Machine Learning

Zusammenfassung

Joel von Rotz / Quelldateien

Inhaltsverzeichnis

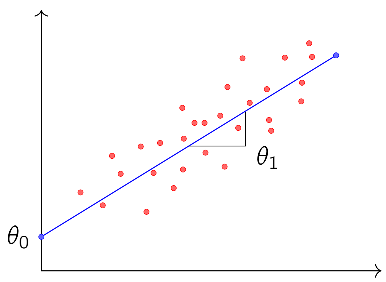
Crash Course

Python

Numpy

Pandas

Linear Regression



$$Y = \theta_0 + \theta_1 \cdot X + \varepsilon$$

X : input data
Y : output data
ε : residual error
θ_0 : offset
θ_1 : slope

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \cdot \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Tipp

Pick the values for (θ_0, θ_1), which minimze some measure of the **total fitting error**

Regression Errors

MAE - Mean Absolute Error

$$MAE = \frac{1}{N} \cdot \sum_{j=1}^{N} |\varepsilon_j| = \frac{1}{N} \cdot \sum_{j=1}^{N} |y_j - \hat{y}_j|$$

N : amount of data points
y_j : j-th output data point
$\hat{y}_j$: j-th output regressed data point

MAPE - Mean Absolute Percent Error

$$MAPE = \frac{1}{N} \cdot \sum_{j=1}^{N} \left| \frac{\varepsilon_j}{y_j} \right| = \frac{1}{N} \cdot \sum_{j=1}^{N} \left| \frac{y_j - \hat{y}_i}{y_j} \right|$$

MSE - Mean Squared Error

$$MAPE = \frac{1}{N} \cdot \sum_{j=1}^{N} \varepsilon_j^2 = \frac{1}{N} \cdot \sum_{j=1}^{N} (y_j - \hat{y}_j)^2$$

OLS - Ordinary Least Squares

Regression Perfomance R^2

! Visualize Data

The *Anscome's quartet* compises four data sets that have nearly identical simple descriptive statistics.

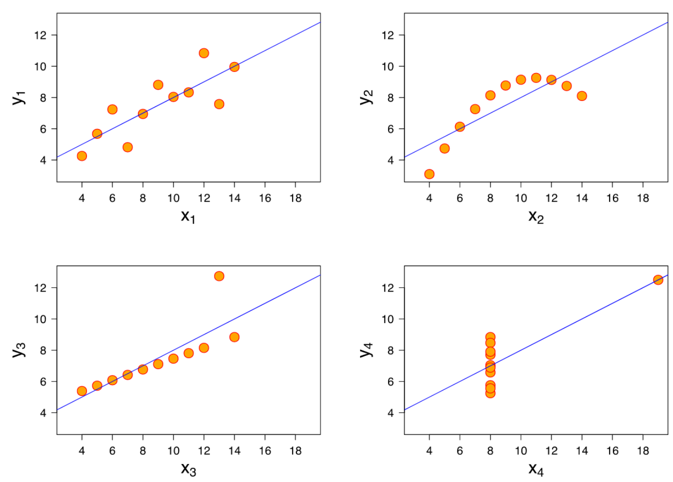


Abbildung 0.1: Anscombe's quartet

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x: s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: s_y^2	4.125	± 0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500 \cdot x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R^2	0.67	to 2 decimal places