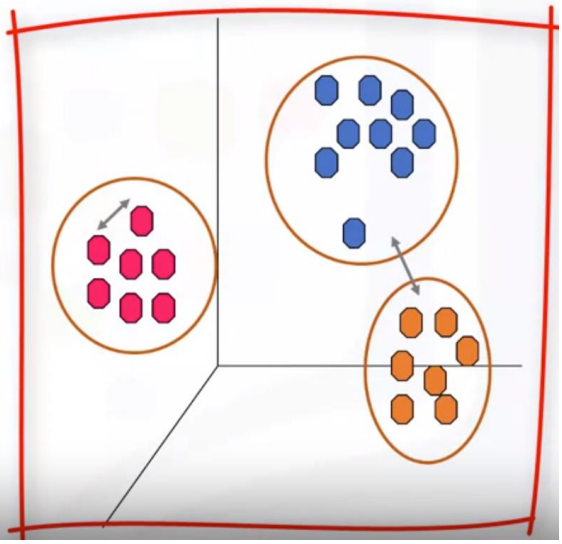
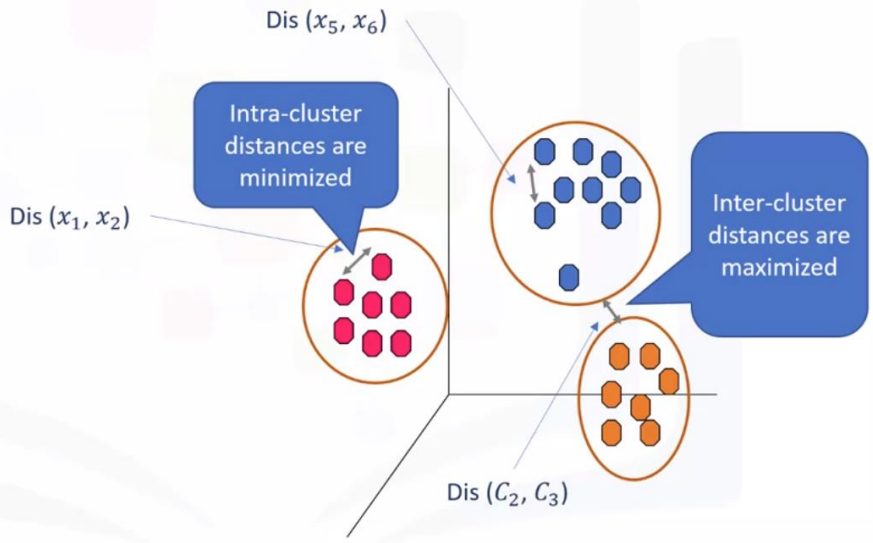
k-Means algorithms

- Partitioning Clustering
- K-means divides the data into non-overlapping subsets (clusters) without any cluster-internal structure
- Examples within a cluster are very similar
- Examples across different clusters are very different

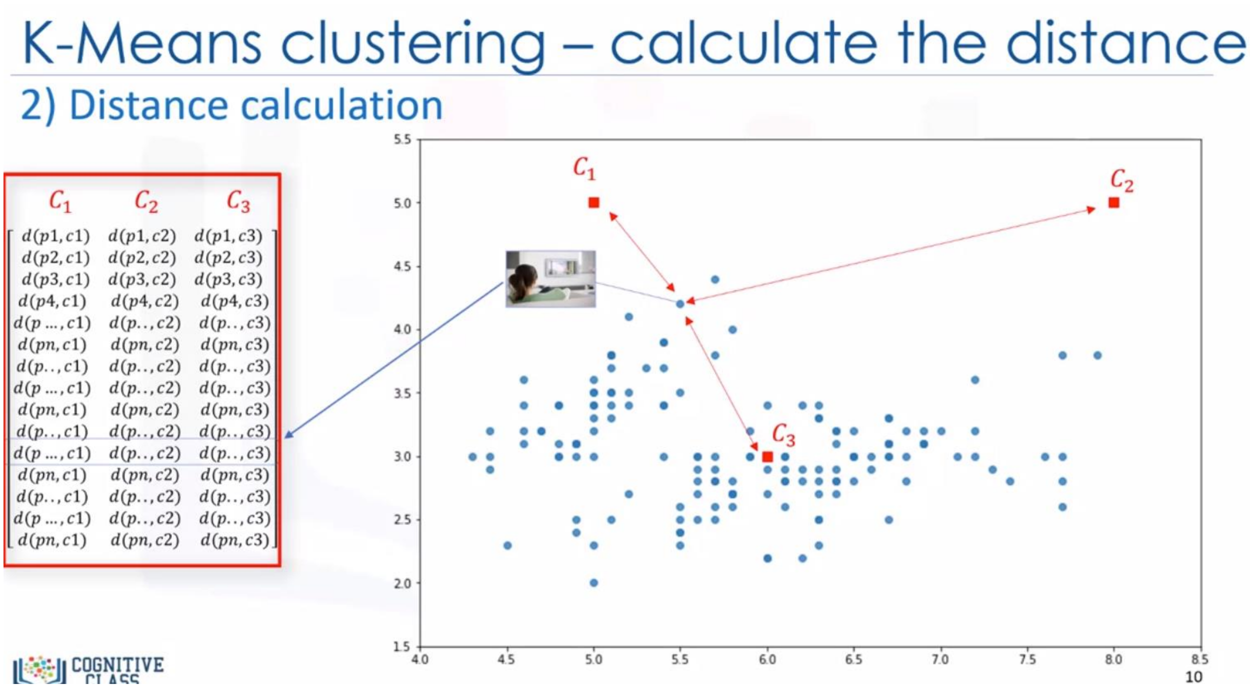


Determine the similarity or dissimilarity

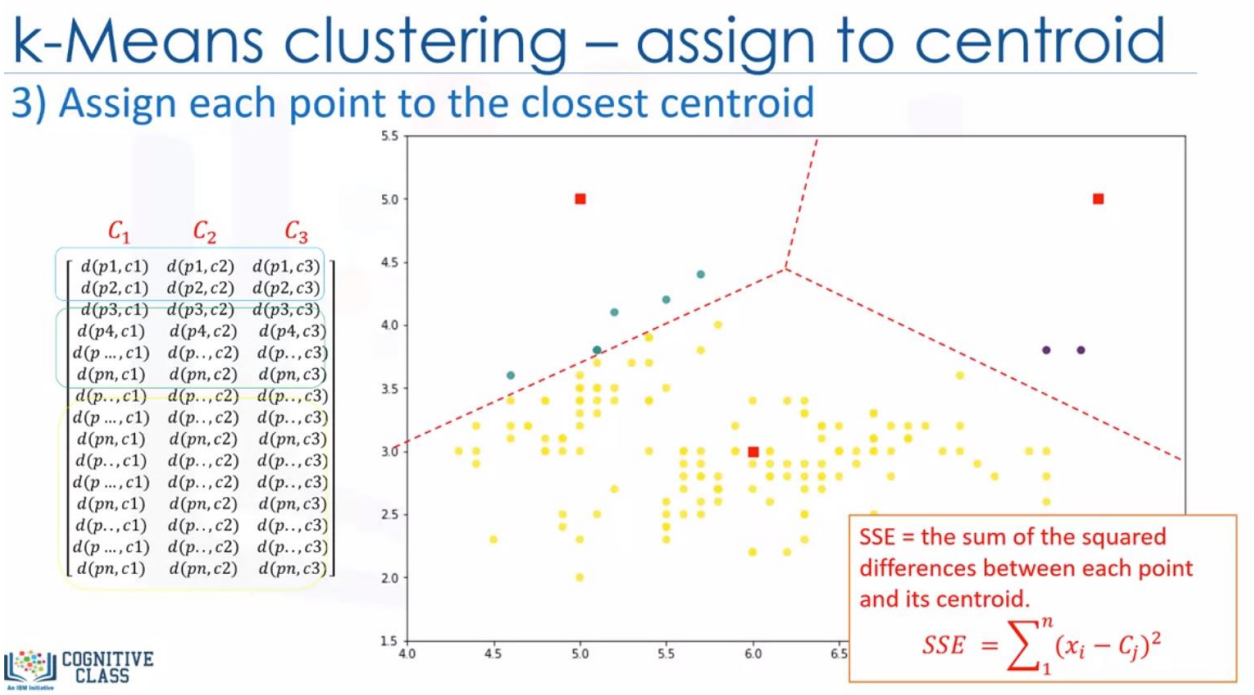


There are 2 approaches to choose these centroids.

1. Randomly choose 3 observations out of the dataset and use these observations as the initial means.
2. We create 3 random points as centroids of the clusters which is our choice that is shown in the plot with red color.



Form a matrix where each row represents the distance of a customer from each centroid. It's called the Distance Matrix.

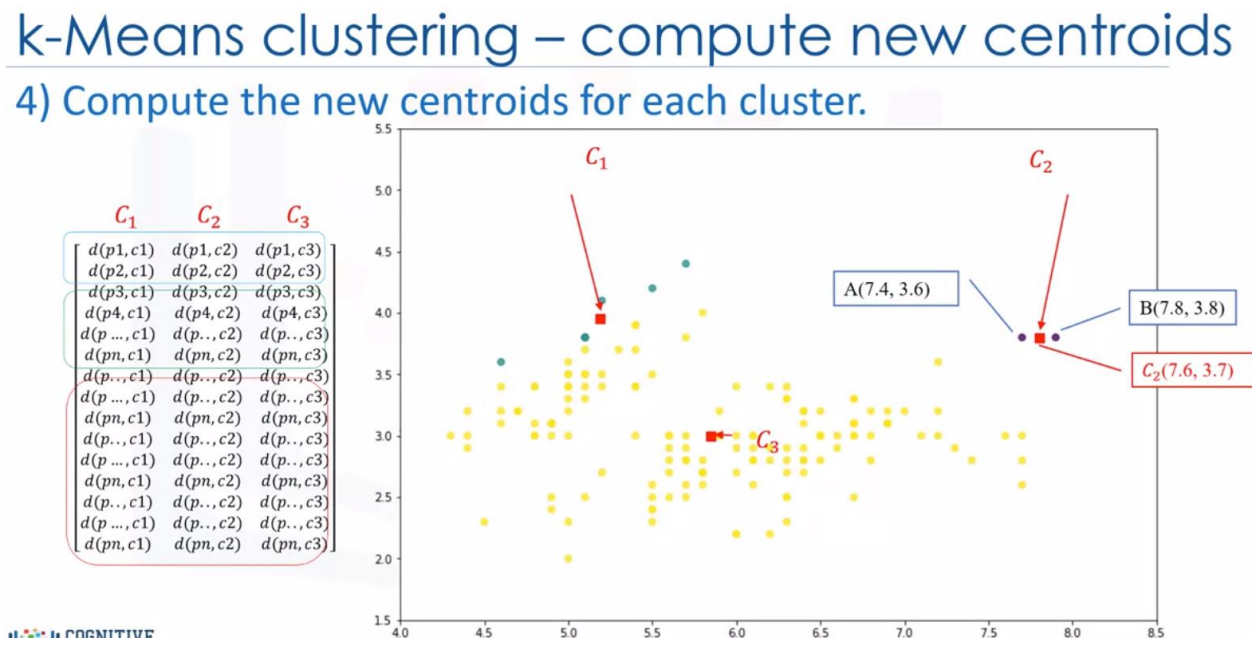


The main objective of K-Means clustering is to minimize the distance of data points from the centroid of this cluster, and maximize the distance from other cluster centroids.

All the customers will fall to a cluster based on their distance from centroids.

We can easily say that it doesn't result in good clusters, because the centroids were chosen randomly from the first. The model would have a high error.

The error is the total distance of each point from its centroid.



The centroid of each of the 3 clusters becomes the new mean.

For example, if point A coordination is 7.4 and 3.6, and B point features are 7.8 and 3.8, the new centroid of this cluster with 2 points would be the average of them, which is 7.6 and 3.7. Now we have new centroids.

Once again we will have to calculate the distance of all points from the new centroids. The points are reclustered and the centroids move again.

This continues until the centroids no longer move.

K-Means is an iterative algorithm and we have to repeat steps 2 to 4 until the algorithm converges.

In each iteration, it will move the centroids, calculate the distances from new centroids and assign data points to the nearest centroid.

It results in the clusters with minimum error or the most dense clusters.

There is no guarantee that it will converge to the global optimum and the result may depend on the initial clusters.

To solve this problem, it's common to run the whole process multiple times with different starting conditions. This means with randomized starting centroids, it may give a better outcome.

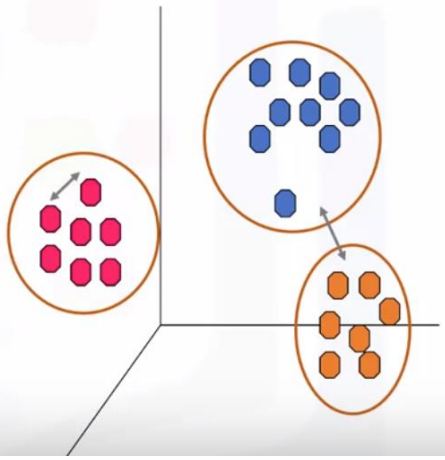
As the algorithm is usually very fast, it wouldn't be any problem to run it multiple times.

k-Means clustering algorithm

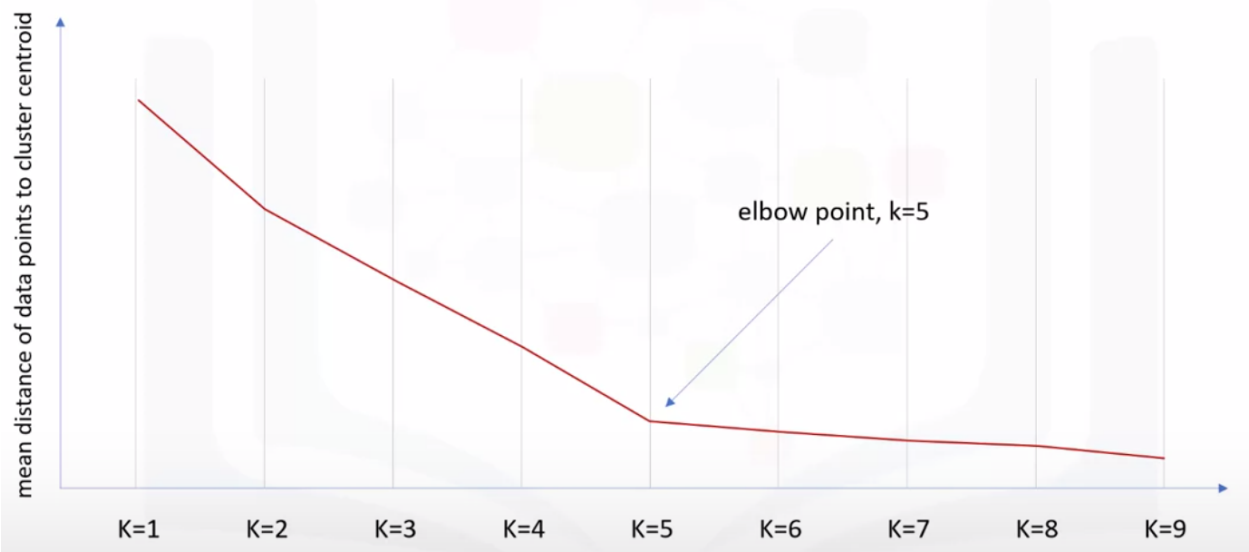
1. Randomly placing *k* centroids, one for each cluster.
2. Calculate the distance of each point from each centroid.
3. Assign each data point (object) to its closest centroid, creating a cluster.
4. Recalculate the position of the *k* centroids.
5. Repeat the steps 2-4, until the centroids no longer move.

k-Means accuracy

- External approach
  - Compare the clusters with the ground truth, if it is available.
- Internal approach
  - Average the distance between data points within a cluster.



Choosing k



In clustering evaluation process, “elbow point” is where the rate of accuracy increase sharply.

k-Means recap

- Med and Large sized databases (*Relatively efficient*)
- Produces sphere-like clusters
- Needs number of clusters (k)