Data science and Machine learning with Python

Designed by Abdur Rahman Joy - MCSD, MCPD, MCSE, MCTS, OCJP, Sr. Technical Trainer for VFX at IDB BISW (Scholarship program), and C#.net, R, Scala, Kotlin, JAVA, Android/IOS/Windows Mobile Apps, SQL server, Azure, Oracle, SharePoint Development, AWS , CEH, KALI Linux, Python, Data Science, Machine Learning ,Software Testing, Graphics, Multimedia and Game Developer at Joy Infosys and other premises like BITM, SkillsJob, PNTL, Leads Training and New Horizon inc , Cell #: +880-1712587348, email: jspaonline@gmail.com. Web URL: http://www.joyinfosys.com/me.

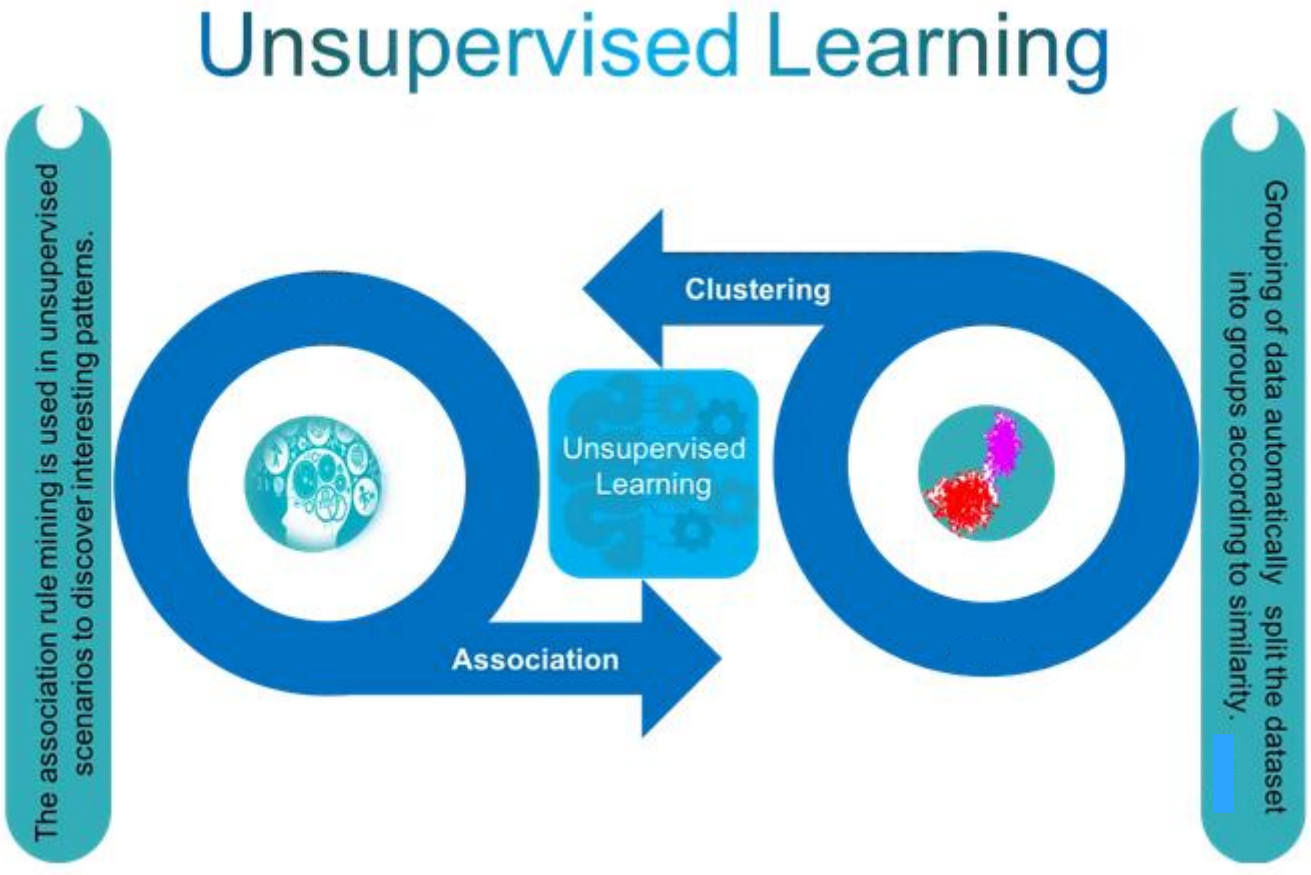
Unsupervised Learning

In Unsupervised Learning, there is no labeled data. The algorithm identifies the patterns within the dataset and learns them. The algorithm groups the data into various clusters based on their density. Using it, one can perform visualization on high dimensional data. One example of this type of Machine learning algorithm is the Principle Component Analysis.

The learning process in Unsupervised Learning is solely on the basis of finding patterns in the data. After learning the patterns, the model then makes conclusions.

Unsupervised Learning; is one of three types of machine learning i.e. Supervised Machine Learning, Unsupervised Machine Learning (UML) and Reinforcement Learning. The most common method in UML is cluster analysis. Cluster analysis is used for exploring hidden patterns or grouping in data behind data analysis. The algorithm used in this to draw inferences from data sets consisting of input data without labels. In short, UML is

- A technique with the idea to explore hidden gems/patterns.
- To find some intrinsic structure in data.
- That something can't be seen with naked eye requires magnifier (UML)



Designed by Abdur Rahman Joy - MCSD, MCPD, MCSE, MCTS, OCJP, Sr. Technical Trainer for VFX at IDB BISW (Scholarship program), and C#.net, R, Scala, Kotlin, JAVA, Android/IOS/Windows Mobile Apps, SQL server, Azure, Oracle, SharePoint Development, AWS , CEH, KALI Linux, Python, Data Science, Machine Learning ,Software Testing, Graphics, Multimedia and Game Developer at Joy Infosys and other premises like BITM, SkillsJob, PNTL, Leads Training and New Horizon inc , Cell #: +880-1712587348, email: jspaonline@gmail.com. Web URL: http://www.joyinfosys.com/me.

In UML systems are not trained by feeding it with the intended answers unlike what we do in supervised learning. Rather we just allow algorithms to infer patterns from a dataset without reference to known, or labelled outcomes. There are mainly 2 ways we achieve unsupervised learning goals

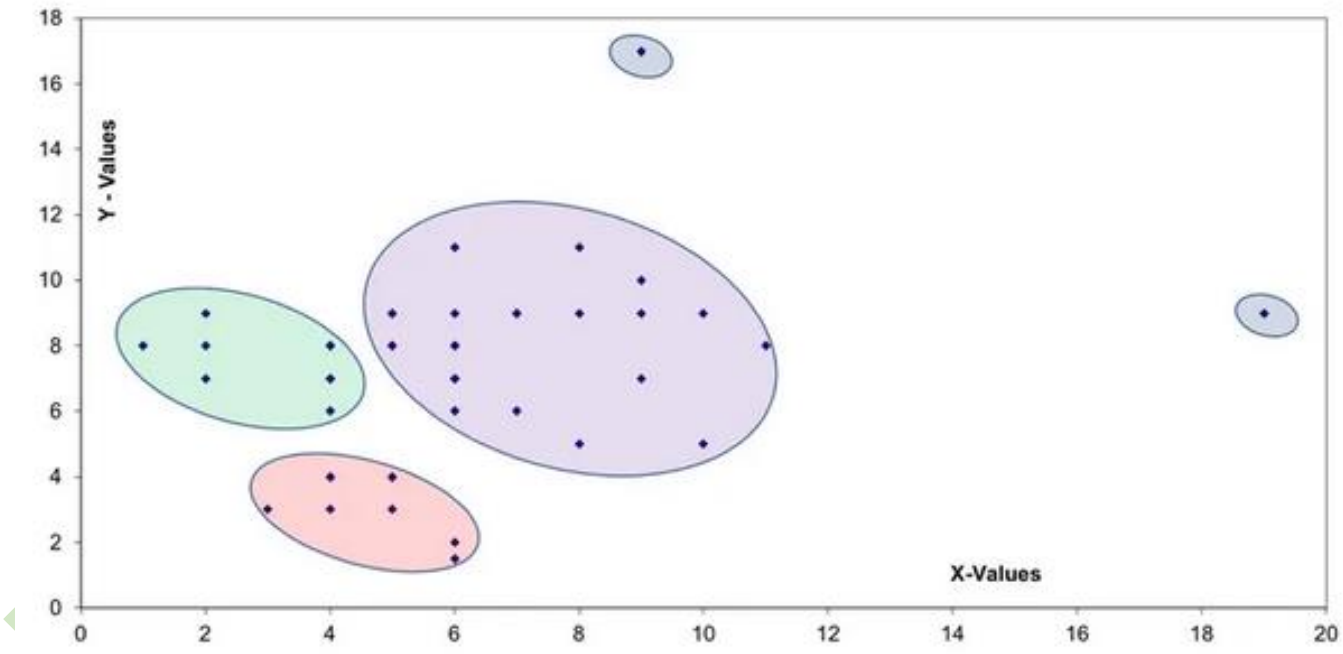
- ***Clustering*** – Partitions data into distinct clusters based on distance to the centroid of a cluster
- ***Association –*** The association rules are used to discover interesting patterns.

UML can't be applied to a regression or a classification problem as there is no idea what the values for the output data might be. Unsupervised learning algorithms get trained completely differently compared to supervise learning. Instead, algorithms here work as secret agents (yeah maybe 007 styles) for discovering the underlying structure of the data.

Clustering The Data

Clustering allows grouping of data points i.e. automatically split the dataset into groups according to similarity. Algorithms in this technique are based on one principle which is similarity/dissimilarity.

When clustering algorithm is used it classify the data points in groups with similar properties & features and underline them as a common reason to group. So each group has data points with similarity while intra-groups feature and properties are dissimilar to each other.



Because of the reason above it often overestimates the similarity between groups. Overestimates brings poor quality thus bad results. While clustering may work well for customer segmentation but do poorly on targeting.

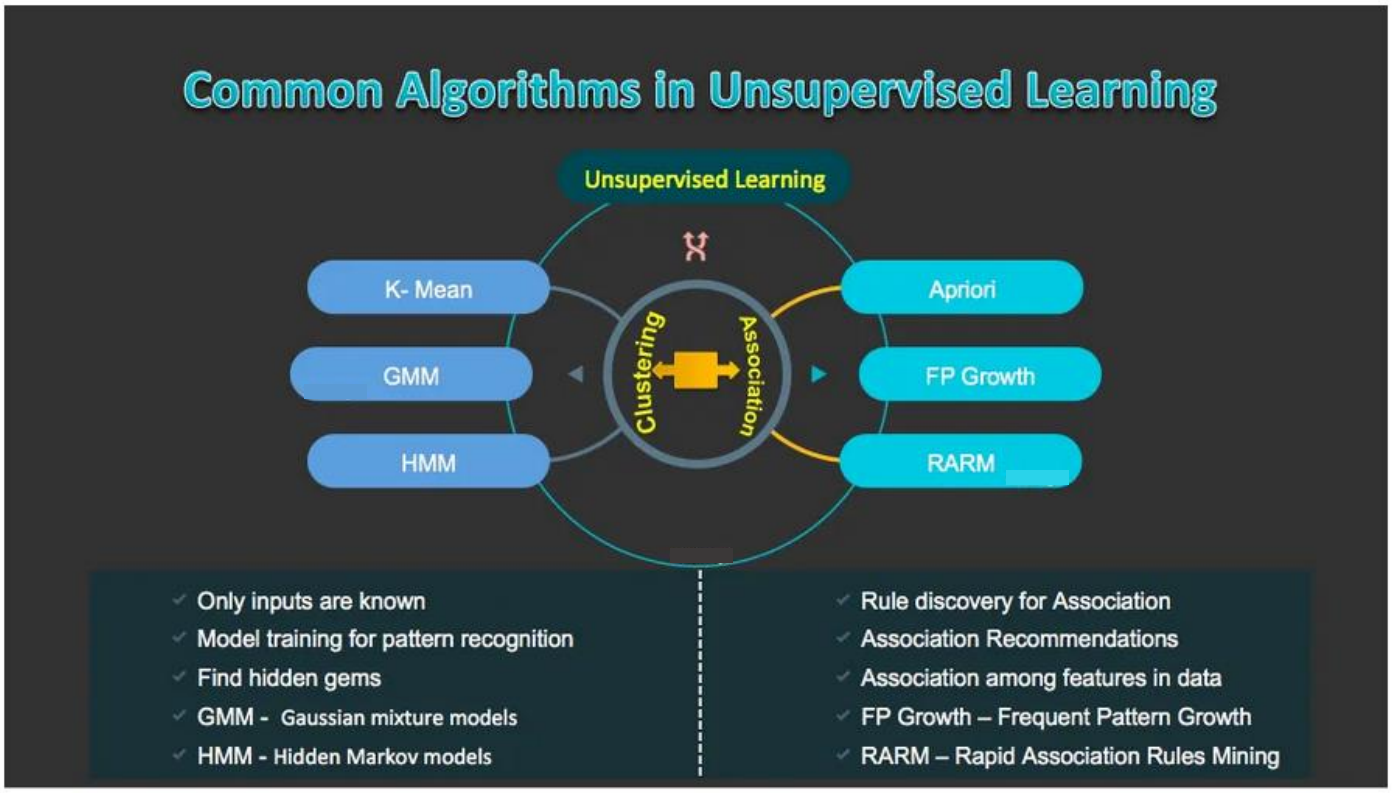
Common Clustering Algorithms

1. ***K-Means Clustering***: The most common and well know clustering algorithm. Super simple to understand, code and implement. It starts from the Centre points of vectors of the same length. It's pretty fast as well.
2. ***Mean-Shift Clustering:*** This algorithm works with baby step strategy. It's a sliding-window-based algorithm for finding its dense areas of data points. It gives freedom from selecting the number of clusters as it automatically discovers.
3. ***DBSCAN Clustering:*** It's a density-based spatial clustering of applications with noise algorithm. Based on density and starts with an arbitrary data point that has not been visited. It does not need a pre-set number of clusters.
4. ***Expectation Maximization:*** This clustering method uses Gaussian Mixture Models (GMM). To distribute GMM parameters for each cluster randomly it starts by selecting the number of clusters. GMMs are a lot more flexible in terms of cluster covariance and can have multiple clusters per data point.
5. ***Agglomerative Hierarchical Clustering:*** Cluster tree is built with the multilevel hierarchy of clusters. No assumptions on the number of clusters
    - *Agglomerative* – In this technique, its start with the points as each cluster as it moves forward; at each step, merge the closest pair of clusters until only one cluster left.
    - *Divisive* – Here its start with one, all-inclusive cluster. At each step, split a cluster until each cluster has a point.

Association Mining In Data

In contrast with sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions. Association rules mining is another key unsupervised data mining method, after clustering, that finds interesting associations (relationships, dependencies) in large sets of data items.

The association rule mining is used in unsupervised scenarios to discover interesting patterns. It also gets used in supervised learning as well. It identifies sets of items that often occur together. Data mining transform data into useful information.



E-commerce companies used this commonly and on an almost everyday basis. It is used for the shopping cart or basket analysis. It is very helpful to analysts to discover bundle goods often purchased at the same time and to develop more effective marketing strategies.

- Apriori Algorithm
- Frequent Pattern (FP) Growth Algorithm
- Rapid Association Rule Mining (RARM)
- ECLAT Algorithm
- Associated Sensor Pattern Mining of Data Stream (ASPMS) – For greater needs of frequent pattern mining algorithms.

An Angle for Unlabeled Data & Secret Labels

Unsupervised learning can be a challenging goal in itself. The training data consists of a set of input vectors x without any corresponding target values; hence known as learning/working without a supervisor.

The system does self-discovery of patterns, regularities and features etc. from the input data and relations for the input data over output data. Discovering similarities and dissimilarities to forms clusters i.e. self-discovery is the main target here.

Examples given to the learner are unlabeled, there is no error or reward signal to check a potential solution. Since no labels are given to the learning algorithm, leaving it on its own to find structure in its input. This distinguishes unsupervised learning from supervised learning and reinforcement learning.

- Pros
    - It can detect what human eyes cannot understand
    - The potential of hidden patterns can be very powerful for the business or even detect extremely amazing facts, fraud detection etc.
    - Output can decide the unexplored territories and new ventures for businesses. Exploratory analytics can be applied to understand the financial, business and working drivers behind what happened.

- Cons
    - As seen in the above explanation unsupervised learning is harder as compared to supervised learning.
    - It can be a costly affair, as we might need external expert look at the results for some time.
    - Usefulness of the results; are of any value or not is difficult to confirm since no answer labels are available.

Guarantee to no guarantee

What is guaranteed in unsupervised learning is; there is no guarantee or assurance that after so much of efforts and hard work of massaging the data we will find anything inspiring or something useful in data.

Since outcomes are known thus there is no way to decide accuracy of it. This makes supervised machine learning more applicable to real-world problems. The best time to use unsupervised machine learning is when you don't have data on desired outcomes, like determining a target market for an entirely new product that your business has never sold before.

Why is Unsupervised Machine Learning important?

One of the biggest advantages of unsupervised machine learning methods is reusability for other learning methods. The patterns uncovering & detection with unsupervised machine learning methods come in handy when implementing supervised machine learning.

- ***Anomaly Detection-*** This is the key feature for automatic discovery of unusual data points in a given dataset. As shown in the above picture the outlier can pinpoint fraudulent transactions/activities. Discovering faulty pieces of hardware or identifying an outlier caused by a human error during data entry are also seen here.

- ***Latent Variable Models-*** The data preprocessing happens in every business every day and most of the time too much similar data with the same features. Latent variable modelling helps in performing dimensional reduction i.e. reducing the number of features in the dataset or decomposing the dataset into multiple components.

Some of the Use Cases

Unsupervised learning is used to find anomalies in data or cluster data items to groups that humans can't assume themselves. Since output variables are unspecified here so algorithms look for structures in the data to describe and hidden distribution or structure of data. Some of the examples here are.

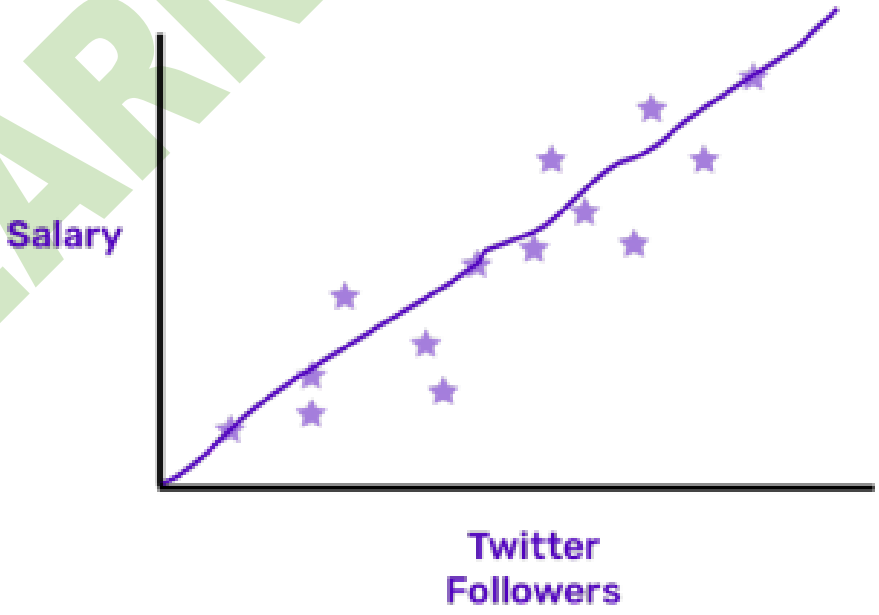
- Customer segmentation in different groups for specific interventions
- Product Targeting
- Market Categorization
- Recommendation Engines

Collecting and labelling a large set of sample patterns can be very expensive. How this type of learning helps business to see some potentials which are usually hidden normally. The goal in such unsupervised learning problems may be to discover groups of similar examples within the data, where it is called clustering, or to decide how the data is distributed in the space, known as density estimation.

Unsupervised machine learning cannot be directly applied to a regression because it is unknown what the output values could be, therefore making it impossible to train the algorithm how you normally would.

The easiest way to understand what's going on here is to think of a test. When you took tests in school, there were questions and answers; your grade was determined by how close your answers were to the actual ones (or the answer key). But imagine if there was no answer key, and there were only questions. How would you grade yourself?

Now apply this framework to machine learning. Traditional datasets in ML have labels (think: the answer key), and follow the logic of "X leads to Y." For example: we might want to figure out if people with more Twitter followers typically make higher salaries. We think that our input (Twitter followers) might lead to our output (salary), and we try to approximate what that relationship is.



Designed by Abdur Rahman Joy - MCSD, MCPD, MCSE, MCTS, OCJP, Sr. Technical Trainer for VFX at IDB BISW (Scholarship program), and C#.net, R, Scala, Kotlin, JAVA, Android/IOS/Windows Mobile Apps, SQL server, Azure, Oracle, SharePoint Development, AWS , CEH, KALI Linux, Python, Data Science, Machine Learning ,Software Testing, Graphics, Multimedia and Game Developer at Joy Infosys and other premises like BITM, SkillsJob, PNTL, Leads Training and New Horizon inc , Cell #: +880-1712587348, email: jspaonline@gmail.com. Web URL: http://www.joyinfosys.com/me.

The stars are data points, and machine learning works on creating a line that explains how the input and outcomes are related. But in unsupervised learning, there are no outcomes! We're *just* looking to analyze in the input, which is our Twitter followers. There is no salary, or Y, involved at all. Just like there not being an answer key for the test.

Maybe we don't have access to salary data, or we're just interested in different questions. It doesn't matter! The important thing is that there is no output to match to, and no line to draw that represents a relationship.

So what exactly is the goal of unsupervised learning then? What do we do when we only have input data without labels?