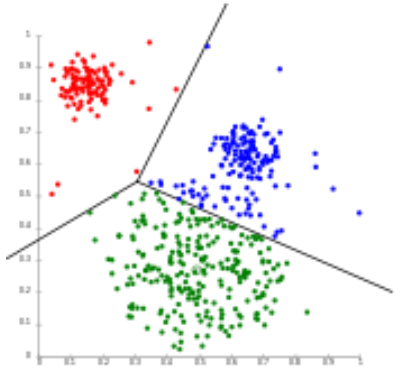
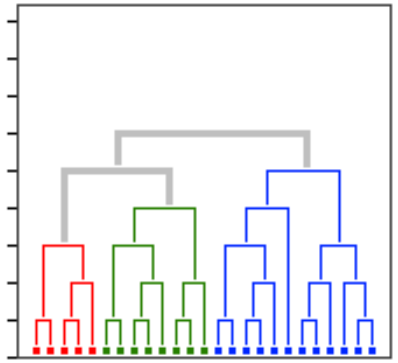
LIVEWIRE
FOR LIVE CAREERS

BUSINESS ANALYTICS

Clustering

- Only for continuous variable
- Will form a cluster with similar characteristic and will compare with other continuous variable or category variable
- Methods of clustering:
  - K-means
  - Hierarchical





Clustering…..

- K-means:

  - Clustering size will be fixed before starting the analysis, based on continuous variable

  - K-mean cluster also called iteration process

  - It will allocate the data point to assign to single group

  - By default, 3 is the cluster size

Note: If we don’t know the group size, we can go for Hierarchical clustering.

Clustering…..

- Hierarchical:
  - Clustering size will be fixed after the analysis, based on need
  - Also called Dendrogram (hierarchical structure)
  - Based on the method the cluster will be formed

Note: In SAS, we cannot do clustering as of now, the option might available in future.

Machine Learning

"Computers the ability to learn without being explicitly programmed"

- Machine learning tasks are typically classifed into three broad categories
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning

Supervised learning

- use an algorithm to learn the mapping function from the input to the output.
- when you have new input data that you can predict the output variables for that data.
- Classification
- Regression

Unsupervised Machine Learning

- where you only have input data and no corresponding output variables.
- Algorithms are left to their own devises to discover and present the interesting structure in the data.
- Unsupervised learning problems can be further grouped into clustering and association problems.
- Clustering
- Association

Classifiers in Machine Learning

- A Supervised function where the learned attribute is categorical
- K-Nearest Neighbors (KNN) algorithm
- Naïve Bayes Classifier

K-Nearest Neighbors (KNN) algorithm

- “Nearest neighbor” learning is also known as “Instance based” learning.
- It is based on Similarity calculation between instances.
- Default method is Euclidean distance
- Requirements:
  - K value based on the square root of sample size
  - Data should be normalized

Naïve Bayes Classifier

- A classification technique based on Bayes' Theorem with an assumption of independence among predictors.
- For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter.

TEXT MINING

- Extraction of hidden, previously unknown, and potentially useful information from (large amount of) textual data.
- Forming the WordCloud

CASE STUDY

- Case study using Insurance data set in R & SAS