K-means Clustering Algorithm

Group Members:

Muhammad Asim Noor

Abdul Basit

Furqan Sadiq

Overview

Problem Statement

Motivation

Explanation

Complexity and Comparison

Improvements

Q / A

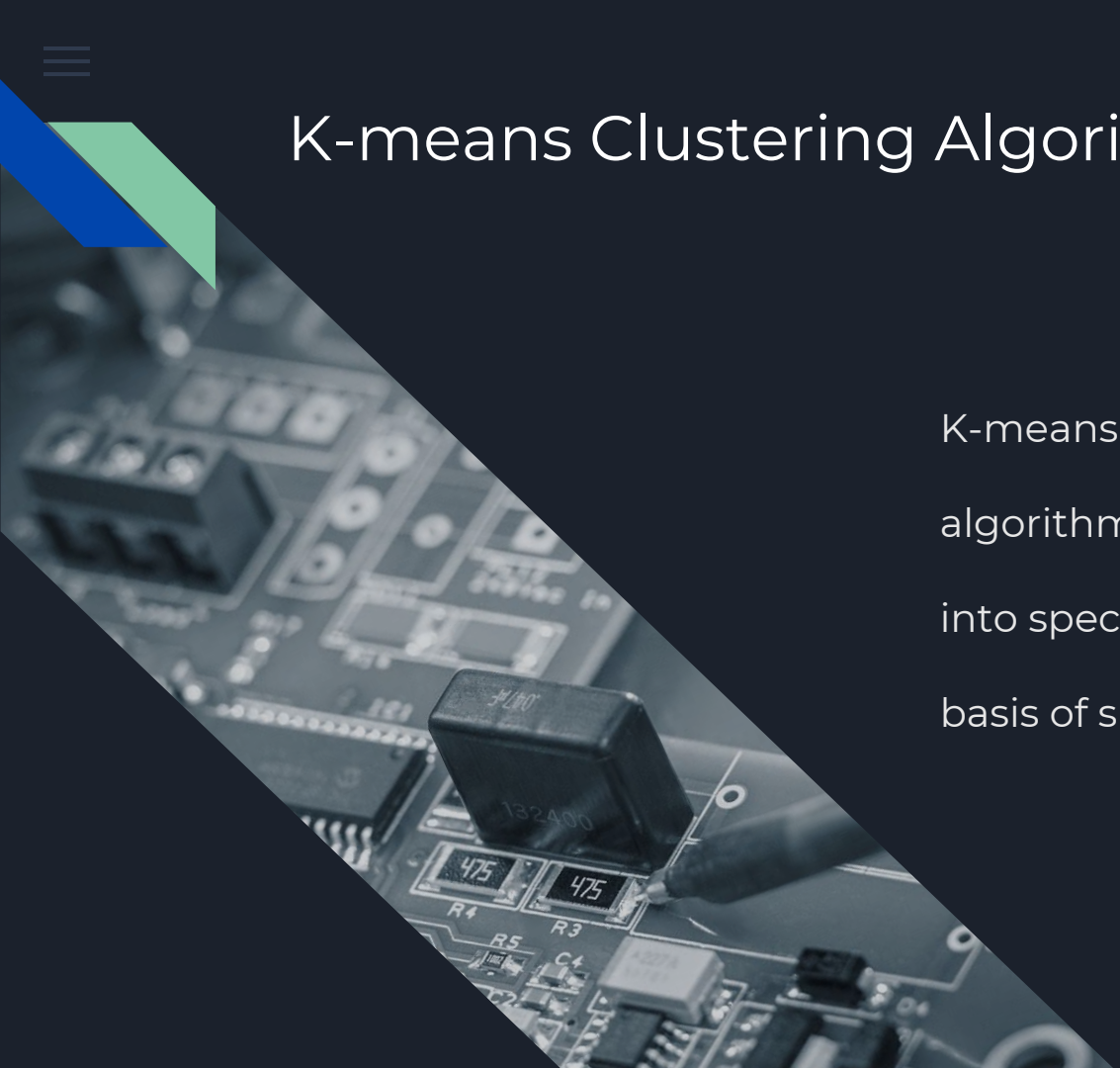
Motivation

- Big Data
- Data Mining
- Machine Learning

Example


International Seminar
on
Peace and Conflict Resolution
October 23-24, 2013
Centre for International Peace & Stability (CIPS)
NUST

K-means Clustering Algorithm

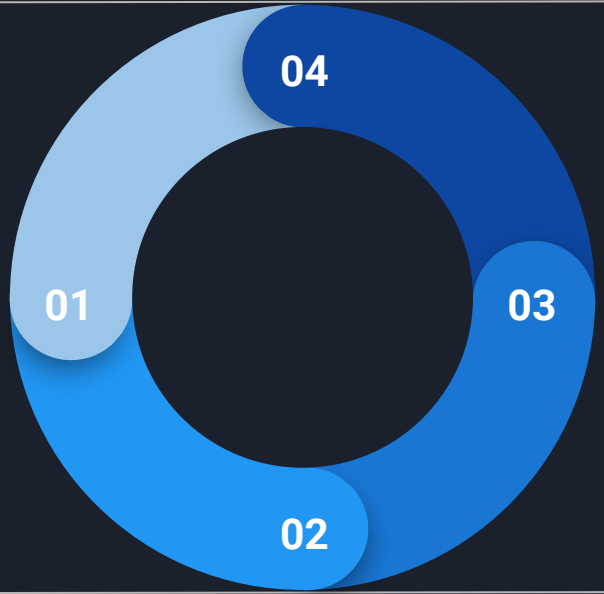


K-means Clustering Algorithm is an algorithm to cluster dataset of n objects into specific number of clusters k on the basis of similarity in their traits.

Steps

Choose number of clusters k

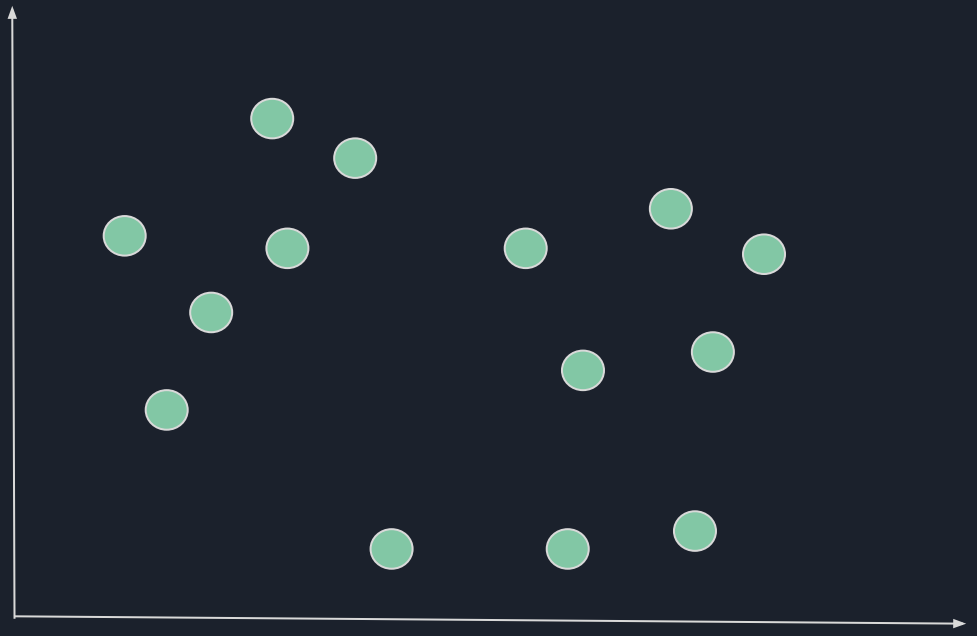
Calculate mean and Euclidean distance



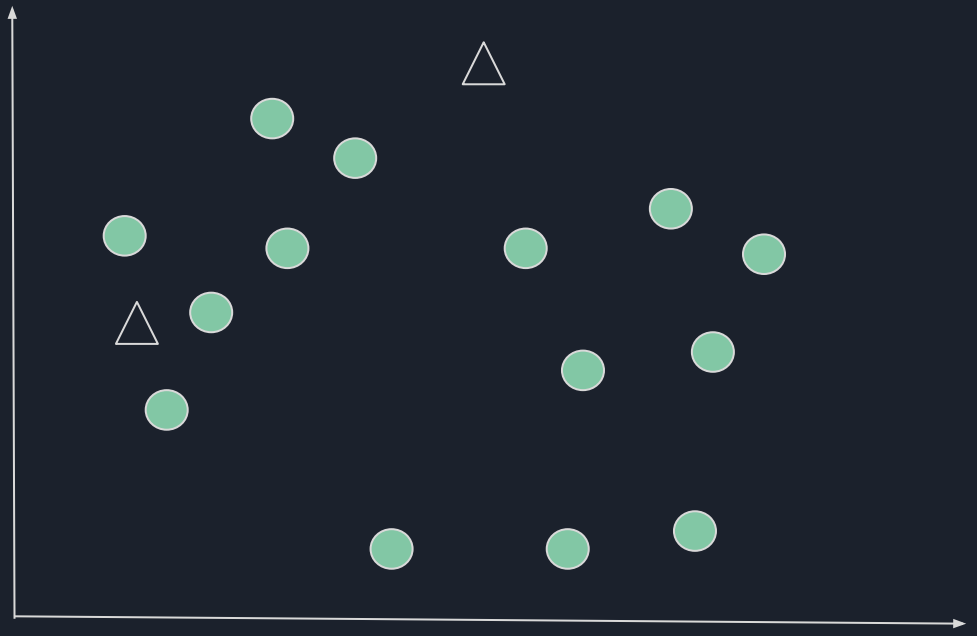
Repeat until data points converge

Assign data point to closest centroid and make clusters

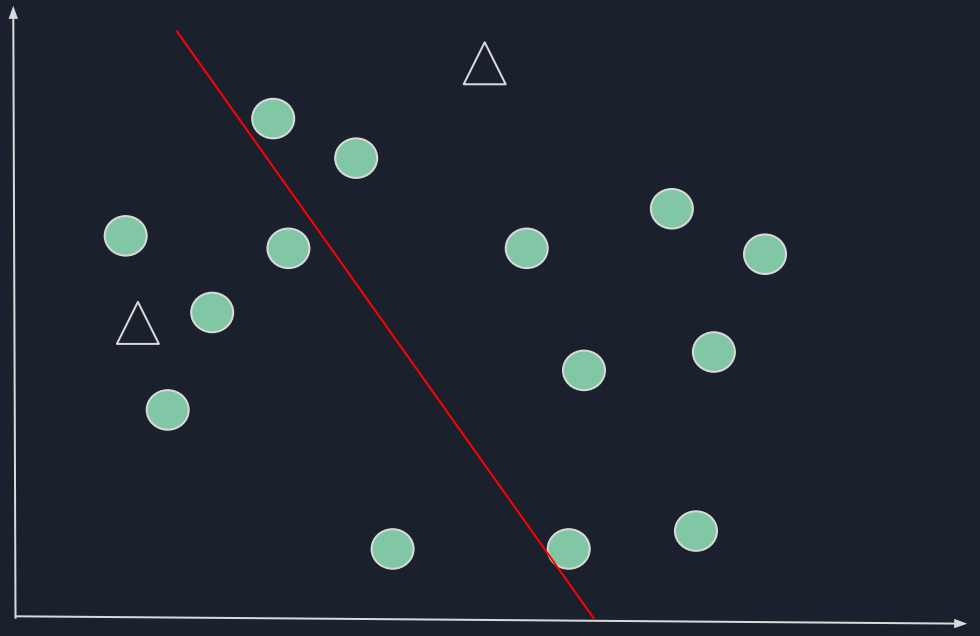
Raw data:



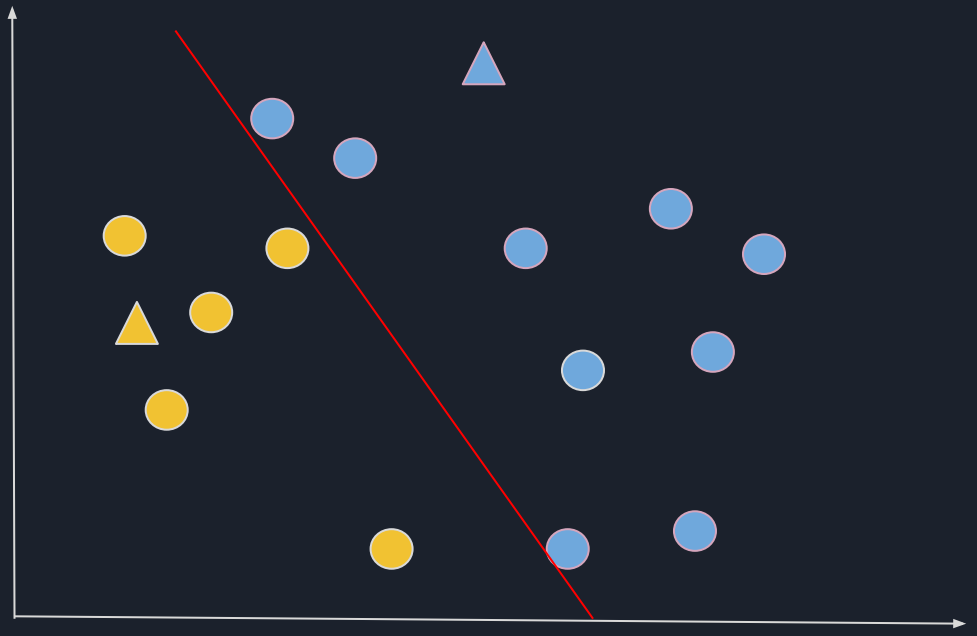
Assigning centroids:



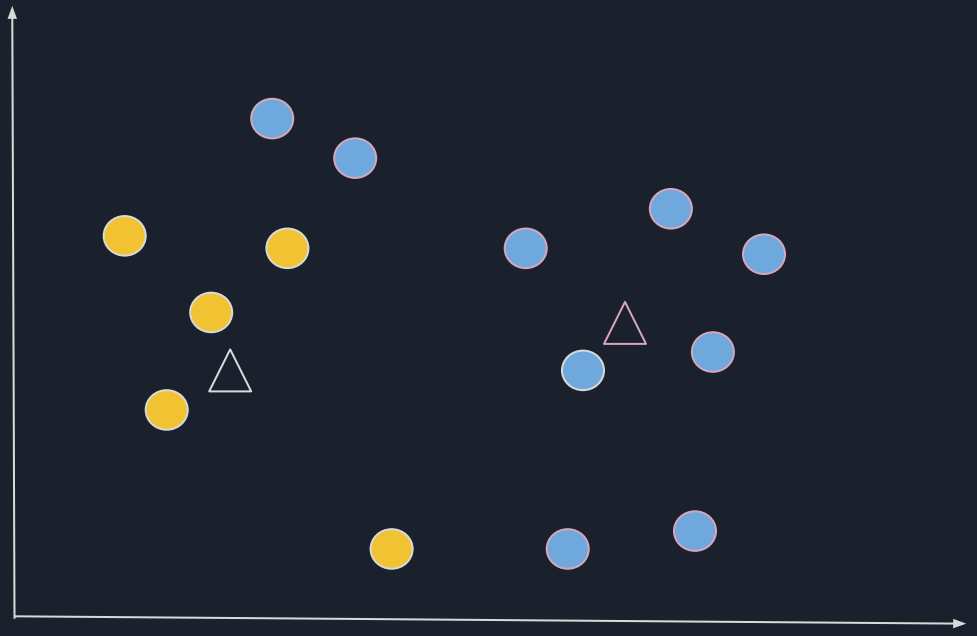
Euclidean distance:



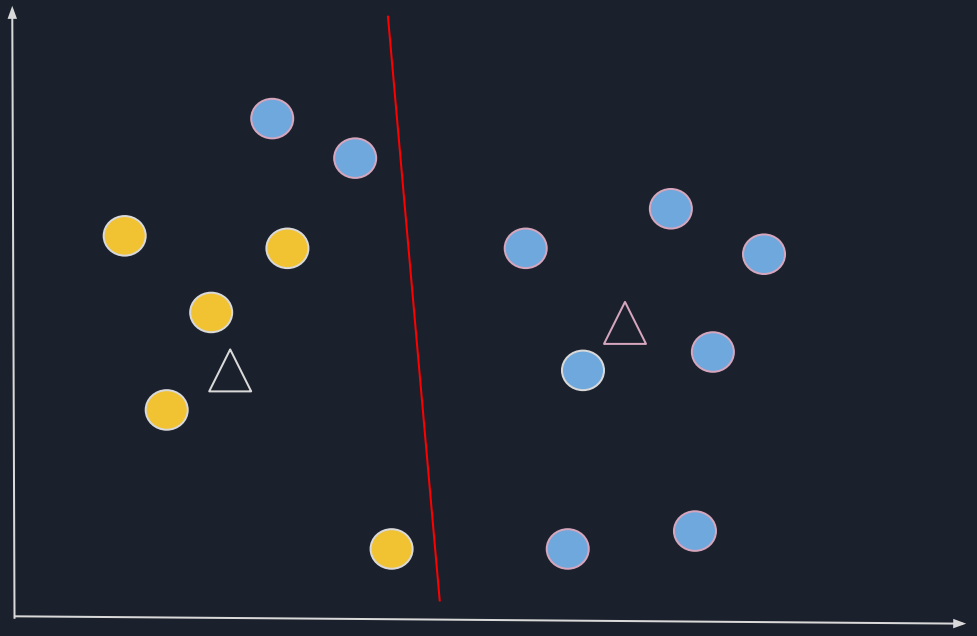
Making clusters of data points:



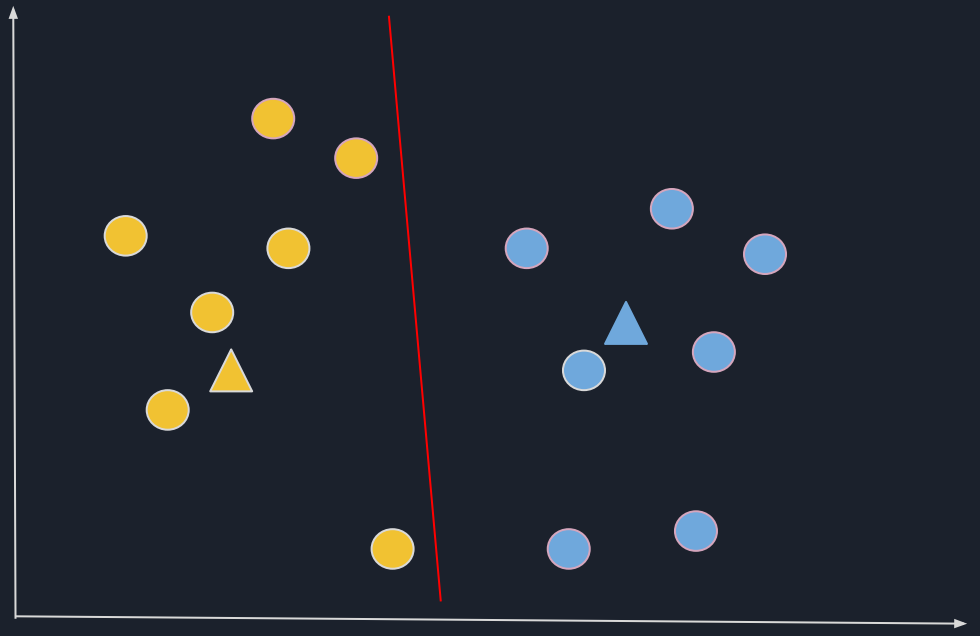
Calculating new centroids:



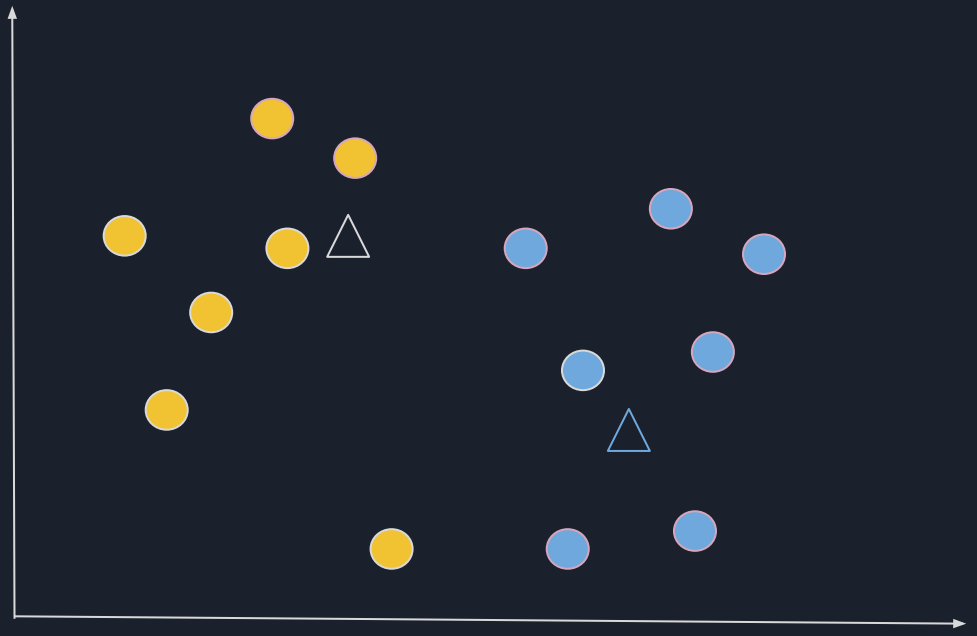
Repeat:



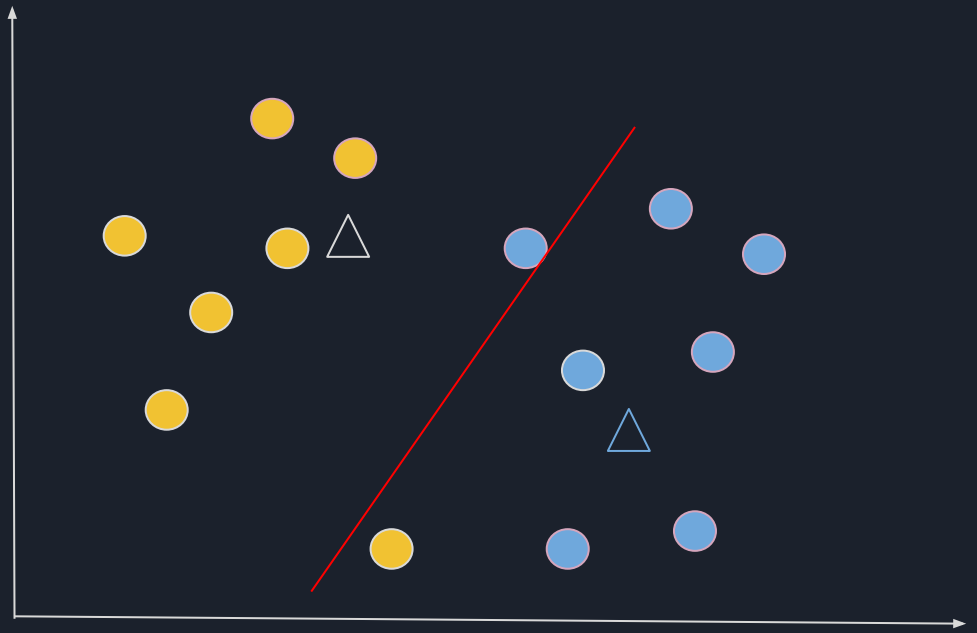
Repeat:



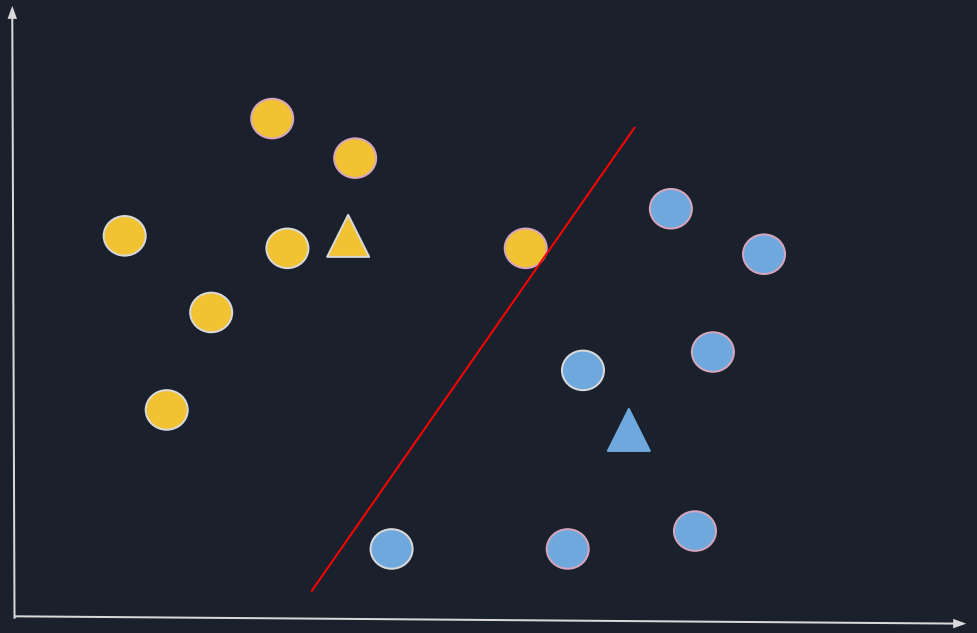
Repeat:



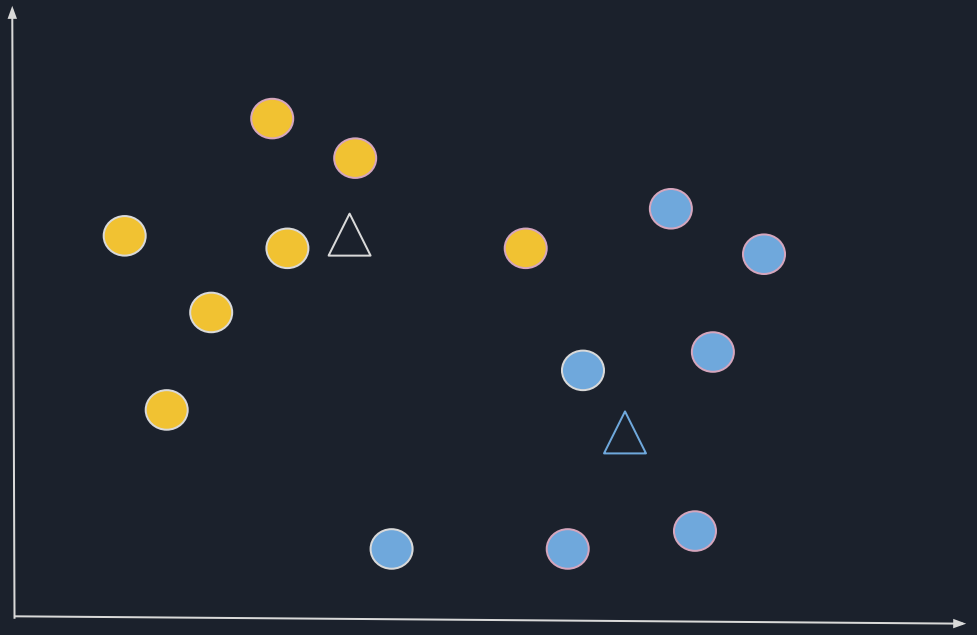
Repeat:



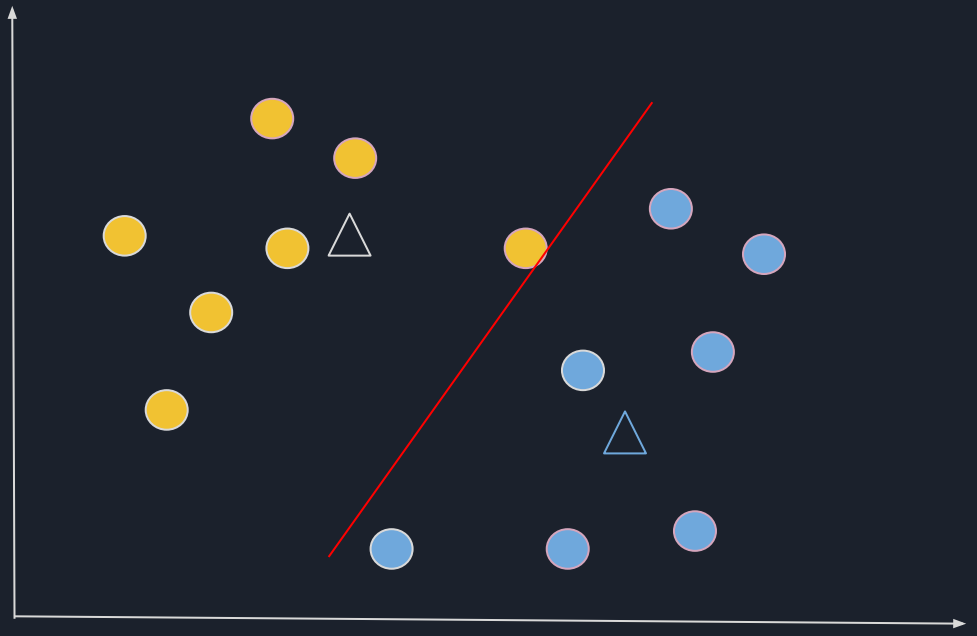
Repeat:



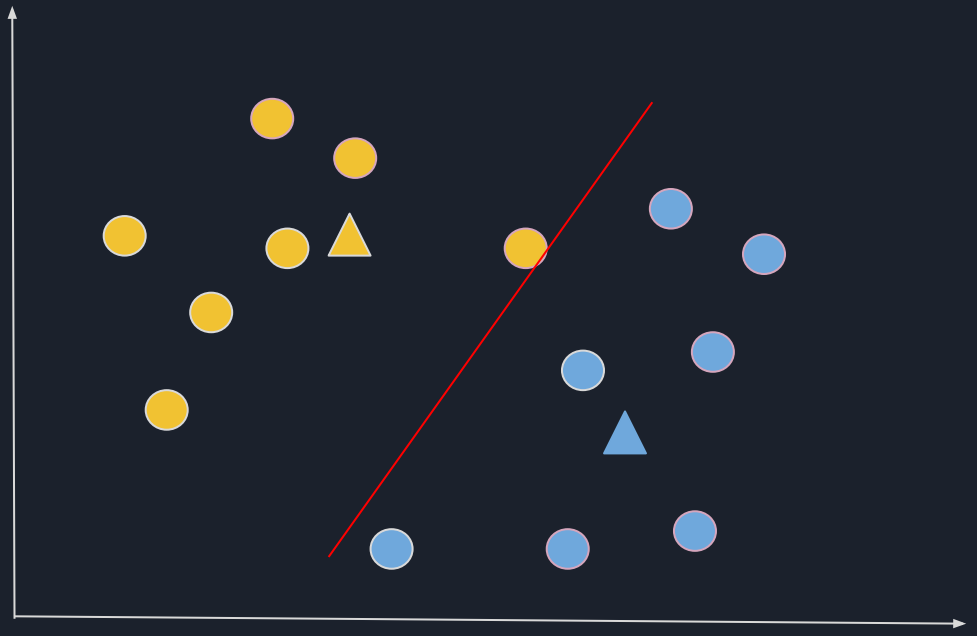
Repeat:



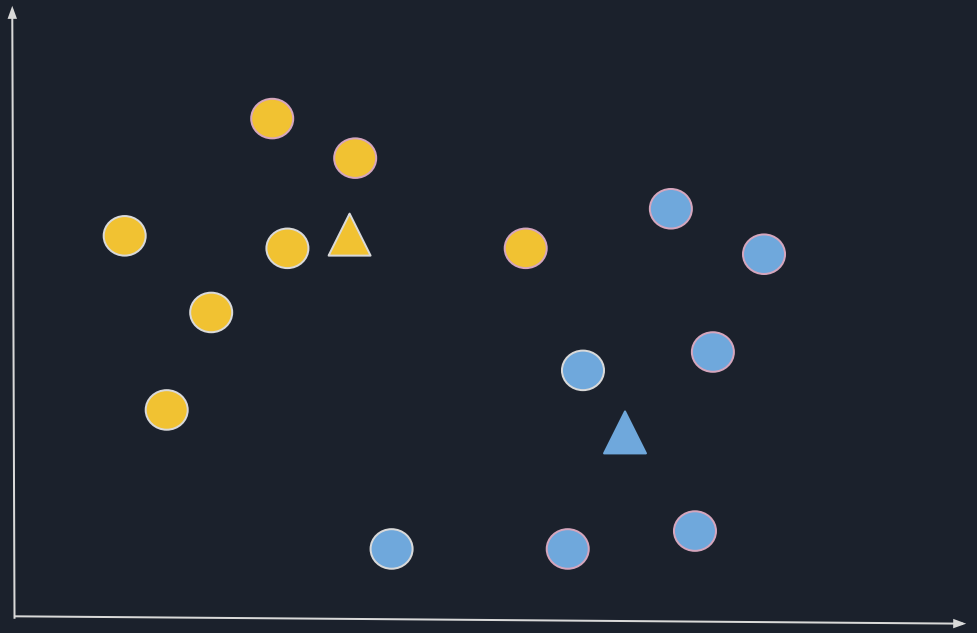
Repeat:



Repeat:

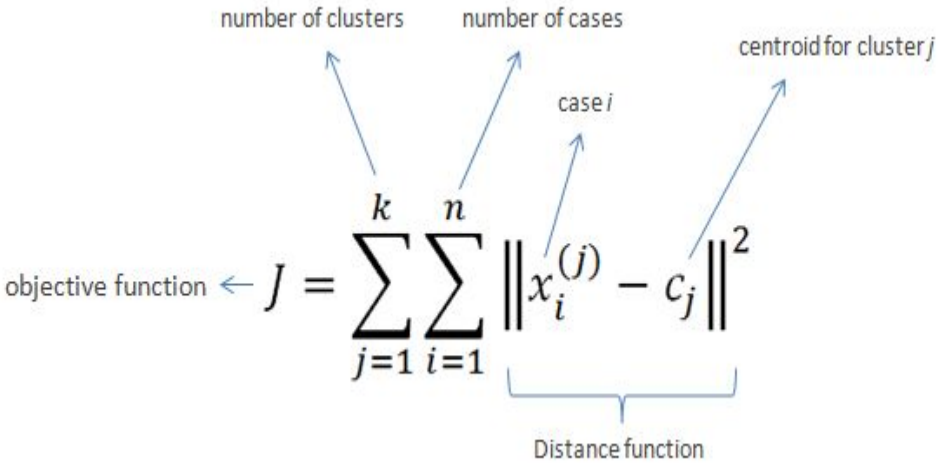


Final Clusters:



Object Function:

This object can be pass to any function to compute the Euclidean distance from all the clusters and put into table.



Euclidean distance for n dimensions:

In Cartesian coordinates,

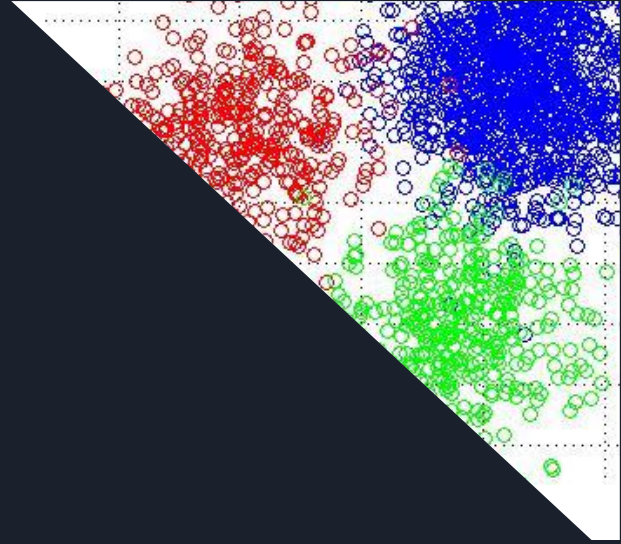
If
p = (p1, p2,..., pn) and
q = (q1, q2,..., qn)

are two points in Euclidean n-space, then the distance (d) from p to q, or from q to p is given by the Pythagorean formula:

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Complexity:



Time Complexity:

O (n * k * i * d)

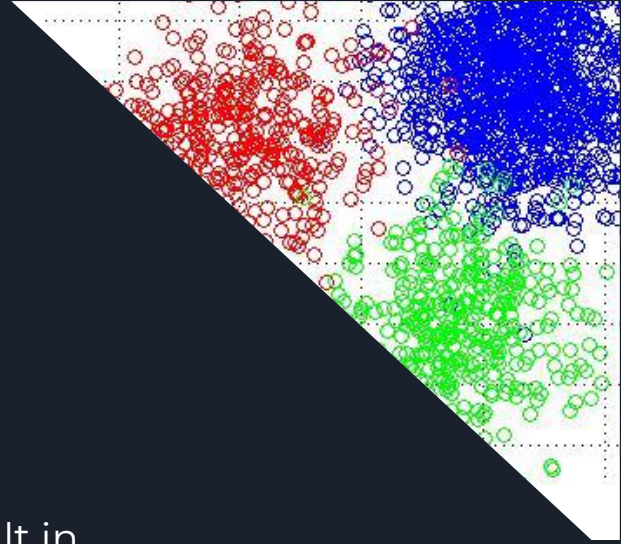
Space Complexity:

O((n+k)* a)

where

n : number of data points
k : number of clusters
i : number of iterations
d : number of attributes

Constraints:

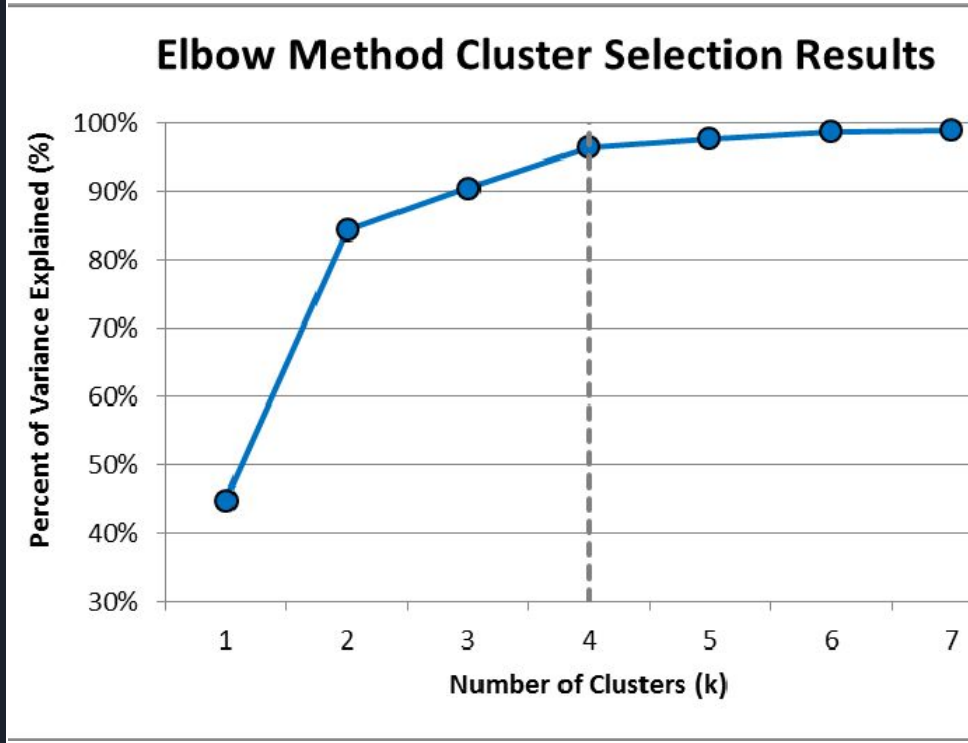


- ➔ Clustering is dependent on k
- ➔ Choosing an appropriate k
- ➔ Can process only numerical data
- ➔ Different selection of initial centroids can result in different final clusters

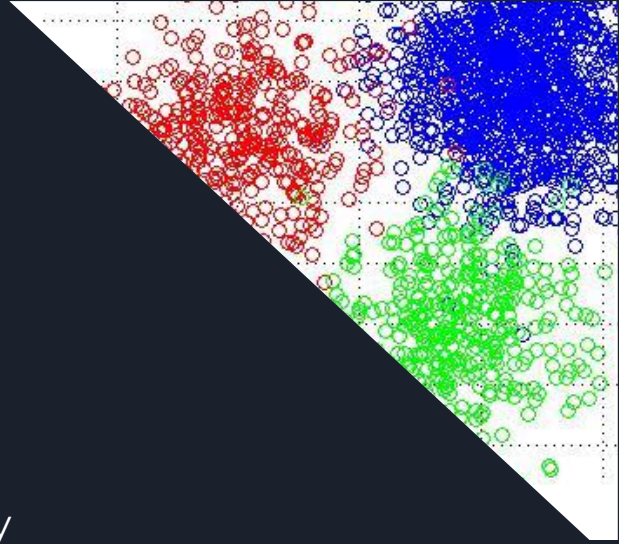
Improvement:

Choose k using elbow-plot:

You can repeat experiment several times and then select the most appropriate k for clustering of data points



Improvement:

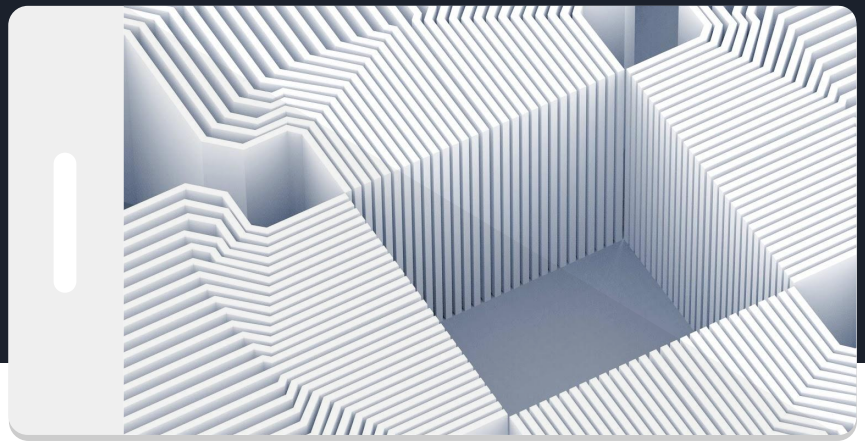


Sampling:

- Take sample of original data
- Cluster based on hierarchical clustering or any other clustering algorithm
- Calculate centroid of each cluster
- Use these centroids as starting points of your algorithm

Applications:

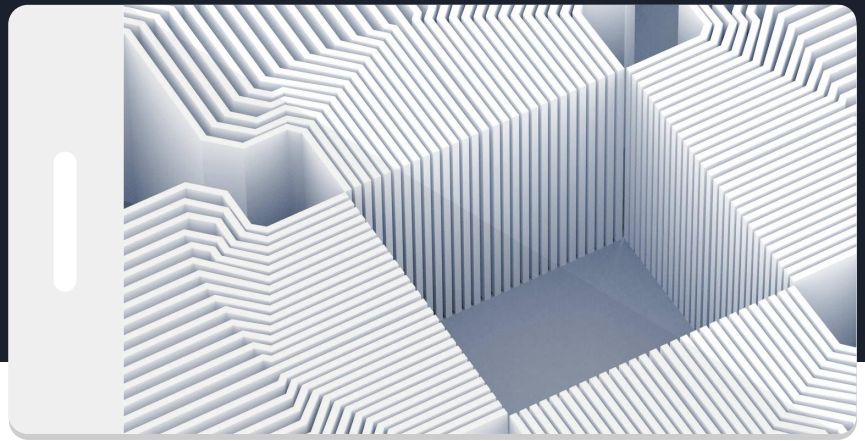
Search Engine:



Clustering algorithm is the backbone behind the search engines. Search engines try to group similar objects in one cluster and the dissimilar objects far from each other. It provides result for the searched data according to the nearest similar object which are clustered around the data to be searched.

Applications:

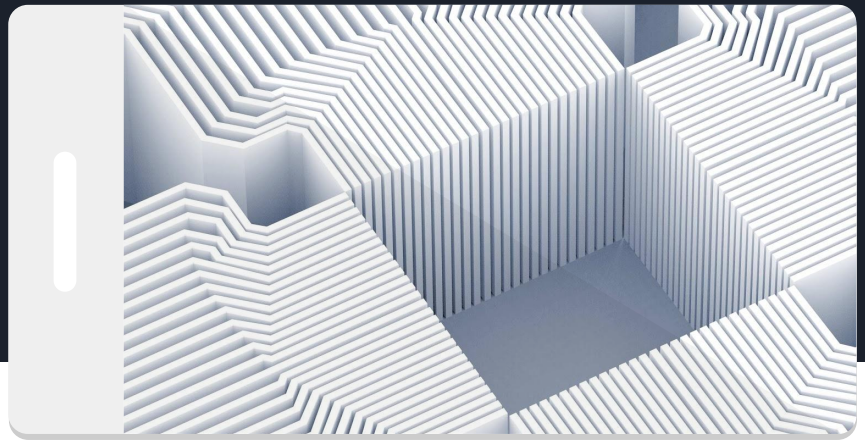
Clustering Algorithm in Academics:



The ability to monitor the progress of students' academic performance has been the critical issue for the academic community of higher learning. Clustering algorithm can be used to monitor the students' academic performance. Based on the students' score they are grouped into different-different clusters

Applications:

Segmentation:



Every organisation wants their customers to be loyal but some time defining loyalty could be a tricky. Some of the common expectations are frequent spend, higher spend, higher ticket size and diversified spend.

Thank you!

QUESTIONS ?