



# Introduction to Machine Learning

## Curve fitting and Cross-validation

---

Nikolay Manchev

September 29, 2016

London Machine Learning Study Group

## **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>

Assumptions in Linear Regression

Polynomial Regression

Model validation

# Assumptions in Linear Regression

---

# Univariate Linear Regression

## Fitting a linear regression model

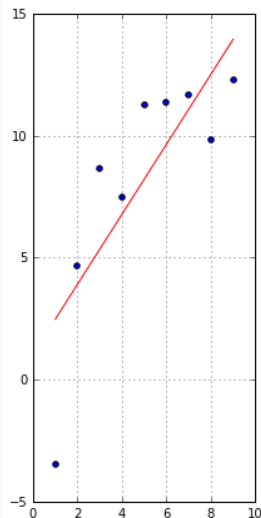
$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}^T$$

$$\mathbf{y} = \{y_1, y_2, \dots, y_N\}^T$$

## Cost function minimisation

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

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



# 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:  $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$

- Cost function

- $J(\mathbf{w}) = \frac{1}{2N} \sum_{i=1}^N (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y})$

- $\frac{\partial}{\partial \mathbf{w}} J(\mathbf{w}) = \frac{\mathbf{X}^T (\mathbf{X}\mathbf{w} - \mathbf{y})}{N}$

- Solving for  $\mathbf{w}$  using normal equations:  $\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

## 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

## How to check the linear relationship assumption


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



## UCI Machine Learning Repository –

[archive.ics.uci.edu/ml](http://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.



**Machine Learning Repository**  
Center for Machine Learning and Intelligent Systems

### Auto MPG Data Set

Download: [Data Folder](#), [Data Set Description](#)



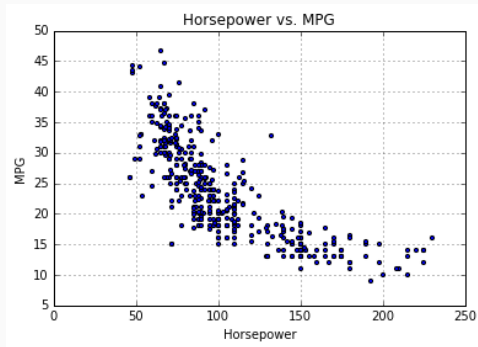
**Abstract:** Revised from CMU StatLib library, data concerns city-cycle fuel consumption

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

# MPG vs Horsepower

## Simple use-case

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



# Polynomial Regression

---

## Definition

- Relationship is modelled using a  $k^{\text{th}}$  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

$$\mathbf{x} = \{x_1, x_2, \dots, x_N\}$$

$$\mathbf{y} = \{y_1, y_2, \dots, y_N\}$$

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

## Polynomial regression with one input variable

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

where  $k$  is the degree of the polynomial.

# Polynomial Regression - Matrices (3/3)

## Univariate regression with one input variable

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

## Polynomial regression with one input variable

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

# 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

$$t = \frac{w_j}{\text{SE}} = \frac{w_j}{\sqrt{C_{jj}}}$$

$$C = \sigma(X^T X)^{-1} = \frac{\text{SSE}}{(N-(k+1))} (X^T X)^{-1}$$

4. We compute the cumulative probability based on the t statistic and compare the  $p$  values to  $\alpha$

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

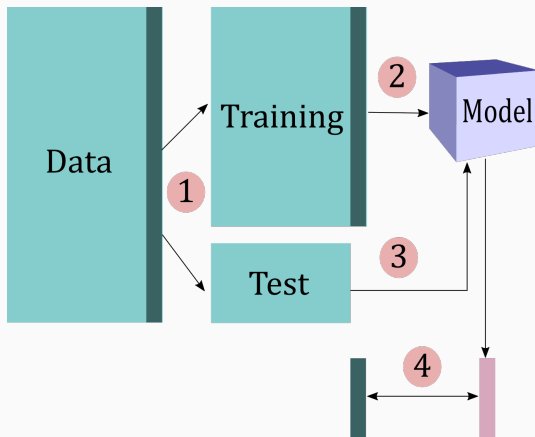


## Model validation

---

# Holdout method

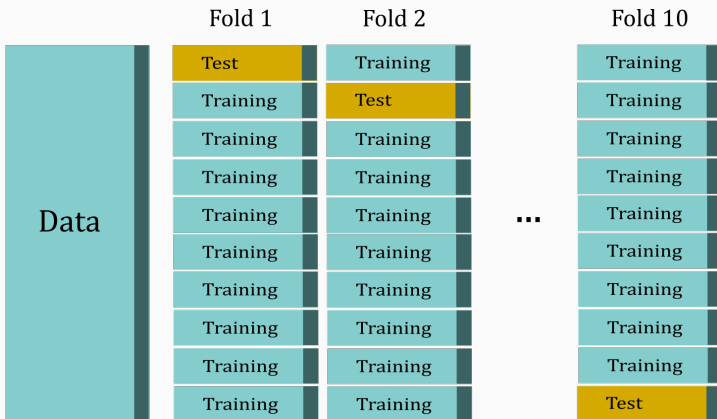
- Simplest validation technique




## k-fold Cross Validation algorithm

1. Divide the dataset  $D$  into  $k$  subsets –  $D_1, D_2, \dots, D_k$
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



-  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.