Machine Learning, Lecture 5: Linear regression

S. Nõmm

[1]Department of Computer Science, Tallinn University of Technology

05.03.2015

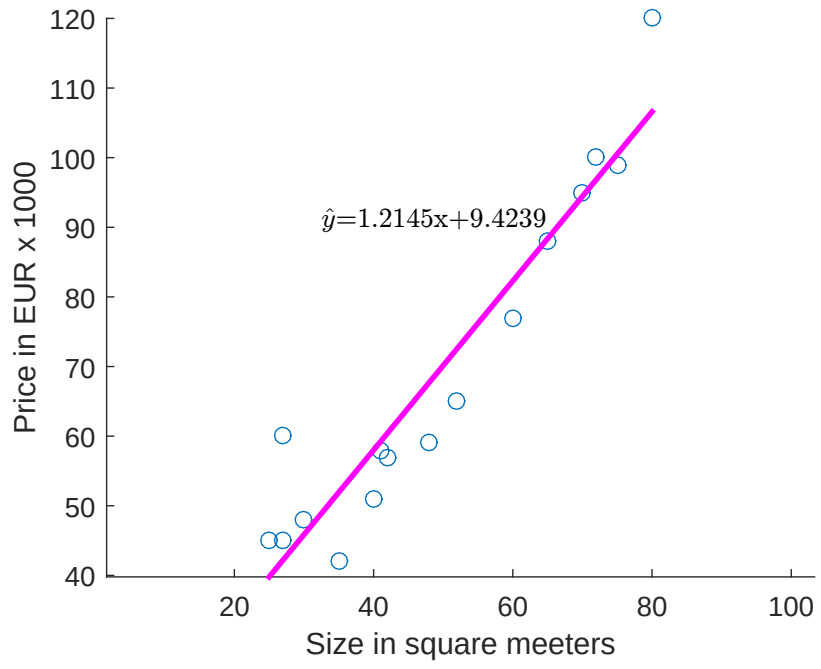
Motivation

- linear regression is frequently referred as the "work horse" of statistics and machine learning. [Machine Learning, K.P. Murphy]
- The goal is to predict continuous values.
- Based on the training data set find parameters of the model (coefficients of the function).
- Use the model (function) to predict (compute) the value of dependent variable for the given value(s) of independent variable(s).

Practical Approach (Very simple case)

Let X represent the size of the apartment in square meeters and y - price of the apartment in thousands of EUR. The goal is to train a model $\hat{y} = ax + b$ able to predict the price of the apartment on the basis of its size.

$$y = \begin{pmatrix} 45 \\ 60 \\ 45 \\ 48 \\ 42 \\ 51 \\ 57 \\ 58 \\ 59 \\ 65 \\ 77 \\ 88 \\ 95 \\ 100 \\ 99 \\ 120 \end{pmatrix} \qquad X = \begin{pmatrix} 25 \\ 27 \\ 27 \\ 30 \\ 35 \\ 40 \\ 42 \\ 41 \\ 48 \\ 52 \\ 60 \\ 65 \\ 70 \\ 72 \\ 75 \\ 80 \end{pmatrix}$$



About the model

- The model coefficients were computed on the basis of a random sample.
- Could it be that for another sample it would be impossible to identify the parameters? Or another sample would result in completely different model?
- If my model is trustworthy how good/precise is it?
- Could the model be improved?
- Just a prediction or some other goals?

Goodness of the model

- Determination coefficient.
- Significance of the model.
- Standard error.
- Significance of each variable.
- Normal probability plot
- Residual plots
- ...

Model building

Let $X = \{x_1, x_2, \ldots, x_n\}$ is the set of independent variables available for the modelled process. The goal is to select the subset X_s which is optimal for predicting variable y.

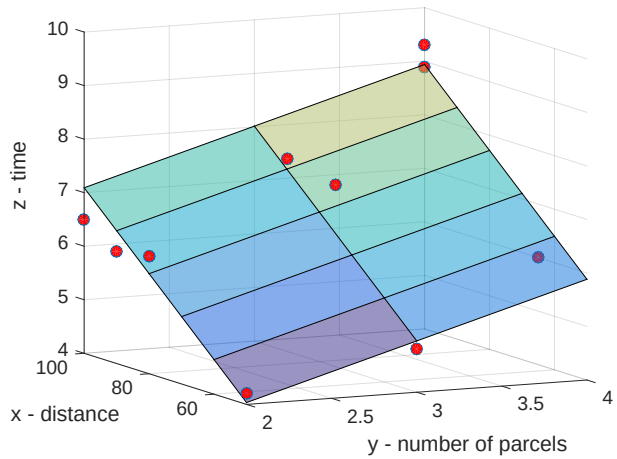
- Using all the available variables may lead:
  - Overparametrization
  - Unnecessary computational complexity.
- Using too few variables may lead:
  - Loss of precision.
  - Inadequate behaviour.

One needs to determine if adding/removing variable effected the model quality.

One more very simple example

Let X be the matrix where the first column contains the information about distances delivery agent has covered to deliver all the parcels and the second column contains the information about the number of parcels. Dependent variable y is the time to complete the assignment.

$$X = \begin{pmatrix} 100 & 4 \\ 50 & 3 \\ 100 & 4 \\ 100 & 2 \\ 50 & 2 \\ 80 & 2 \\ 75 & 3 \\ 65 & 4 \\ 90 & 3 \\ 90 & 2 \end{pmatrix}; \quad y = \begin{pmatrix} 9.3 \\ 4.8 \\ 8.9 \\ 6.5 \\ 4.2 \\ 6.2 \\ 7.4 \\ 6 \\ 7.6 \\ 6.1 \end{pmatrix}$$



Example

Let us construct the model with just one variable describing distance:

$$\begin{aligned} \hat{y} &= 0.067826087x + 1.273913043 \\ R^2 &= 0,664071312 \\ S &= 1,001791873 \\ F &= 15,81457814 \\ SSR &= 8,028695652 \end{aligned}$$

Let us now construct the model with two variables:

$$\begin{aligned} \hat{y} &= 0.061134599x_1 + 0.923425367x_2 - 0.868701467 \\ R^2_{adj} &= 0,876300111 \\ S &= 0,573142152 \\ F &= 32,87836743 \\ SSR &= 2,299443486 \end{aligned}$$

Could one conclude that the second model is better?

Example

The null hypothesis is that there is no change in quality and the alternative hypothesis is the opposite:

$$\begin{aligned} H_0 : SSR_1 - SSR_2 &= 0 \\ H_1 : SSR_1 - SSR_2 &\neq 0 \end{aligned}$$

For the level of significance $\alpha = 0.05$ Rejection rule is **Reject** H_0 if $F > 5.5914$. For this particular case $F = 17,59536505$. This rejects the hypothesis H_0. $SSR_1 - SSR_2 \neq 0$ which in turn means that adding variable x_2 - parcels number has improved the model quality.

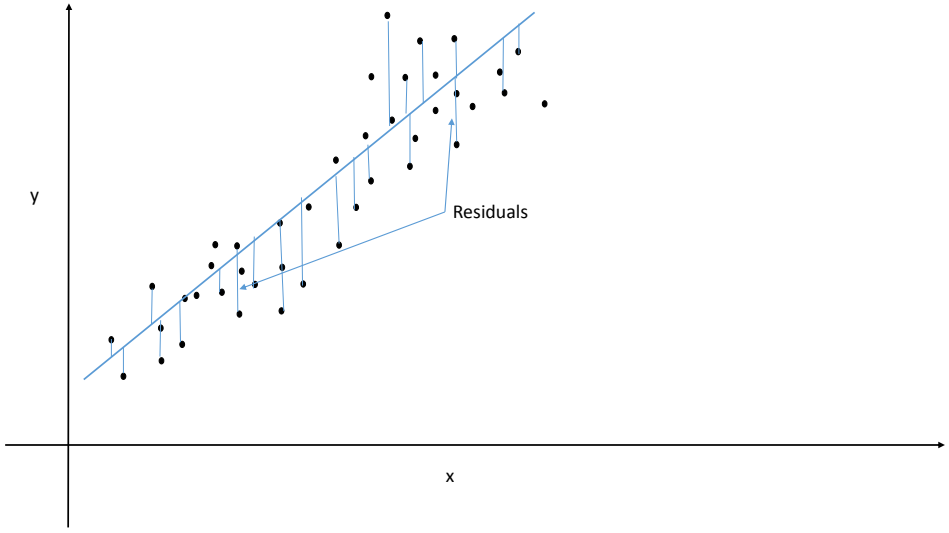
- More examples of practical implementation will be given during the practice session.
- What kind of method(s) was/were used to find the coefficients of the model?

Least squares method

- The goal is to find the values for the coefficients a_i and intercept b that would minimize sum of squared residuals:

$$SSR = \sum r_i^2 = \sum (y_i - \hat{y}_i)^2$$

- Why does it work? Are there any other methods?



Formal approach

- The goal is to find the parameters of the linear function that best fits the data

$$y(x) = \boldsymbol{w}^T \boldsymbol{x} + \epsilon = \sum_{j=1}^{D} \omega_j x_j + \epsilon$$

- It is often assumed that ϵ has a gaussian distribution

$$\epsilon \sim \mathcal{N}(\mu, \sigma^2).$$

- One may rewrite the model in the following form

$$p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N} + \big(y \mid \mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x})\big).$$

$\mu = (\boldsymbol{w}^T \boldsymbol{x}$ and $\sigma^2(x) = \sigma^2$ in this case $\theta = ((w), \sigma^2)$

  - Example: the case of one dimensional input is $\mu(\boldsymbol{x}) = \omega_0 + \omega_1 x = \boldsymbol{w}^T \boldsymbol{x}$.

Maximal likelihood

- Squared cost function leads convex objective function, which have only one optimum, which is global.
- Compute MLE and find parameters which maximize log likelihood function or minimize negative log likelihood function.

$$\ell(\boldsymbol{\theta}) = \log(p \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \log p(y_i \mid \boldsymbol{x_i}, \theta)$$

- The log likelihood of the defined model is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \left[\left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} \left(y_i - \boldsymbol{\theta}^T \boldsymbol{x}_i \right)^2 \right) \right]$$

$$= \frac{-1}{2\sigma^2} SSR(\boldsymbol{\theta}) - \frac{N}{2} \log(2\pi\sigma^2)$$

where $SSR = \sum_{i=1}^{N} (y_i \boldsymbol{\theta}^T \boldsymbol{x_i})^2$

Gradient descend

- Iterative technique, parameters are updated on each iteration.
- Initialize parameters randomly.
- Each parameter is updated in the direction of its negative gradient.
- For each $\boldsymbol{\theta}_i$ repeat in parallel until converge

$$\theta_j^{(k+1)} = \theta_j^{(k)} \alpha \frac{\partial J(\boldsymbol{\theta})}{\partial \theta_j}$$

  - α is a learning rate.
  - this is first order algorithm.

Gradient descent for the least squares

- Find the derivative of the objective function

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \theta_j} = \frac{\partial}{\partial \theta_j}\frac{1}{2}(\hat{y} - y)^2 = (\hat{y} - y)\frac{\partial}{\partial \theta_j}\hat{y} = (\hat{y} - y)x_j$$

- For the entire data set

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \theta_j} = \sum_{i=1}^{m}(\hat{y}_i - y_i)x_{i,j}$$

- Update rule:

$$\theta_j^{k+1} = \theta_j^k - \alpha\sum_{i=1}^{m}(\hat{y}_i - y_i)x_{i,j}.$$

Non-linear functions and linear regression

- Replace $\boldsymbol{x}$ with some nonlinear functions $\phi(\boldsymbol{x})$

$$\hat{y} = \boldsymbol{\theta}^T \phi(\boldsymbol{x})$$

- This operation is called *basis function expansion*.
- Example: *polynomial regression*

$$\phi(x) = [1, x, x^2, \ldots, x^d]$$

- While the function is non-linear, it is still linear in its parameters.

Polynomial regression

- On the one hand polynomial function allows to fit the data with a very high precision, which is achieved by large positive and negative values of the coefficients.
- On the other hand small changes in the data will lead greater changes in the coefficients.
- Makes it problematic to model noisy data.

Encouraging small er values of the parameters

- Use a zero-mean Gaussian prior:

$$P(\boldsymbol{\theta}) = \prod_j \mathcal{N}(\theta_j \mid 0, \tau^2)$$

- Corresponding log likelihood function:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{m} \log \mathcal{N}(y_i \mid \boldsymbol{\theta}^T \boldsymbol{x}_i, \sigma^2) \sum_{j=1}^{n} \log \mathcal{N}(\theta_j \mid 0, \tau^2)$$

Ridge regression

- Regularized objective function:

$$J(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{m}(y_i - \boldsymbol{\theta}^T \boldsymbol{x}_i)^2 \frac{\lambda}{2} \mid\mid \boldsymbol{\theta} \mid\mid_2^2$$

  where $\lambda = \sigma^2/\tau^2$
- The regularized linear regression is called *ridge regression*
- Ridge regression normal equation normal equation is given by

$$\boldsymbol{\theta}_{ridge} = (\lambda \boldsymbol{I} - \boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T Y$$

- Adding Gaussian prior to parameters is called ℓ_2 regulaarization