Week4:
K-Means

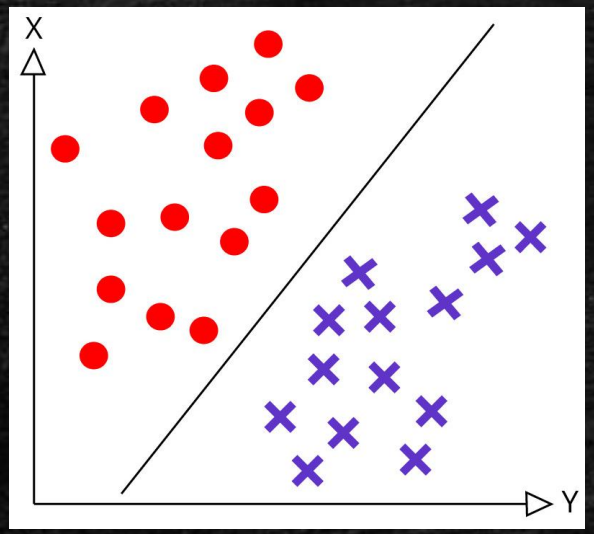
Shayan Dadman, PhD candidate
UiT, Narvik

What is K-Means?

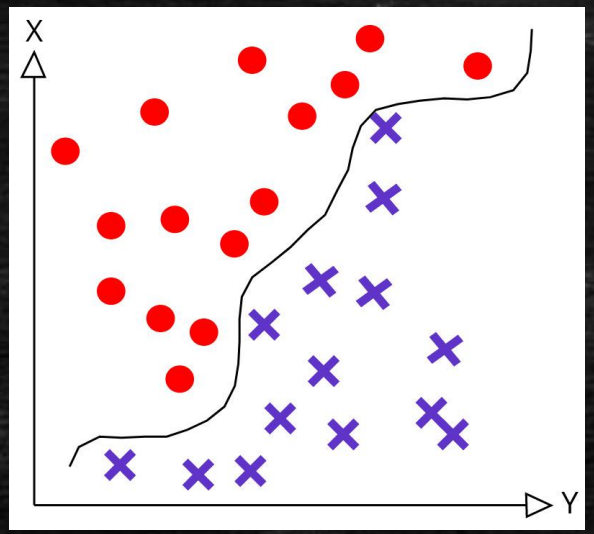
- K-Means is a clustering unsupervised learning algorithm.
- It has a wide range of applications such as recommendation systems, customer segmentation, document clustering, and image segmentation.
- K-Means groups the attributes in the unlabeled dataset into clusters.
  - "K" refers to the number of clusters.
  - "Means" relates to the cluster centroid, which determines by the average of the cluster content.
- The main goal of the K-Means algorithm is to minimize the sum of the distances between the data points and their corresponding clusters.

What is clustering?

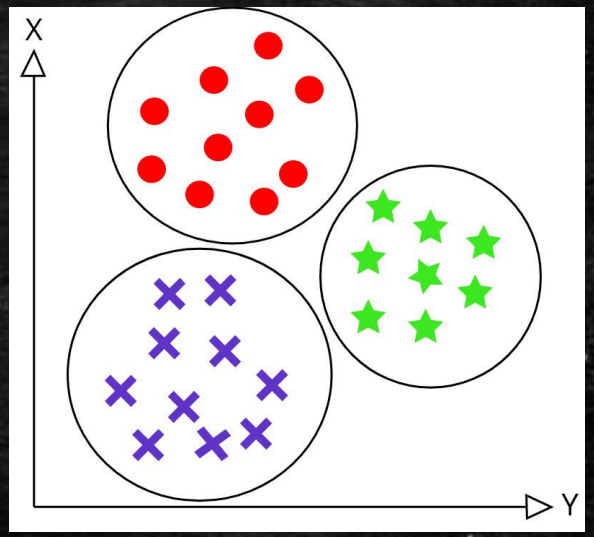
- Clustering divides the entire data into groups (clusters) based on the similarities between the data points.



Linear classification



Non-linear classification



Clustering

What is feature vector?

- The feature is a list of values, for example, age, name, and height.
- The feature vector is an n-dimensional vector of features that represent a particular object or observation.
- For example, in the table below, columns are the features, and each row is a feature vector.

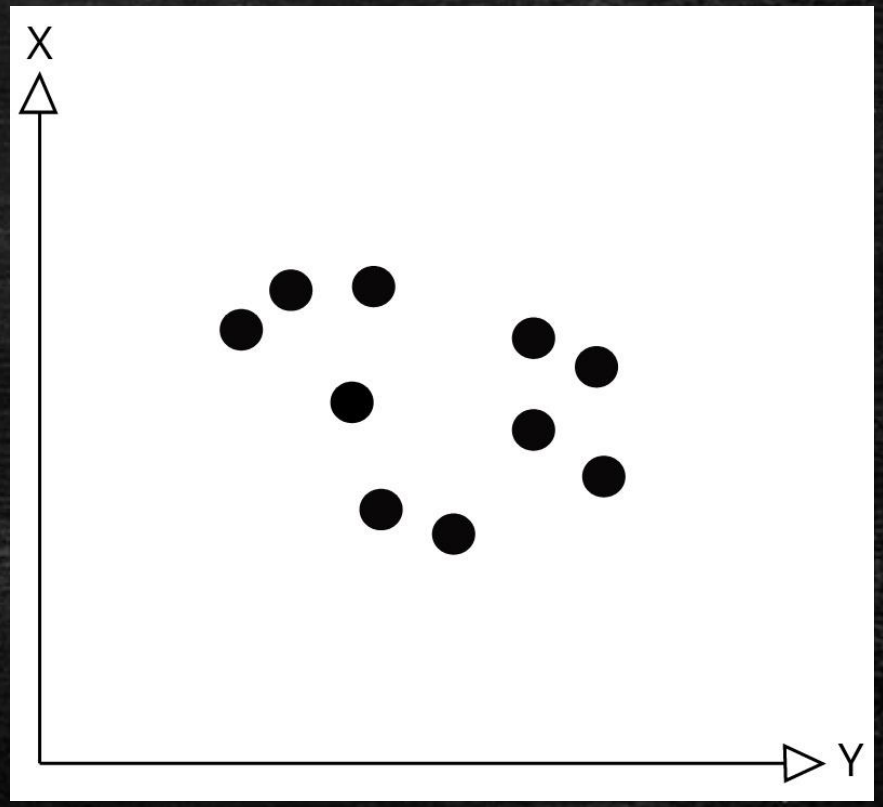
ID	First Name	Last Name	Email	Year of Birth
1	John	Johnson	john.johnson@university.edu	1992
2	Jack	Knife	jack.knife@university.edu	1982
3	Chris P.	Bacon	Chrispbacon@university.edu	1994
4	David	Letty	David.letty@university.edu	1976

How does the algorithm work?

- The overall procedure is:
  - taking the unlabeled dataset
  - Dividing them into k-number of clusters
  - Iterating until cannot find the best clusters
- We choose a cluster for a given feature vector by calculating the Euclidean distance between the feature vector and the available cluster centroids.
  - By assigning a new feature to a cluster, the centroid of that cluster is also changed.
  - We update the centroids with each new entry.
- During the process, the feature vectors can move from one cluster to another.

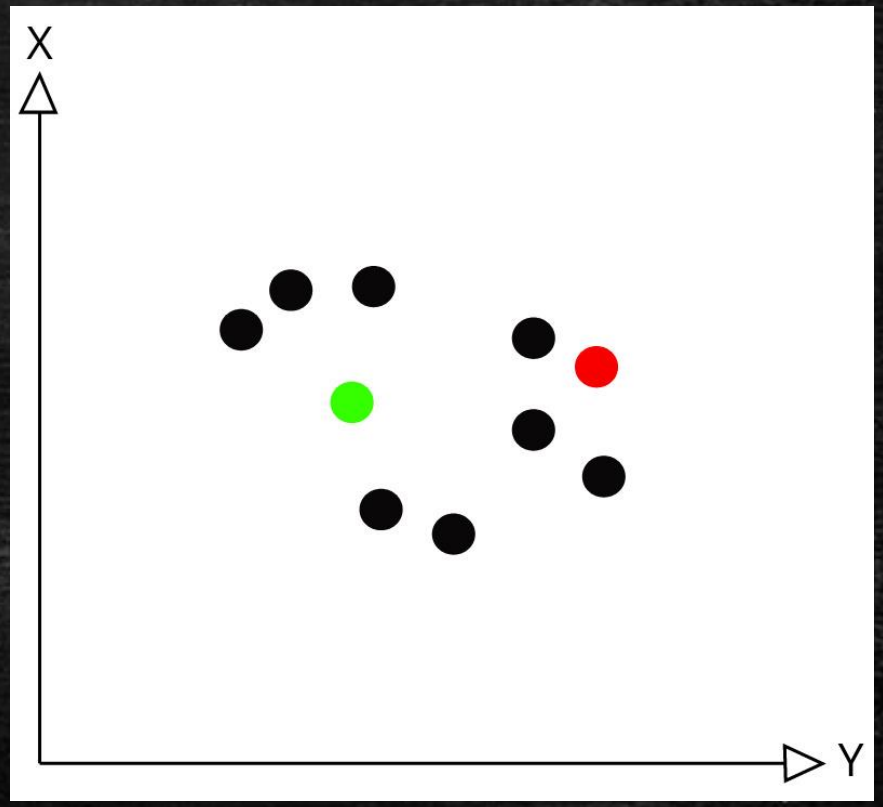
How does the algorithm work?

- Here we have a plot of our two dimensional feature space data points.
- First we pick our number of clusters. In this case we choose two clusters.



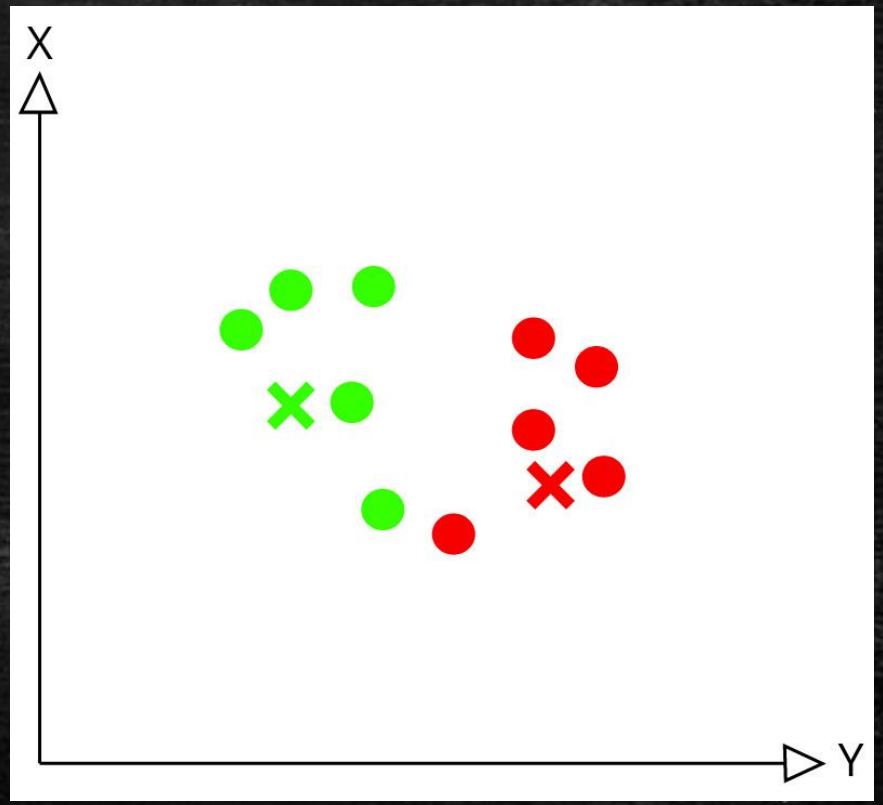
How does the algorithm work?

- Since our K is equal to two, we select two centroids randomly.
- Green dot is the centroid 1 and red dot is the centeroid 2.
- We calculate the distance and assign the data points to the nearest cluster.



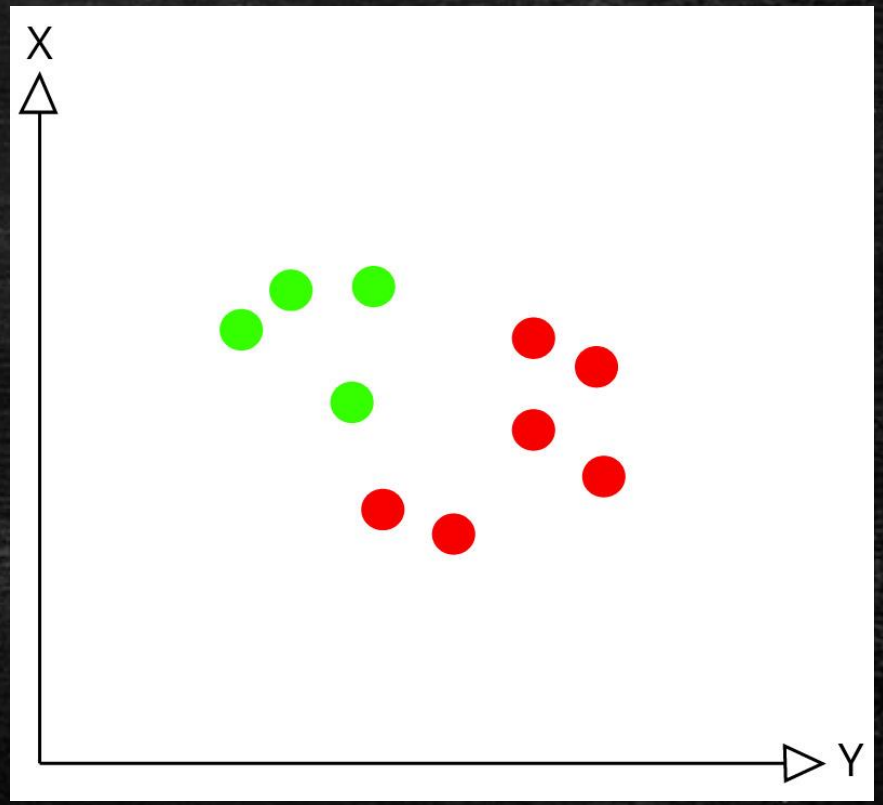
How does the algorithm work?

- In the next step, we recalculate the centroids of the clusters.
- The green and red crosses represent the new centroids.



How does the algorithm work?

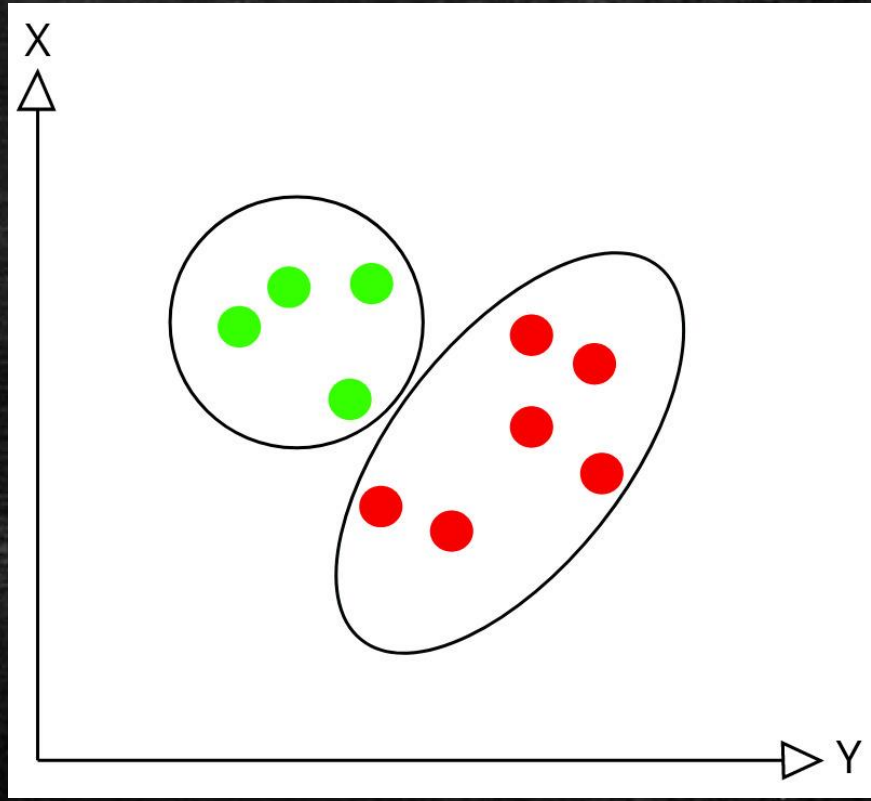
- Having the new centroids, we repeat the process by:
  - recalculating the distances between the data points and the centroids.
  - Updating the cluster's centroids.



How does the algorithm work?

- We stopped since the clusters are formed. However, to stop the algorithm, we need to define criteria.

- There are three main criteria:
  - The centroids do not change after the clusters are updated
  - The data points remain in the same cluster
  - The process reached a certain number of iterations



K-Means pseudo code

1. Load the data
2. Initialize the value of K to decide the number of clusters
3. Select random K points or centroids
4. Assign each data point to their closest centroid, which will form the predefined K clusters.
5. Calculate the variance and update a new centroid of each cluster.
6. Repeat the fourth step, which means reassign each data point to the new closest centroid of each cluster.
7. Stop if the criteria are met; else, go to the fifth step.

Questions

1. Email me at Shayan.Dadman@uit.no
2. My office D3430