Introduction to Machine Learning
Curve fitting and model validation

Nikolay Manchev

29 September 2016

London Machine Learning Study Group

Housekeeping

Next events

http://www.meetup.com/London-Machine-Learning-Study-Group

Follow me

https://twitter.com/nikolaymanchev

Slides and code

Available at https://github.com/nmanchev/MachineLearningStudyGroup

Outline

Assumptions in Linear Regression

Polynomial Regression

Model validation

Assumptions in Linear Regression

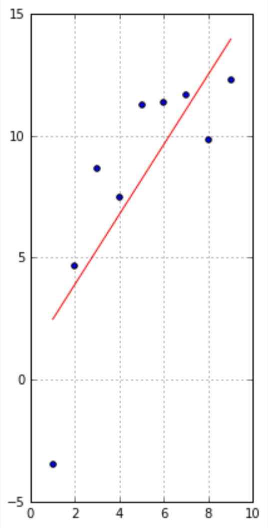
Fitting a linear regression model

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}^T$$
$$\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}^T$$

Cost function minimisation

Minimising the cost function leads us to the coefficients of the best fitting line.

$$J(\boldsymbol{w}) = \frac{1}{2N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$



Matrix Notation

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1D} \\ 1 & x_{21} & x_{22} & \dots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{N1} & x_{N2} & \dots & x_{ND} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix}$$

- Hypothesis: $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{w}$
- Cost function
  - $J(\boldsymbol{w}) = \frac{1}{2N} \sum_{i=1}^{N} (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^T (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})$
  - $\frac{\partial}{\partial w} J(\boldsymbol{w}) = \frac{\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})}{N}$
- Solving for $\boldsymbol{w}$ using normal equations: $\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

Linear Regression Assumptions

Four assumptions of multiple linear regression you should always test [OW02]

- Linear relationship between the independent and dependent variable(s)
- Variables are normally distributed
- Variables are measured reliably
- Assumption of homoscedasticity

Linear Relationship Assumption

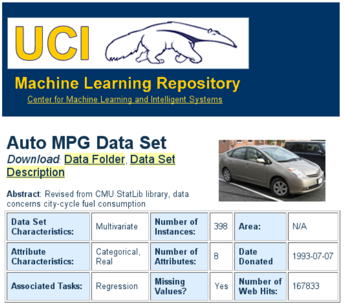
How to check the linear relationship assumption

- Use previous research / domain knowledge
- Visual inspection of the variables
- Examination of plots

Data Set Example

UCI Machine Learning Repository – `archive.ics.uci.edu/ml`

- Great resource for Machine Learning data sets
- Over 330 freely available sets
- Auto MPG Data Set
  - Fuel consumption in MPG
  - Attributes: mpg, cylinders, displacement, horsepower, weight, acceleration etc.


UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Auto MPG Data Set

Download: Data Folder, Data Set Description

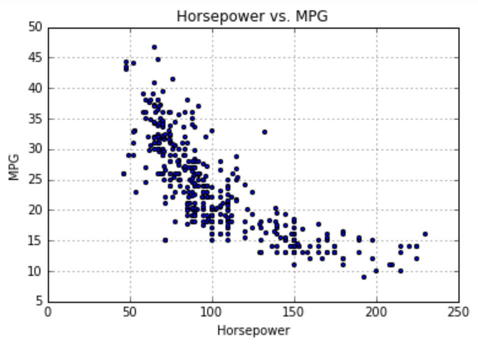
Abstract: Revised from CMU StatLib library, data concerns city-cycle fuel consumption

Data Set Characteristics:	Multivariate	**Number of Instances:**	398	**Area:**	N/A
Attribute Characteristics:	Categorical, Real	**Number of Attributes:**	8	**Date Donated**	1993-07-07
Associated Tasks:	Regression	**Missing Values?**	Yes	**Number of Web Hits:**	167833

MPG vs Horsepower

Simple use-case

- Auto MPG Data Set
- Predicting *MPG* based on *Horsepower*



Polynomial Regression

Polynomial Regression (1/3)

Definition

- Relationship is modelled using a k^{th} degree polynomial
- Curvilinear response function (fits non-linear relationship between x and y)
- Polynomial models can approximate a complex non-linear relationship

Polynomial Regression (2/3)

Univariate regression with one input variable

$$\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$$

$$\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}$$

$$\hat{y}(x_i) = w_0 + w_1 x_i$$

Polynomial regression with one input variable

$$\hat{y}(x_i) = w_0 + w_1 x_i + w_2 x_i^2 + \cdots + w_k x_i^k$$

where k is the degree of the polynomial.

Polynomial Regression - Matrices (3/3)

Univariate regression with one input variable

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}$$

Polynomial regression with one input variable

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^k \\ 1 & x_2 & x_2^2 & \dots & x_2^k \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^k \end{bmatrix}$$

Significance of the coefficients

- We want to keep the order of the model as low as possible
- We can use Student's t-test to test the significance of individual regression coefficients
    1. We establish a null and an alternative hypothesis:
       $H_0 : w_j = 0$
       $H_0 : w_j \neq 0$
    2. We decide on a significance level ($\alpha = 0.05$)
    3. We compute the test statistic
       $t = \frac{w_j}{\mathsf{SE}} = \frac{w_j}{\sqrt{C_{jj}}}$
       $C = \sigma(X^T X)^{-1} = \frac{\mathsf{SSE}}{(N-(k+1))}(X^T X)^{-1}$
    4. We compute the cumulative probability based on the t statistic and compare the p values to α

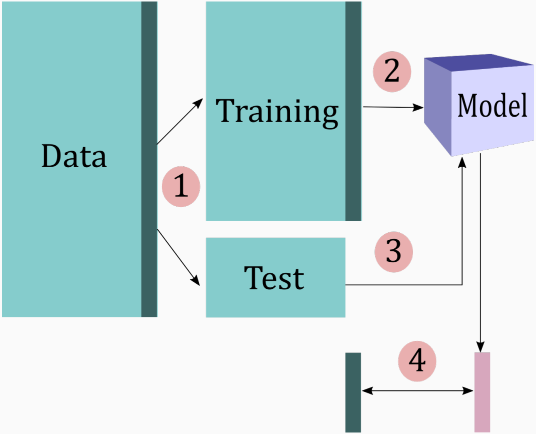
Generalisation

- Fitting a model is about minimising the error on the training data
- "Learning" is about generalising the knowledge

Model validation

Holdout method

- Simplest validation technique

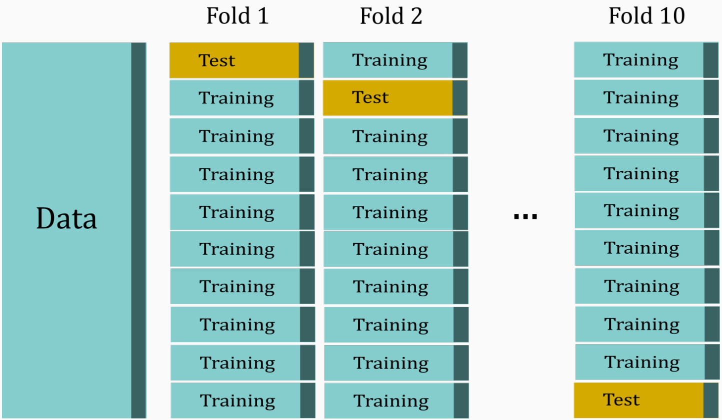


Cross Validation

k-fold Cross Validation algorithm

1. Divide the dataset D into k subsets – D_1, D_2,..., D_k
2. For each D_i $(1 \leq i \leq k)$:
    2.1 Train a model with $D - D_i$
    2.2 Test the model with D_i
3. Calculate the average error

10-fold Cross Validation



References I

Jason W. Osborne and Elaine Waters, *Four assumptions of multiple regression that researchers should always test*, Practical Assessment Research and Evaluation **8** (2002), no. 2.

