Cognizant

Welcome
to the course

Unsupervised Learning

Unsupervised Learning

What is Unsupervised Learning?

In machine learning, the challenge that unsupervised learning faces is that of trying to find hidden structures in unlabeled data.

Unsupervised Learning

Objectives

By the end of this module, you will be able to:

- Describe the concept of Unsupervised Learning
- Describe Clustering
- Define K means Clustering

Unsupervised Learning

What is Unsupervised Learning?

Let's take the example of Google to understand the concept of Unsupervised learning.

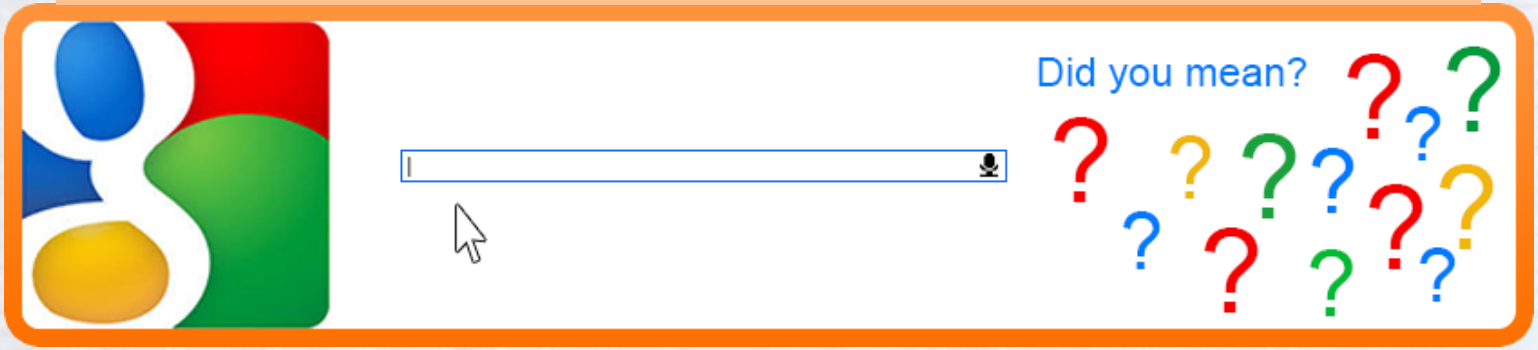
Have you noticed that while searching on Google, if one makes a typological error, Google provides the message: 'Did you mean_____?'

Let's take a look at how Google provides such suggestions.

Unsupervised Learning

What is Unsupervised Learning?

The shown message is the output of one Google's Machine learning algorithms. The algorithm keeps a tab of all the keywords that are being used in the search bar.



Unsupervised Learning

What is Unsupervised Learning?

The algorithm detects all the searches made in a couple of seconds after making the first one.

t
te
tel
tele
teles
telesc
telesco
telescop
telescope

Unsupervised Learning

What is Unsupervised Learning?

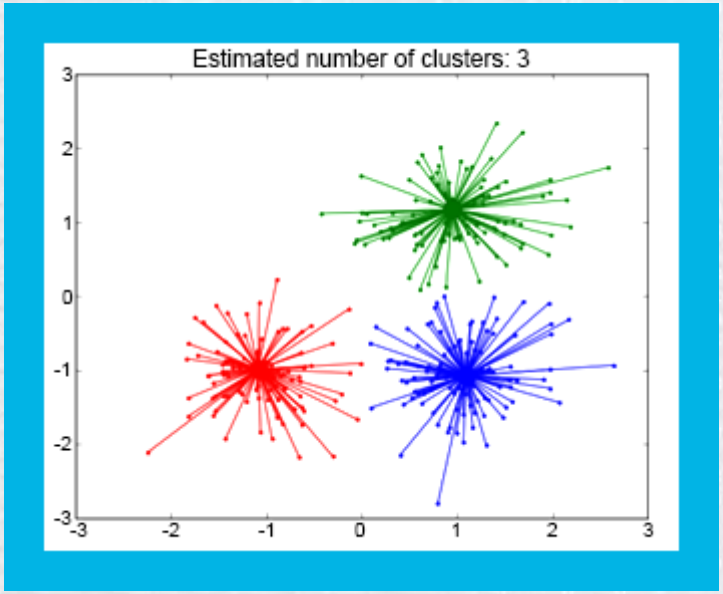
While searching for ‘telescope’ on Google, if somebody accidentally types ‘teliscope’, the algorithm would recognize the searching patterns and log it for future users who would make a similar typological error.

Unsupervised Learning

What is Clustering?

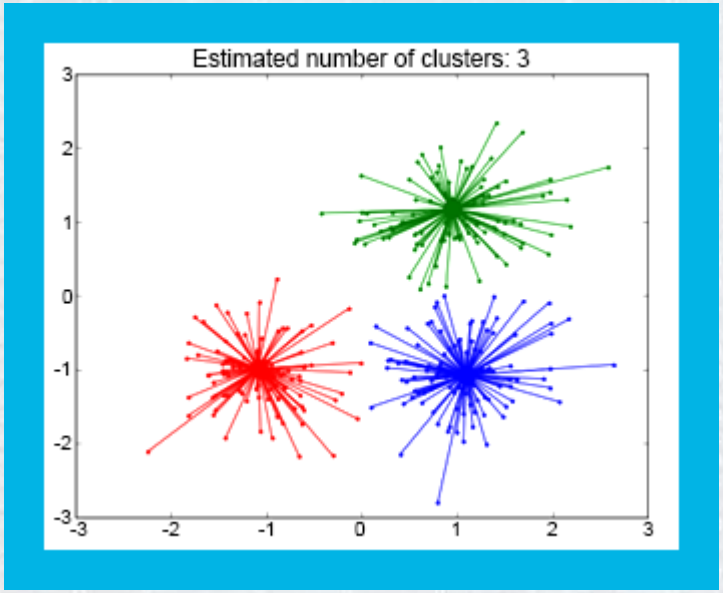
Clustering in machine learning is used to classify similar data elements into groups by analyzing the hidden data structure. K Means clustering is one of the most widely used clustering techniques.

Let us take a look at a more relatable example.



Unsupervised Learning

What is Clustering?



Cluster analysis or clustering can also be defined as the task of grouping a set of objects in such a way that objects in the same group called a cluster are more similar to each other than to those in other groups or clusters.

It is a common technique used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Unsupervised Learning

Clustering

Clustering is the most naturally occurring phenomenon. It enables identification of patterns in everyday life.

It also helps us find new friends or contacts in social networking sites like Facebook and LinkedIn.

In unsupervised learning, the machine or the algorithm analyses the underlying data pattern of any number of inputs using mathematical approaches. The mathematical approach displays the data pattern in the form of naturally occurring clusters and then the machine takes the appropriate decision

Unsupervised Learning

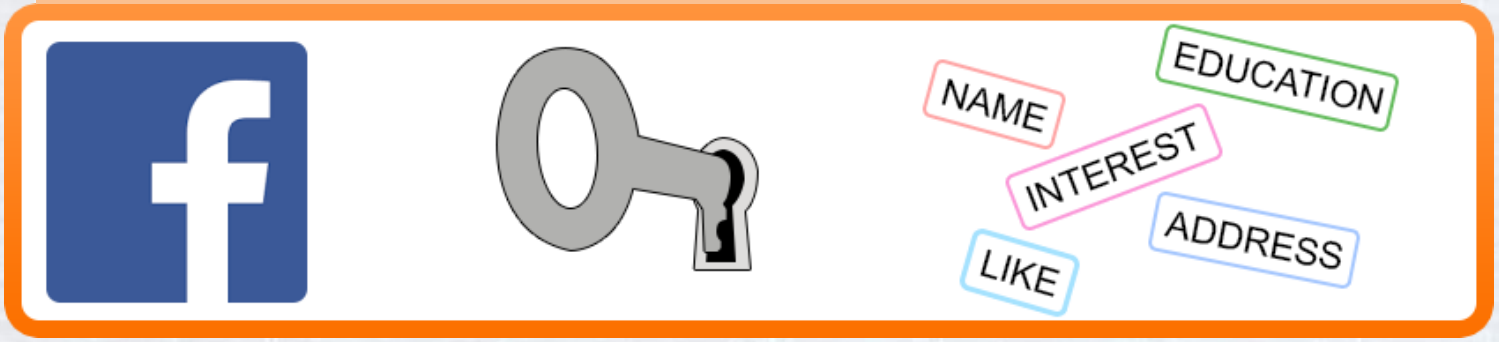
Clustering

Let's take the example of a feature of Facebook to understand the concept of clustering.

Unsupervised Learning

Clustering

When a user creates a profile for the first time, he has to provide some information to the application in the form of Name, Address, Education, Likes, Interests, etc.



Unsupervised Learning

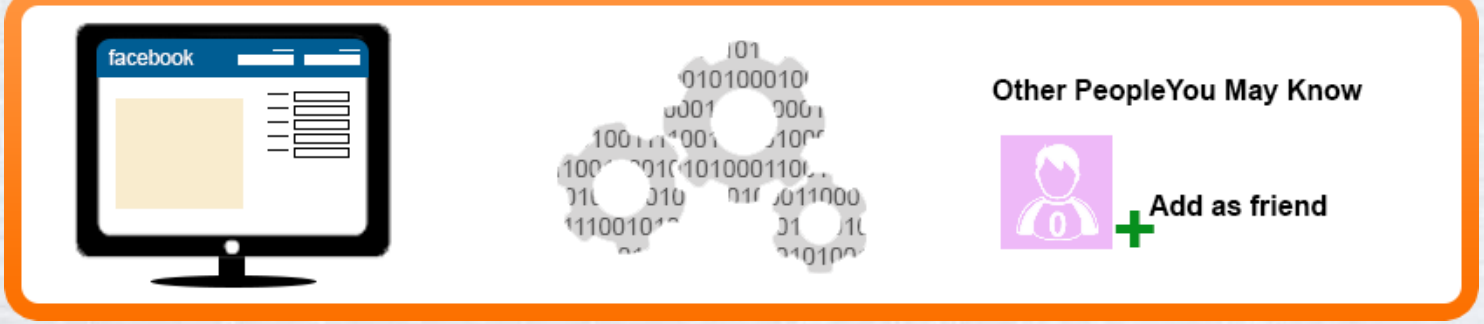
Clustering

Based on the information provided by the user, the algorithm forms clusters. The clusters are populated with people who fit the parameters like the same area of residence, similar educational background, similar taste in co-curricular activities. The algorithm also finds other people who fit the naturally occurring clusters.

Unsupervised Learning

Clustering

Based on the data pattern formed by the clusters, the algorithm takes the appropriate decision—whom to display in the 'Other People You May Know' list.



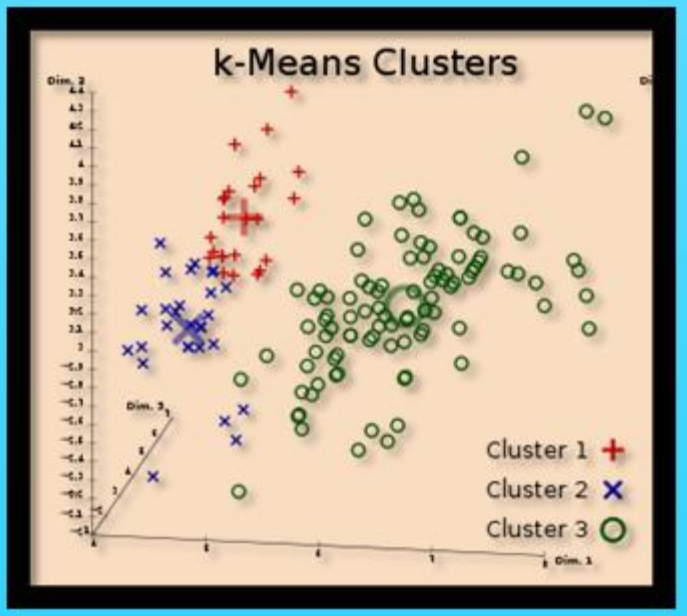
Unsupervised Learning

K-Means Clustering

The main objective of K-means is to partition or group '*n*' number of observations into '*k*' number of clusters. It works best on continuous variables.

$$d=\sqrt{(q_1-p_1)^2+(q_3-p_3)^2+\cdots+(q_n-p_n)^2}$$

The Euclidean distance measures the distance between two given points or coordinates (p1,q1) and (p2,q2). It can also be extended to measure distance between "n" number of data points.



Unsupervised Learning

Centroids

Centroids reflect the number of clusters and is the key input to the algorithm. Centroids are formulated based on logical inputs or inputs validated by trial and error method. The logical inputs differ, based on the context of the set of data.

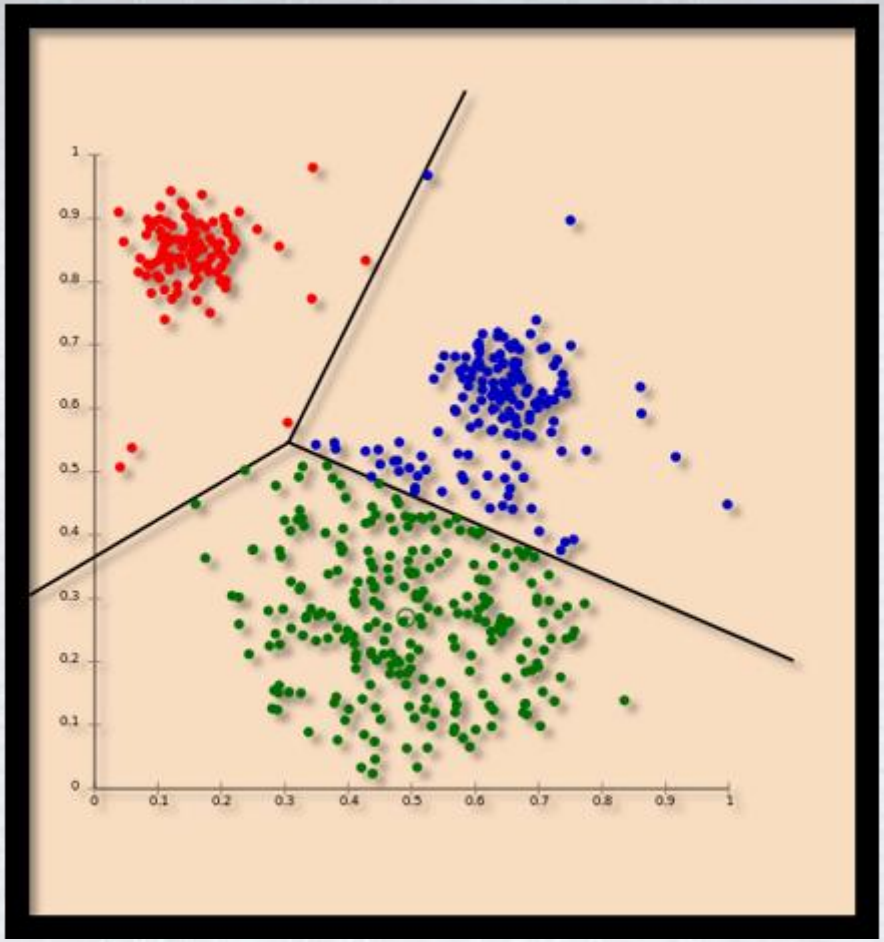
For Example: In a dataset comprising of apparels, the number of centroids can be the sizes of the T-shirts such as XS, S, M, L, XL and XXL. So, in the dataset comprising of apparels the number of centroids is 6.

Unsupervised Learning

K-Means Algorithm

In the K-means algorithm method, the steps followed are:

1. Select a random number of Centroids.
2. Place the centroids in a random space among the data points.
3. Find the closest centroid for each data point using the Euclidean distance formula.
4. Reposition the centroids by taking the mean of all the data points attached to the respective centroid.
5. Repeat steps 3 and 4 till the centroids have reached an optimal location, where repositioning does not yield any changes to the centroids.
6. Repeat Step 1 to increment the number of centroids.
7. Identify the optimal number of centroids through the incremental process.



Unsupervised Learning

Iterations

Let us look at a Use Case based approach, like Organizing Computing Clusters taking only two variables into account—Memory Consumption and CPU Cycles.

Three centroids are taken and marked '*x*'. The centroids have been assigned random values. The distance between the data points, marked in red, blue and green, are calculated. Let's take a look at the snapshots of the iterations performed.

Unsupervised Learning

Identifying Centroids

The graph shows the data points plotted by CPU cycles and Memory consumption.

The objective is to cluster computers based on memory consumption and CPU cycles. We want to identify the following clusters:

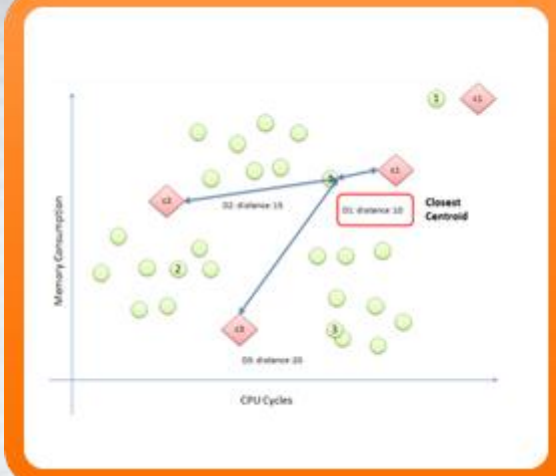
- Low memory and High CPU
- High memory and Low CPU
- Low memory and Low CPU

To achieve this goal, we have to first identify the centroids.

Unsupervised Learning

Identifying Centroids

Step 1

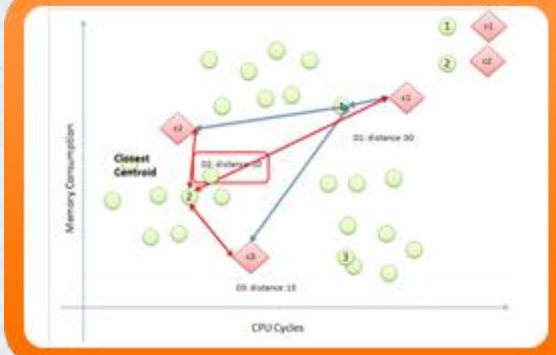


In K-Means clustering:

1. The first step is providing the number of centroids. In our case; we need 3 clusters, so we have our centroid count as 3.
2. To start positioning the centroids—three random data points—c1, c2 and c3, are chosen and those are considered as our initial centroids.
3. Then the distance between each data point and the three centroids—c1, c2 and c3—is calculated. For example, the distance between data point 1 and c1 is 10, data point 1 and c2 is 15, data point c1 and c3 is 20.

The data point 1 is closer to c1 with a minimum distance of 10

Step 2



1. Now, the objective is to find the nearest centroids to each data point.
2. The distance between data point 2 and c1 is 30, data point 2 and c2 is 10, data point 2 and c3 is 15.
3. The data point 2, is closer to c2 with minimum distance of 10.

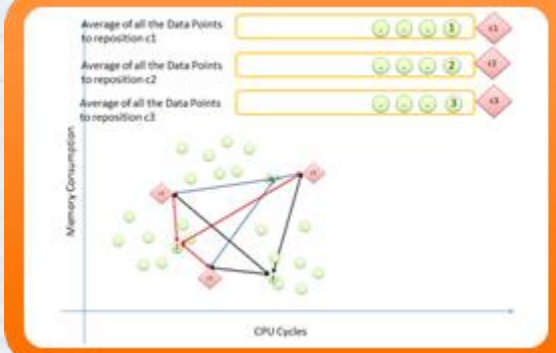
Unsupervised Learning

Identifying Centroids

Step 3

1. The nearest centroid for each data point is located. The minimum distance between a data point and the centroid is calculated.
2. Data points are assigned to the respective centroid; consider this a group.

Step 4



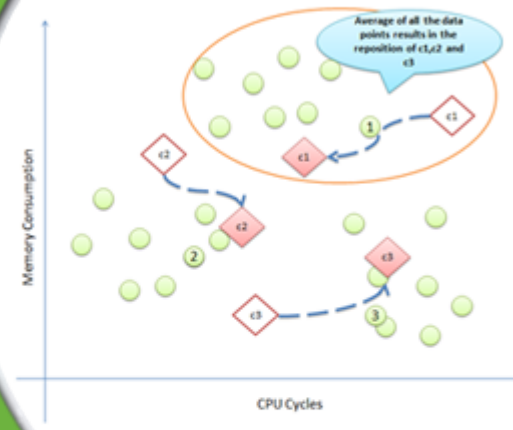
1. When no point is remaining, the step is complete.
2. In each centroid's group, the average of all the Data Points is calculated. The resulting mean is the new centroid for that group.

Unsupervised Learning

Optimizing Centroids

The positions of the centroids have been recalculated by taking a mean of all the data points assigned to them. The image shows the movement of the centroids to the new coordinates.

Once the centroids are recalculated, the next objective would be to optimize these.



Iterations

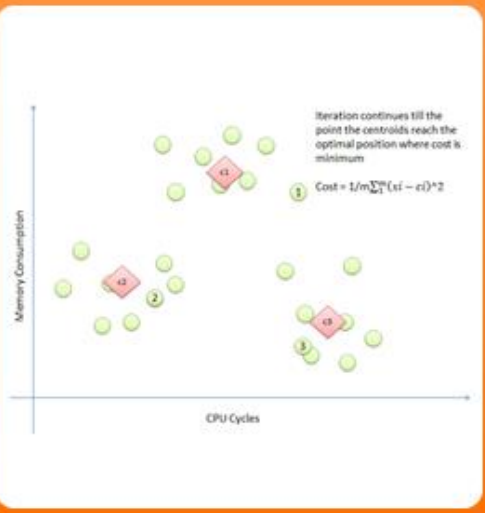
Reaching Optimal Clustering

Elbow Curve

Unsupervised Learning

Optimizing Centroids

Iterations



Iteration continues till the centroids reach the optimal position where cost is minimum. The formula of cost is given below.

The centroids have reached their optimal coordinates at certain junctures. The recalculation of the centroid coordinates does not yield any new group formation – the data points no longer move toward other centroids.

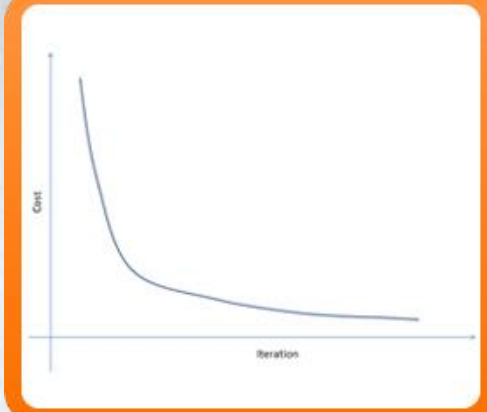
This represents the hidden clusters in the data set. From our example, each of these cluster belong to one of the following groups:

- Low Memory and High CPU
- High Memory and Low CPU
- Low Memory and Low CPU

Unsupervised Learning

Optimizing Centroids

Reaching Optimal Clustering



As the number of iterations increases, the cost decreases. It reaches a point where cost doesn't change much between iterations. This indicates that we have reached optimal clustering.

CPU Cycles

Elbow Curve

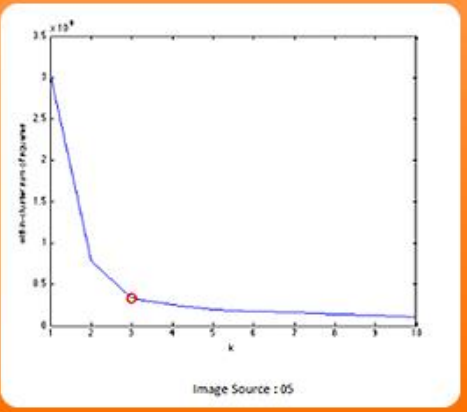


Image Source : 05

The Elbow Curve is the criterion through which the optimal number of centroids is identified. Initially, the algorithm starts with a random number of centroids, e.g. 2 and then iterates till a configured max data point e.g. 10.

For each iteration and the centroid selection, the cost is calculated, and when plotted it shows the elbow curve as shown here. The red circle on 3 indicates the optimal centroid after which the cost flattens out.

Unsupervised Learning

Applications of Unsupervised Learning

Unsupervised learning finds use in the following fields:

1. Market Segmentation
2. Social Network Analysis
3. Organize Computing Clusters
4. Astronomical Data Analysis
5. Image Compression

Unsupervised Learning

Summary

From this module, you have learned the following:

- Clustering in machine learning is used to place data elements into related groups without advance knowledge of the group definitions.
- K-means Clustering is the method of grouping or quantifying data in clusters.

Cognizant

Thank You