

Big Data Analytics Techniques

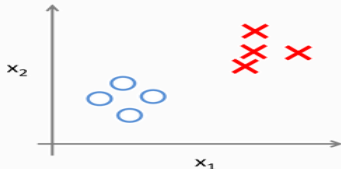
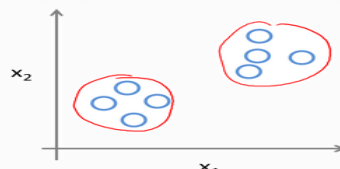
Presented by
Dr. Amany AbdElSamea



Outline

- Supervised and Unsupervised Learning
- Clustering
- Applications for Cluster Analysis
- Hard Clustering vs. Soft Clustering
- Types of Clustering
- K-mean clustering
- Association Rules
- Linear Regression
- Logistic Regression

Supervised and Unsupervised Learning

Supervised	Unsupervised
Input Data is labelled	Input Data is Unlabelled
Uses training Dataset	Uses just input dataset
Data is classified based on training dataset	Uses properties of given data to classify it.
Used for prediction	Used for Analysis
Divided into two types Regression & Classification	Divided into two types Clustering & Association
Known number of classes	Unknown number of classes
	
Use off-line analysis of data	Use Real-Time analysis of data

Big Data Techniques

Problem to solve	Category of techniques	Example
I want to group items by similarity	Clustering	K-mean clustering
I want to discover relationships between actions or items	Association rules	Apriori
I want to determine the relationship between the outcome and the input variables	Regression	Linear regression Logistic regression
I want to analyze my text data	Text analysis	Term-Frequency-Inverse Document-Frequency (TF-IDF)
I want to assign known labels to objects	Classification	Naïve Bayes Decision trees
I want to forecast the behavior of a temporal process	Time series analysis	ARIMA

Clustering

- **Clustering** is the process of dividing the datasets into groups, consisting of similar data-points.
- It is unsupervised machine learning technique
- Points in the same group are as similar as possible.
- Points in different group are as dissimilar as possible.

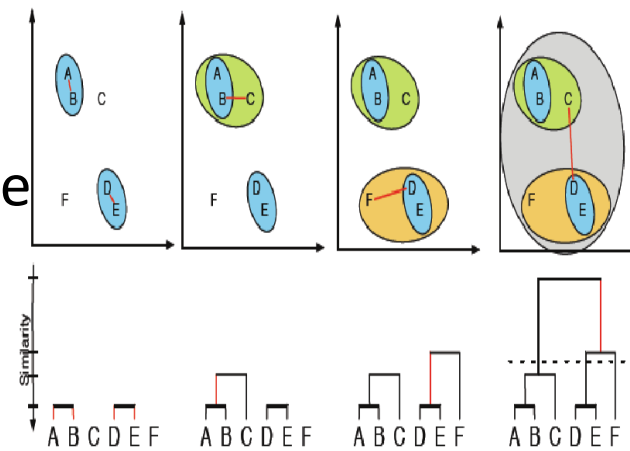


Applications for Cluster Analysis

- Marketing: discover distinct groups in customer bases, and develop targeted marketing programs.
- Biology: plant and animal taxonomies, genes functionality
- City planning: identify groups of houses according to their house type, value, and geographical location
- Also used for pattern recognition, data analysis, and image processing

Types of Clustering

- Exclusive Clustering:
 - ✓ Hard Clustering:
 - ✓ Data point/ Item belongs exclusively to one cluster
 - ✓ For example: k-Means Clustering
- Overlapping Clustering:
 - ✓ Soft Cluster
 - ✓ Data Point/Item belongs to multiple cluster
 - ✓ For example: Fuzzy/ C-Means Clustering
- Hierarchical Clustering:
 - ✓ The hierarchy of clusters is developed in the form of a tree in this technique, and this tree-shaped structure is known as the dendrogram.



K-mean clustering

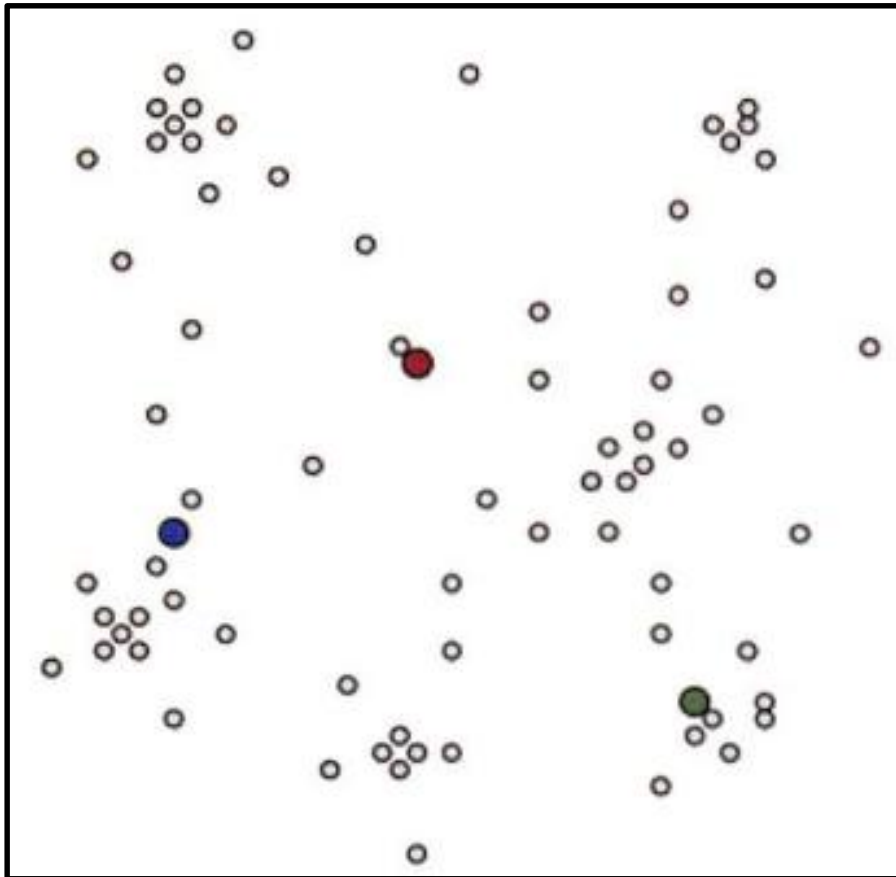
- It is a type of unsupervised learning used when you have unlabeled data
- Aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.
- **Input:** Numerical. There must be a distance metric defined over the variable space
 - Euclidian distance
- **Output:** The centers of each discovered cluster, and the assignment of each input datum to a cluster.
 - Centroid

K-mean Steps

1. Choose the value of k and the initial guesses for the centroids
2. Compute the distance from each data point to each centroid, and assign each point to the closest centroid
3. Compute the centroid of each newly defined cluster from step 2
4. Repeat steps 2 and 3 until the algorithm converges (no changes occur)

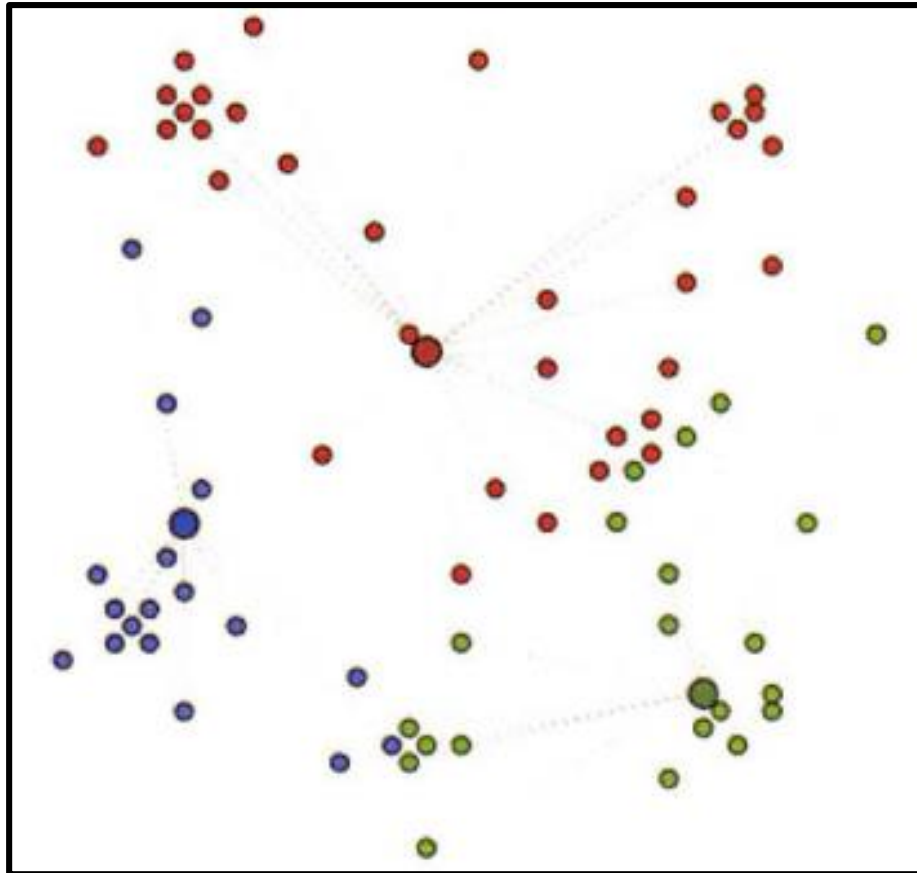
Step 1

Set $k = 3$ and initial clusters centers



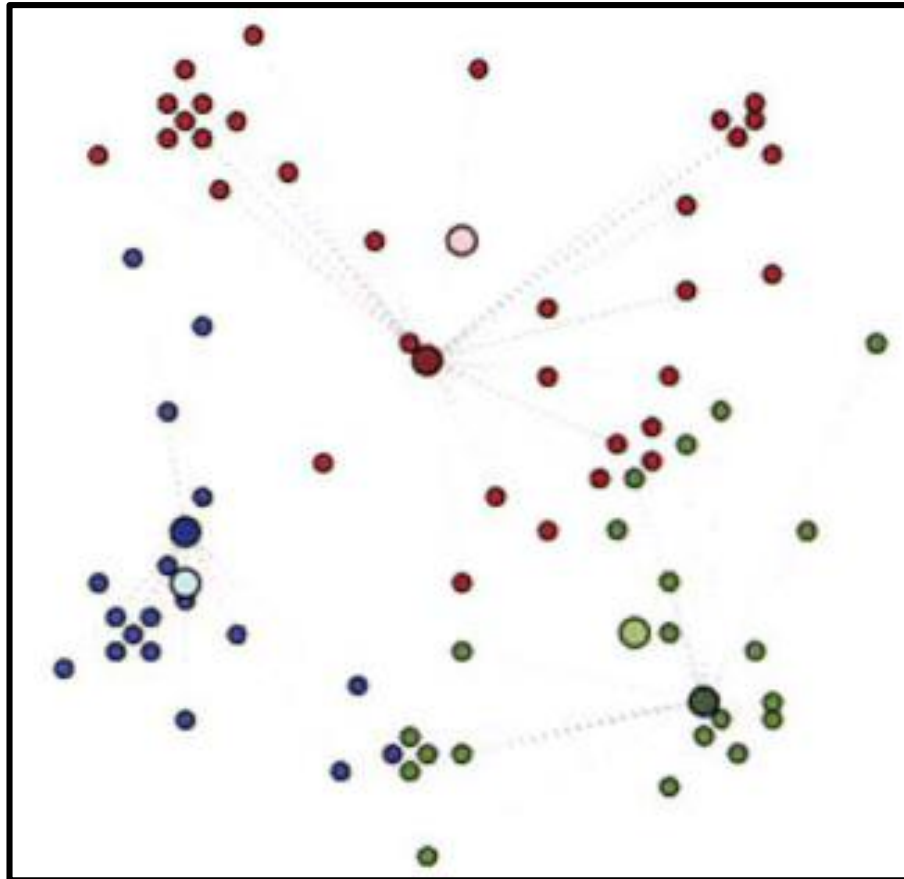
Step 2

Points are assigned to the closest centroid



Step 3

Compute centroids of the new clusters



Step 4

- Repeat steps 2 and 3 until convergence
- Convergence occurs when the centroids do not change or when the centroids oscillate back and forth
 - This can occur when one or more points have equal distances from the centroid centers

Association Rule

- Association rules is another unsupervised learning method.
- Not a predictive method. There is no prediction performed, but this method is used to discover relationships within the data.
- Help identify interesting patterns and connections among sets of items:
 - Rules take the form of “If X is observed, then Y is also observed”
- Use case: Understand customer buying habits by finding associations between the different items that customers place in their “shopping basket”
 - Known as market basket analysis
 - Example Apriori algorithm

Regression

- Regression focusses on the relationship between an outcome and its input variables.
 - Provides an estimate of the outcome based on the input values
 - Models how changes in the input variables affect the outcome
- Regression can find the input variables having the greatest statistical influence on the outcome
 - Then, can try to produce better values of input variables
 - E.g. – if 10-year-old reading level predicts students' later success, then try to improve early age reading level
- Approaches: Linear regression and Logistic regression

Linear Regression

- Models the relationship between several input variables and a continuous outcome variable
 - Assumption is that the relationship is linear
 - Various transformations can be used to achieve a linear relationship
- Linear regression models are probabilistic
 - Involves randomness and uncertainty
 - Not deterministic like Ohm's Law ($V=IR$)

Model Description

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{p-1} x_{p-1} + \epsilon$$

where:

y is the outcome variable

x_j are the input variables, for $j = 1, 2, \dots, p - 1$

β_0 is the value of y when each x_j equals zero

β_j is the change in y based on a unit change in x_j , for $j = 1, 2, \dots, p - 1$

ϵ is a random error term that represents the difference in the linear model and a particular observed value for y

Logistic Regression

- In linear regression modeling, the outcome variable is continuous – e.g., income ~ age and education
- In logistic regression, the outcome variable is categorical, like true/false, pass/fail, or yes/no

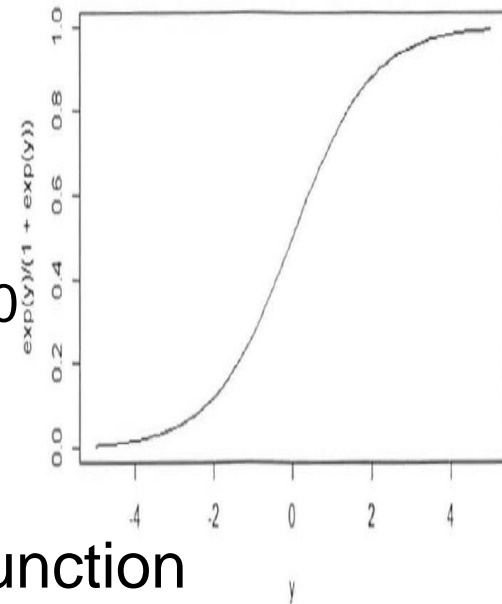
Logistic Regression

Model Description

- Logical regression is based on the logistic function

$$f(y) = \frac{e^y}{1 + e^y} \quad \text{for } -\infty < y < \infty$$

- As $y \rightarrow \infty$, $f(y) \rightarrow 1$; and as $y \rightarrow -\infty$, $f(y) \rightarrow 0$



- With the range of $f(y)$ as $(0,1)$, the logistic function models the probability of an outcome occurring

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots + \beta_{p-1} x_{p-1}$$

$$p(x_1, x_2, \dots, x_{p-1}) = f(y) = \frac{e^y}{1 + e^y} \quad \text{for } -\infty < y < \infty$$

In contrast to linear regression, the values of y are not directly observed; only the values of $f(y)$ in terms of success or failure are observed.

Questions