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:

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

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

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

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?

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 \( y_i \) 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:

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

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)?