Unsupervised Learning

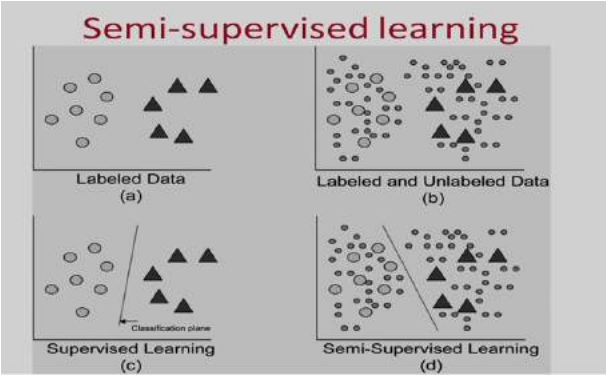
Wei Wu

May 11, 2022

Contents

1 Overview of machine learning

Supervised learning: Input is provided as a labelled dataset, a model can learn from it to provide the result of the problem easily.
Unsupervised learning: There is no complete and clean labelled dataset.
Reinforcement learning: The algorithms learn to react to an environment on their own.
Semi- supervised learning: It is an approach to machine learning that combines a small amount of labeled data with a large amount of unlabeled data during training.



2 What is Unsupervised Learning?

In supervised learning, we aim to train a model to be capable of mapping an input to output after learning some features, acquiring a generalization ability to correctly classify never-seen samples of data. But sometimes we don't know what is the output, as we only have the input data and we can't define an output label for each input sample. Let's suppose we're working for a company that sells clothes and we have data from previous customers: how much they spent, their ages and the day that they bought the product. Our task is to find a pattern or relationship between the variables in order to provide the company with useful information so they can create marketing strategies, decide on which type of client they should focus on to maximize the profits or which customer segment they can put more effort to expand in the market.

In this case, the output will not be a label since we can't model our needs in this way. Instead, our program should be able to group the customers accordingly to what makes them similar or unique. This grouping will be conducted from the features learned during the training phase, and in this case, we have unsupervised learning approach since there is no supervisor to provide labels for the inputs to map them to the output.

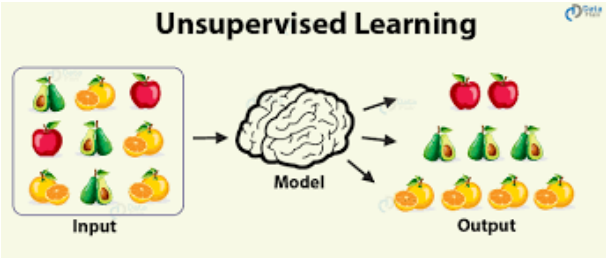
2.1 Definition

Unsupervised machine learning is the process of inferring underlying hidden patterns from historical data. Within such an approach, a machine learning model tries to find any similarities, differences, patterns, and structure in data by itself. No prior human intervention is needed.

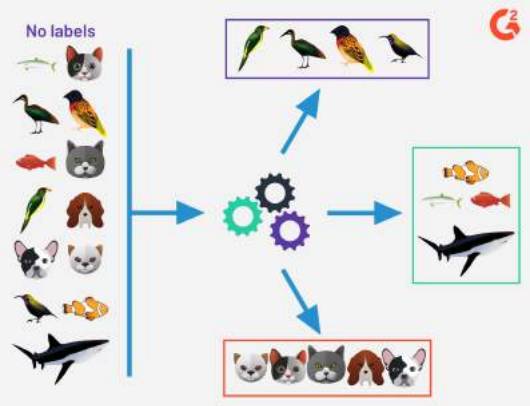
2.2 Examples

1.Picture a toddler. The child knows what the family cat looks like (provided they have one) but has no idea that there are a lot of other cats in the world that are all different. The thing is, if the kid sees another cat, he or she will still be able to recognize it as a cat through a set of features such as two ears, four legs, a tail, fur, whiskers, etc.

2.We have the following items and want to cluster them with unsupervised learning. They are clustered by some different characteristics.

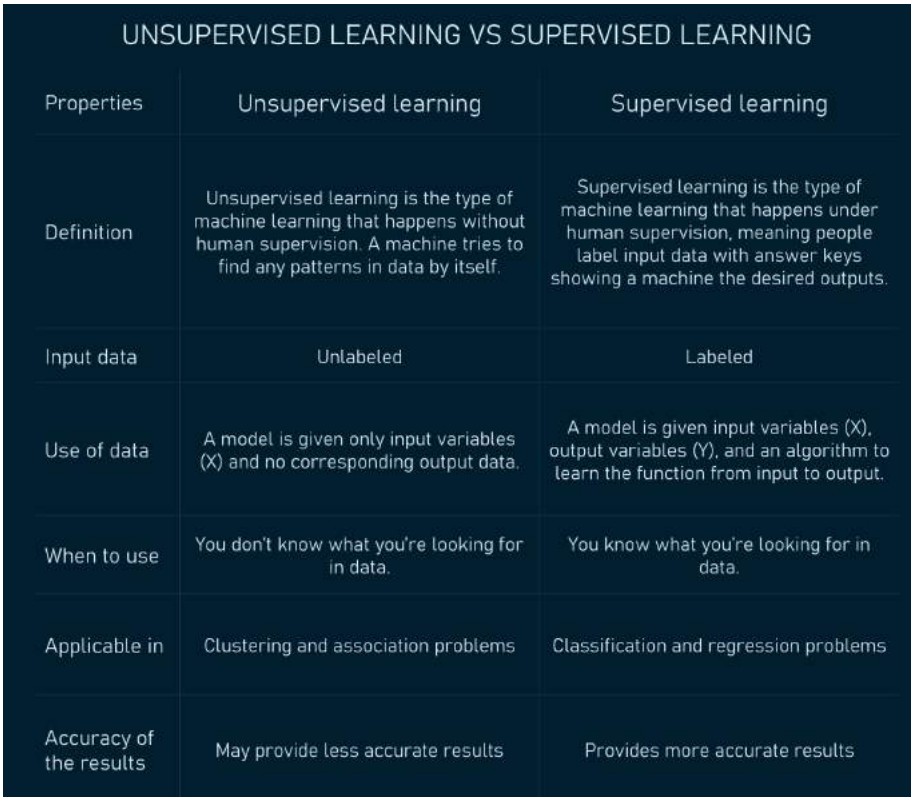


3.We have the following animals and try to cluster them with the help of unsupervised learning.



2.3 Differences compare with supervised learning

The key difference is that with supervised learning, a model learns to predict outputs based on the labeled dataset, meaning it already contains the examples of correct answers carefully mapped out by human supervisors. Unsupervised learning, on the other hand, implies that a model swims in the ocean of unlabeled input data, trying to make sense of it without human supervision.



2.4 Advantages

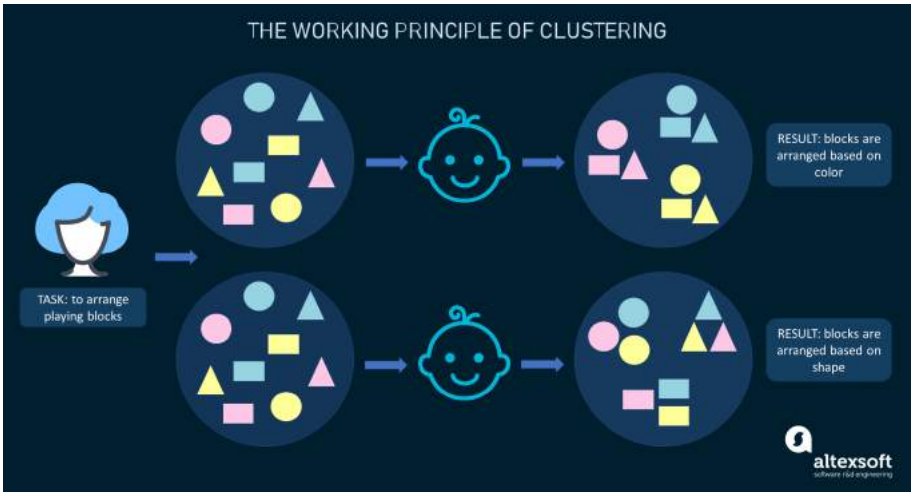
Why are we using Unsupervised Learning?

- Unsupervised learning is helpful for data science teams that don't know what they're looking for in data. It can be used to search for unknown similarities and differences in data and create corresponding groups. For example, user categorization by their social media activity.
- The given method doesn't require training data to be labeled, saving time spent on manual classification tasks.
- Unlabeled data is much easier and faster to get.
- It reduces the chance of human error and bias, which could occur during manual labeling processes.

3 Types Of Algorithms(commonly used)

3.1 Clustering

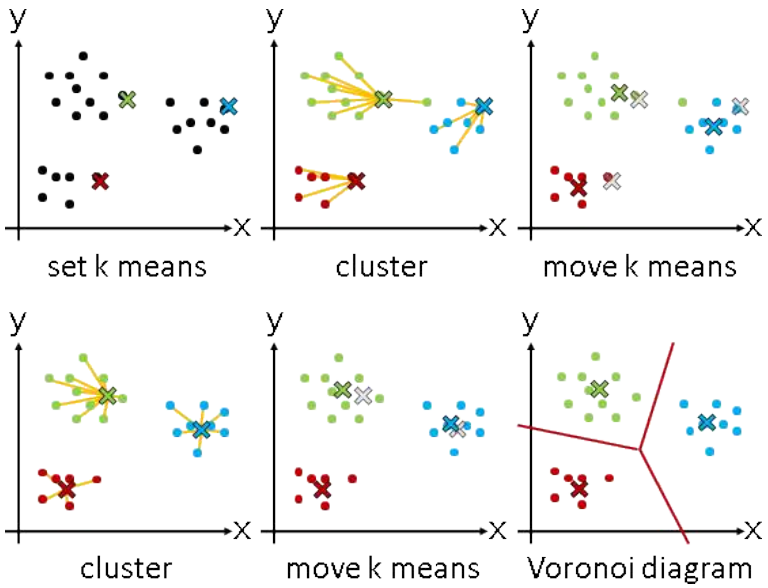
Clustering automatically categorizes data into groups according to similarity criteria.
For example, We need to arrange the following blocks of shapes and colors. According to different characteristics of this blocks, we could have two different results.



3.1.1 K-means Clustering

An iterative algorithm that categorizes data into a predefined number of groups or clusters.
Given an inital set of k means $m_1^{(1)}$,...,$m_k^{(1)}$,the algorithm proceeds by alternating between two steps:
Assignment Step: It measures distances between k centroids and individual data points.
$S_i^{(t)} = \{x_p : ||x_p - m_i^{(t)}||^2 \leq ||x_p - m_j^{(t)}||^2 \quad \forall j, 1 \leq j \leq k\}$,
and each x_p is assigned to exactly one $S^{(t)}$.
Update Step: Then, it updates the position of each centroid as the mean of respective data points belonging to each cluster.
$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$

It repeats the process until no centroid moves more than a given threshold.

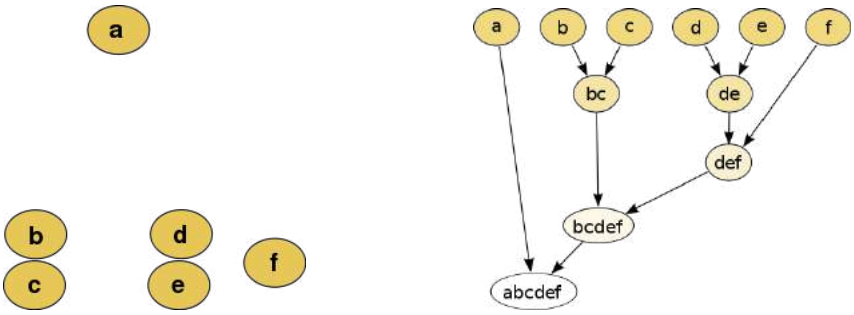


The algorithm has converged when the assignments no longer change.It is often presented as assigning objects to the nearest cluster by distance.
Advantages: Guarantees convergence. Can warm-start the positions of centroids. Generalizes to clusters of different shapes and sizes, such as elliptical clusters.
Disadvantages: Firstly, you have to select how many groups/classes there are. This isn't always trivial and ideally with a clustering algorithm we'd want it to figure those out for us because the point of it is to gain some insight from the data. K-means also starts with a random choice of cluster centers and therefore it may yield different clustering results on different runs of the algorithm. Thus, the results may not be repeatable and lack consistency.

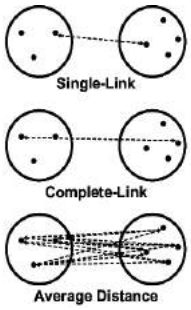
3.1.2 Hierarchical Clustering

An iterative algorithm that categorizes data into hierarchical groupings. It has 2 categories: top-down(divisive) or bottom-up(agglomerate).
(1) Agglomerative Hierarchical Clustering: Bottom-up algorithms treat each data point as a single cluster at the outset and then successively merge pairs of clusters until all clusters have been merged into a single cluster that contains all data points.

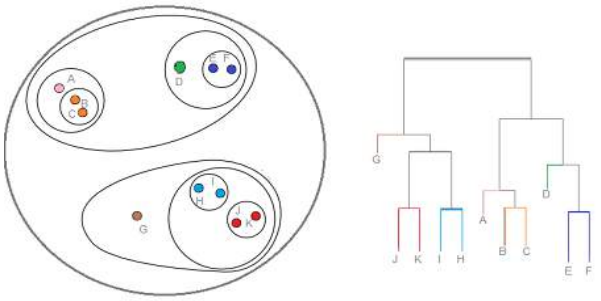
Example:



The example shows how six different clusters(data points) are merged step by step on distance until they all create one large cluster.
How to compute the distance between clusters that contain more than one data object? We have four methods below:
Centroid: Distance between the centroids of the two clusters.
Average Linkage: Average Distance between all pairs of points of the two clusters.
Single Linkage: Distance between the two most similar data objects of the two clusters.
Complete Linkage: Distance between the two most dissimilar data objects of the two clusters.



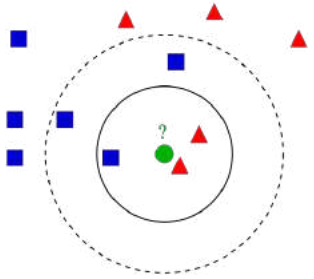
(2) Divisive Hierarchical Clustering: starts with the whole data set as a single cluster and then divides clusters step by step into small clusters.



The advantage of Hierarchical Clustering is we don't have to pre-specify the clusters. However, it doesn't work very well on vast amounts of data or huge datasets. And there are some disadvantages of the Hierarchical Clustering algorithm that it is not suitable for large datasets.

3.1.3 K nearest neighbors

It is the simplest of all machine learning classifiers. It works very well when there is a distance between examples. The learning speed is slow when the training set is large, and the distance calculation is nontrivial.
Example:



The test sample (green dot) should be classified either to blue squares or to red triangles.
If k = 3 (solid line circle) it is assigned to the red triangles because there are 2 triangles and only 1 square inside the inner circle.
If k = 5 (dashed line circle) it is assigned to the blue squares (3 squares vs. 2 triangles inside the outer circle).

Advantages of KNN:
1. No Training Period: KNN is called Lazy Learner. There is no training period for it. It stores the training dataset and learns from it only at the time

of making real time predictions. So it is much faster than other algorithms.
2.Since the KNN algorithm requires no training before making predictions, new data can be added seamlessly which will not impact the accuracy of the algorithm.
Disadvantages of KNN:
1. Does not work well with large dataset: In large datasets, the cost of calculating the distance between the new point and each existing points is huge which degrades the performance of the algorithm.
2. Does not work well with high dimensions: The KNN algorithm doesn't work well with high dimensional data because with large number of dimensions, it becomes difficult for the algorithm to calculate the distance in each dimension.

3.2 Association Analysis

Association Analysis is the task of finding interesting relationships in large datasets. And there are two forms to look for the relationship: Association rules and Frequent Itemset .

3.2.1 Association Rules

Association rules allow you to establish associations among data objects inside large databases. It is descriptive not the predictive method, generally used to discover the relationships between variables in large databases. For example, people that buy a new home most likely to buy new furniture. In addition, Amazon will recommend you some sport clothes if the last search was related to sports.
The strength of a given association rule is measured by two main parameters: support and confidence. Support refers to how often a given rule appears in the database being mined. Confidence refers to the amount of times a given rule turns out to be true in practice.

Support and Confidence
Definition 1: Support of itemset A, denoted support(A)= frequency an Itemset A appears in the dataset T:
support(A)= $\frac{|\{t \in T : A \subseteq t\}|}{|T|}$, t is an itemset.
Definition 2: Let A and B be two itemsets. An association rule $A \rightarrow B$

asserts that if a transaction contains A, it is also likely to contain B.
Definition 3: The support of an association rule $A \rightarrow B$ is $|AB|$.
Definition 4: The Confidence is the percentage of all transactions satisfying X that also satisfy Y.
Definition 5: The confidence of an association rule $A \rightarrow B$ is $\frac{|AB|}{|A|}$:
$\text{conf}(A \Rightarrow B) = \frac{supp(A \cap B)}{supp(A)} = \frac{\text{number of transactions containing A and B}}{\text{number of transactions containing A}}$

Example:
We have Itemsets: A={apple}, B={peach and apple}, C={peach and banana}, D={strawberry}, E={banana}
Dataset T={A,B,C,D,E}
Then we have the following information:
support(A)=$\frac{2}{5}$
$\text{conf}(A \Rightarrow B)$=$\frac{1}{2}$

Advantages of association rules : There are many benefits of using Association rules like finding the pattern that helps understand the correlations and co-occurrences between data sets.

Disadvantages of association rules :
(1) It requires high computation if the itemsets are very large and the minimum support is kept very low.
(2) The entire database needs to be scanned.

3.2.2 Apriori Algorithm

Apriori Algorithm uses frequent itemsets to generate association rules and it is based on the Apriori principle mentioned above. Apriori algorithms are designed to process databases containing transactional information content (e.g., lists of items purchased by customers).

Frequent itemset
Frequent itemset is an itemset whose support value is greater than threshold value . Which means: support(X)$\geq$ minsup.

Example on finding Frequent itemsets:

Consider the given dataset with given transactions.

Transaction ID	Items
1	{A,C,D}
2	{B,C,E}
3	{A,B,C,E}
4	{B,E}
5	{A,B,C,E}

We can try to calculate the support value of different itemsets.
support(A)=$\frac{3}{5}$, support(D)=$\frac{1}{5}$, ...
Let minsup=$\frac{2}{5}$, then we have :

Itemset	Support	F/I	Itemset	Support	F/I
A	3/5	F	ABD	0/5	I
B	4/5	F	ABE	2/5	F
C	4/5	F	ACD	1/5	I
D	1/5	I	ACE	2/5	F
E	4/5	F	ADE	0/5	I
AB	2/5	F	BCD	0/5	I
AC	3/5	F	BCE	3/5	F
AD	1/5	I	BDE	0/5	I
AE	2/5	F	CDE	0/5	I
BC	3/5	F	ABCD	0/5	I
BD	0/5	I	ABCE	2/5	F
BE	4/5	F	ABDE	0/5	I
CD	1/5	I	ACDE	0/5	I
CE	3/5	F	BCDE	0/5	I
DE	0/5	I	ABCDE	0/5	I
ABC	2/5	F			

Total itemsets: 31 (= 2^5 - 1), 5 is the amount of single item, which means A,B,C,D,E in this example.
Amount of Frequent itemsets: 15
Amount of Infrequent itemsets: 16

We can try to avoid considering all the items based on Apriori Principle:
(1) If an itemset is infrequent then all its supersets are infrequent.
(2) If an itemset is frequent then all its subsets are frequent.
(What is superset? - The opposite of subset
A is a superset of another set B if all elements of the set B are elements of the set A. The superset relationship is denoted as $A \supset B$. Opposing we say, B is the subset of A.)
With the help of these two principles, we can try to make the algorithm easier, for example:
Itemset D is infrequent, so all its supersets are infrequent. (Which means all the itemsets with element D are infrequent, so we can ignore it and keep looking for frequent itemsets.)
Besides frequent itemsets, two more important itemsets will be introduced below:

Closed itemset
An itemset X is frequent and no immediate superset of X has same support as X.

Maximal itemset
An itemset X is frequent and no immediate supersets of X are frequent.

Itemset	Support	F/I	C/NC	M/NM
A	3/5	F	NC	NM
B	4/5	F	NC	NM
C	4/5	F	C	NM
D	1/5	I	NC	NM
E	4/5	F	NC	NM
AB	2/5	F	NC	NM
AC	3/5	F	C	NM
AE	2/5	F	NC	NM
BC	3/5	F	NC	NM
BE	4/5	F	C	NM
CE	3/5	F	NC	NM
ABC	2/5	F	NC	NM
ABE	2/5	F	NC	NM
ACE	2/5	F	NC	NM
BCE	3/5	F	C	NM
ABCE	2/5	F	C	M

In this example we can find out that for the frequent itemset C with supp(C)=$\frac{4}{5}$, none of its supersets have same support as C, so itemset C is closed.
And the frequent itemset $\{A, B, C, E\}$ is the only maximal itemset because it doesn't have (frequent) supersets.
In conclusion, F=15, C=5, M=1 .
So the relationship between these three itemsets is: $M \subset C \subset F$.

So how does Apriori Algorithm actually work? - It will be shown in the following example:
We have a dataset T with 4 transactions, let minsup = $\frac{2}{4} = \frac{1}{2}$, mincof = $\frac{1}{2}$.

Transaction ID	Itemsets
1	A,C,D
2	B,C,E
3	A,B,C,E
4	B,E

Step 1: Scan T, calculate the support for each candidates, then we can get C_1 (the candidate set of level k) and L_1 (the frequent dataset).

Itemset	Support
A	2/4
B	3/4
C	3/4
D	1/4
E	3/4

C1

Itemset	Support
A	2/4
B	3/4
C	3/4
E	3/4

L1

Step 2: Combine the itemsets in L_1 to generate C_2.

Itemset
A,B
A,C
A,E
B,C
B,E
C,E

Step 3: Scan C_2 in order to get L_2.

Itemset	Support
A,B	1/4
A,C	2/4
A,E	1/4
B,C	2/4
B,E	3/4
C,E	2/4

C2

Itemset	Support
A,C	2/4
B,C	2/4
B,E	3/4
C,E	2/4

L2

Step 4: Combine itemsets in L_2 to get C_3:

Itemset
A,B,C
B,C,E

Step 5: Scan C_3.

Itemset	Support
A,B,C	1/4
B,C,E	2/4

C3

Itemset	Support
B,C,E	2/4

L3

$\Rightarrow$ STOP, the algorithm can't run anymore because no new frequent itemsets are identified.
$\Rightarrow \{B, C, E\}$ is the final frequent itemset.
$\Rightarrow$ Calculate the confidence and compare with the mincof to find out which association rules are possible.

When $\{B, C, E\}$ is the frequent itemset. We have the following possibilities for the association rules:
(1) $B \wedge C \rightarrow E$, $E \rightarrow B \wedge C$
(2) $B \wedge E \rightarrow C$, $C \rightarrow B \wedge E$
(3) $C \wedge E \rightarrow B$, $B \rightarrow C \wedge E$
Then we need to find out which possibilities are acceptable.

(1) conf($B \wedge C \rightarrow E$)= $\frac{2}{2} = 1$, conf($E \rightarrow B \wedge C$) = $\frac{2}{3}$
(2) conf($B \wedge E \rightarrow C$) = $\frac{2}{3}$, conf($C \rightarrow B \wedge E$) = $\frac{2}{3}$
(3) conf($C \wedge E \rightarrow B$) = $\frac{2}{2} = 1$, conf(is $B \rightarrow C \wedge E$) = $\frac{2}{3}$
All confidences are higher than minconf, so all of the possibilities are acceptable.

Advantages of Ariori Algorithm : Easy to implement. Use large itemset property.

Disadvantages of Apriori Algorithm : Requires many database scans. Very slow.

4 Applications

Applications of Apriori Algorithm :
1.In Education Field: Extracting association rules in data mining of admitted students through characteristics and specialties.
2.In the Medical field: Analyzing the patient's database.
3.In Forestry: Analyzing the probability and intensity of forest fire with the forest fire data.
4.Apriori is used by many companies like Amazon in the Recommender System and by Google for the auto-complete feature.

Application of association rules :
Medicine uses Association rules to help diagnose patients. When diagnosing patients there are many variables to consider as many diseases will share similar symptoms. With the use of the Association rules, doctors can determine the conditional probability of an illness by comparing symptom relationships from past cases.

5 Summary

• Unsupervised learning is a machine learning technique, where you do not need to supervise the model.
• Unsupervised machine learning helps you to finds all kind of unknown patterns in data.
• Clustering and Association are two types of Unsupervised learning.
• Three types of clustering methods are 1) k-means 2) Agglomerative 3) k nearest neighbors.
• Association rules allow you to establish associations amongst data objects inside large databases.
• In Supervised learning, Algorithms are trained using labelled data while in Unsupervised learning Algorithms are used against data which is not labelled.
• The biggest drawback of Unsupervised learning is that you cannot get precise information regarding data sorting.

6 Sources

1.https://www.altexsoft.com/blog/unsupervised-machine-learning/

2.https://en.wikipedia.org/wiki/K-meansclustering

3.https://en.wikipedia.org/wiki/Hierarchicalclustering

4.https://www.youtube.com/watch?v=FvFsjdMwANU

5.https://www.guru99.com/unsupervised-machine-learning.html

6.https://medium.com/@manilwagle/association-rules-unsupervised-learning-in-retail-69791aef99a

7.https://datascienceguide.github.io/association-rule-mining

8.https://en.wikipedia.org/wiki/Associationrulelearning

9.https://www.softwaretestinghelp.com/apriori-algorithm/

10.https://www.youtube.com/watch?v=nSpajfE5Ujc

11.https://bainingchao.github.io/2018/09/27/Apriori/

12.https://www.youtube.com/watch?v=rZRxCpLdNrg

13.https://www.aitude.com/supervised-vs-unsupervised-vs-reinforcement/

14.http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-knn.html

15.https://www.google.com.hk/search?q=kclient=safarisxsrf=APq-WBsErrH HV9eY41bPBOnp43bCVA51Mw:1649873895399source=lnmstbm=ischsa=Xved =2ahUKEwiA89Wb05H3AhUMgP0HHbYnAuoQAUoAXoECAIQAwbiw=883bih =708dpr=2imgrc=wz72rE1JY6tXM

16.https://www.iguazio.com/glossary/unsupervised-ml/?utmsource=adword sutmterm=utmcampaign=GGl-DSA-ROWutmmedium=ppcutmcontent= MachineLearninghsacam=11602556551hsaver=3hsamt=hsasrc=ghsagrp= 108334939330hsakw=hsanet=adwordshsaacc=3049108309hsaad=479289328973 hsatgt=dsa-392284169515gclid=CjwKCAjwo8-SBhAlEiwAopc9W6okrua9onhT8F zBIuHRY6mxp7ASfxXKujZA9Z26JmQysAPzfSDOGBoCncAQAvDBwE

17.https://www.baeldung.com/cs/examples-supervised-unsupervised-learning