Ludwig-Maximilians-Universitaet Muenchen 17.06.2015
Institute for Informatics
Prof. Dr. Volker Tresp
Gregor Jossé
Johannes Niedermayer

Machine Learning and Data Mining

Summer 2015

Exercise Sheet 8

Presentation of Solutions to the Exercise Sheet on the 10.06.2015

Exercise 8-1 Human Height

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

$$P_{\mathbf{w}}(\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(\mathbf{x}_i - \mu)^2}{2\sigma^2}\right)$$

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

$$\begin{aligned} P_{\mathbf{w}}(\mathbf{x}_1, \dots, \mathbf{x}_N) &= \prod_{i=1}^{N} P_{\mathbf{w}}(\mathbf{x}_i) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{x}_i; \mu, \sigma^2) = \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\mathbf{x}_i - \mu)^2\right) \end{aligned}$$

a) Determine the maximum likelihood estimator of $P_{\mathbf{w}}(\mathbf{x}_1, \dots, \mathbf{x}_N)$.

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)

$$l(\mu,\sigma) = \log P_{\mathbf{w}}(\mathbf{x}_1,\ldots,\mathbf{x}_N) = \overbrace{\log 1}^{=0} - \log\left(2\pi\sigma^2\right)^{\frac{N}{2}} + \left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\mathbf{x}_i-\mu)^2\right)$$

$$\frac{\partial l(\mu,\sigma)}{\partial \mu} = \frac{\partial\left(-\log(2\pi\sigma^2)^{\frac{N}{2}} - \left(\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\mathbf{x}_i-\mu)^2\right)\right)}{\partial \mu} =$$

$$= 0 - \left(\frac{1}{2\sigma^2}\sum_{i=1}^{N} 2\cdot(\mathbf{x}_i-\mu)\cdot(-1)\right) = \frac{1}{\sigma^2}\sum_{i=1}^{N}(\mathbf{x}_i-\mu) =$$

$$= \frac{1}{\sigma^2}\left(\left(\sum_{i=1}^{N}\mathbf{x}_i\right) - N\cdot\mu\right)$$

$$\frac{\partial l(\hat{\mu}^{\mathrm{ML}},\sigma)}{\partial \hat{\mu}^{\mathrm{ML}}} \stackrel{!}{=} 0 \Rightarrow \hat{\mu}^{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

$$\frac{\partial l(\mu,\sigma)}{\partial \sigma} = \frac{\partial\left(-\frac{N}{2}\log(2\pi\sigma^2) - \left(\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\mathbf{x}_i-\mu)^2\right)\right)}{\partial \sigma} =$$

$$= -\frac{N}{2}\frac{1}{2\pi\sigma^2}\cdot 4\pi\sigma - \left(\frac{1}{2}(-2)\frac{1}{\sigma^3}\sum_{i=1}^{N}(\mathbf{x}_i-\mu)^2\right) =$$

$$= -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(\mathbf{x}_i-\mu)^2$$

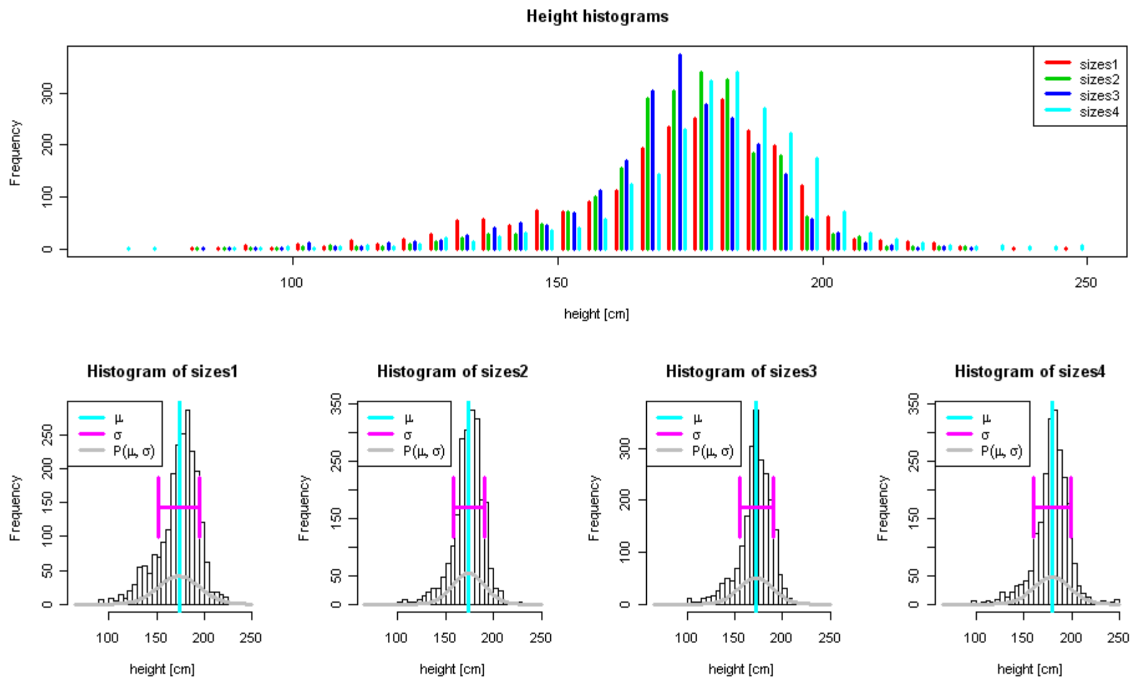
$$\frac{\partial l(\mu,\hat{\sigma}^{\mathrm{ML}})}{\partial \hat{\sigma}^{\mathrm{ML}}} \stackrel{!}{=} 0 \Rightarrow N = \frac{1}{(\hat{\sigma}^{\mathrm{ML}})^2}\sum_{i=1}^{N}(\mathbf{x}_i-\mu)^2 \Rightarrow \left(\hat{\sigma}^{\mathrm{ML}}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\hat{\mu}_{\mathrm{ML}})^2$$

b) All values in cm:

$$\hat{\mu}_i^{\mathrm{ML}} = (161.5536, 153.7481, 154.5920)\ ,\ \hat{\mu}^{\mathrm{ML}} = 156.6312$$
$$\hat{\sigma}_i^{\mathrm{ML}} = (34.67525, 35.48248, 36.18142)\ ,\ \hat{\sigma}^{\mathrm{ML}} = 35.61861$$

Estimator does not really help to understand the data.

Possible Solution:



Also the intervals $0 \pm \sigma;\ 0 \pm 2\sigma;\ 0 \pm 3\sigma$ have counterintuitive frequencies.

Exercise 8-2 Lineare 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 by which depends linearly on $\mathbf{X}$. Due to erroneous measurements the inputs the inputs are noisy, i.e.:

$$y_i = x_i^T \mathbf{w} + \epsilon_i \ ,$$

where ϵ_i is the noise of data point i. Furthermore, assume ϵ 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 $\mathbf{X}$ and the model $\mathbf{w}$, we can then model the distribution of $\mathbf{y}$ as follows:

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

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

You may assume that the $\mathbf{w}$ are distributed independently of the input data $\mathbf{X}$.

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

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

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

Possible Solution:

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

$$L(\mathbf{w}) = P(\mathbf{y}|\mathbf{X}, \mathbf{w}) \cdot P(\mathbf{X}) \ .$$

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

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

$$\begin{aligned} L(\mathbf{w}) &= \prod_{i=1}^{N} P(y_i, x_i|\mathbf{w}) = \prod_{i=1}^{N} P(y_i|x_i, \mathbf{w}) \cdot P(x_i) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - x_i^T\mathbf{w})^2} \cdot P(x_i) \ . \end{aligned}$$

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 $e^{a+b} = e^a \cdot e^b$).

$$\begin{aligned} l(\mathbf{w}) = \ln L(\mathbf{w}) &= \ln\left(\prod_{i=1}^{N} P(y_i|x_i, \mathbf{w}) \cdot P(x_i)\right) = \sum_{i=1}^{N} \ln\left(P(y_i|x_i, \mathbf{w}) \cdot P(x_i)\right) = \\ &= \sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - x_i^T\mathbf{w})^2} \cdot P(x_i)\right) = \\ &= \underbrace{\sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi\sigma^2}}}_{=\underbrace{N \ln 1}_{=0} - N \ln\sqrt{2\pi\sigma^2}} + \sum_{i=1}^{N} \underbrace{\ln e^{-\frac{1}{2\sigma^2}(y_i - x_i^T\mathbf{w})^2}}_{\ln e^{f(x)} = f(x)} + \sum_{i=1}^{N} \ln P(x_i) = \\ &= -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - x_i^T\mathbf{w})^2 + \sum_{i=1}^{N} \ln P(x_i) \ . \end{aligned}$$

$$\begin{aligned} \frac{\partial l(\mathbf{w})}{\partial \mathbf{w}} &= -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (-x_i) \cdot 2 \cdot (y_i - x_i^T\mathbf{w}) = \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{N} \underbrace{x_i}_{M \times 1} \cdot \underbrace{(y_i - x_i^T\mathbf{w})}_{1 \times 1} \end{aligned}$$

b.w.

Possible Solution:

We set this term equal to 0.

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

$$\begin{aligned}
\Leftrightarrow 0 &= \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}^{\mathrm{ML}} \\
\Leftrightarrow \mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}^{\mathrm{ML}} &= \mathbf{X}^T\mathbf{y} \\
\Leftrightarrow \hat{\mathbf{w}}^{\mathrm{ML}} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}$$

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

Alternatively directly by matrix solution:

$$\begin{aligned}
L(\mathbf{w}) &= P(\mathbf{X}) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{w})^2} = \\
&= P(\mathbf{X}) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w})} = \\
&= P(\mathbf{X}) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y}-\mathbf{w}^T\mathbf{X}^T\mathbf{y}-\mathbf{y}^T\mathbf{X}\mathbf{w}+\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w})} = \\
&= P(\mathbf{X}) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y}-2\mathbf{w}^T\mathbf{X}^T\mathbf{y}+\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w})} .
\end{aligned}$$

Derivative:

$$\begin{aligned}
\frac{\partial l(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial \ln L(\mathbf{w})}{\partial \mathbf{w}} &= -\frac{1}{2\sigma^2}(0 - 2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{w}) = \\
&= \frac{1}{\sigma^2}(\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\mathbf{w})
\end{aligned}$$

Rest is as before

b.w.

Possible Solution:

b)

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

Log-Likelihood:

$$
\begin{aligned}
l(\mathbf{w}) &= \ln L(\mathbf{w}) = \ln P(\mathbf{w}) + \ln P(\mathbf{y}|\mathbf{X},\mathbf{w}) + \ln P(\mathbf{X}) = \\
&= \ln\left(\frac{1}{\sqrt{2\pi\alpha^2}^M} e^{\left(-\frac{1}{2\alpha^2}\mathbf{w}^T\mathbf{w}\right)}\right) + \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y}-2\mathbf{w}^T\mathbf{X}^T\mathbf{y}+\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w})}\right) + \ln P(\mathbf{X}) = \\
&= \ln\frac{1}{\sqrt{2\pi\alpha^2}^M} - \frac{1}{2\alpha^2}\mathbf{w}^T\mathbf{w} + \ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y} - 2\mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}) + \ln P(\mathbf{X})\,.
\end{aligned}
$$

Derivative:

$$
\begin{aligned}
\frac{\partial l(\mathbf{w})}{\partial \mathbf{w}} &= -\frac{1}{2\alpha^2}2\mathbf{w} - \frac{1}{2\sigma^2}(-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{w}) = \\
&= -\frac{1}{\alpha^2}\mathbf{w} + \frac{1}{\sigma^2}(\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\mathbf{w})
\end{aligned}
$$

Set equal to 0:

$$
\begin{aligned}
\frac{\partial l(\hat{\mathbf{w}}^{\text{MAP}})}{\partial \hat{\mathbf{w}}^{\text{MAP}}} &= 0 \\
0 &= \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y} - \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}^{\text{MAP}} - \frac{1}{\alpha^2}\hat{\mathbf{w}}^{\text{MAP}} \\
\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}^{\text{MAP}} + \frac{1}{\alpha^2}\hat{\mathbf{w}}^{\text{MAP}} &= \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y} \\
\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\alpha^2}I\right)\hat{\mathbf{w}}^{\text{MAP}} &= \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y} \\
\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\alpha^2}I\right)^{-1}\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\alpha^2}I\right)\hat{\mathbf{w}}^{\text{MAP}} &= \left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\alpha^2}I\right)^{-1}\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y} \\
\hat{\mathbf{w}}^{\text{MAP}} &= \left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\alpha^2}I\right)^{-1}\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y} \\
\hat{\mathbf{w}}^{\text{MAP}} &= \frac{1}{\frac{1}{\sigma^2}}\left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\alpha^2}I\right)^{-1}\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y} \\
\hat{\mathbf{w}}^{\text{MAP}} &= \left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\alpha^2}I\right)^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}
$$

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

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