Barcelona Technology School

Classical Data Analysis

Master in Big Data Solutions 2020-2021

Victor Pajuelo

victor.pajuelo@bts.tech

Project Lifecycle

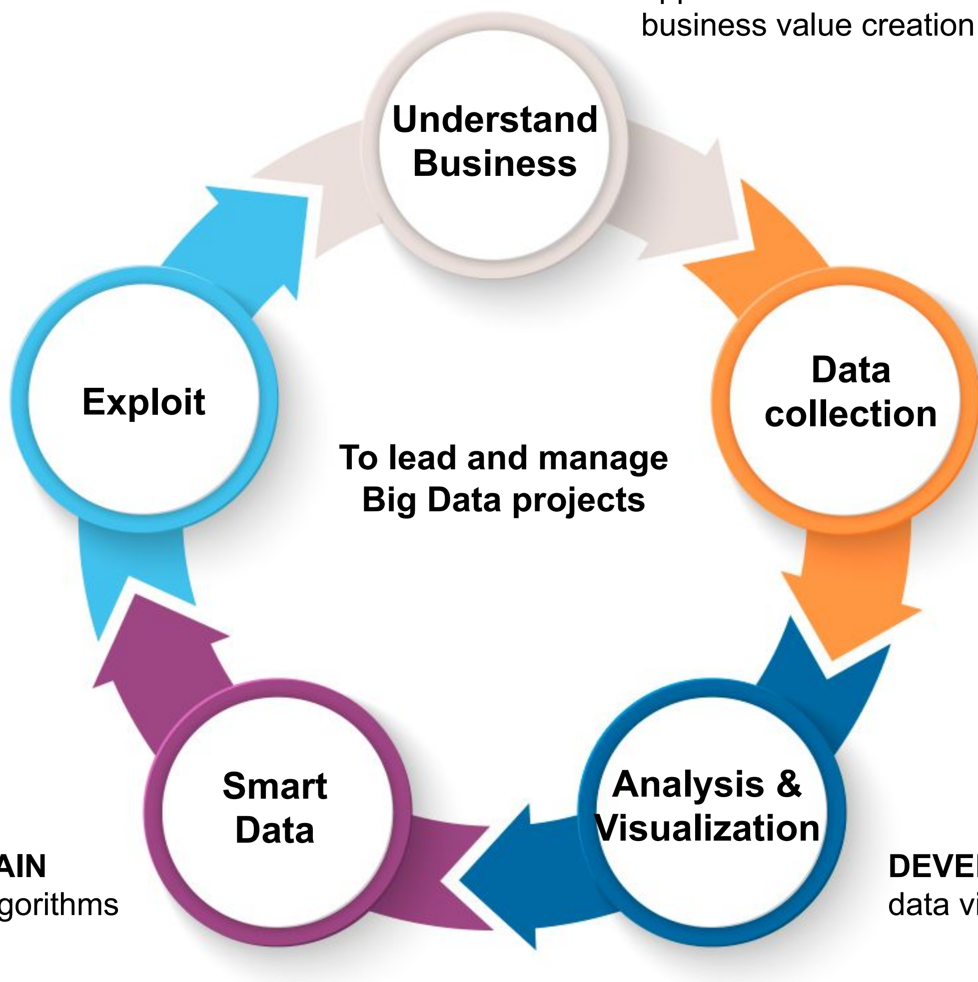
DEFINE data-driven opportunities based on business value creation

IDENTIFY, transform and validate data

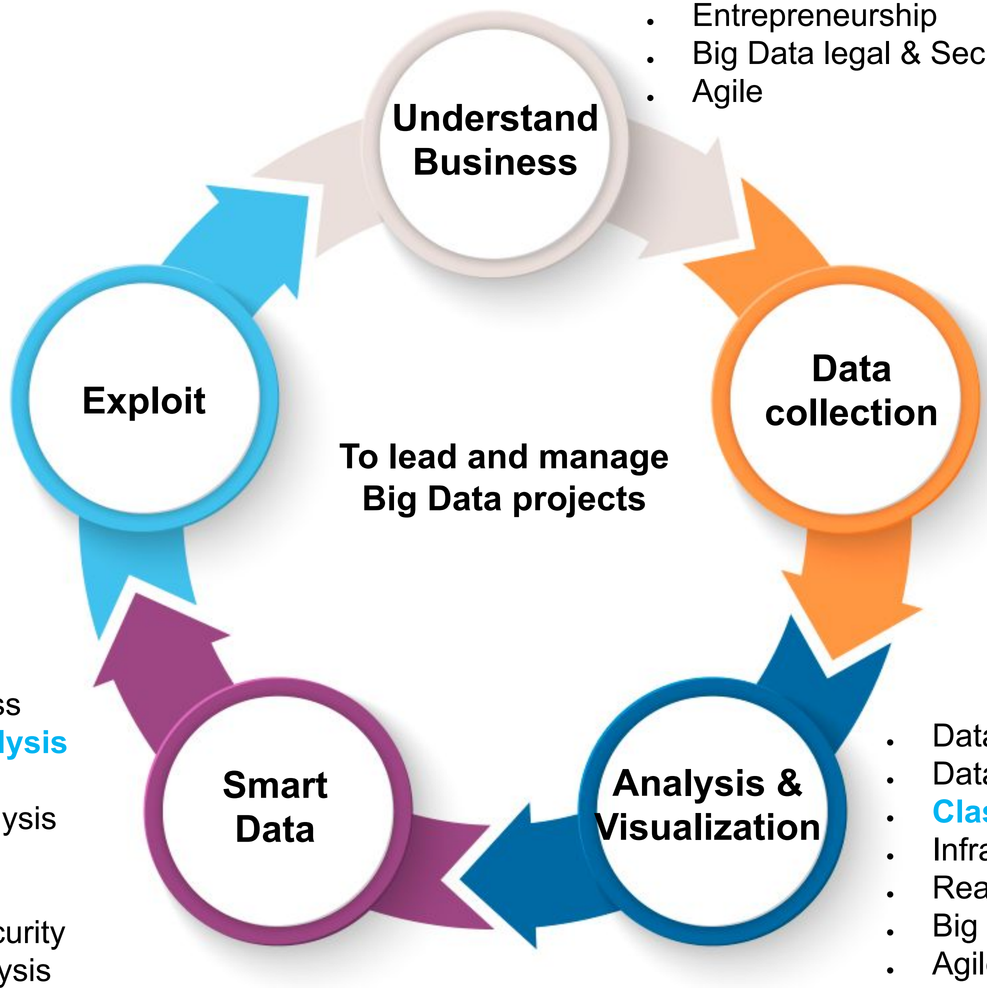
DEVELOP analytical and data visualization techniques

DEVELOP & TRAIN intelligent data algorithms

IMPLEMENT & OPTIMIZE data-driven business models



Project Lifecycle



Understand Business

- Data-Driven Business
- Creative Technology
- Entrepreneurship
- Big Data legal & Security
- Agile

Data collection

- Data-Driven Business
- Data Science Foundations
- Infrastructure
- Real-time Data Analysis
- Big Data legal & Security
- Agile

Analysis & Visualization

- Data-Driven Business
- Data Science Foundations
- **Classical Data Analysis**
- Infrastructure
- Real-time Data Analysis
- Big Data legal & Security
- Agile

Smart Data

- Data-Driven Business
- **Classical Data Analysis**
- Data Visualization
- Advanced Data Analysis
- Artificial Intelligence
- Infrastructure
- Big Data legal & Security
- Real-time Data Analysis
- Agile

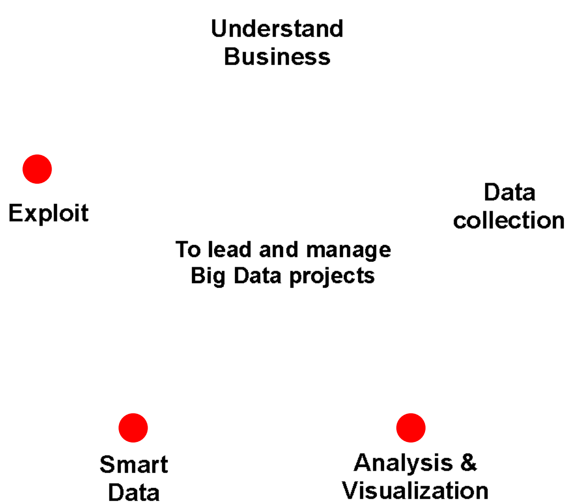
Exploit

- Data-Driven Business
- **Classical Data Analysis**
- Advanced Data Analysis
- Infrastructure
- Real-time Data Analysis
- Big Data legal & Security
- Agile

Classical Data Analysis

What we will learn

Apply machine learning concepts - regression, classification and clustering to an analytic or business intelligence strategy. Use programming languages skills for data analysis.



We will learn:

- Linear regression with one variable.
- Linear regression with multiple variables.
- Logistic regression.
- Supervised classification.
- Unsupervised classification / clustering.

Curriculum review

- Introduction to data analysis
- Linear Regression
- Logistic Regression
- Regression analysis (Polynomial, Ridge, Lasso, etc.)
- Neural Networks (MLPs in depth)
- Support Vector Machines (SVM)
- Decision Trees
- Ensemble Methods (Random Forests, Bagging, etc.)
- Other Classifier (KNN, Naïve Bayes)
- K-Means
- PCA
- Hierarchical Clustering

Curriculum review

Lecture organization and teaching methodology

Theoretical Background	Practical Training	Assignment

Methodology:

- Lectures and analysis of teaching notes.
- Lecture-demonstration by teacher.
- Analysis and discussion of case studies in small groups as well as in class.
- Presentations by student panels from the class: class invited to participate.
- Individual reading and research.
- Reading assignments in journals, monographs, etc.

Curriculum review

Evaluation

- 30%:
  - Classroom participation (participation and involvement during lectures)
  - Presentation of group projects
    - Final projects algorithm development
- 60%:
  - Assignments (Coding and discussions about the used methodologies).
- 10%:
  - Multiple choice tests

Today's Objective

- Introduction to data mining algorithms
  - Supervised versus unsupervised algorithms
  - Regression versus classification
- Linear regression
  - Linear regression theory
  - Linear regression with Python

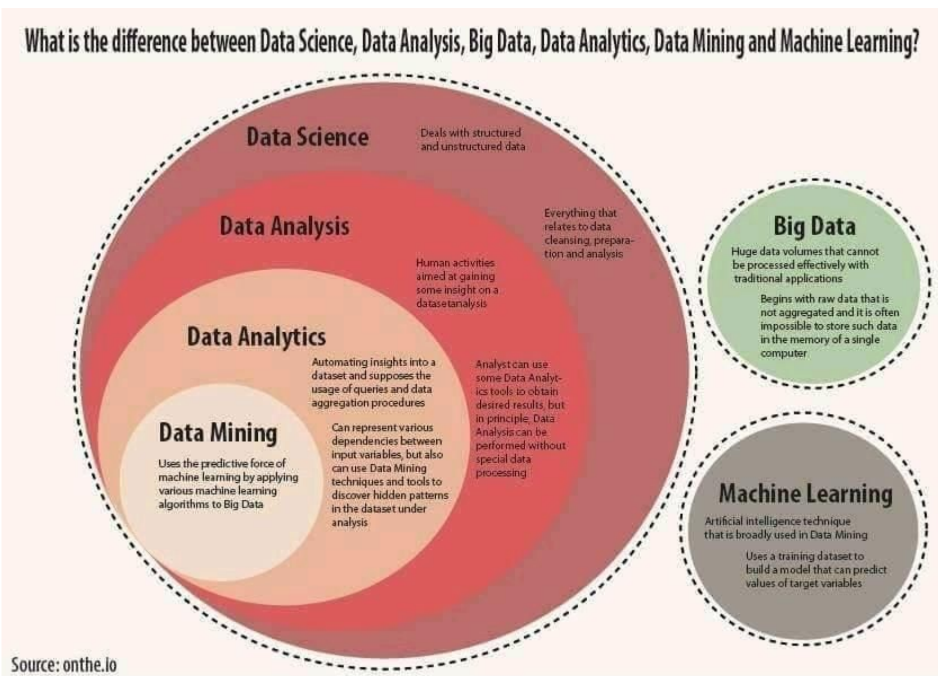
Contents

- Introduction to data analysis
- Linear Regression
- Logistic Regression
- Regression analysis (Polynomial, Ridge, Lasso, etc.)
- Neural Networks (MLPs in depth)
- Support Vector Machines (SVM)
- Decision Trees
- Ensemble Methods (Random Forests, Bagging, etc.)
- Other Classifier (KNN, Naïve Bayes)
- K-Means
- PCA
- Hierarchical Clustering

Session 1
What is Data Analysis

What is Data Analysis

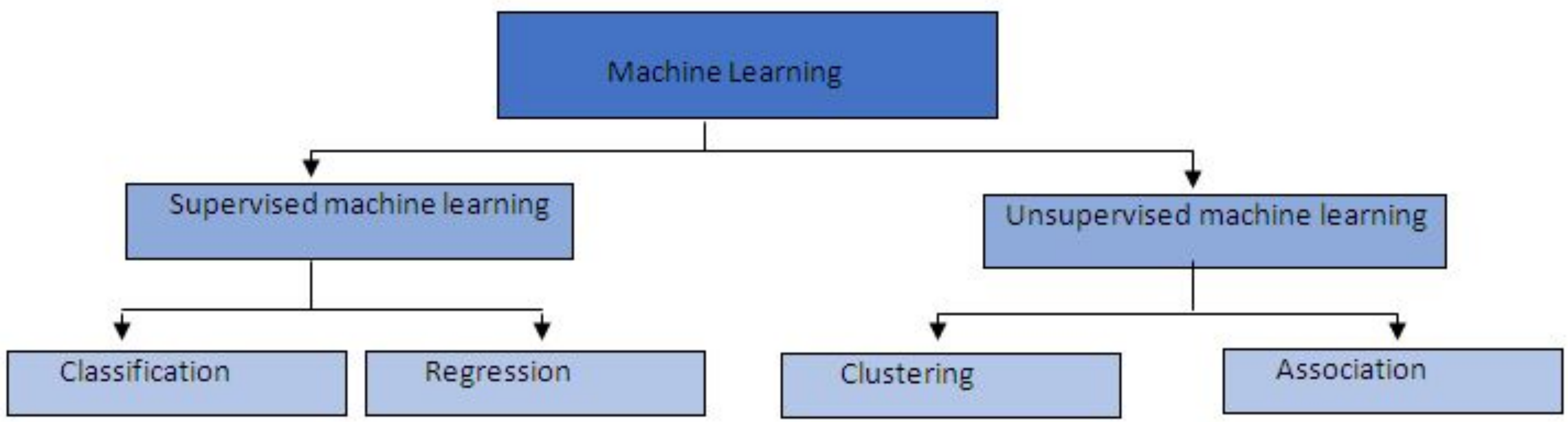
”Extracting, cleaning, transforming, modeling and visualization of data with an intention to uncover meaningful and useful information that can help in deriving conclusion and take decisions.”



Data Analysis is characterized by a wide use of **Data Mining algorithms**

What is Data Analysis

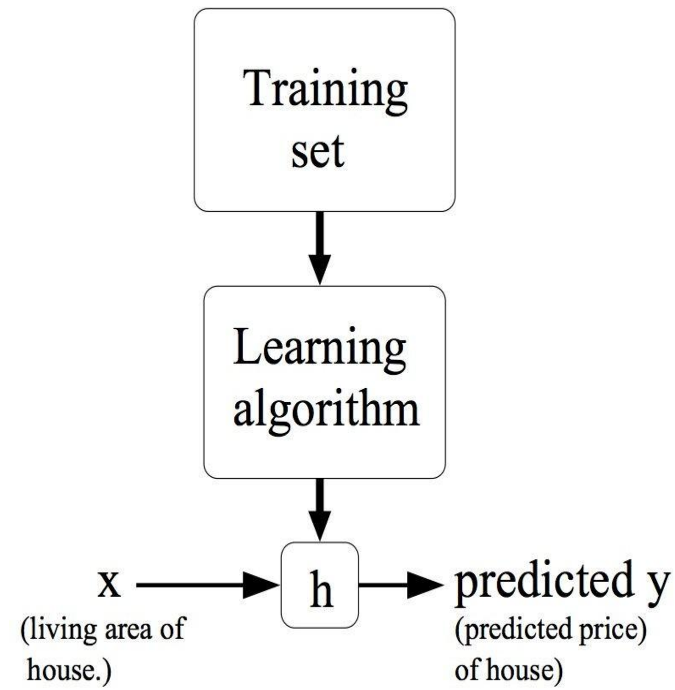
Data mining algorithms



Supervised learning algorithms

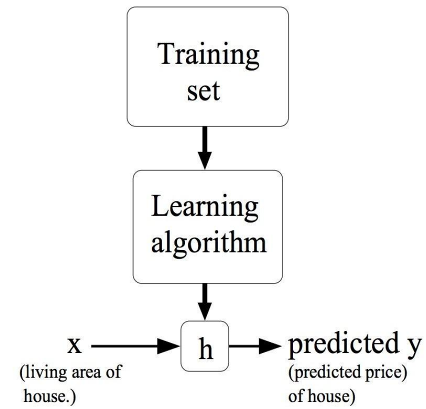
- **Formal Definition:** given a training set, to learn a function **h : X → Y** (h is also called hypothesis function) so that **h(x)** is a "good" predictor for the corresponding value of y, where:

  - X denote the space of input values
  - Y the space of output values.



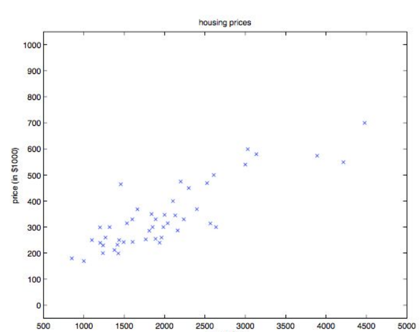
Supervised learning algorithms

- **Formal Definition:** given a training set, to learn a function **h : X → Y** (h is also called hypothesis function) so that **h(x)** is a "good" predictor for the corresponding value of y, where:



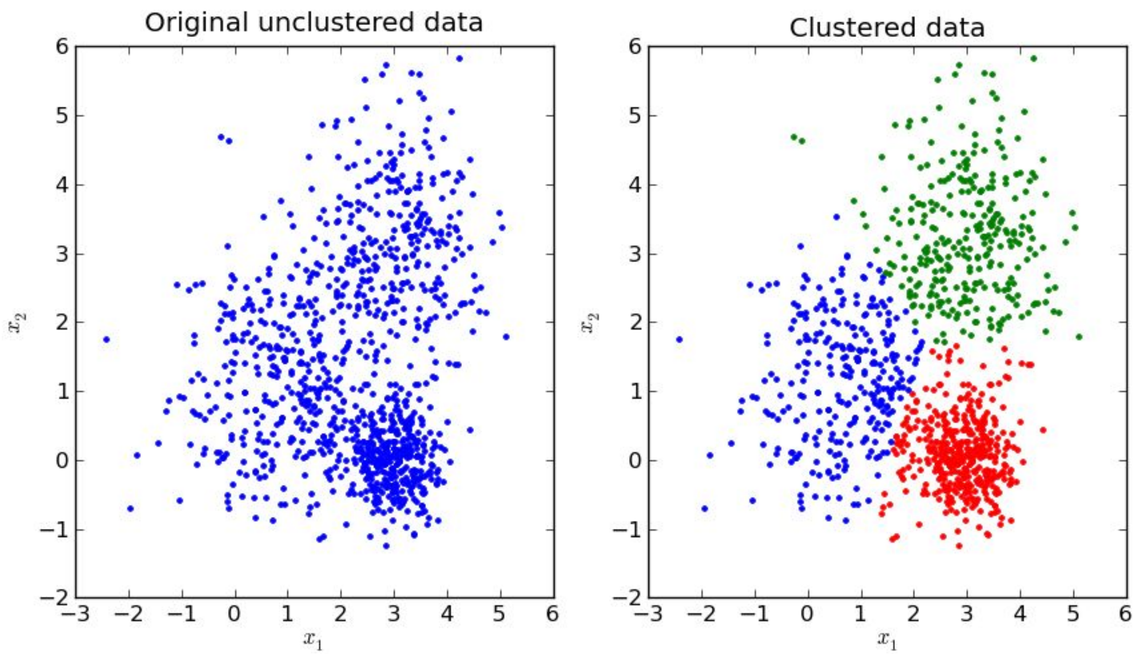
- X denote the space of input values
- Y the space of output values.

Living area ($feet^2$)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮



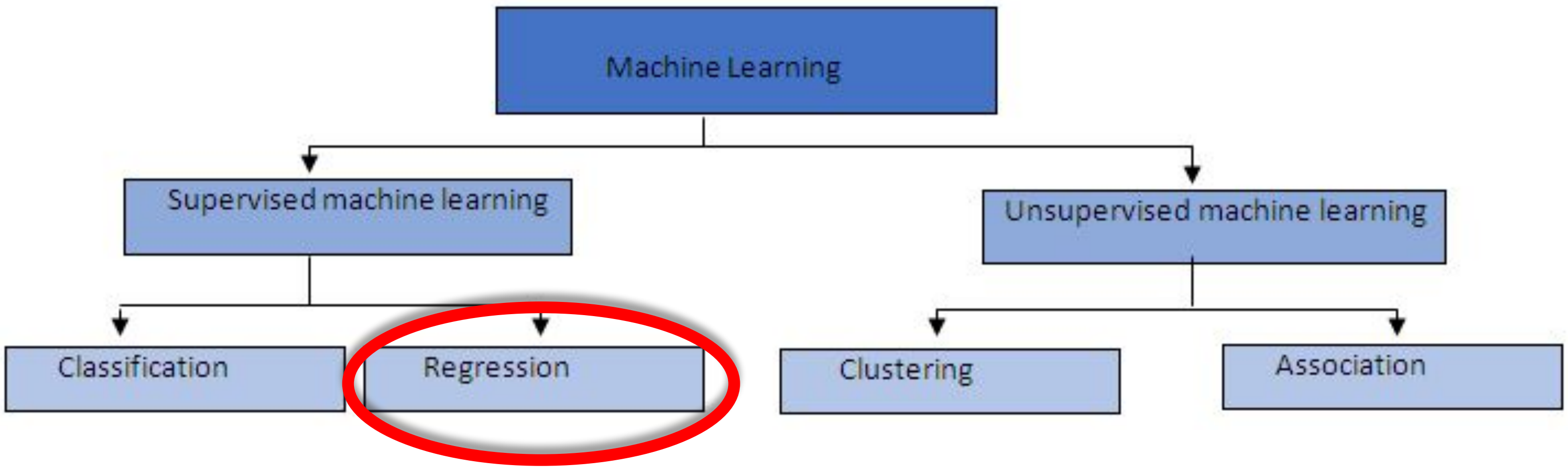
Unsupervised learning algorithms

▪**Example (Clustering):** Take a collection of 1 Milion different travelers and find a way to automatically group these travelers in clusters that are somehow similar based on different aspects, such as trip distances, durations, purpose, etc.



Session 1
Linear Regression

Supervised learning: Regression



Supervised learning: Regression

- **Regression**: Predict continuous valued output
- Examples:
  - Predict stock market index based on other indicator
  - Predict the total amount of sales of a company based on the total budget spent for advertising
  - Predict the price of a house based based on its characteristic
  - Other examples?

Supervised learning: Linear Regression

Given a dataset

x_1 Living area (feet^2)	x_2 #bedrooms	y Price (1000\$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮

- x_1 and x_2 are the explanatory variables (aka indipendent variables). They can be either discrete or continuous.
- y is the target (aka dependent variable). It **must be** continuous.

- Linear Regression performs the task to predict a dependent variable value (y) based on a given independent variable (x).So it finds out a linear relationship between x(input) and y(output).
- The goal is to find the **function that best fits the data** minimizing the error.
- The error in a regression task is the difference between the prediction of the regression model h(X) and the actual target value y. It can be expressed as:

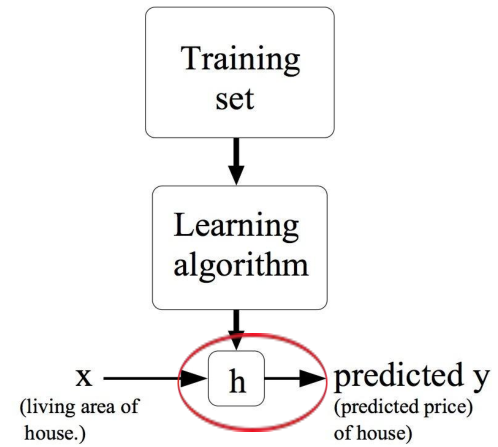
Absolute error: $\sum_i | y_i - h_\vartheta(x_i) |$

Squared error: $\sum_i (y_i - h_\theta(x_i))^2$

Supervised learning: Linear Regression

Training Set

Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
...	...



We assume a linear model $h_\vartheta(x) = \vartheta_0 + \vartheta_1\, x$

Given some estimates of the **coefficients** θ_0 and θ_1 we predict future observations using:

$$h_\vartheta(x) = \hat{y} = \hat{\vartheta}_0 + \hat{\vartheta}_1\, x$$

we want to come up with values for the **parameters θ_0 and θ_1 so that $h_\theta(x)$ (the predicion) is close to y (the actual value) for our training set (X,Y).**

Supervised learning: Linear Regression

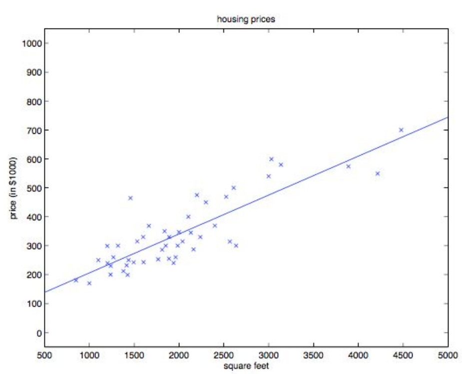
Univariate VS Multivariate Linear regression

Univariate (Simple LR)

Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

Hypothesis:

$$h_\theta(x) = \theta_0 + \theta_1 x$$

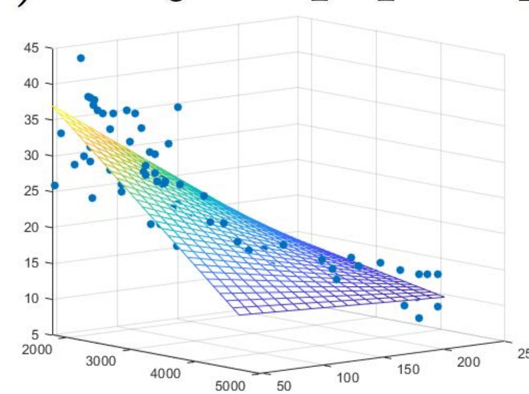


Multivariate

Living area (feet²)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮

Hypothesis:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$



Supervised learning: Linear Regression

Univariate (simple) Linear regression

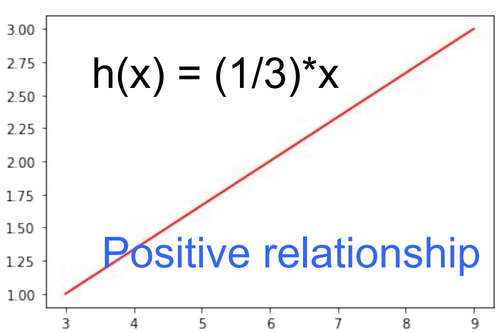
Hypothesis:

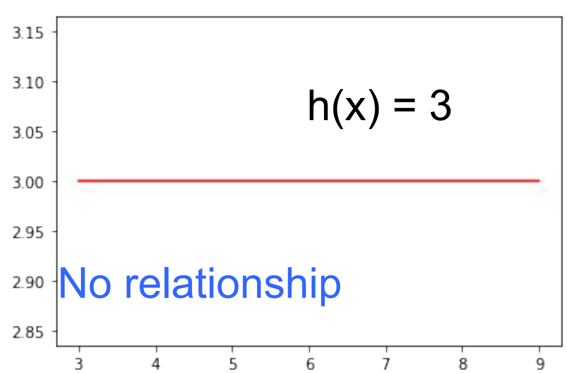
$$h_\theta(x) = \theta_0 + \theta_1 x$$

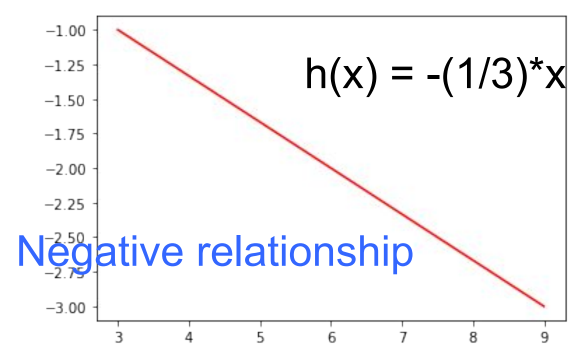
θ_0: is the intercept of the line. It is the expected value of y when x=0

θ_1: is the slope of the line. A value very close to 0 indicates little to no relationship; large positive or negative values indicate large positive or negative relationships, respectively.

Examples







Supervised learning: Linear Regression

Linear regression Interpretation

The coefficient value signifies how much the mean of the dependent variable changes given a one-unit shift in the independent variable while holding other variables in the model constant.

Examples

h(x) = -3 +2x
h(4) = 5

If we change x of 1 unit, which will be the value of y?

h(5) = ?

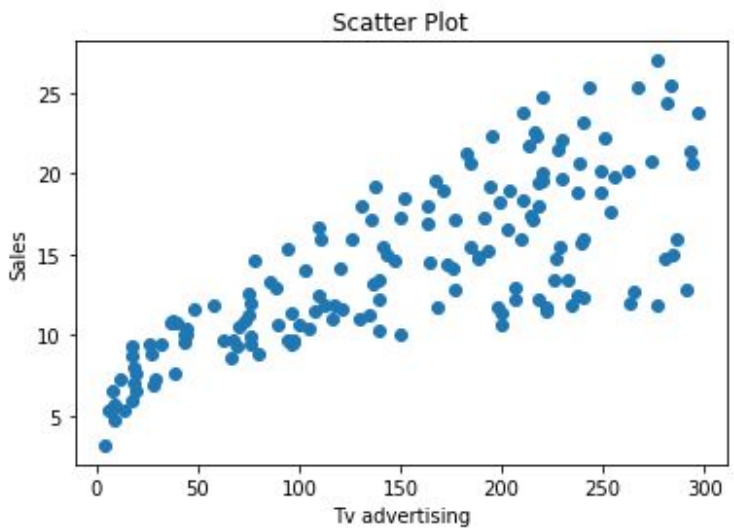
h(5) = -3 +2*5 = 7
Changing 1 unit in x we obtained a change of 2 (the slope) units in the dependent variable

Supervised learning: Linear Regression

Assumptions Of Linear Regression Algorithm

Hypothesis:

1. Linear Relationship between the features (x) and target (y)



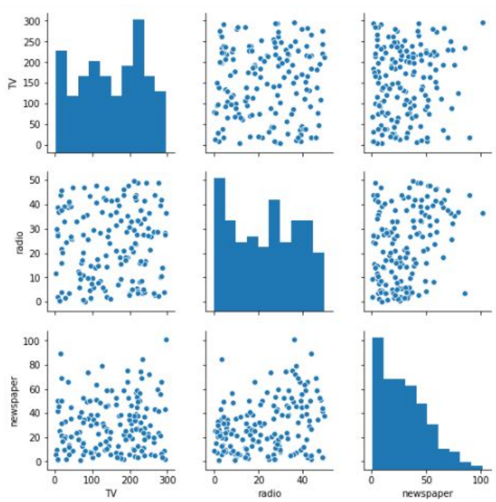
It can be validated by plotting a scatter plot between the features and the target.

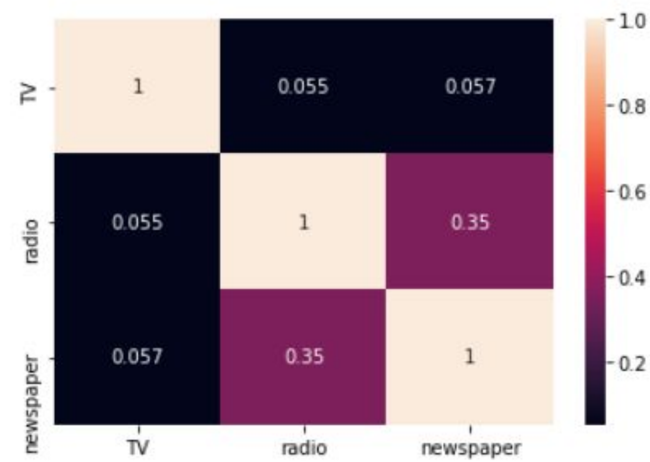
Supervised learning: Linear Regression

Assumptions Of Linear Regression Algorithm

Hypothesis:

2. Little or no Multicollinearity between the features, i.e., very high inter-correlations or inter-associations among the independent variables





Pair plots and heatmaps(correlation matrix) can be used for identifying highly correlated features.

Supervised learning: Linear Regression

Evaluation Metrics

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

RMSE (Root Mean Square Error) – Particularly used because it is differentiable

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

MAE (Mean Absolute Error) - is a linear score

$$\hat{R}^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

R Squared - The maximum is 1 but minimum can be negative infinity (even if it is unlikely scenario, usually the minimum is 0)

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1}\right]$$

Adjusted R Squared

https://medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-models-part-1-a99d7d7414e4

Supervised learning: Linear Regression

Learning the parameters of the model

Training Set

Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
...	...

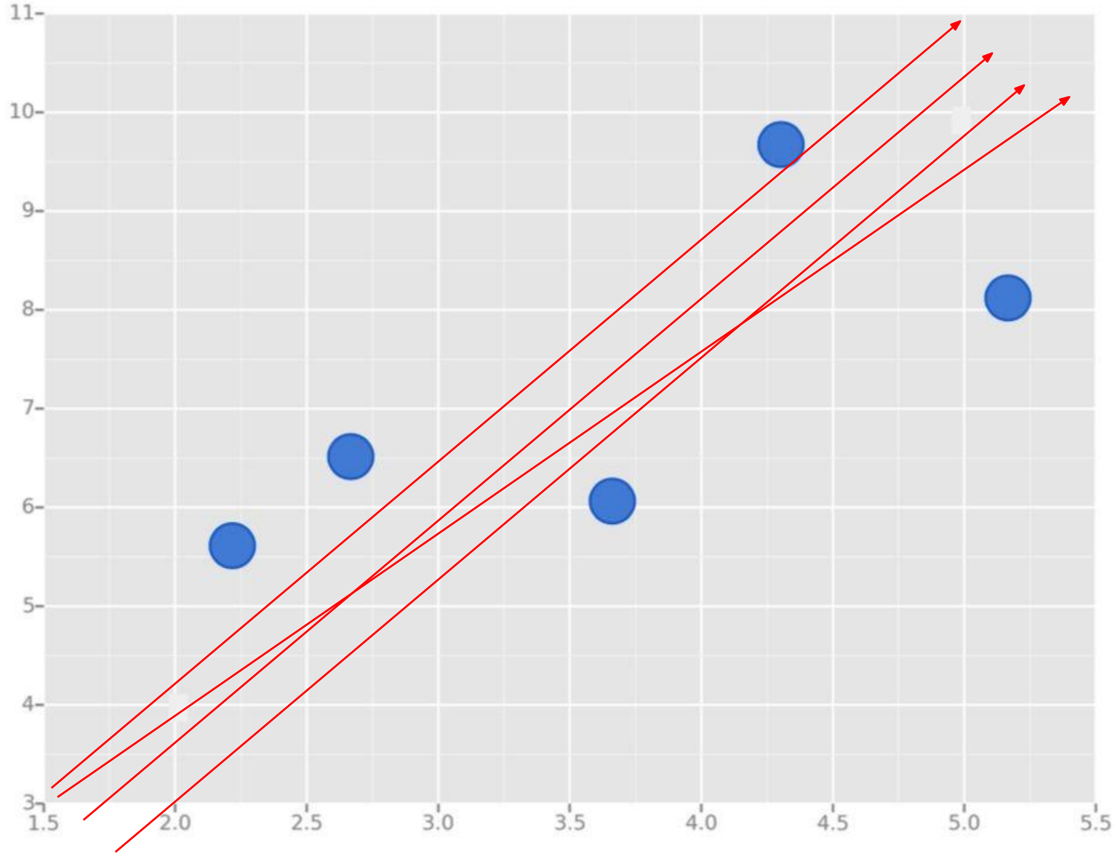
$h(x) = \theta_0 + \theta_1 x$

How to choose θ_0 and θ_1 to minimize the distance between actual values (Y) and predictions(h(x))?

- Ordinary least squares
- Gradient Descent

Supervised learning: Linear Regression

Ordinary least squares



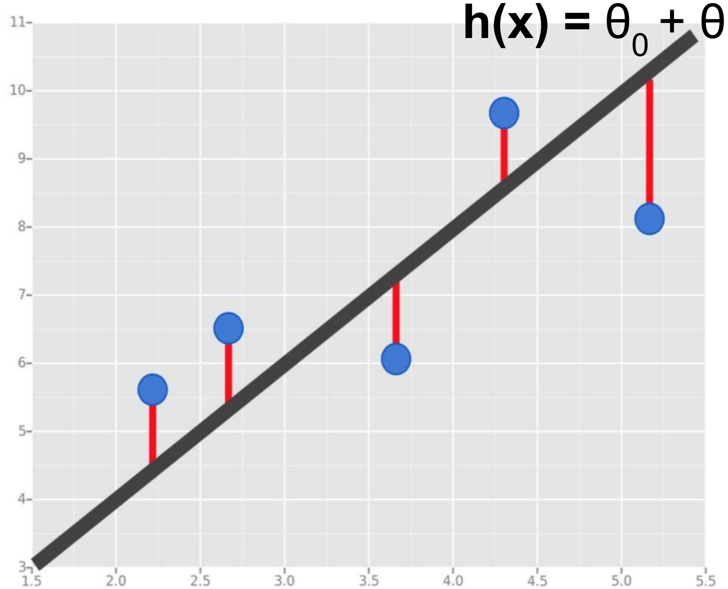
Supervised learning: Linear Regression

Ordinary least squares

we want to come up with values for the **parameters θ_0 and θ_1 so that $h_\theta(x)$ (the predicion) is close to y (the actual value) for our training set (X,Y).**

Linear Model:

h(x) = $\theta_0 + \theta_1 x$



The least squares approach chooses θ_0 and θ_1 to minimize the RSS (residual sum of squares).

$$\theta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\theta_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

Supervised learning: Linear Regression

Gradient Descent

The gradient descent algorithm is:

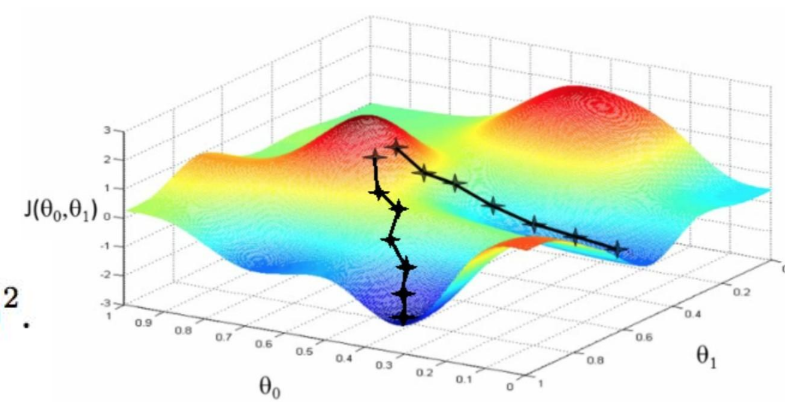
$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

j=0,1 represents the feature index number.

At each iteration, one should simultaneously update the parameters $\theta_1, \theta_2, \ldots$.

Model: $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

Cost function: $J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2.$

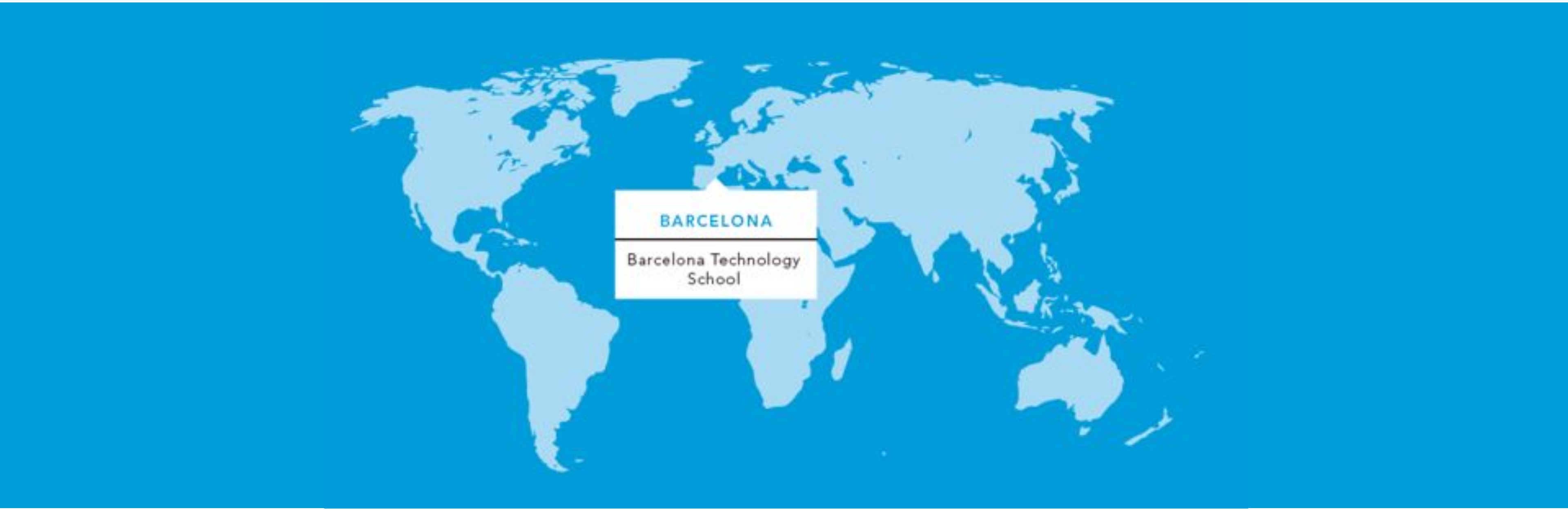


Let’s go to notebook

Resources

- Tan, Pang-Ning. *Introduction to data mining*. Pearson Education India, 2006.

- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. New York: Springer series in statistics, 2001.


BARCELONA
Barcelona Technology School