Visual solution: \( \{ p_1, p_3, p_6 \} \) are the support vectors, thus:

\[
\begin{align*}
s_1 &= \begin{pmatrix} 1.25 \\ 2.75 \end{pmatrix}, \quad s_2 = \begin{pmatrix} 0.75 \\ 1 \end{pmatrix} \\
\text{linear equation:} & \quad h = 0 = w_0 + w_1 \cdot x_1 + w_2 \cdot x_2
\end{align*}
\]

\( \Rightarrow \) I): \( w_0 + w_1 \cdot 1.25 + w_2 \cdot 2.75 = 0 \)

II): \( w_0 + w_1 \cdot 0.75 + w_2 \cdot 1 = 0 \)

I) – II): \( w_1 \cdot 0.5 + w_2 \cdot 1.75 = 0 \quad \Rightarrow \quad w_1 = -3.5 \cdot w_2 \)

define \( w_2 = 1 \Rightarrow w_1 = -3.5 \cdot w_2 \) in I)

\[ w_0 = 3.5 \cdot 1.25 - 1 \cdot 2.75 = 1.625 \]

Margin-condition: \( y_i (x_i^T w_{opt}) \geq 1 \). Test for correct classification:

\[
y_i \cdot \sum_{j=0}^2 w_j x_{i,j} \overset{p_1}{\overset{p_3}{\overset{p_6}{=}}} 1 \cdot (1.625 - 3.5 \cdot 2 + 1 \cdot 4) = -1.375 \not\geq 1.
\]

Thus the signs of the vector have to be multiplied with \((-1)\) \( \Rightarrow w = \begin{pmatrix} -1.625 \\ 3.5 \\ -1 \end{pmatrix} \).

Now we want to get the minimal \( w \), since the margin \( C = \frac{1}{||w_{opt}||} \) is supposed to be maximal.

Furthermore the margin-condition has to hold and \( y_i (x_i^T w_{opt}) = 1 \) for the support vectors. So far \( \min_{p_1,p_3,p_6} y_i (x_i^T w) = 1.375 \), thus it’s enough to divide by 1.375 to get the optimal separating hyperplane:

\[ w_{opt} = \frac{w}{y_i (x_i^T w_{opt})} = \frac{w}{1.375} = \begin{pmatrix} -1.18 \\ 2.54 \\ -0.72 \end{pmatrix} \]

Possible Solution:

The margin is defined as

\[
C = \frac{1}{||w_{opt}||} = \frac{1}{||w_{opt,1} w_{opt,2}||^T} = \frac{1}{\sqrt{\frac{2.54^2}{1.375^2} + \frac{2.54^2}{1.375^2}}} \approx 1/7.0082 \approx 0.142689.
\]

Two-class dataset

![Two-class dataset diagram](diagram.png)