Introduction to Machine Learning

Lecture 3: Regression

Alexis Zubiolo
alexis.zubiolo@gmail.com

Data Science Team Lead @ Adcash

November 3, 2016

Regression in Machine Learning

This lecture is about regression in Machine learning.

Reminder: In regression, the output y is **continous**.

Example:

- **Price estimation**: $y =$ price (*e.g.* 50000 BGN for a house)
- **Predicting the future** (*e.g.* weather forecast): $y =$ temperature

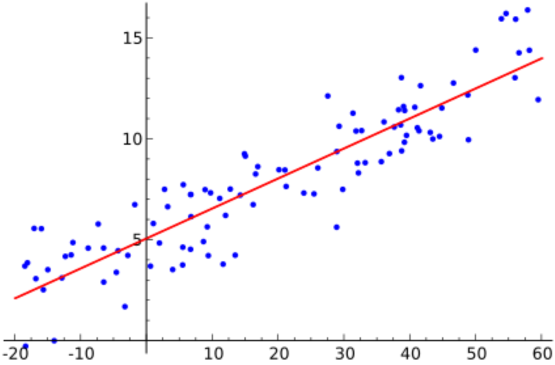
Regression in Machine Learning: Applications

Domains of application:

- Price estimation/prediction
- Weather forecast
- Production quantity estimation
- Stock option price prediction
- Fit statistical model to data
- Physics & chemistry
- ... and others

Linear and polynomial regression

Purpose of regression: **approximate solutions** of **overdetermined systems**.



In this course, we will see

- Linear regression
- Polynomial regression

Linear regression

Linear regression

Principal components:

- Old problem (least-squares method usually credited to Carl Friedrich Gauss in 1795)
- Several ways to approximate the data
  - Linear model
  - Polynomial model (remember kernels from SVMs)
  - Fit a distribution
  - ...
- Several ways to formulate the problem
  - Least Squares
  - Support Vector regression
  - ...
- Several ways to solve the problem

Linear regression with ordinary least-squares

Linear regression: Estimate y as a **linear** function of x:

$$\hat{y} = w^T x$$

living area (m^2)	**# bedrooms**	price (1000's euros)
50	**1**	30
76	**2**	48
26	**1**	12
102	**3**	90

living area (m^2)	**# bedrooms**	price (1000's euros)
50	**1**	30
76	**2**	48
26	**1**	12
102	**3**	90
61	**2**	?

Variable standardisation

Variables have various magnitudes. Example:

- Living area: Up to a few hundreds m^2
- Price: Up to a few 100 000s BGN (and even more)

Variable standardisation

Variables have various magnitudes. Example:

- Living area: Up to a few hundreds m^2
- Price: Up to a few 100 000s BGN (and even more)

It is possible to calculate the **standard score** z of a variable x

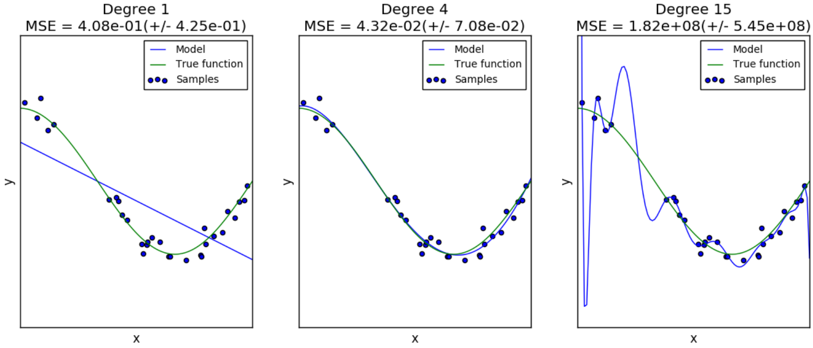
$$z = \frac{x - \mu}{\sigma}$$

where

- μ is the mean of the variable
- σ is its standard deviation

Overfitting and underfitting

$$y = \cos\left(\frac{3\pi}{2}x\right)$$



Parameter selection

Parameter selection plays a huge role in the regression performance.

Fitting a distribution

Thank you! Questions?