LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

Carlos J. Costa

MACHINE LEARNING

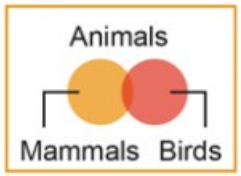
Machine Learning

- It is as a subset of artificial intelligence.
- It is the scientific study of algorithms that computer systems use to perform a specific task without using explicit instructions
- study and construction of algorithms that can learn from and make predictions on data

Machine Learning

Tribe	Origins	Master Algorithm
Symbolists	Logic, philosophy	Inverse deduction
Connectionists	Neuroscience	Backpropagation
Evolutionaries	Evolutionary biology	Genetic programming
Bayesians	Statistics	Probabilistic inference
Analogizers	Psychology	Kernel machines

Symbolists



Use symbols, rules, and logic to represent knowledge and draw logical inference

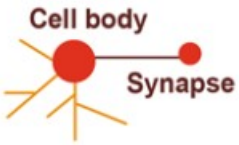
Favored algorithm
Rules and decision trees

Bayesians

Assess the likelihood of occurrence for probabilistic inference

Favored algorithm
Naive Bayes or Markov

Connectionists



Recognize and generalize patterns dynamically with matrices of probabilistic, weighted neurons

Favored algorithm
Neural networks

Evolutionaries

Generate variations and then assess the fitness of each for a given purpose

Favored algorithm
Genetic programs

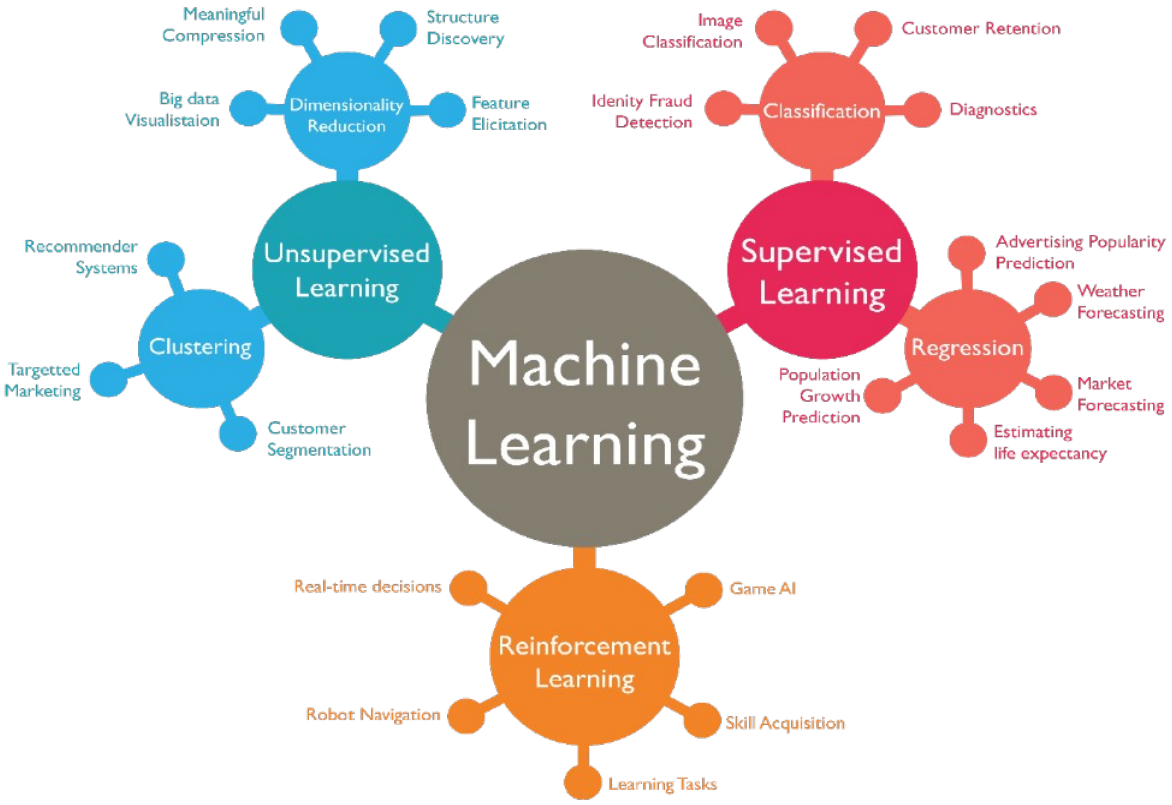
Analogizers

Optimize a function in light of constraints ("going as high as you can while staying on the road")

Favored algorithm
Support vectors

Source: Pedro Domingos, *The Master Algorithm*, 2015

Machine Learning



Machine Learning

- Supervised learning
- It is the machine learning task of learning a function that maps an input to an output based on example input-output pairs (Hinton & Sejnowski,1999)
- Classification
- Regression

Machine Learning

- Unsupervised learning
- The goal of unsupervised learning is to extract an efficient internal representation of the statistical structure implicit in the inputs. (Hinton & Sejnowski,1999)
- Clustering
- Dimensional Reduction

Machine Learning

- Reinforcement Learning (RL)
- There are 3 main components:
  - Agent,
  - Environment
  - Actions (performed by the agent)
- The purpose of RL is to train an intelligent agent that is capable of navigating its environment and performing actions that arrives at the end goal.
- Actions changes the state of the environment and the agent receives rewards or punishments
- The challenge of the agent is to maximize the ammount of rewars at the end of a specific period

Machine Learning

- Train- Validate-Test
- Step 1: Making the model examine data.
- Step 2: Making the model learn from its mistakes.
- Step 3: Making a conclusion on how well the model performs

Machine Learning

- Data Processing and Machine Learning
  - Libraries: Numpy, Pandas, statsmodels, sklearn, networkx
  - Tools: IDE – Jupiter

IDE: Integrated Development Environment

LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

Carlos J. Costa

REGRESSIONS

Regression

- Is a set of statistical processes for estimating the relationships among variables.
- Dependent variable,outcome variable
- Independent variables, predictor, covariates, or features

Regression

- simple regression/multivariate regression

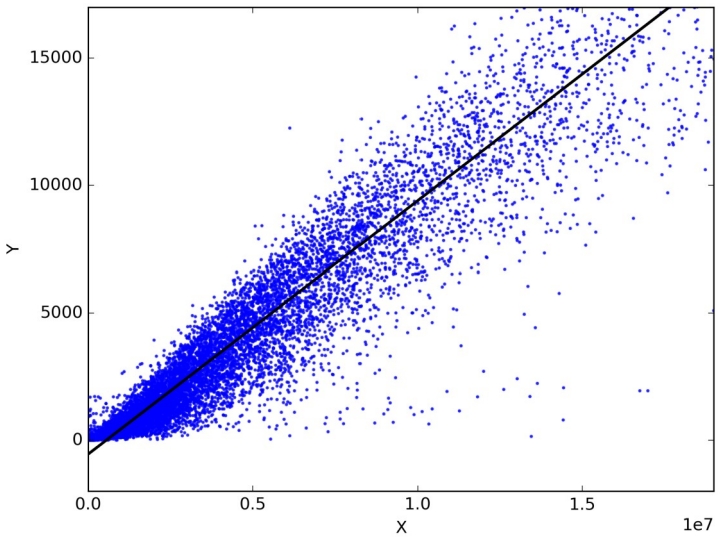
$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i.$$

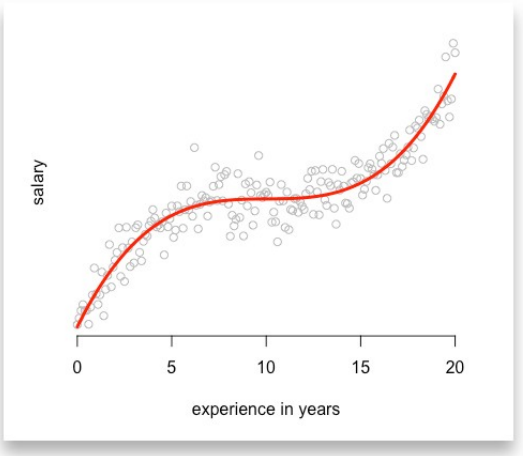
Regression

- .Linear/non linear

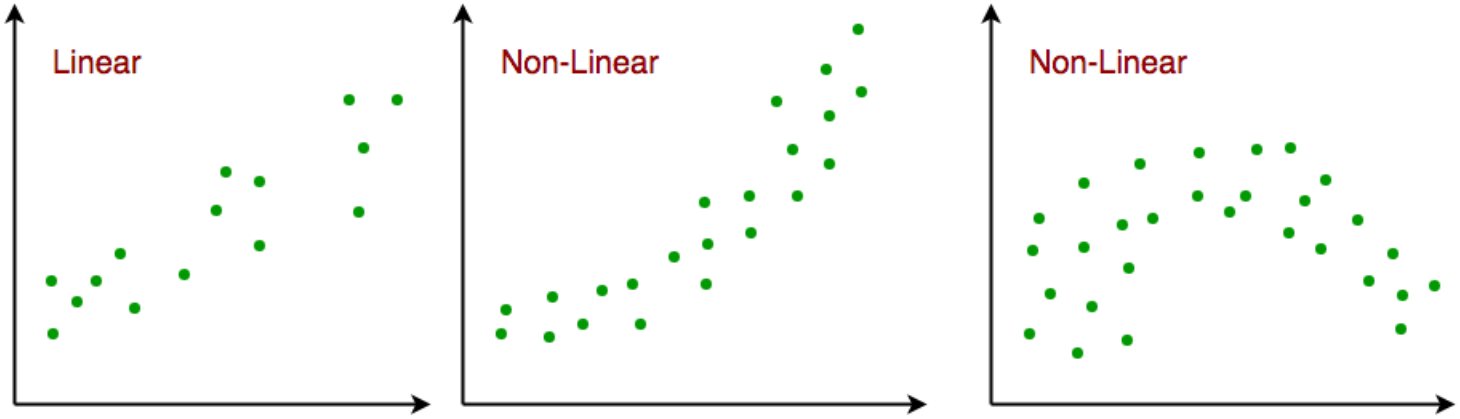
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$



$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \; i = 1, \ldots, n.$$



Regression



LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

Carlos J. Costa

CLASSIFICATION

Classification

- Supervised learning approach
- Categorizing some unknown items into discrete set of categories or “classes”
- The target attribute is a categorical variable
- To solve a classification problem
  - identify the target or class, which is the variable to predict.
  - the target balancing is mandatory
  - choose the best training strategy to train classification models.

Classification

- Churn (not churn rate) depends from several characteristics of the client, product and communication.

age	address	income	ed	employ	equip	callcard	wireless	churn
33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes
33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes
30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No
35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No

age	address	income	ed	employ	equip	callcard	wireless	churn
35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?

Classification

- What is the best drug according to specific characteristics of the patient

Age	Sex	BP	Cholesterol	Na	K	Drug
23	F	HIGH	HIGH	0.793	0.031	drugY
47	M	LOW	HIGH	0.739	0.056	drugC
47	M	LOW	HIGH	0.697	0.069	drugC
28	F	NORMAL	HIGH	0.564	0.072	drugX
61	F	LOW	HIGH	0.559	0.031	drugY
22	F	NORMAL	HIGH	0.677	0.079	drugX
49	F	NORMAL	HIGH	0.79	0.049	drugY
41	M	LOW	HIGH	0.767	0.069	drugC
60	M	NORMAL	HIGH	0.777	0.051	drugY
43	M	LOW	NORMAL	0.526	0.027	drugY

Categorical Variable

Age	Sex	BP	Cholesterol	Na	K	Drug
36	F	LOW	HIGH	0.697	0.069	

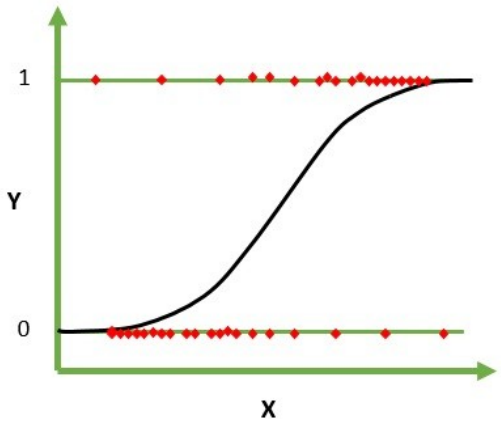
Classification

Classification algorithms in machine learning:

- Decision Trees
- Naive Bayes
- Linear Discriminate Analysis
- K -Near Neighbor (KNN)
- Logistic Regression
- Neural Networks
- Support Vector Machines (SVM)

Logistics Regression

- A regression that having binary dependent variable
- in its basic form, uses a logistic function to model a binary dependent variable



Random Forest

- are an ensemble learning method for classification, regression and other tasks
- operates by constructing a multitude of decision trees at training time
- outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

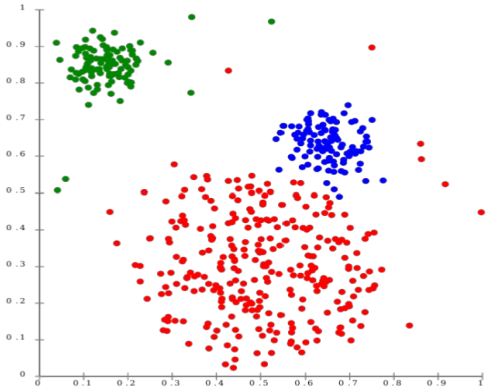
LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

Carlos J. Costa

CLUSTERS ANALYSIS

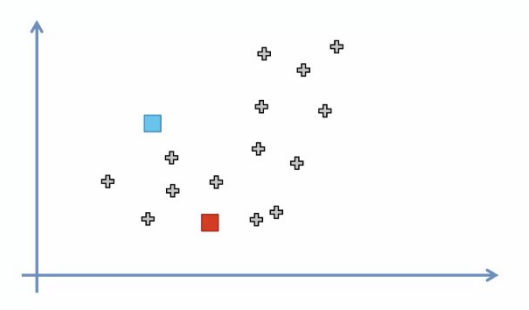
Cluster Analysis

- Cluster analysis is a multivariate method
- aims to classify a sample of subjects (or objects) into several different groups such that similar subjects are placed in the same group
- based on a set of measured variables

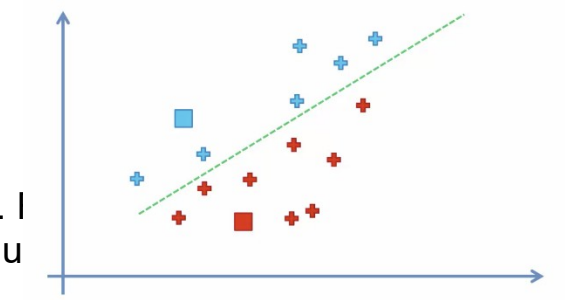


K-means Clustering

- 1.Select K (i.e. 2) random points as cluster centres called centroids

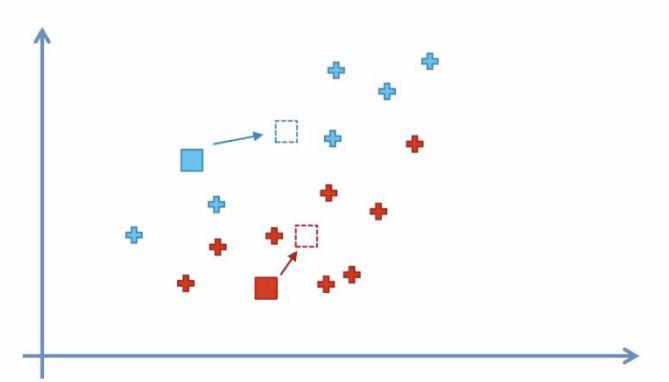


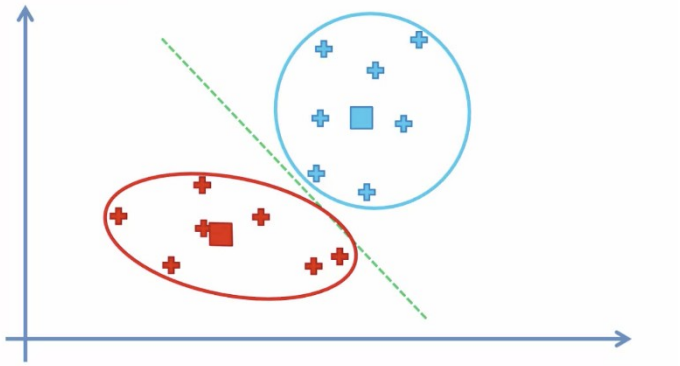
- 2. Assign each data point to the closest cluster by calculating its distance with respect to each centroid



- 4. I of the clu

- 3. Determine the new cluster centre by computing the average of the assigned points



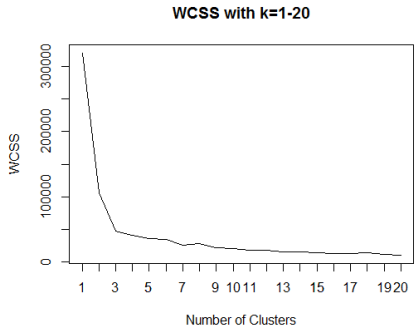


WCSS

- Within-Cluster-Sum-of-Squares (WCSS)- Implicit **objective function in k-Means** measures sum of distances of observations from their cluster centroids.

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$

Yi is centroid for observation Xi.



- Given that k-Means has no in-built preference for right number of clusters, following are some of the common ways k can be selected:
  - Domain Knowledge
  - Rule of Thumb
  - Elbow-Method using WCSS
  - Cluster Quality using Silhouette Coefficient

References

- Albon, Ch. (2018) *Machine Learning with Python Cookbook*. O'Reilly
- Domingos, P. (2015) *The Master Algorithm*, Penguin Books
- Hinton, J.; Sejnowski, T.(1999). *Unsupervised Learning: Foundations of Neural Computation*. MIT Press
- Morgan; P. (2019) *Data Science from Scratch with Python*, AI Science
- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective* (1 edition). Cambridge, MA: The MIT Press.
- Otte, E.; Rousseau, R. (2002). "Social network analysis: a powerful strategy, also for the information sciences". *Journal of Information Science*. 28 (6): 441–453. doi:10.1177/016555150202800601.
- Stuart J. R., Norvig, P. (2010) *Artificial Intelligence: A Modern Approach*, Third Edition, Prentice Hall ISBN 9780136042594.