Machine Learning –p4

DAVE3625

Umair Mehmood Imam

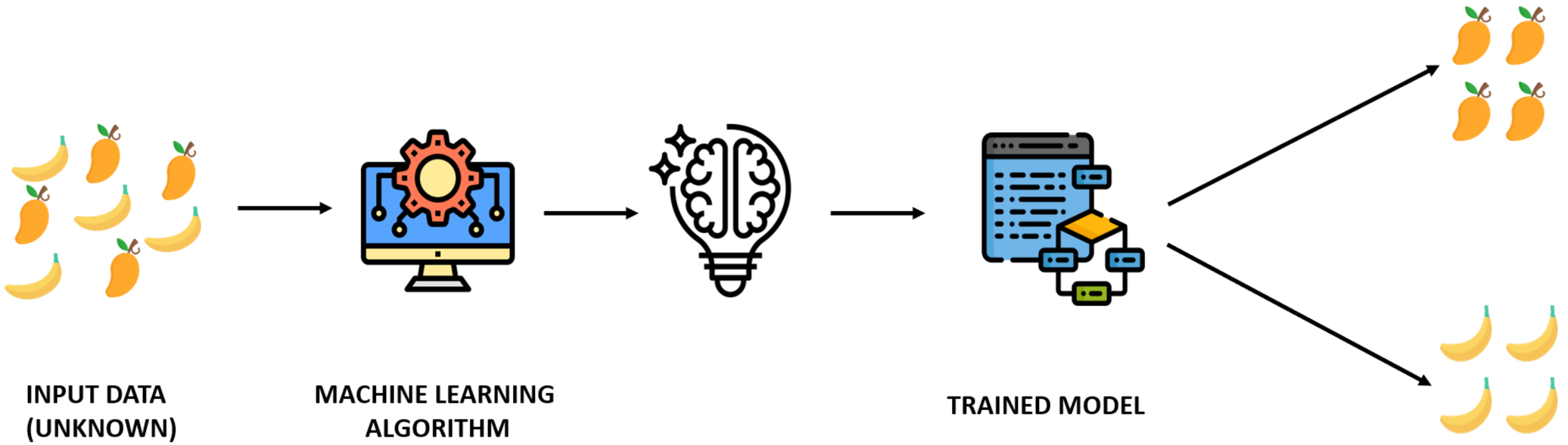
DAVE3625- INTRO TO A.I. BY UMAIR M.I

OSLO METROPOLITAN UNIVERSITY
STORBYUNIVERSITETET

Unsupervised machine learning

Looks for undetected patterns in a data set (with no labels and minimum human supervision)

Unsupervised learning



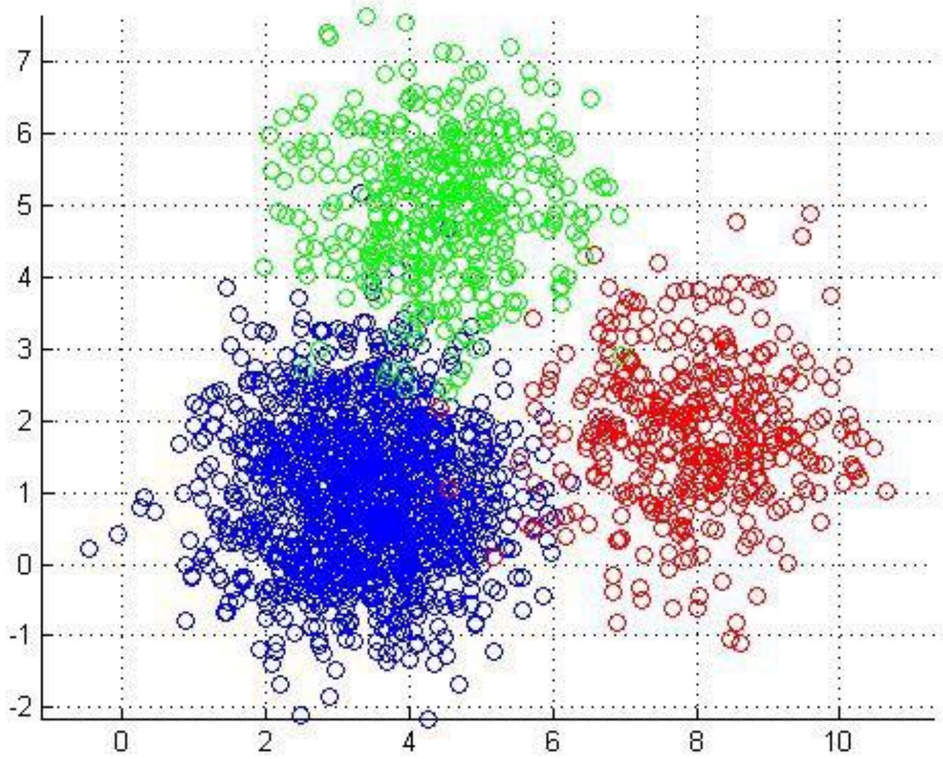
1. Clustering

1. Clustering

- Where we want to discover the groupings in data

sample

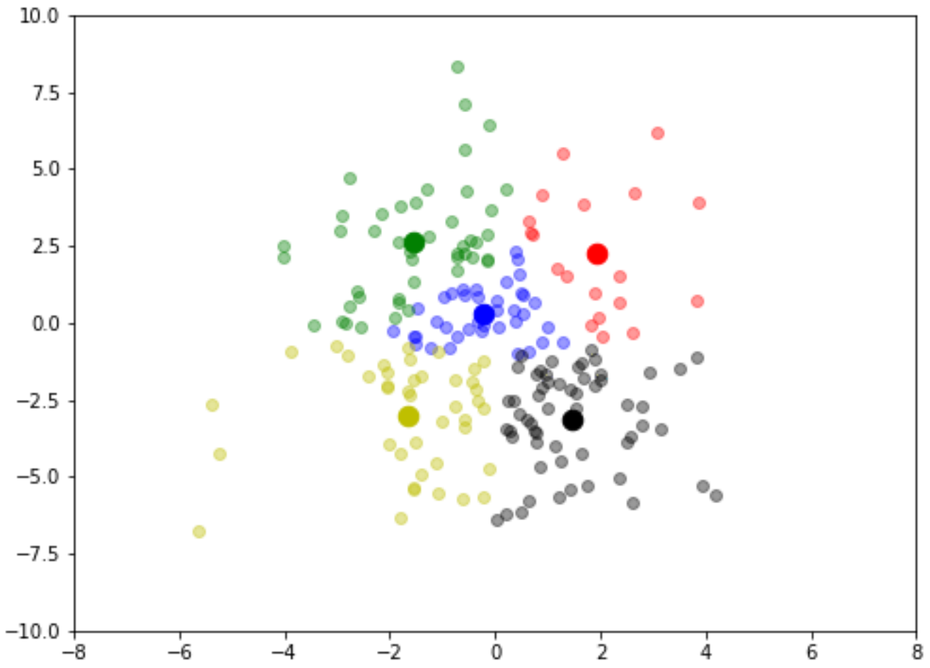
Cluster/group

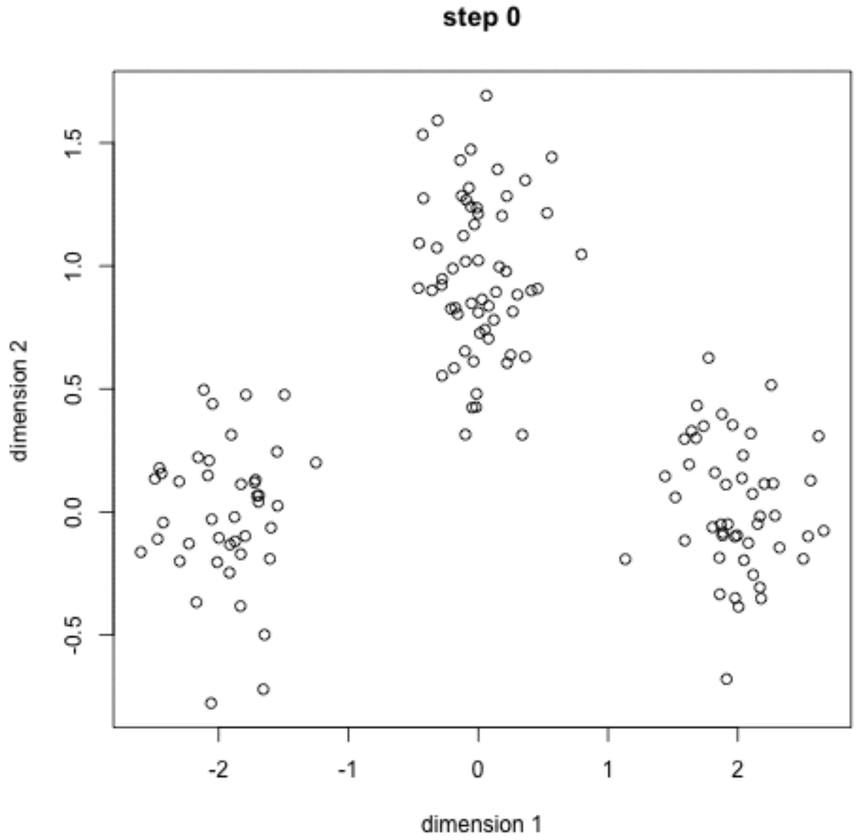


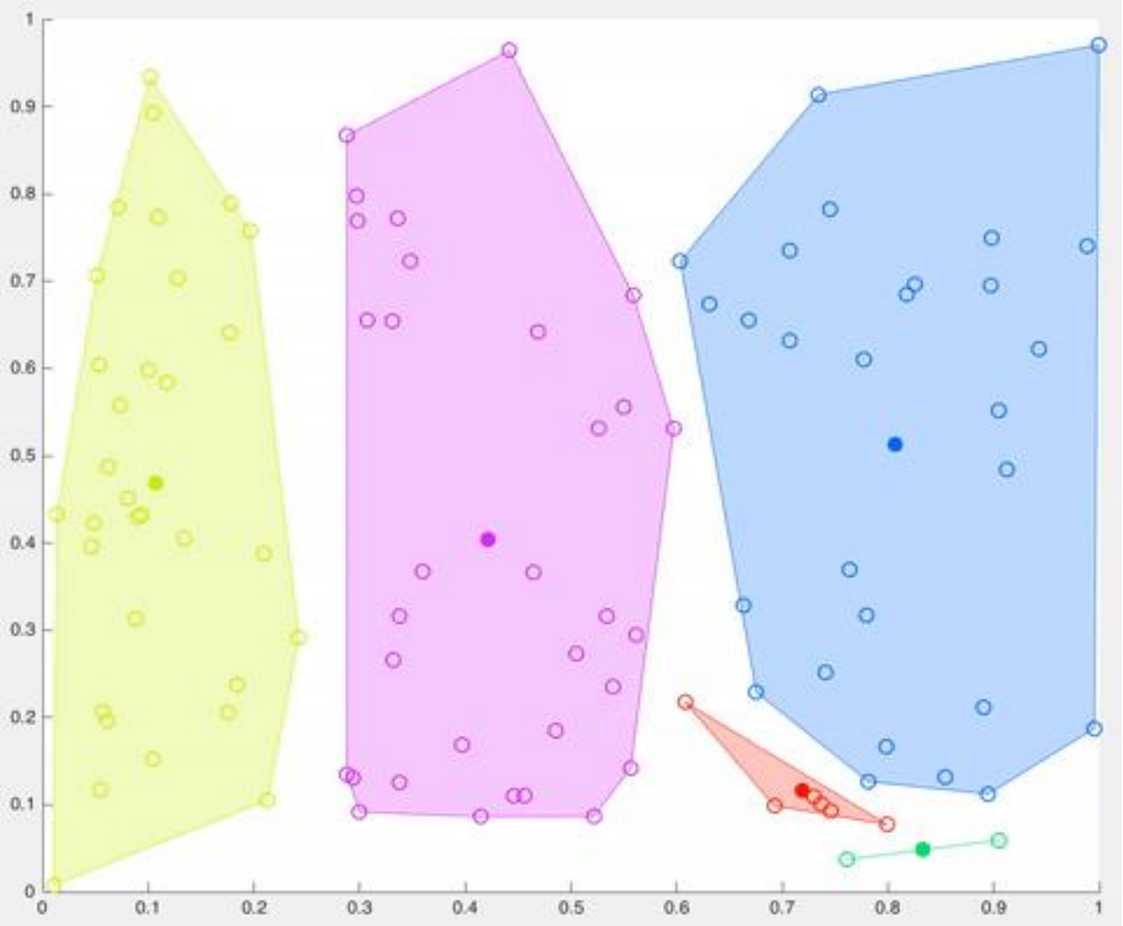
https://miro.medium.com/max/561/0*ff7kw5DRQbs_uixR.jpg

K-Means Clustering

K-means is a distance based algorithm where we calculate distances between data points to assign a point to the cluster


10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−8
−6
−4
−2
0
2
4
6
8


step 0
dimension 2
1.5
1.0
0.5
0.0
-0.5
-2
-1
0
1
2
dimension 1


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

It will stop when

- The centroids have been stabilized – there is no change in their values since clustering has been successful
  - A centroid is the imaginary or real location representing the center of the cluster.
- The defined number of iterations have been reached.

Advantages of K-means

- Very simple to run. (choose k and run it a number of times)
- Most projects donot need quality sensitive clusters

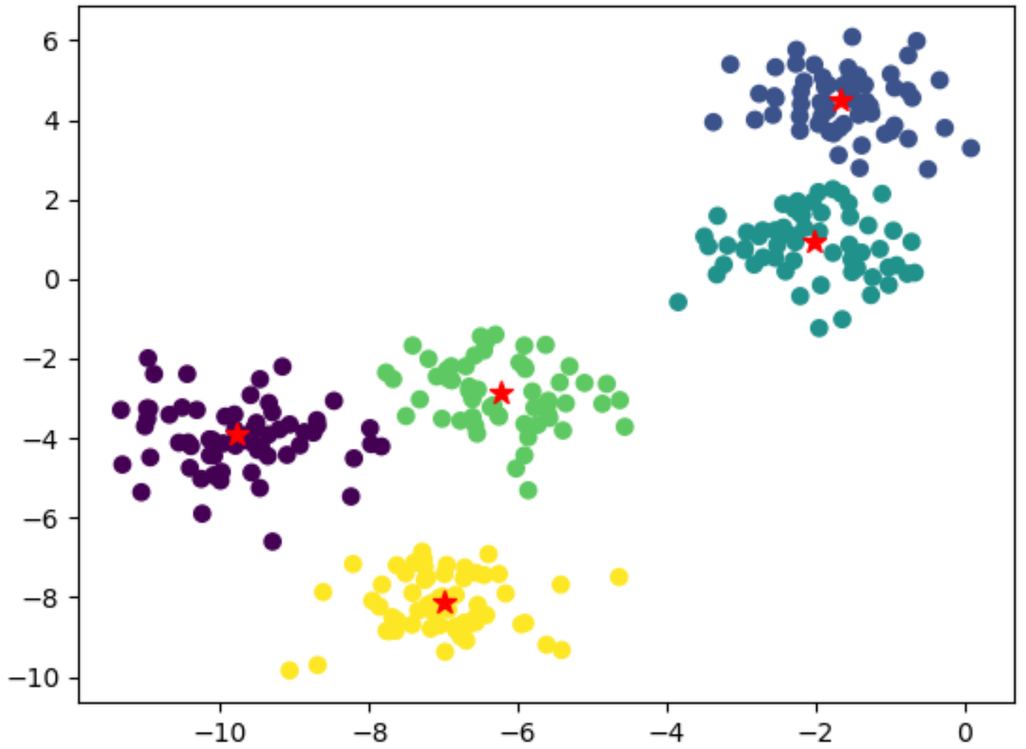
Uses of K-means

- Document classification
- Customer segmentation
- Fraud detection (insurance n bank)
- Ride share data analysis (uber etc)
- Detection of anomalies
- Sorting sensor measurements

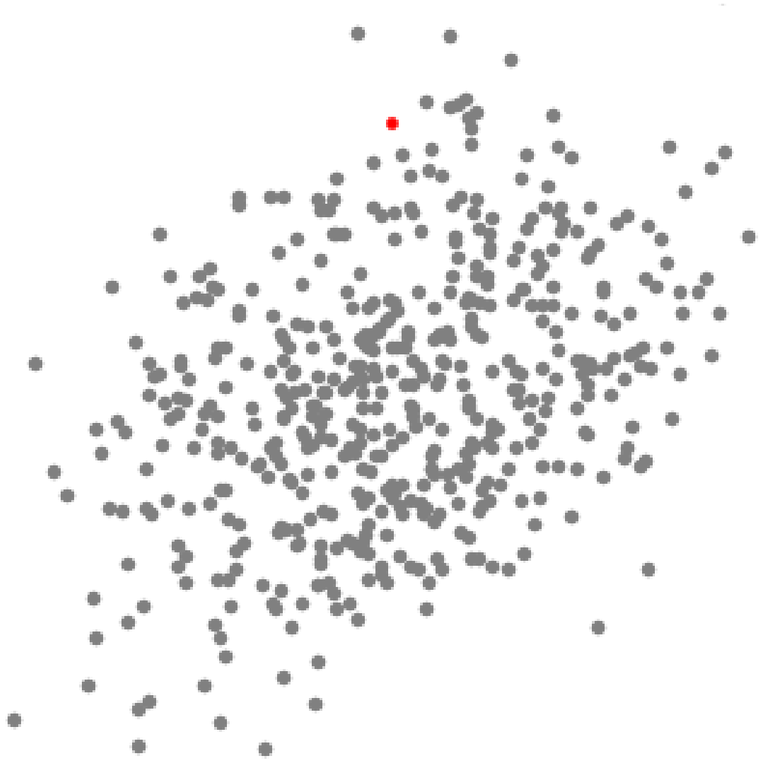
Mean shift clustering algorithm

- It locates the heavy density clusters in a data

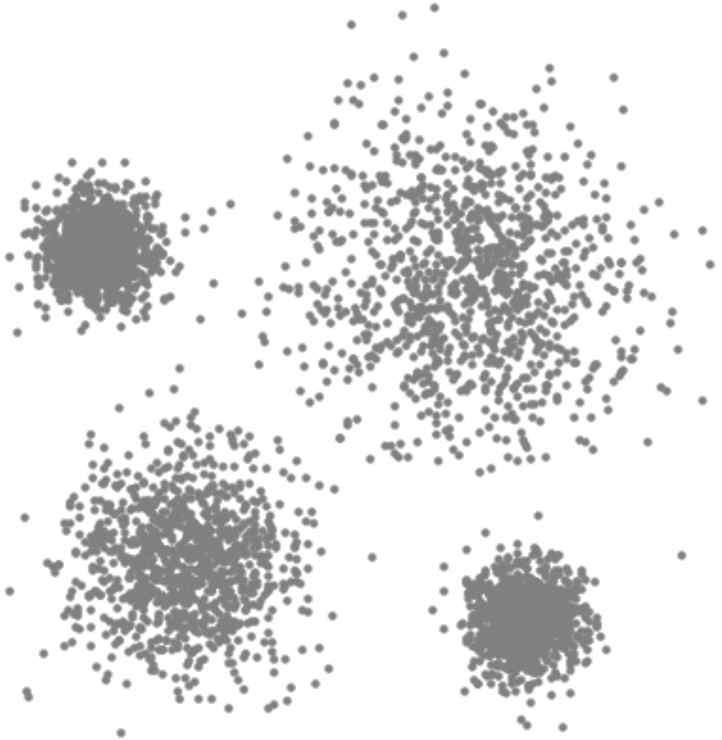
- Uses:
- Computer vision
- Image processing



Example 1



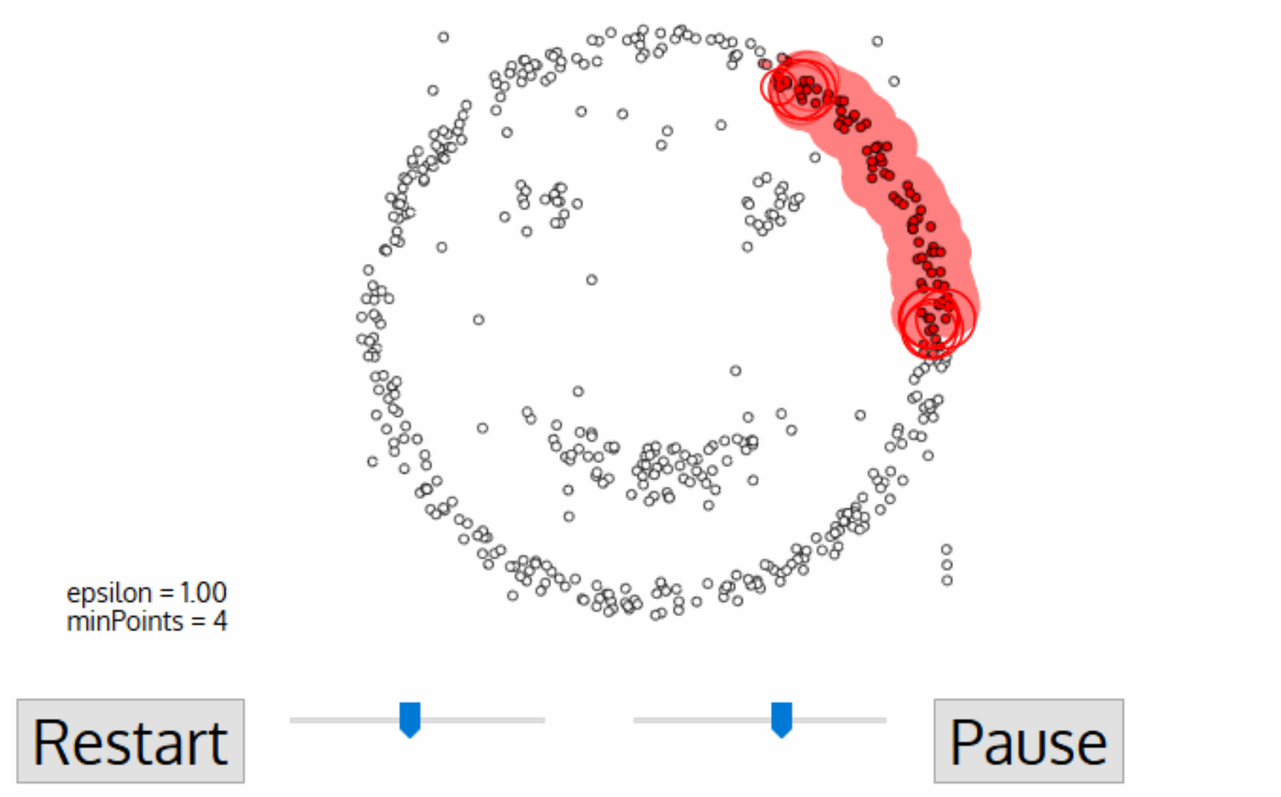
Example 2



HTTPS://TOWARDSDATASCIENCE.COM/THE-5-CLUSTERING-ALGORITHMS-DATA-SCIENTISTS-NEED-TO-KNOW-A36D136EF68

DBSCAN Algorithm

- Stands for Density-Based Spatial Clustering of Applications with Noise
- This is also a density based algorithm
- It separates regions by areas of low-density so that it can detect outliers between the high-density clusters.
- Uses two parameters:
  - minPts: the minimum number of data points that need to be clustered together for an area to be considered high-density
  - Eps: the distance used to determine if a data point is in the same area as other data points

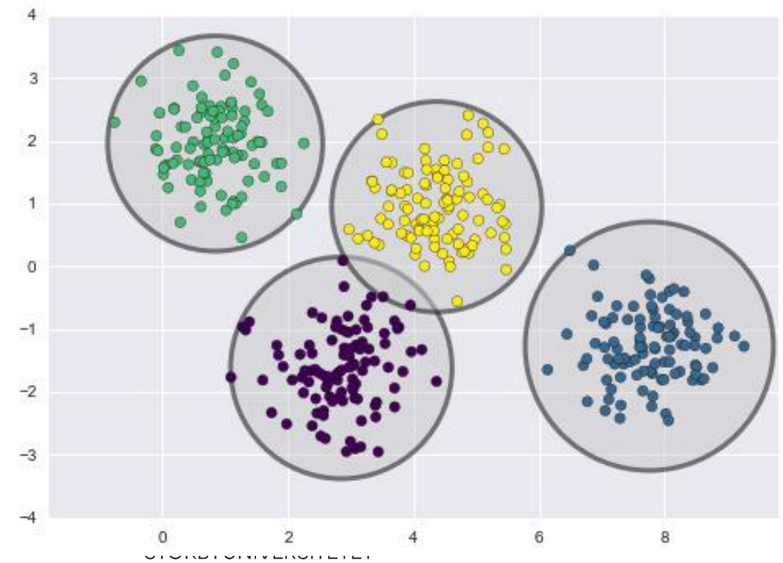


HTTPS://TOWARDSDATASCIENCE.COM/THE-5-CLUSTERING-ALGORITHMS-DATA-SCIENTISTS-NEED-TO-KNOW-A36D136EF68

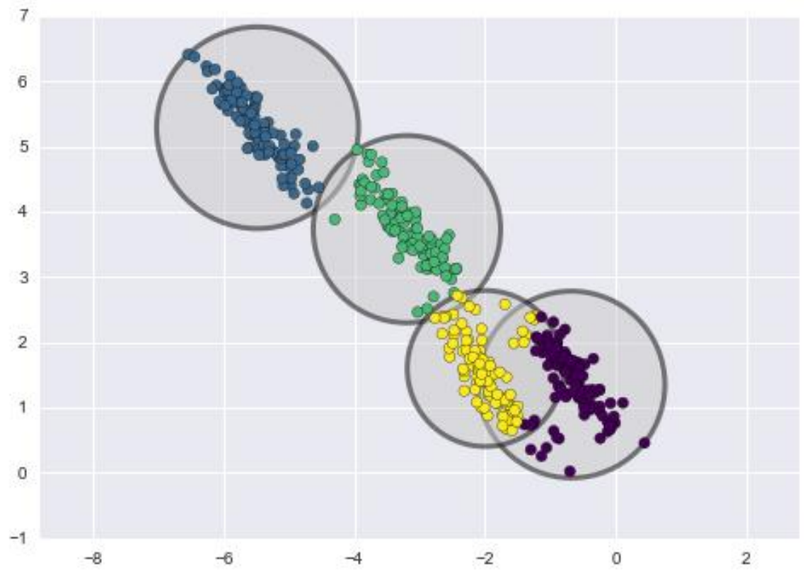
Gaussian mixture model

- Very similar to K-Means HOWEVER
  - K-menas follows a circular format
  - Gaussian can take on any format

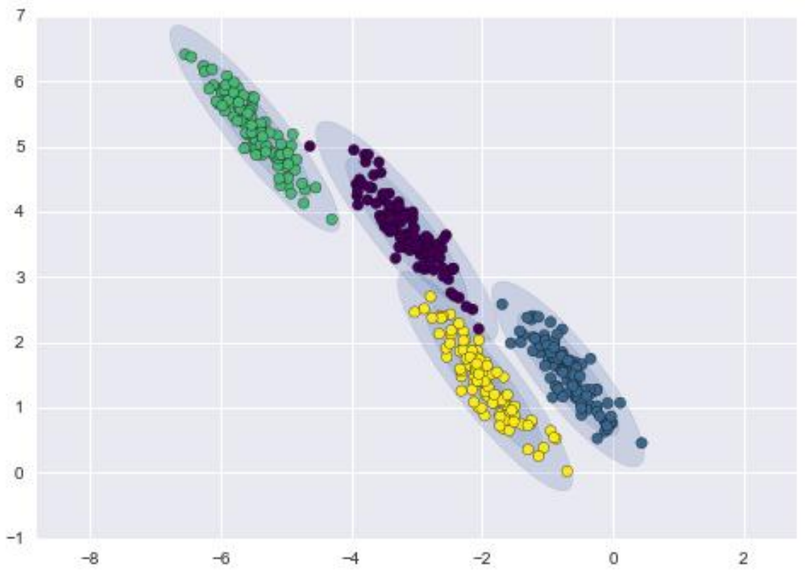
K-means



K-means



Gaussian



https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e

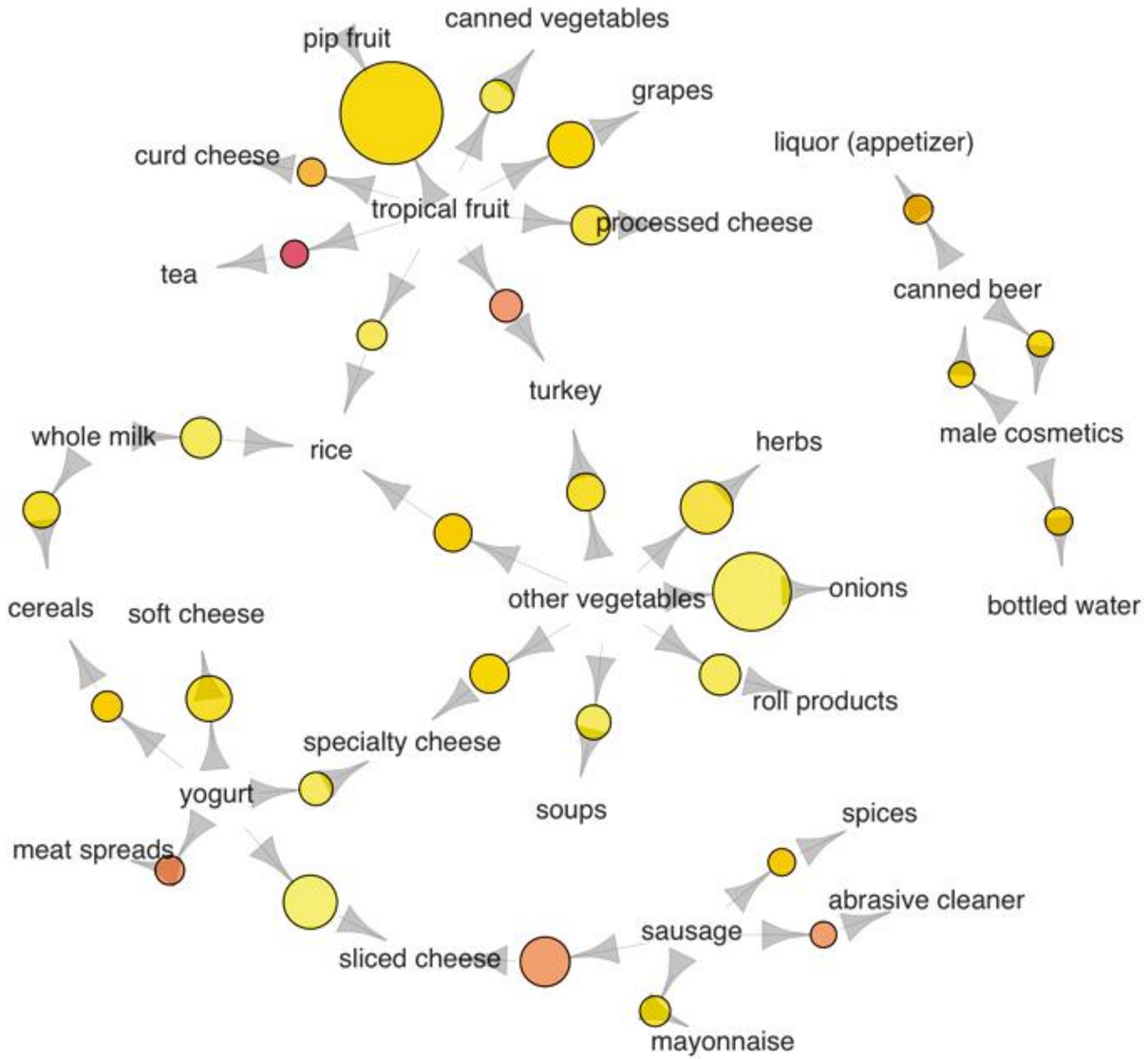
Few more clusstering algorithms

- BIRCH algorithm
- Affinity Propagation clustering algorithm
- OPTICS algorithm
- Agglomerative Hierarchy clustering algorithm
- etc etc

Types of unsupervised learning

- Clustering
- Association

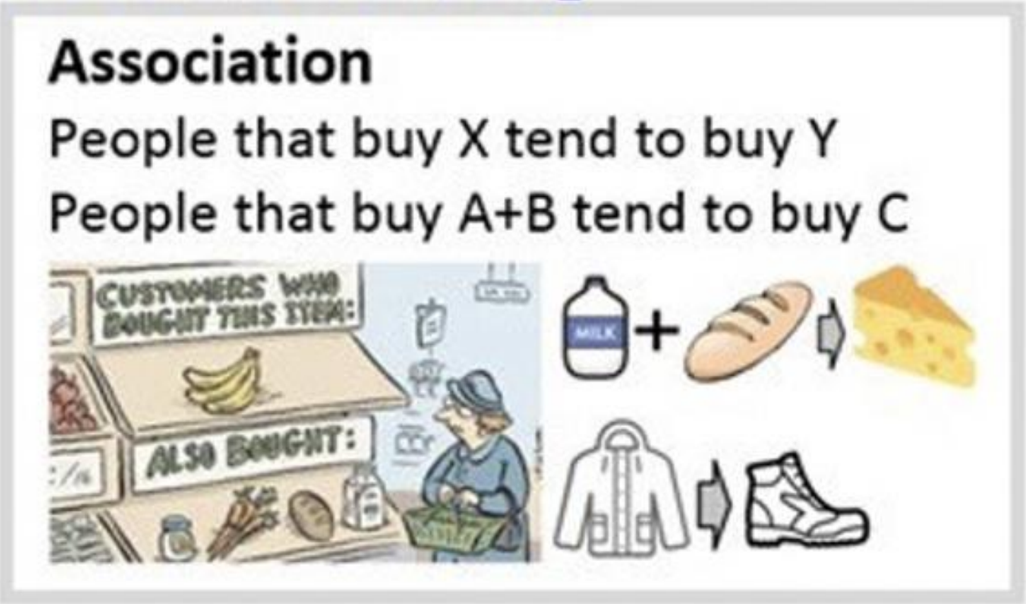
Association algorithms



https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html

2. Association

- When we want to discover rules that describe our data.



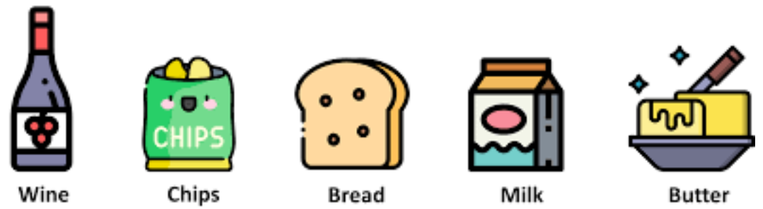
https://miro.medium.com/max/1324/0*LHiIYYxuVUd4qYFW.jpeg

Transaction number	Items
0	soy milk, lettuce
1	lettuce, diapers, wine, chard
2	soy milk, diapers, wine, orange juice
3	lettuce, soy milk, diapers, wine
4	lettuce, soy milk, diapers, orange juice

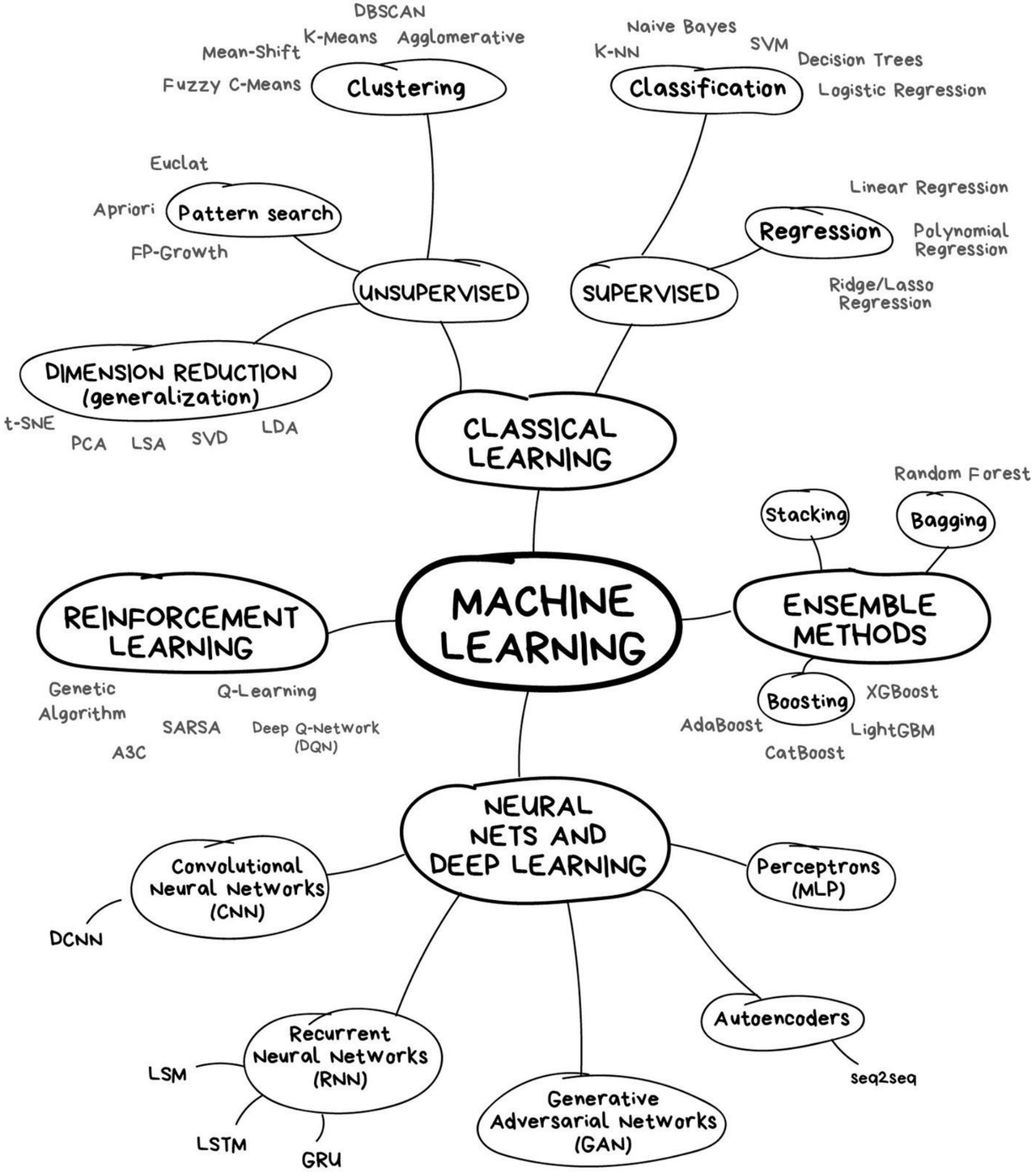
Apriori algorithm

- is used for mining frequent itemsets and devising association rules.
- It is created to operate on a database containing a lot of transactions, for instance, items brought by customers in a store.

- This is the algorithm behind: “You may also like”



OSLOMET



https://towardsdatascience.com/what-is-machine-learning-885aa35db58b