Ludwig-Maximilians-Universitaet Muenchen 06.05.2015
Institute for Informatics
Prof. Dr. Volker Tresp
Gregor Jossé
Johannes Niedermayer

Machine Learning and Data Mining

Summer 2015

Exercise Sheet 2

Presentation of Solutions to the Exercise Sheet on the 04.05.2015

Exercise 2-2 Linear Regression

Let X be a variable providing the data and its occurrences Y:

x	3	4	5	6	7	8
y	150	155	150	170	160	175

a) Presume the model exhibits the following linear relation:
$y_i = \beta_0 + \beta_1 x_i = x^T w$
Use the least squares-estimator introduced in the lecture to determine w.

b) Now, presume the non-linear relation
$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 = x^T w$
and, again, determine w.

c) How could the empiric quadratic error between model and data be visualized? Explain and sketch your suggestion in two as well as in three dimensions on arbitrary data.

d) Which of the models a) and b) is better? Compute the average quadratic error and evaluate the models. How could a better model be realized?

Hint: Matrix arithmetic need not be done manually. You can use `R`, Maple, Octave or Python.

Possible Solution:

The quadratic error is defined as:
$J_N(w) = \sum (y_i - f(x_i,w))^2$
We derive $J_N(w)$ w.r.t. w and set the derivative $= 0$. From this we obtain the following estimator:

$$\hat{w}_{LS} = (X^T X)^{-1} X^T y \quad \text{mit} \quad X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,m} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,m} \\ \vdots & & & & \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,m} \end{bmatrix}$$

Assuming a linear model, we get: $y_i = \beta_0 + \beta_1 x_i = x^T w$

$$\text{mit} \quad x = \begin{pmatrix} 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{pmatrix} \quad \text{folgt} \quad X = \begin{bmatrix} 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \\ 1 & 8 \end{bmatrix}. \text{ Additionally } \; y = \begin{pmatrix} 150 \\ 155 \\ 150 \\ 170 \\ 160 \\ 175 \end{pmatrix} \quad \text{and} \quad \hat{w}_{LS} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

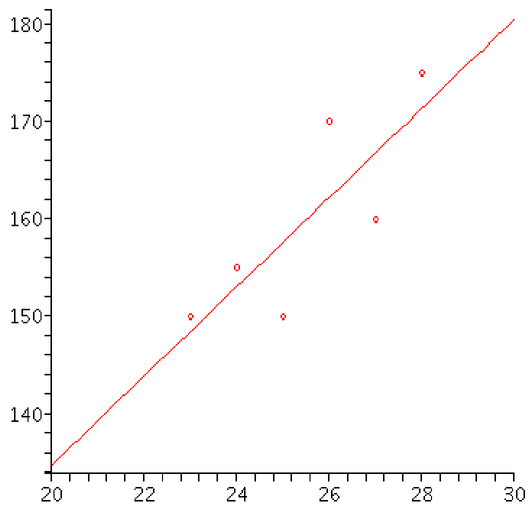
The 1 in the first column represents the translation along the y-axis, i.e., the constant bias of the neuron. The second column represents the data of the input variable x.

$(X^T X)^{-1} X^T$ is a bit more complicated to compute by hand, but easily done by machine:

$$\hat{w}_{LS} = (X^T X)^{-1} X^T y \approx \begin{bmatrix} 0.95 & 0.64 & 0.32 & 0.01 & -0.30 & -0.62 \\ -0.14 & -0.09 & -0.03 & 0.03 & 0.09 & 0.14 \end{bmatrix} \begin{pmatrix} 150 \\ 155 \\ 150 \\ 170 \\ 160 \\ 175 \end{pmatrix}$$

$$\hat{w}_{LS} \approx \begin{bmatrix} 134.86 \\ 4.57 \end{bmatrix}$$

Hence, the linear model
$y_i = \beta_0 + \beta x_i = x^T w$
corresponds to the straight line:
$y_i = \beta_1 + \beta_1 x_i = x^T w = 134.86 + 4.57 x_i$



Possible Solution:

Assuming a non-linear model, we get: $y_i = a + \beta_1 x_i + \beta_2 x_i^2 = x^T w$

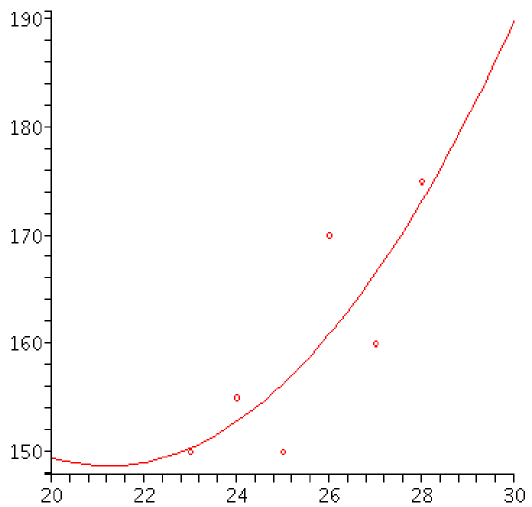
Thus: $X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2}^2 \\ 1 & x_{2,1} & x_{2,2}^2 \\ \vdots & & \\ 1 & x_{n,1} & x_{n,2}^2 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \\ 1 & 7 & 49 \\ 1 & 8 & 64 \end{bmatrix}$

$y = \begin{pmatrix} 150 \\ 155 \\ 150 \\ 170 \\ 160 \\ 175 \end{pmatrix}$ und $\hat{w}_{LS} = \begin{pmatrix} a \\ \beta_1 \\ \beta_2 \end{pmatrix}$

$$\hat{w}_{LS} = (X^T X)^{-1} X^T y \approx \begin{bmatrix} 3.39 & 0.15 & -1.63 & -1.94 & -0.79 & 1.82 \\ -1.13 & 0.11 & 0.76 & 0.81 & 0.28 & -0.84 \\ 0.09 & -0.02 & -0.07 & -0.07 & -002 & 0.09 \end{bmatrix} \begin{pmatrix} 150 \\ 155 \\ 150 \\ 170 \\ 160 \\ 175 \end{pmatrix}$$

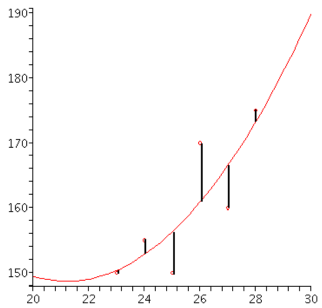
$$\hat{w}_{LS} \approx \begin{bmatrix} 149.5 \\ -1.321 \\ 0.536 \end{bmatrix}$$

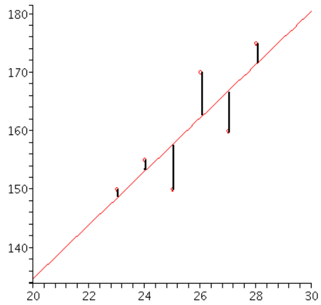
Hence, the non-linear model:
$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 = x^T w$
corresponds to the 2nd order polynomial:
$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 = x^T w = 149.5 - 1.321 x_i + 0.536 x_i^2$

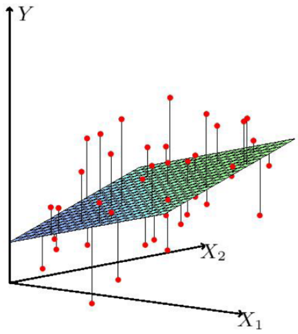


Possible Solution:

c) Visualization of the quadratic error: The error is the sum of deviations from the straight line (or hyperplane).







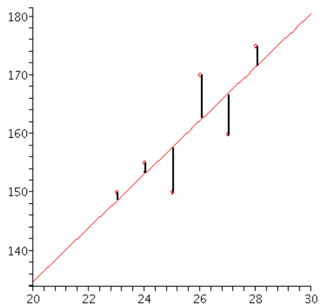
Possible Solution:

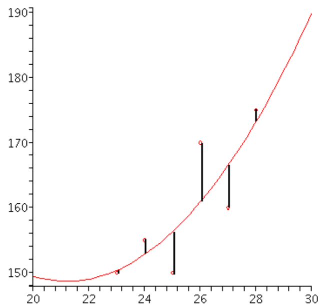
d) Computation of the mean squared error (MSE):
Linear model: $y_i = \beta_0 + \beta_1 x_i = x^T w = 134.86 + 4.57 x_i$
Non-linear model: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 = x^T w = 149.5 - 1.321 x_i + 0.536 x_i^2$
MSE: $MSE(f,g) = E||f(X) - g(X)||^2 = E||f(X) - f(\hat{X})||^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
One may reduce the error by extending the model X, e.g., by employing higher order polnymonials. However, this may cause overfitting.

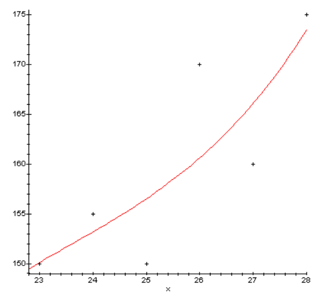
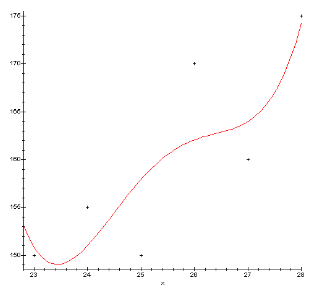
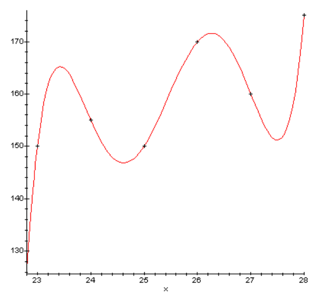
$y_i = \beta_0 + \beta_1 x_i$
$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$
$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$
$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4$
$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \beta_5 x_i^5$

Werte:

x	3	4	5	6	7	8	
y	150	155	150	170	160	175	
$f_{poly1}(x)$	148,57	153,14	157,71	162,29	166,86	171,42	MSE: 30,71
$f_{poly2}(x)$	150,36	152,79	156,28	160,86	166,5	173,21	MSE: 28.93
$\epsilon_{poly3}(x)$							MSE: 28,84
$\epsilon_{poly4}(x)$							MSE: 26,45
$\epsilon_{poly5}(x)$							MSE: 0





Possible Solution: