Exercise 8-1  Human Height

Assume that the height of a human from a finite population is a Gaussian random variable:

\[ P_w(x_i) = \mathcal{N}(x_i; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(x_i - \mu)^2}{2\sigma^2} \right) \]

For independent \( x_i \in \mathbb{R} \) from such a population \( w = (\mu, \sigma)^T \in \mathbb{R}^2 \) holds

\[ P_w(x_1, \ldots, x_N) = \prod_{i=1}^{N} P_w(x_i) = \prod_{i=1}^{N} \mathcal{N}(x_i; \mu, \sigma^2) = \]
\[ = \frac{1}{(2\pi \sigma^2)^\frac{N}{2}} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2 \right) \]

a) Determine the maximum likelihood estimator of \( P_w(x_1, \ldots, x_N) \).

b) Compute the corresponding estimators for the four height datasets in the file `body_sizes.txt` and visualize the respective distributions. How does the estimator reflect the understanding of the underlying data?
Possible Solution:

a)

\[ l(\mu, \sigma) = \log P_w(x_1, \ldots, x_N) = \log \frac{1}{\sqrt{2\pi \sigma^2}} \sum_{i=1}^{N} (x_i - \mu)^2 \]

\[ \frac{\partial l(\mu, \sigma)}{\partial \mu} = \frac{\partial}{\partial \mu} \left( -\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2 \right) = 0 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} 2 \cdot (x_i - \mu) \cdot (-1) = \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu) = \frac{1}{\sigma^2} \left( \sum_{i=1}^{N} x_i \right) - N \cdot \mu \]

\[ \frac{\partial l(\hat{\mu}^{ML}, \sigma)}{\partial \hat{\mu}^{ML}} \bigg|_0 = \hat{\mu}^{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i \]

\[ \frac{\partial l(\mu, \hat{\sigma}^{ML})}{\partial \hat{\sigma}^{ML}} \bigg|_0 = N = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu}^{ML})^2 \Rightarrow (\hat{\sigma}^{ML})^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu}^{ML})^2 \]

\[ \frac{\partial l(\mu, \sigma)}{\partial \sigma} = \frac{\partial}{\partial \sigma} \left( -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2 \right) = -\frac{N}{2} \cdot \frac{1}{2\pi\sigma^2} \cdot 4\pi\sigma - \left( \frac{1}{2} (-2) \frac{1}{\sigma^3} \sum_{i=1}^{N} (x_i - \mu)^2 \right) = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{N} (x_i - \mu)^2 \]

b) All values in cm:

\[ \hat{\mu}^{ML} = (161.5536, 153.7481, 154.5920) \]
\[ \hat{\sigma}^{ML} = (34.67525, 35.48248, 36.18142) \]

Estimator does not really help to understand the data.
Possible Solution:

Exercise 8-2  Lineare Regression with Gaussian Noise

Let \( D_i = (x_{i1}, \ldots, x_{iM}, y_i)^T \), be a dataset of size \( N \) with \( M \) features and an output \( y \) which depends linearly on \( X \). Due to erroneous measurements the inputs 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:

\[
P(w) = \frac{1}{(2\pi \alpha^2)^\frac{M}{2}} e^{-\frac{1}{2\alpha^2} \sum_{j=0}^{M-1} w_j^2}
\]

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 \( \ln(ab) = \ln a + \ln b \) (which is not the same as \( e^{a+b} = e^a \cdot e^b \)).

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

\[
\frac{\partial l(w)}{\partial w} = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (-x_i) \cdot 2 \cdot (y_i - x_i^T w) =
\]
\[
= \frac{1}{\sigma^2} \sum_{i=4}^{N} x_i \cdot \underbrace{(y_i - x_i^T w)}_{1 \times 1}
\]

b.w.
Possible Solution:

We set this term equal to 0.

\[ \frac{\partial l(\hat{w}^{\text{ML}})}{\partial \hat{w}^{\text{ML}}} = 0 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \frac{1}{\sigma^2} \begin{pmatrix} X^T \\ y - \hat{w}^{\text{ML}} \end{pmatrix}_{M \times N}^{N \times 1} \]

\[ \Leftrightarrow 0 = X^T y - X^T \hat{w}^{\text{ML}} \]

\[ \Leftrightarrow X^T \hat{w}^{\text{ML}} = X^T y \]

\[ \Leftrightarrow \hat{w}^{\text{ML}} = (X^T X)^{-1} X^T y \]

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

Alternatively directly by matrix solution:

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

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

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

\[ \text{Derivative:} \]

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

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

Rest is as before
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:

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

Derivative:

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

Set equal to 0:

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

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:

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