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

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

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

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?

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)?