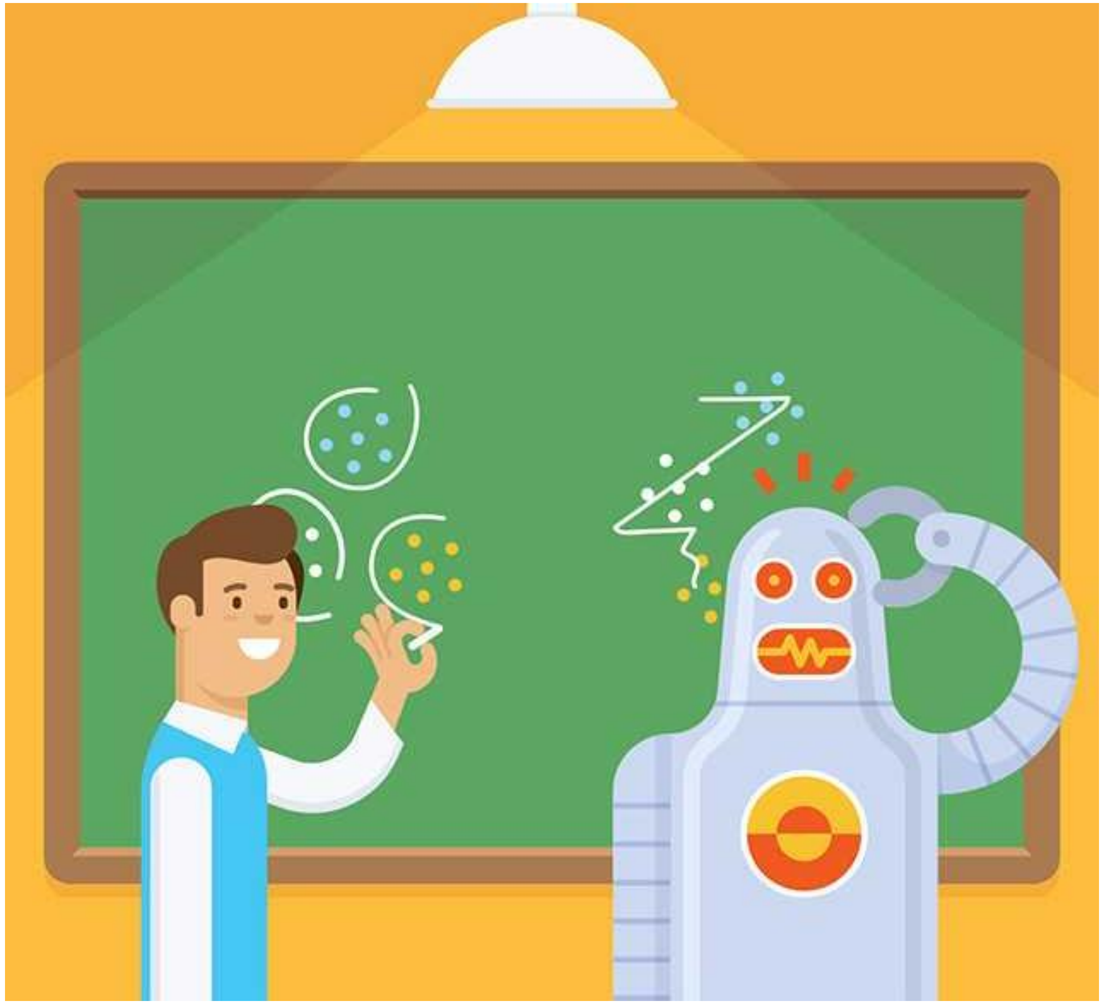
code unnati

edunet foundation

K-Means Clustering

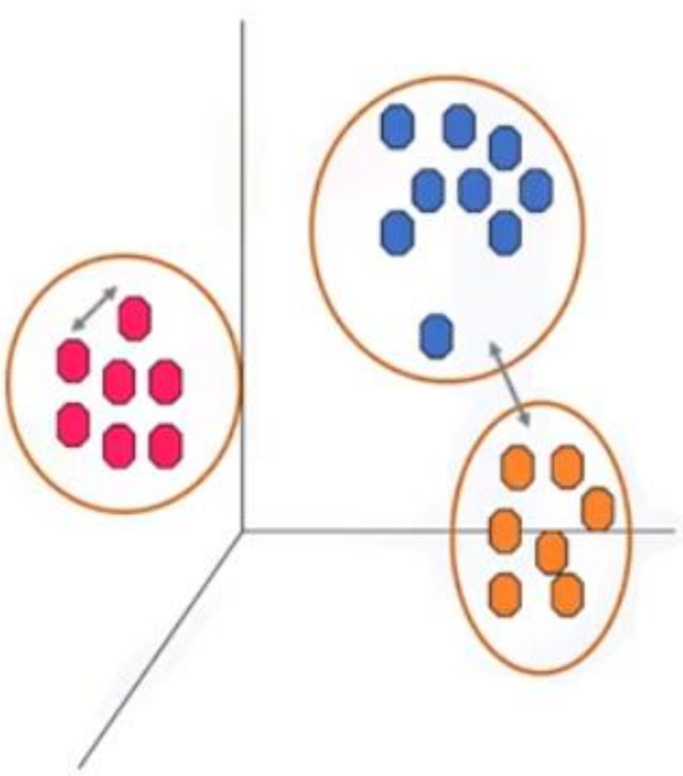


Objective

- Introduction
- Calculating dissimilarity or distance between two cases
- Dissimilarity Measures
- Scatter Plot
- Clustering
- Cluster Initiation
- Assign data points to Clusters
- Distance Matrix
- Cluster Shapes
- Iterative Process
- Summary

Introduction

- Various types of clustering algorithms available, such as partitioning, hierarchical, or density-based clustering.
- We would consider k-Means, which is a type of partitioning clustering.
- Divides the data into k non-overlapping subsets (or clusters) without any cluster-internal structure, or labels.
- It’s an unsupervised algorithm.
- Objects within a cluster are very similar and objects across different clusters are very different or dissimilar.



Reference : https://medium.com/analytics-vidhya/k-means-clustering-cdcba44a1d51

Introduction

- Imagine we have a customer dataset, and we need to apply customer segmentation on this dataset.
- As we already know, customer segmentation is the practice of partitioning a customer base into groups of individuals that have similar characteristics.
- For customer segmentation, we can choose k-Means clustering.
- k-Means can group data only “unsupervised,” based on the similarity of customers to each other.

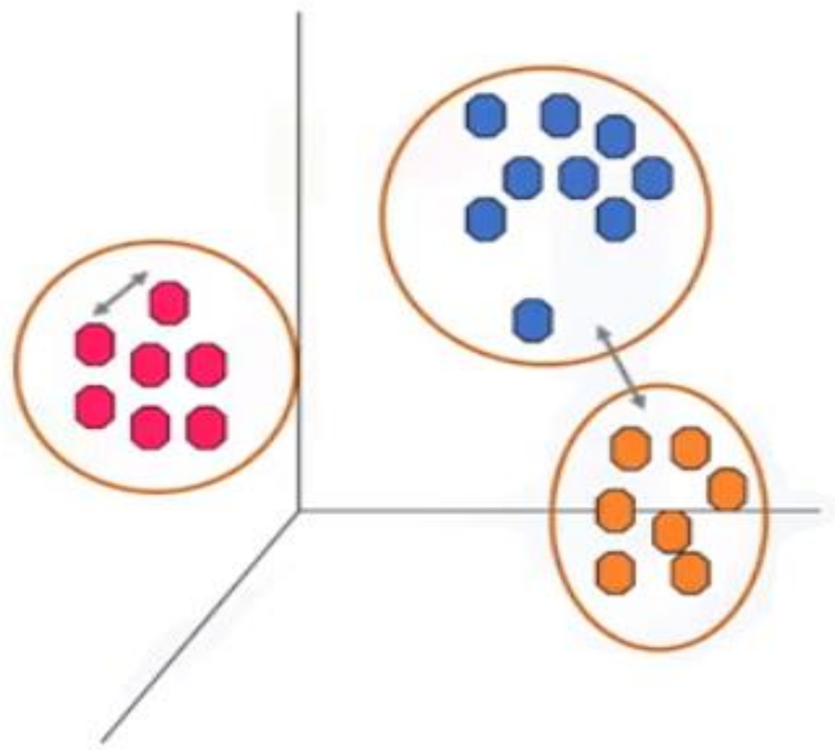
Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

Reference : https://medium.com/analytics-vidhya/k-means-clustering-cdcba44a1d51

Introduction

- Conventionally, in K-means, the distance of samples from each other is used to shape the clusters.

- So, we can say, k-Means tries to minimize the “intra-cluster” distances and maximize the “inter-cluster” distances with help of “dissimilarity matrix”



Reference : https://medium.com/analytics-vidhya/k-means-clustering-cdcba44a1d51

Calculating dissimilarity or distance between two cases

- Assume that we have two customers viz. customer 1 and customer 2, which have only one feature “Age”.
- Customer1: Age = 54 Customer2: Age = 50
- Euclidean distance can be used to measure distance between two customers

$$Dis(x_1, x_2) = \sqrt{\sum_{i=0}^{n} (x_{1i} - x_{2i})^2}$$

$$Dis(x_1, x_2) = \sqrt{\sum_{i=0}^{n} (54 - 50)^2} = 4$$

Similarly for other features, in case of multi-dimensional vectors

Dissimilarity Measures

- We must normalize our feature set to get the accurate "dissimilarity measure".

- For example, you may use Euclidean distance, cosine similarity, average distance, and so on.

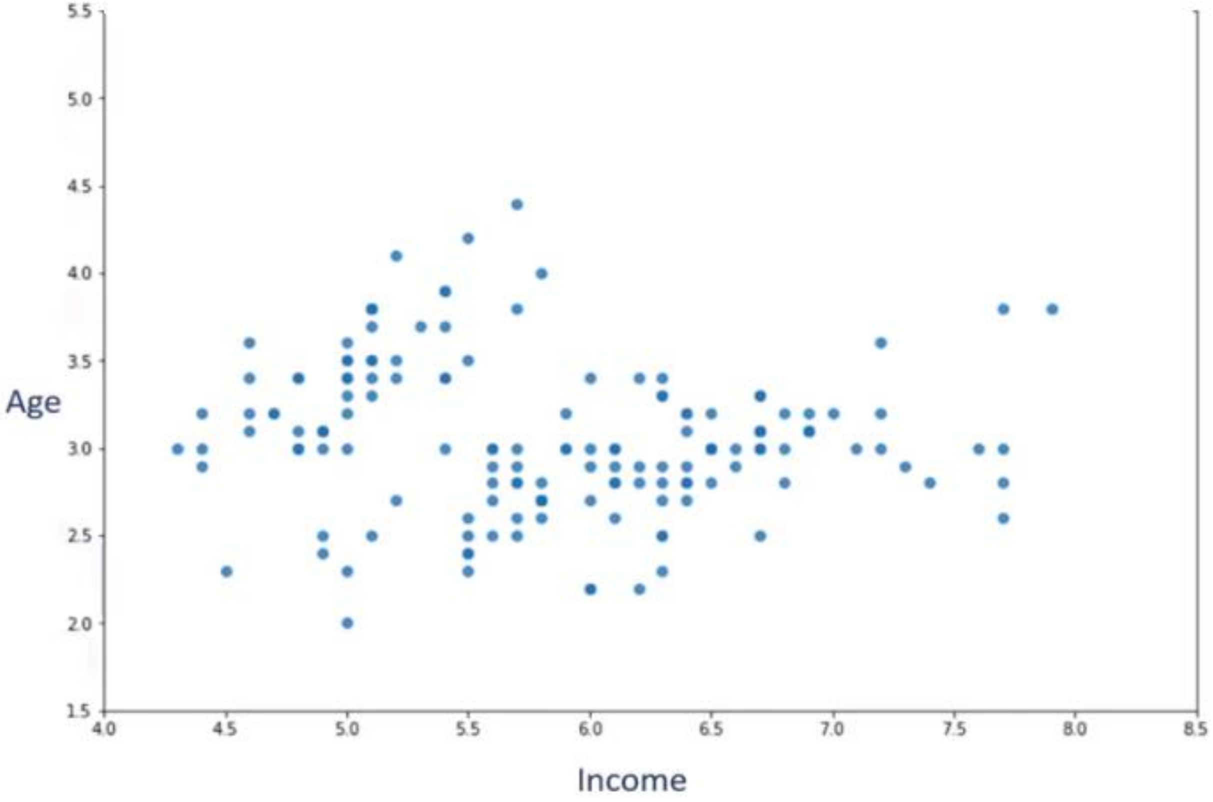
- "Similarity measure" highly controls how the clusters are formed, so it is recommended to understand the domain knowledge of your dataset, and data type of features, and then choose the meaningful distance measurement.

Scatter Plot

- To keep things simple, let's assume that our dataset has only two features, the age and income of customers.

- This means, it's a 2-dimentional space.

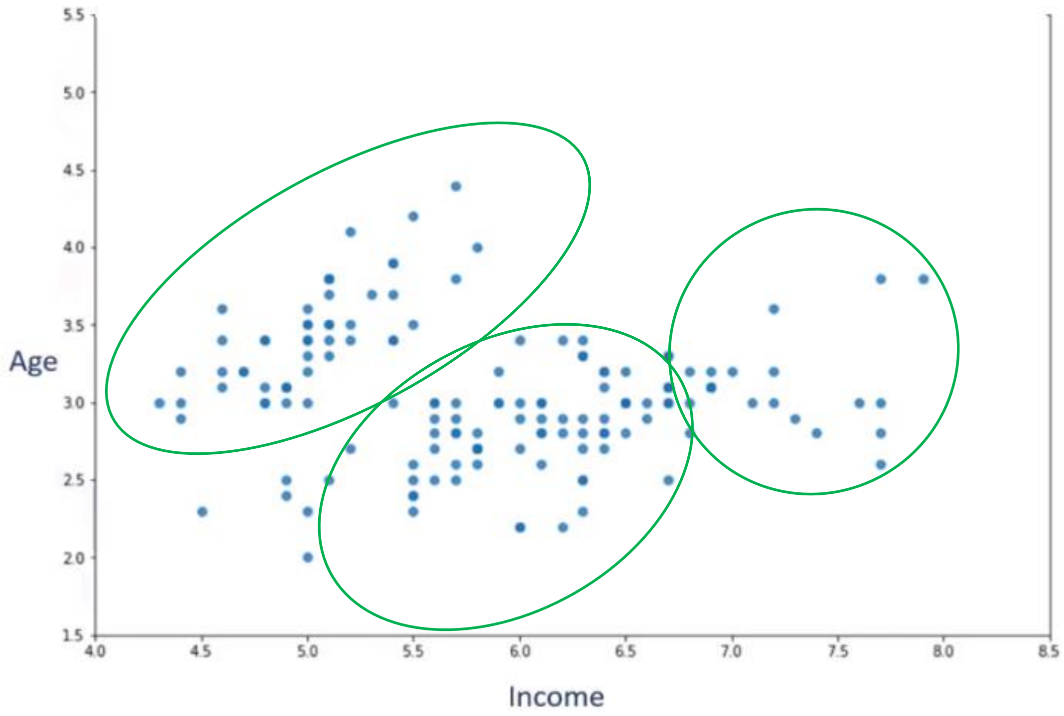
- We can show the distribution of customers using a scatterplot.



Clustering

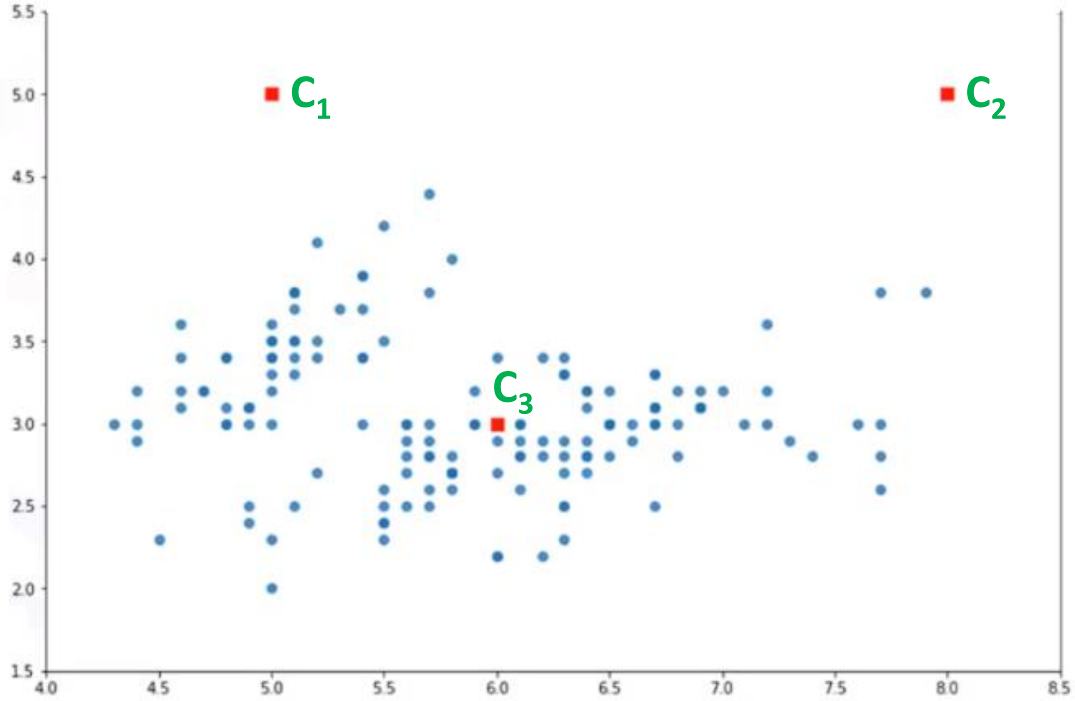
- We try to cluster the customer dataset into distinct groups (or clusters) based on these two dimensions.
- In the first step, we should determine the number of clusters.

Customer ID	Age	Income
1	3	4.4
2	2.3	4.5
3	2	5
….	….	….



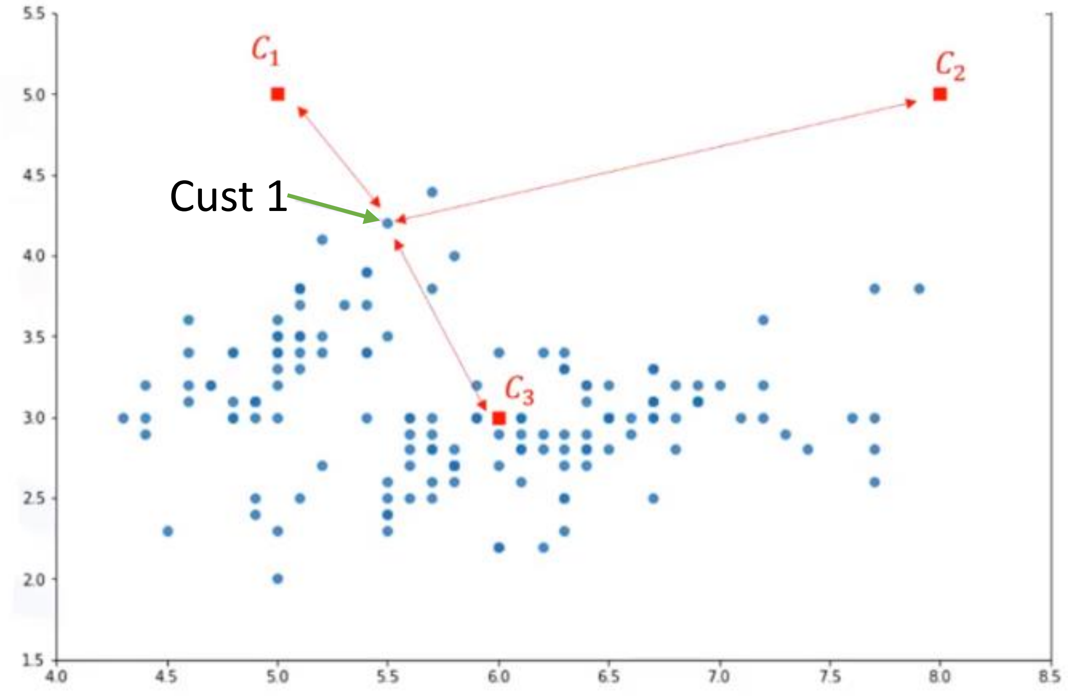
Clusters Initiation

- The key concept of the k-Means algorithm is that it randomly picks a center point for each cluster.
- It means, we must initialize k, which represents "number of clusters."
- Essentially, determining the number of clusters in a data set, or k, is a hard problem so we randomly take k=3 for our dataset.
- These 3 data points are called “centroids of clusters”, and should be of same feature size of our customer feature set.



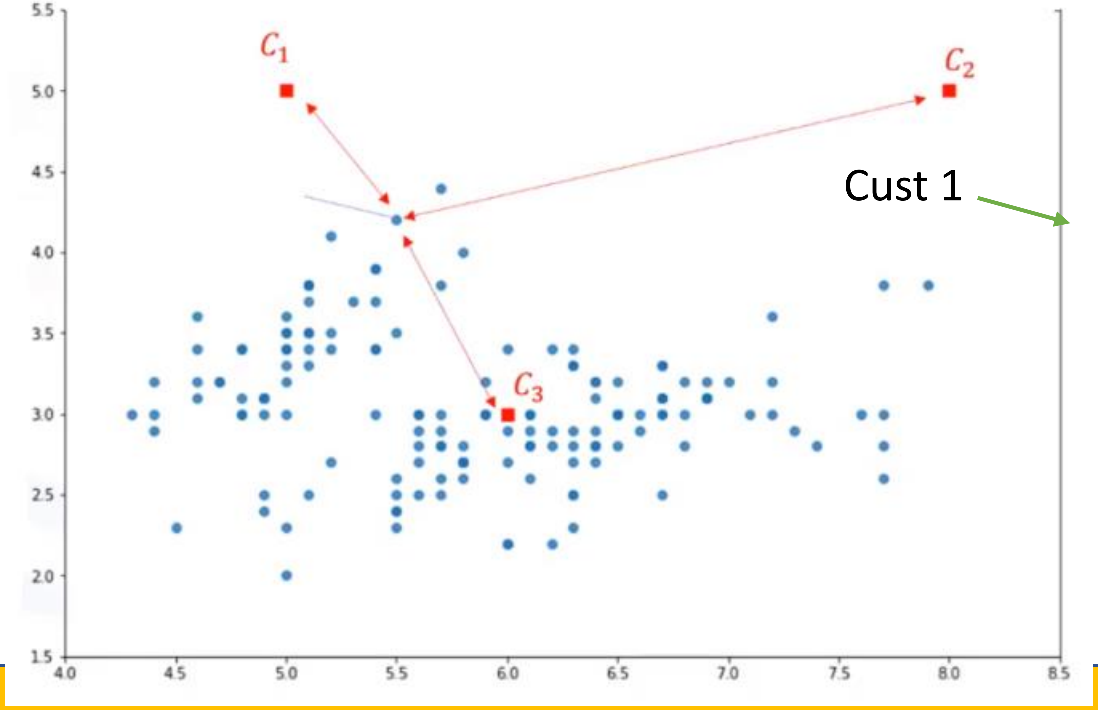
Assign data points to Clusters

- Next step is to assign each customer to the closest center so that we can find the closest centroid to each data point.



Distance Matrix

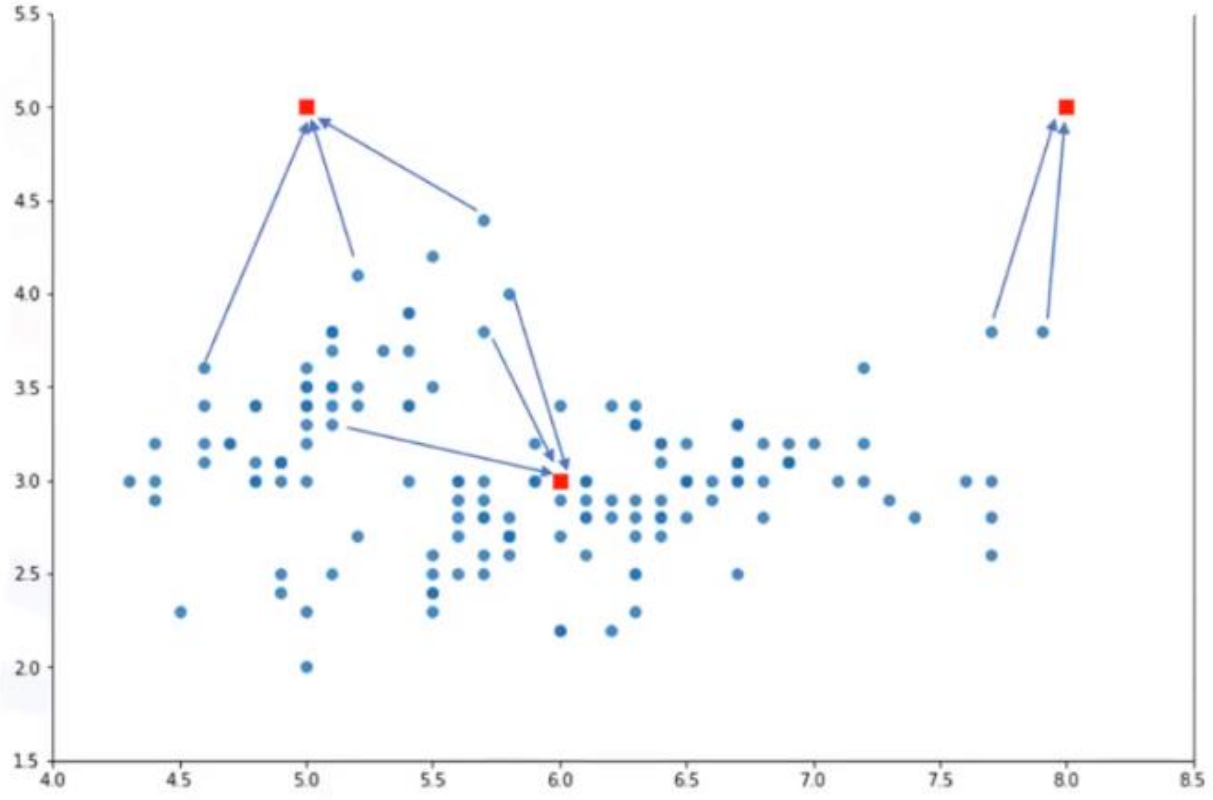
- Therefore, you will form a matrix where each row represents the distance of a customer from each centroid, called the "distance-matrix.“
- The main objective of k-Means clustering is to minimize the distance of data points from the centroid of its cluster and maximize the distance from other cluster centroids.



Assign all data points

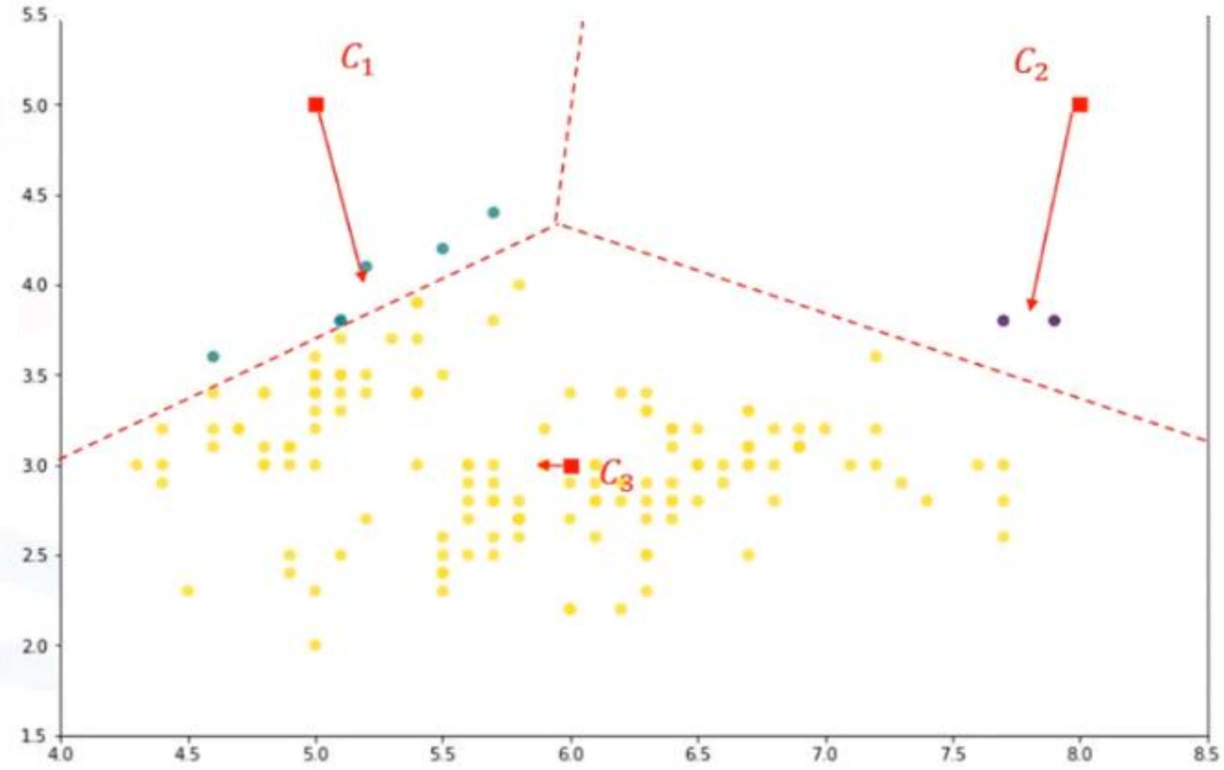
- We can use the distance-matrix to find the nearest centroid to data points.

- After finding the closest centroids for each data point, we assign each data point to that cluster.



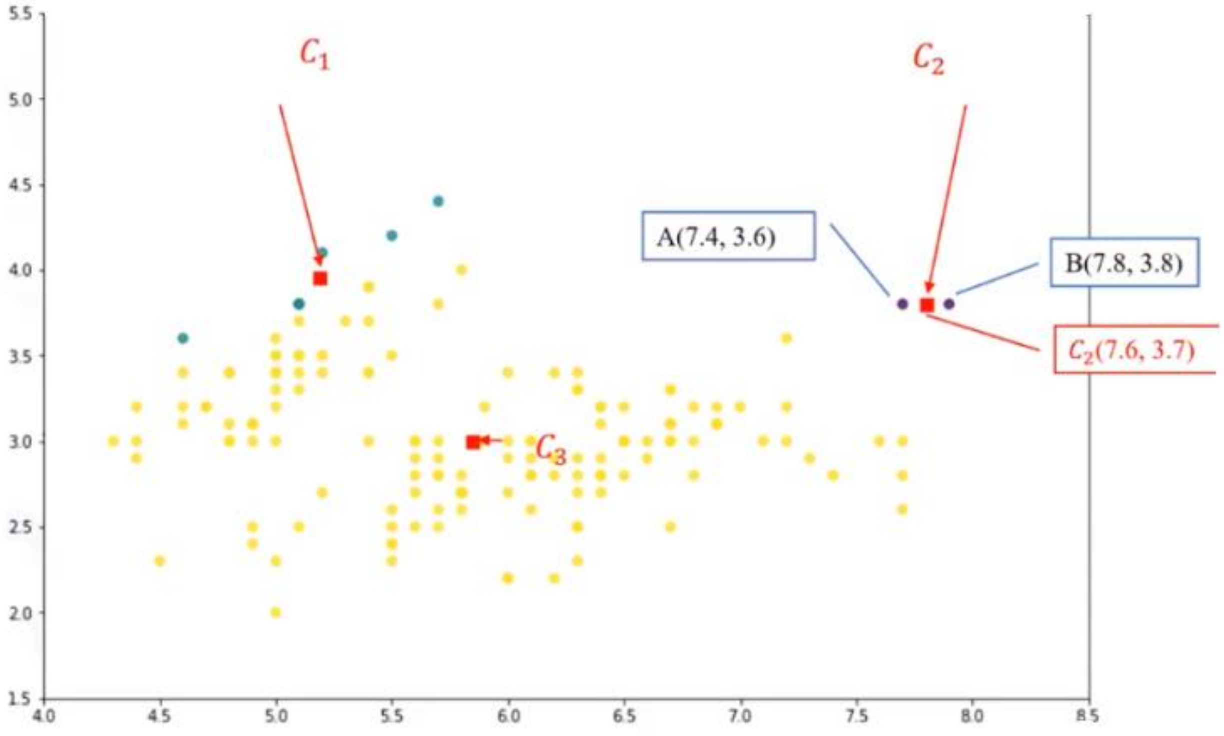
Cluster Shapes

- To reduce this error, we should shape clusters in such a way that the total distance of all members of a cluster from its centroid be minimized.
- Take average of data points in each cluster
- Shift the cluster centre to new location



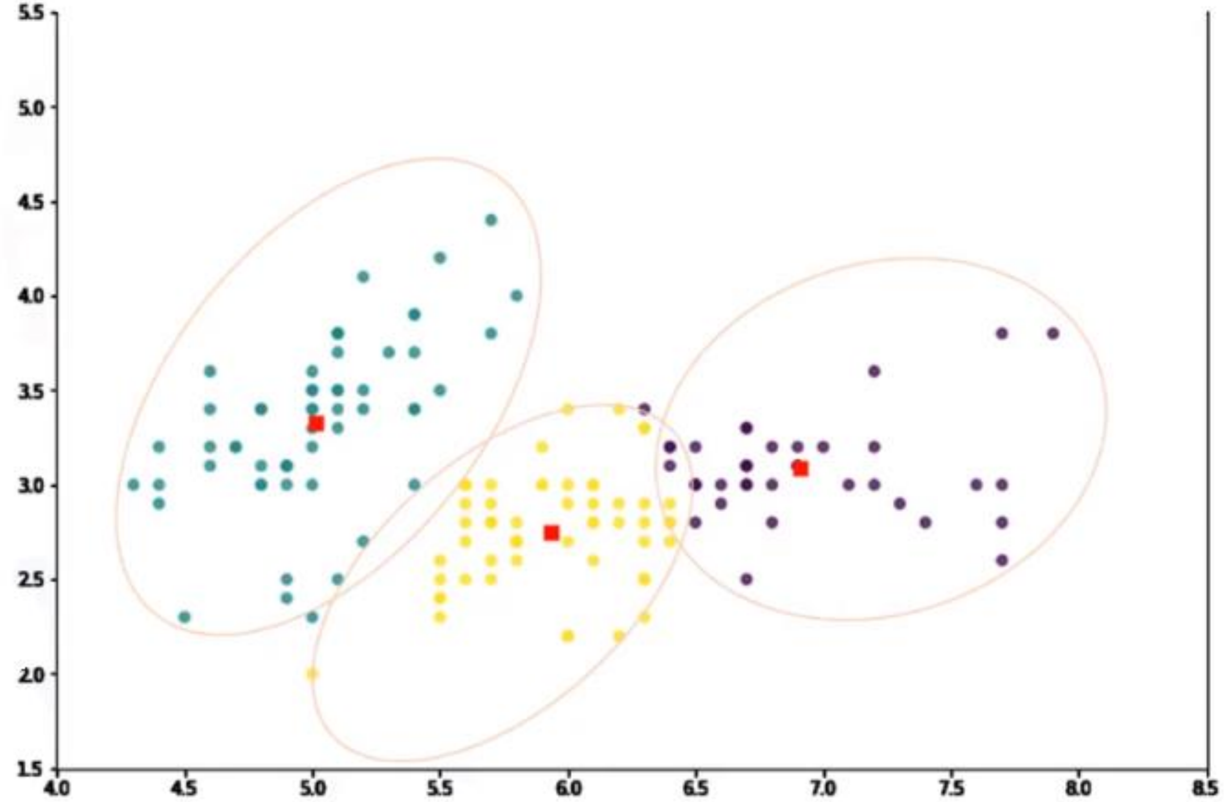
Handle new centroids

- Now we have new centroids.
- We will have to calculate the distance of all points from the new centroids.
- The points are re-clustered and the centroids move again.
- This continues until the centroids no longer move.
- Please note that whenever a centroid moves, each point's distance to the centroid needs to be measured again.



Iterative Process

- k-Means is an iterative algorithm, and we have to repeat steps 2 to 4 until the algorithm converges.
- It results in the clusters with minimum error, or the most dense clusters.
- However, as it is a heuristic algorithm, there is no guarantee that it will converge to the global optimum, and the result may depend on the initial clusters.
- It means this algorithm is guaranteed to converge to a result, but the result may be a local optimum (i.e. not necessarily the best possible outcome).



K-Mean Clustering Algorithm: Summary

Steps:

1. Randomly placing centroid for each cluster.
    - Clusters should be placed as far as possible.
2. Calculate the distance of each points from each centroids
    - Generally, Euclidian distance is used to measure distance from the object to the centroid.
3. Assign each point to the closest centroid creating a cluster.
4. After assigning each data points to a cluster, recalculate the position of k centroids for each cluster.
    - The new centroid position is determined by the mean of all points in the cluster.
5. Repeat Step 2-3-4 until the centroids no longer moves

Determining ‘K’ in K-Means

- Determining the number of clusters in a data set, or k, as in the k-Means algorithm, is a frequent problem in data clustering.

- The correct choice of k is often ambiguous, because it’s very dependent on the shape and scale of the distribution of points in a data set.

- There are some approaches to address this problem, but one of the techniques that is commonly used, is to **run the clustering across the different values of K, and looking at a metric of accuracy for clustering.**

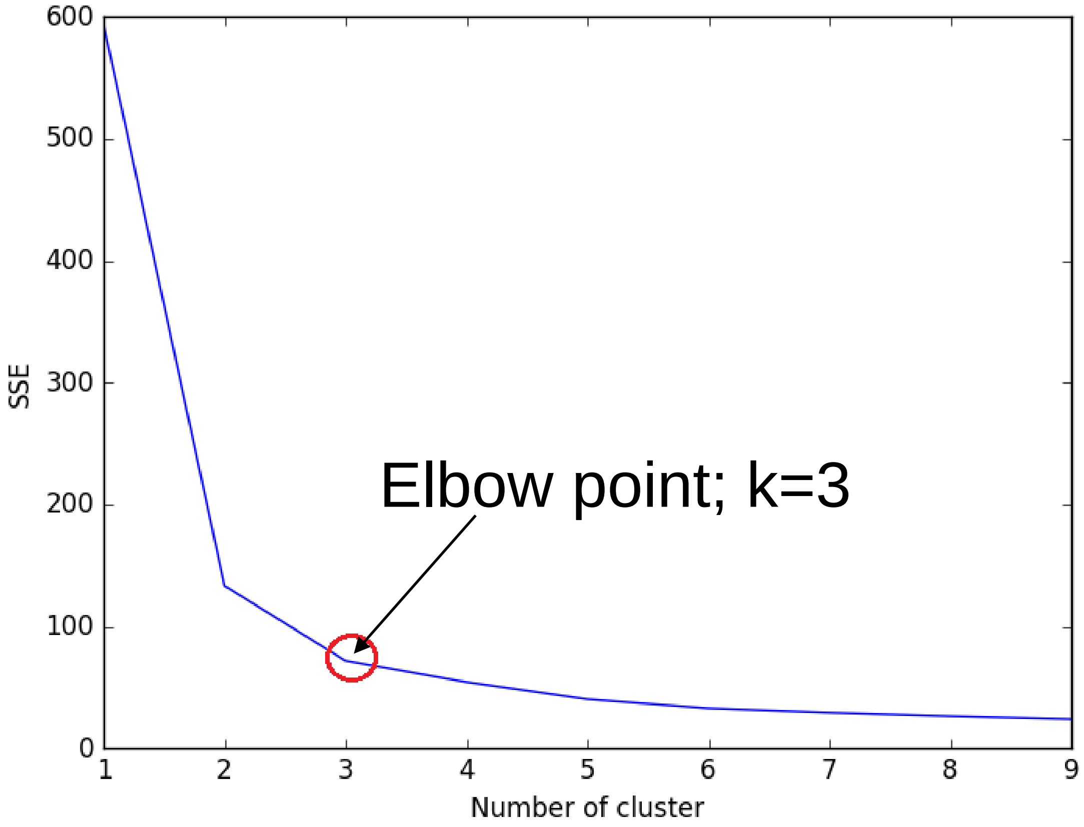
Elbow method

- This metric can be “mean distance between data points and their cluster centroid,” or SSE (Sum of Squared Error) which indicate how dense our clusters are, or to what extend we minimized the error of clustering.

- Then looking at the change of this metric, we can find the best value for k.

- **But the problem is that with increasing the number of clusters, the <u>distance of centroids to data points will always reduce.</u>**

Elbow Point



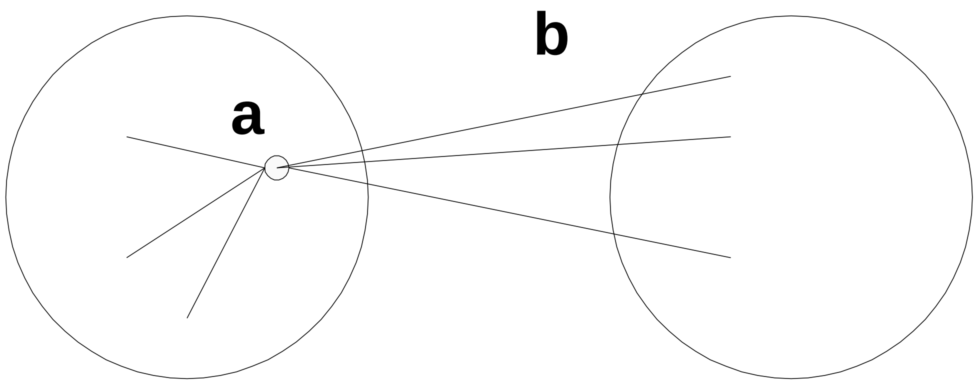
600
500
400
300
200
100
0
SSE
Elbow point; k=3
1
2
3
4
5
6
7
8
9
Number of cluster

Silhouette Coefficient

- Silhouette Coefficient combines ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings

- For an individual point, I
  - a = average distance of i to the points in the same cluster
  - b = min (average distance of i to points in another cluster)
  - silhouette coefficient of i:

    s = 1 – a/b   if a < b



  - Typically between 0 and 1.
  - The closer to 1 the better.

Hands On

Thank You