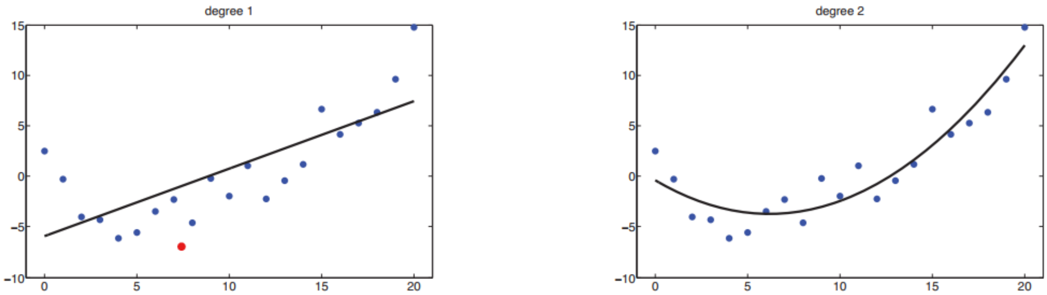
Machine Learning and Data Mining

11. May 2021.

Introduction to Machine Learning Algorithms: Linear Regression

Linear regression

- an approach to modelling the relationship between the independent variable x and the dependent scalar variable y
- if the variable x is a scalar, it is a simple linear regression.
- if the variable x is a vector, it is called multiple linear regression



(a)

(b)

Figure 1.7 (a) Linear regression on some 1d data. (b) Same data with polynomial regression (degree 2). Figure generated by `linregPolyVsDegree`.

Principal Components Analysis (PCA)

Principal Components Analysis (PCA)

In case you want a higher-dimensional space. You need to select a basis for that space and only the 200 most important scores of that basis. This base is known as a principal component. The subset you select constitutes a new space that is small in size compared to the original space. It maintains as much of the complexity of data as possible.

Uncorrelated random variables

- If two variables are uncorrelated, there is no linear relationship between them.
- In probability theory and statistics, two real-valued random variables X, Y, are said to be uncorrelated if their covariance,$\operatorname{cov}[X, Y] = \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y]$, is zero.
- Uncorrelated random variables have a Pearson correlation coefficient of zero
- If X and Y are independent, with finite second moments, then they are uncorrelated. However, not all uncorrelated variables are independent.

Principal Components Analysis (PCA)

Principal Components Analysis (PCA)

- a linear dimensionality reduction technique
- an Unsupervised dimensionality reduction technique,
- you can cluster the similar data points based on the feature correlation between them without any supervision (or labels)

Principal Components Analysis (PCA) Application

- Data Visualization
- Speeding Machine Learning (ML) Algorithm

Principal Components Analysis (PCA) Application

- Principal Components captures (or holds) most of the variance (information) of your data.
- Principal components have both direction and magnitude

Understanding the Data

Understanding the Data

- The Breast Cancer data set is a real-valued multivariate data that consists of two classes
- The malignant class has 212 samples, whereas the benign class has 357 samples.
- $https : //archive.ics.uci.edu/ml/datasets/Breast + Cancer + Wisconsin + (Diagnostic)$
- an easy way is by loading it with the help of the sklearn library.

Object Oriented Programming Terminology

Object Oriented Programming Terminology

- Class
- Class variable
- Data member
- Function overloading
- Instance variable
- Inheritance
- Instance
- Method
- Object
- Operator overloading