Introduction to Machine Learning

25th

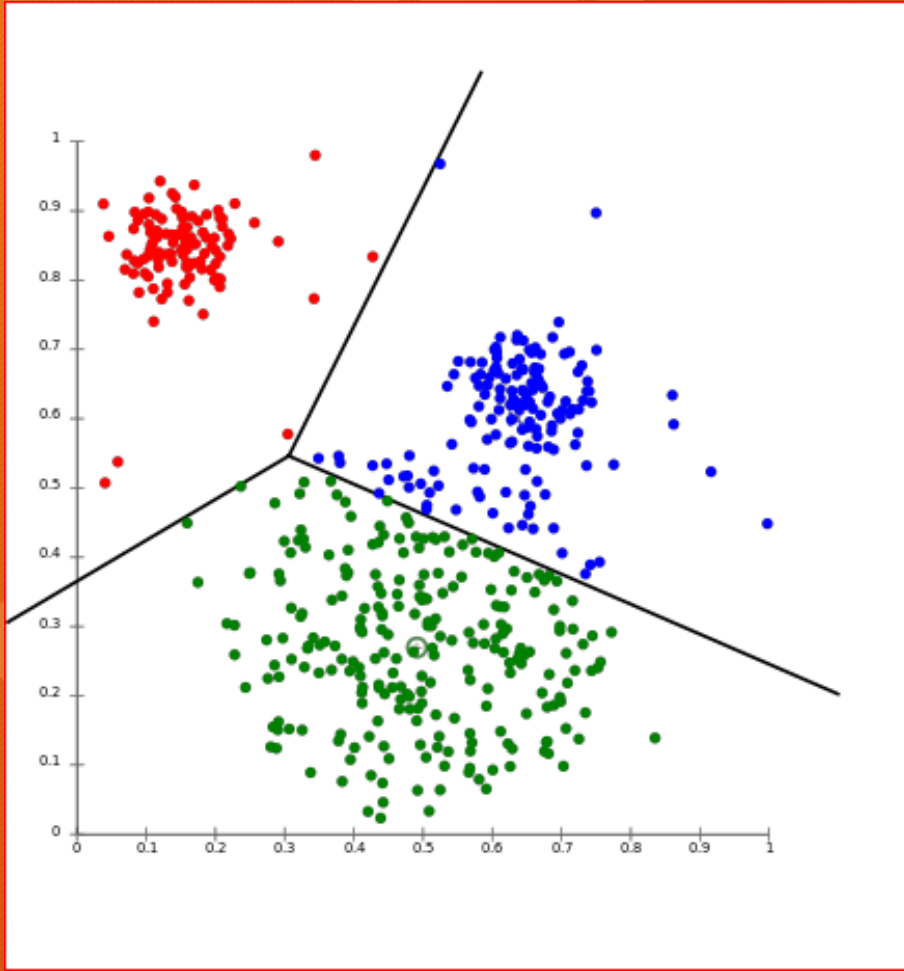
By Ayub Odhiambo

http://www.a4ayub.me/

Objectives

- K-Means Clustering
- Hierarchical Clustering
- Association Rule Learning
- Reinforcement Learning
- Natural Language Processing
- Lets Point out Practical Problems from the Supermarket Dataset
- Practical Jupyter Notebook

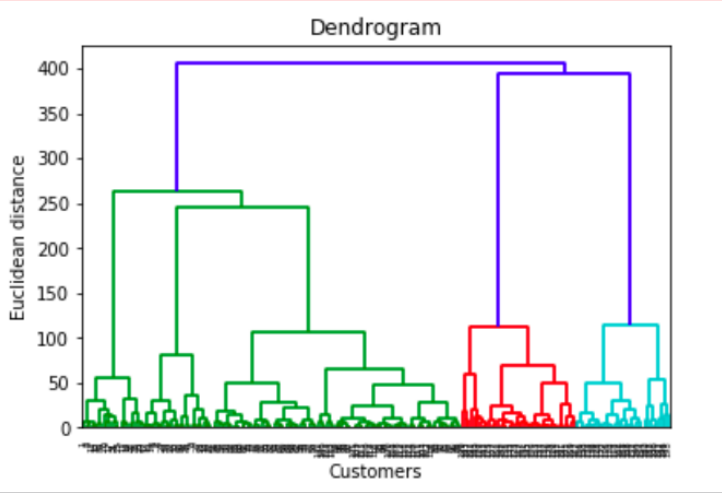
K-Means Clustering



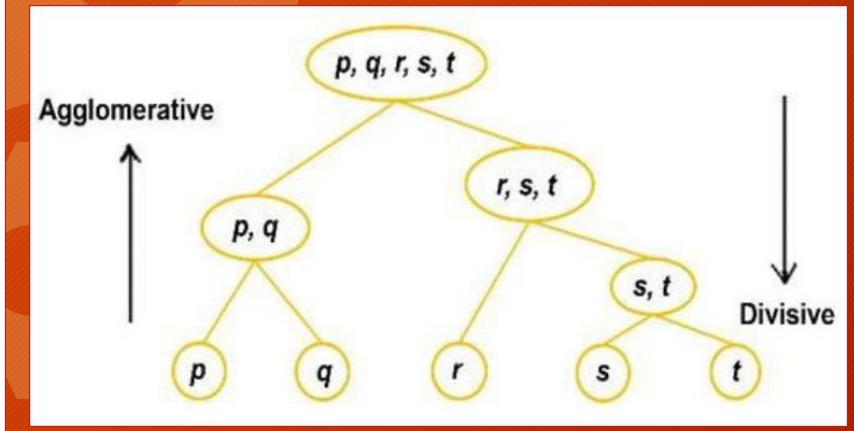
- Un-Supervised i.e. Make inference from datasets using only input vectors without referring to known, or labelled, outcomes
- Groups similar data-points together and discover underlying patterns.
- Define k which is the number of centroids you need in a dataset. A centroid is an imaginary or real location representing the center of a cluster
- Every data-point is allocated to each of the clusters through reducing the in-cluster sum of squares. (Within-cluster Sum of Squares)

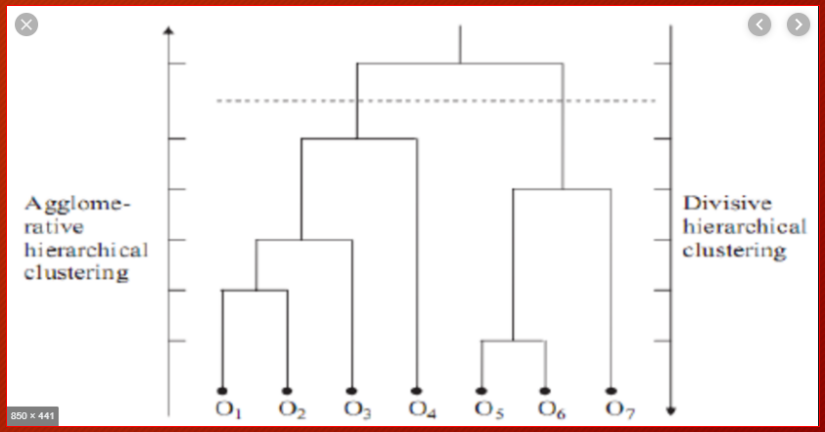
$$WSS = \sum_{i=1}^{m} (x_i - c_i)^2$$

Hierarchical Clustering



- It is a data mining technique used to group sets of objects in a way that objects in the same cluster are more similar to each other than those in other clusters
- **Agglomerative** – assign each point to a single cluster, then at each iteration, we merge the closest pair of clusters and repeat this step until a single cluster is left
  - **Single Linkage**
  - **Complete Linkage**
- **Divisive** – This works in the opposite way. Instead of starting with n clusters, we start with a single cluster and assign all data points to that cluster and in each iteration we separate the farthest point in the cluster until each cluster has a single point





Association Rule

TID	ITEMS
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- Data mining technique that finds interesting associations and relationships among large sets of data items.
- The rule shows how frequently an itemset occurs in a transaction e.g. Market Basket Analysis.
- Given a set of transactions we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

- **Support** – Indicates how frequently an item or item-set appears in the dataset. Tells us if the rule is significant.
- **Confidence** – Indicates how often a rule is found to be true. Tells us how likely the items will occur together
- **Lift** – This gives us an idea on the extent to which the occurrence of one item or item-set increases the occurrence of the other item or itemset

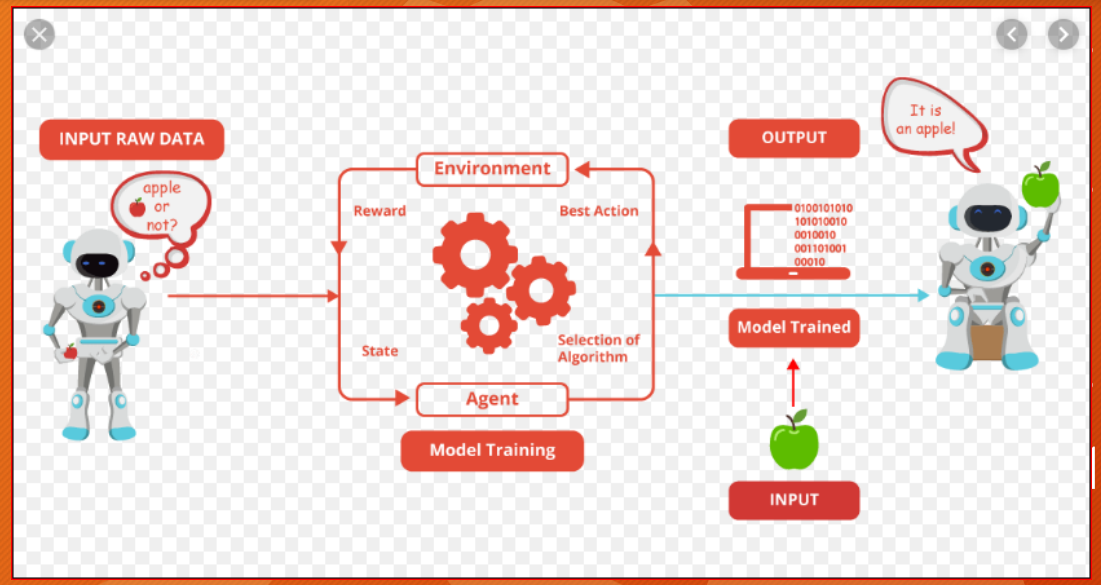
$$Rule: X \Rightarrow Y$$

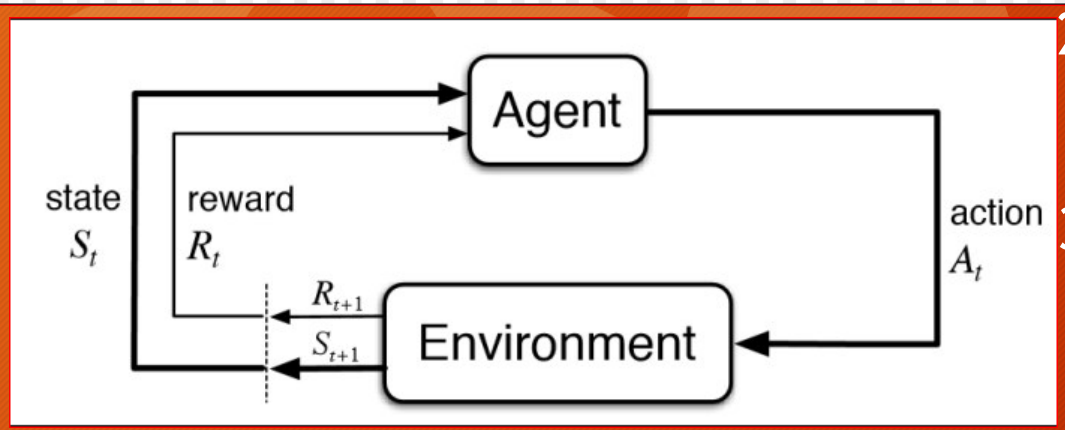
$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Reinforcement Learning





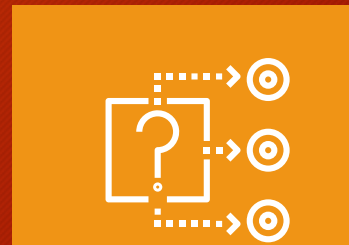
The agent learns to perform certain actions in an environment which leads it to maximum reward.

It does so through explorations of knowledge it learns by repeated trials of maximizing the reward

Three approaches:

1. Value Based -One tries to maximize a value function
2. Policy Based – one comes up with a policy such that the action performed at each state is optimal to gain maximum reward in the future
3. Model Based – One creates a virtual model for each environment, and the agent learns to perform in that specific environment

Regression Jupyter Notebooks



Hands-On practical sessions on 6 Regressors.