Unsupervised Learning.

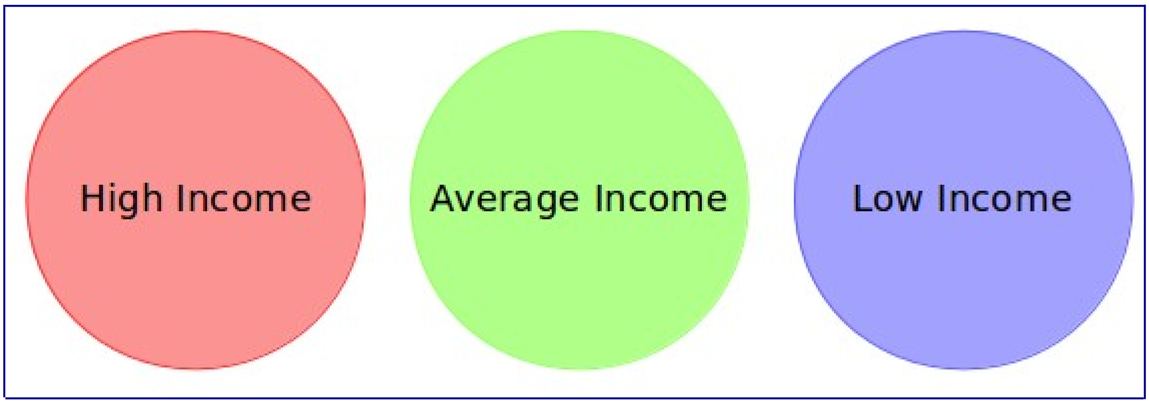
Unsupervised learning is another branch in machine learning, here the response y/outcome/label is not available in the dataset. SInce there is no response variable Y, the aim is to identify the clusters with in the data based on similarity with in the cluster members.

What is Clustering?

Let's kick things off with a simple example. A bank wants to give credit card offers to its customers. Currently, they look at the details of each customer and based on this information, decide which offer should be given to which customer.

Now, the bank can potentially have millions of customers. Does it make sense to look at the details of each customer separately and then make a decision? Certainly not! It is a manual process and will take a huge amount of time.

So what can the bank do? One option is to segment its customers into different groups. For instance, the bank can group the customers based on their income:



Can you see where I'm going with this? The bank can now make three different strategies or offers, one for each group. Here, instead of creating different strategies for individual customers, they only have to make 3 strategies. This will reduce the effort as well as the time.

The groups shown above are known as clusters and the process of creating these groups is known as clustering. Formally, we can say that:

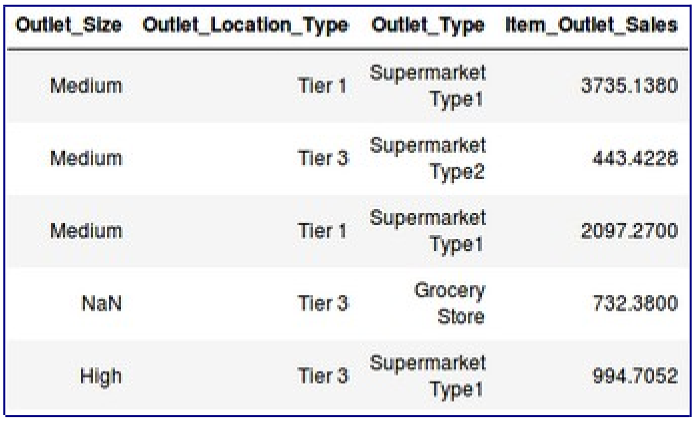
> *Clustering is the process of dividing the entire data into groups (also known as clusters) based on the patterns in the data.*

Can you guess which type of learning problem clustering is? Is it a supervised or unsupervised learning problem?

Think about it for a moment and make use of the example we just saw. Got it? Clustering is an unsupervised learning problem!

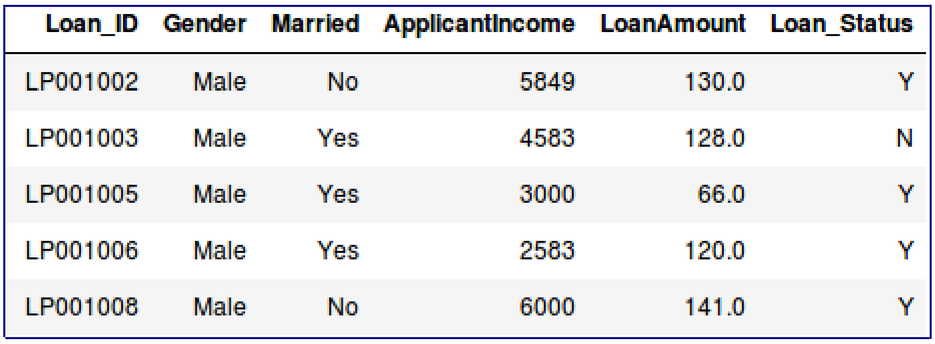
How is Clustering an Unsupervised Learning Problem?

Let's say you are working on a project where you need to predict the sales of a big mart:



Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
Medium	Tier 1	Supermarket Type1	3735.1380
Medium	Tier 3	Supermarket Type2	443.4228
Medium	Tier 1	Supermarket Type1	2097.2700
NaN	Tier 3	Grocery Store	732.3800
High	Tier 3	Supermarket Type1	994.7052

Or, a project where your task is to predict whether a loan will be approved or not:



Loan_ID	Gender	Married	ApplicantIncome	LoanAmount	Loan_Status
LP001002	Male	No	5849	130.0	Y
LP001003	Male	Yes	4583	128.0	N
LP001005	Male	Yes	3000	66.0	Y
LP001006	Male	Yes	2583	120.0	Y
LP001008	Male	No	6000	141.0	Y

We have a fixed target to predict in both of these situations. In the sales prediction problem, we have to predict the *Item_Outlet_Sales* based on *outlet_size, outlet_location_type*, etc. and in the loan approval problem, we have to predict the

Loan_Status depending on the *Gender*, *marital status*, *the income of the customers*, *etc*.

> So, when we have a target variable to predict based on a given set of predictors or independent variables, such problems are called supervised learning problems.

Now, there might be situations where we do *not* have any target variable to predict.

> Such problems, without any fixed target variable, are known as unsupervised learning problems. In these problems, we only have the independent variables and no target/dependent variable.

In clustering, we do not have a target to predict. We look at the data and then try to club similar observations and form different groups. Hence it is an unsupervised learning problem.

We now know what are clusters and the concept of clustering.

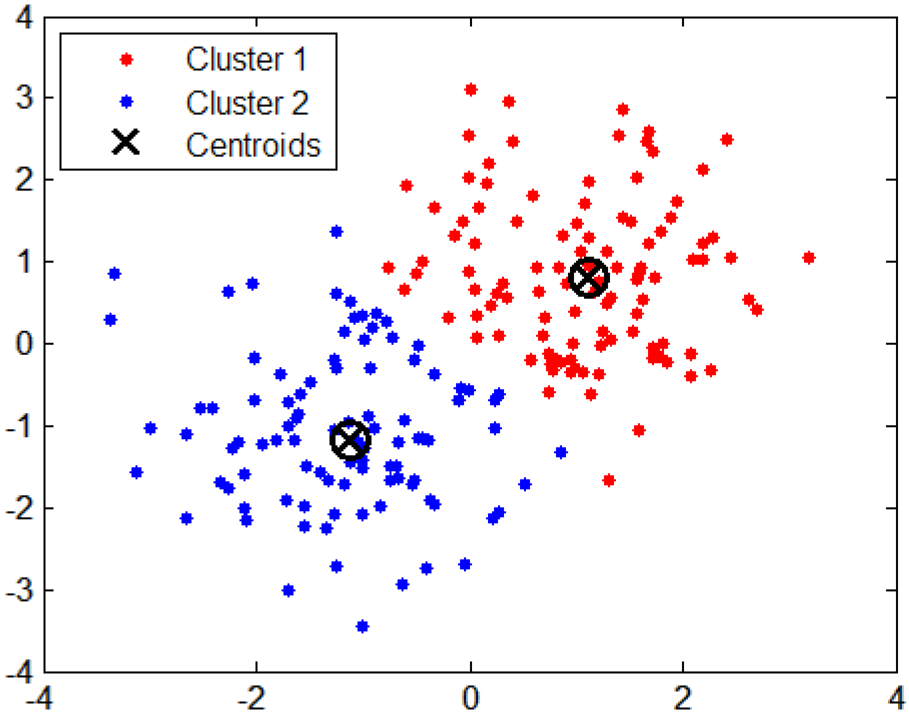
The k-means algorithm is one of the oldest and most commonly used clustering algorithms. It is a great starting point for new ML enthusiasts to pick up, given the simplicity of its implementation. As part of this post, we will review the origins of this algorithm and typical usage scenarios.

The History

The term "k-means" was first used by James MacQueen in 1967 as part of his paper on "Some methods for classification and analysis of multivariate observations". The standard algorithm was also used in Bell Labs as part of a technique in pulse code modulation in 1957. It was also published by In 1965 by E. W. Forgy and typically is also known as the Lloyd-Forgy method.

What Is K-Means?

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. The goal of the k-means algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. In the reference image below, K=2, and there are two clusters identified from the source dataset.



Reference

The outputs of executing a k-means on a dataset are:

- K centroids: Centroids for each of the k clusters identified from the dataset.
- Complete dataset labeled to ensure each data point is assigned to one of the clusters.

Where Can I Apply K-Means?

k-means can typically be applied to data that has a smaller number of dimensions, is numeric, and is continuous. Think of a scenario in which you want to make groups of similar things from a randomly distributed collection of things; k-means is very suitable for such scenarios.

Here is a list of ten interesting use cases for k-means.

1. **Document Classification**

Cluster documents in multiple categories based on tags, topics, and the content of the document. This is a very standard classification problem and k-means is a highly suitable algorithm for this purpose. The initial processing of the documents is needed to represent each document as a vector and uses term frequency to identify commonly used terms that help classify the document. The document vectors are then clustered to help identify similarity in document groups. Here is a sample implementation of the k-means for document clustering.

2. **Delivery Store Optimization**

Optimize the process of good delivery using truck drones by using a combination of k-means to find the optimal number of launch locations and a genetic algorithm to solve the truck route as a traveling salesman problem. Here is a whitepaper on the same topic.

3. **Identifying Crime Localities**

With data related to crimes available in specific localities in a city, the category of crime, the area of the crime, and the association between the two can give quality insight into crime-prone areas within a city or a locality. Here is an interesting paper based on crime data from Delhi FIRs.

4. **Customer Segmentation**

Clustering helps marketers improve their customer base, work on target areas, and segment customers based on purchase history, interests, or activity monitoring. Here is a white paper on how telecom providers can cluster pre-paid customers to identify patterns in terms of money spent in recharging, sending SMS, and browsing the internet. The classification would help the company target specific clusters of customers for specific campaigns.

5. **Fantasy League Stat Analysis**

Analyzing player stats has always been a critical element of the sporting world, and with increasing competition, machine learning has a critical role to play here. As an interesting exercise, if you would like to create a fantasy draft team and like to identify similar players based on player stats, k-means can be a useful option. Check out this article for details and a sample implementation.

6. **Insurance Fraud Detection**

Machine learning has a critical role to play in fraud detection and has numerous applications in automobile, healthcare, and insurance fraud detection. Utilizing past historical data on fraudulent claims, it is possible to isolate new claims based on its proximity to clusters that indicate fraudulent patterns. Since insurance fraud can potentially have a multi-million dollar impact on a company, the ability to detect frauds is crucial. Check out this white paper on using clustering in automobile insurance to detect frauds.

7. **Rideshare Data Analysis**

The publicly available Uber ride information dataset provides a large amount of valuable data around traffic, transit time, peak pickup localities, and more. Analyzing this data is useful not just in the context of Uber but also in providing insight into urban traffic patterns and helping us plan for the cities of the future. Here is an article with links to a sample dataset and a process for analyzing Uber data.

8. **Cyber-Profiling Criminals**

Cyber-profiling is the process of collecting data from individuals and groups to identify significant co-relations. The idea of cyber profiling is derived from criminal profiles, which provide information on the investigation division to classify the types of criminals who were at the crime scene. Here is an interesting white paper on how to cyber-profile users in an academic environment based on user data preferences.

9. **Call Record Detail Analysis**

A call detail record (CDR) is the information captured by telecom companies during the call, SMS, and internet activity of a customer. This information provides greater insights about the customer's needs when used with customer demographics. In this article, you will understand how you can cluster customer activities for 24 hours by using the unsupervised k-means clustering algorithm. It is used to understand segments of customers with respect to their usage by hours.

10. **Automatic Clustering of IT Alerts**

Large enterprise IT infrastructure technology components such as network, storage, or database generate large volumes of alert messages. Because alert messages potentially point to operational issues, they must be manually screened for prioritization for downstream processes. Clustering of data can provide insight into categories of alerts and mean time to repair, and help in failure predictions.