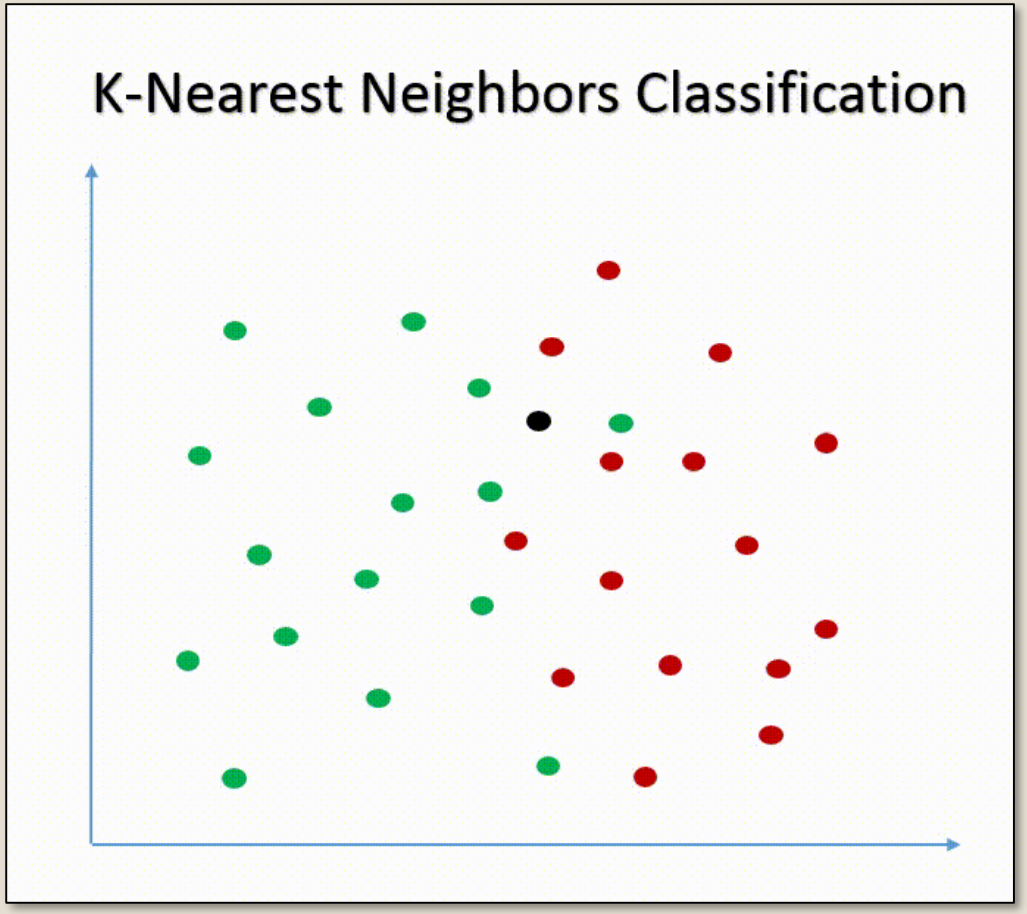
DATA SCIENCE FOR ENGINEERS

Week 8

Session Co-Ordinator : Abhijit Bhakte

Classification with KNN

- KNN stands for **K-Nearest Neighbour**
- KNN explains a **categorical value** using **majority votes** of nearest neighbour
- KNN is **nonparametric algorithm**
- There is **no any training phase** in KNN
- **Scaling is important** in KNN



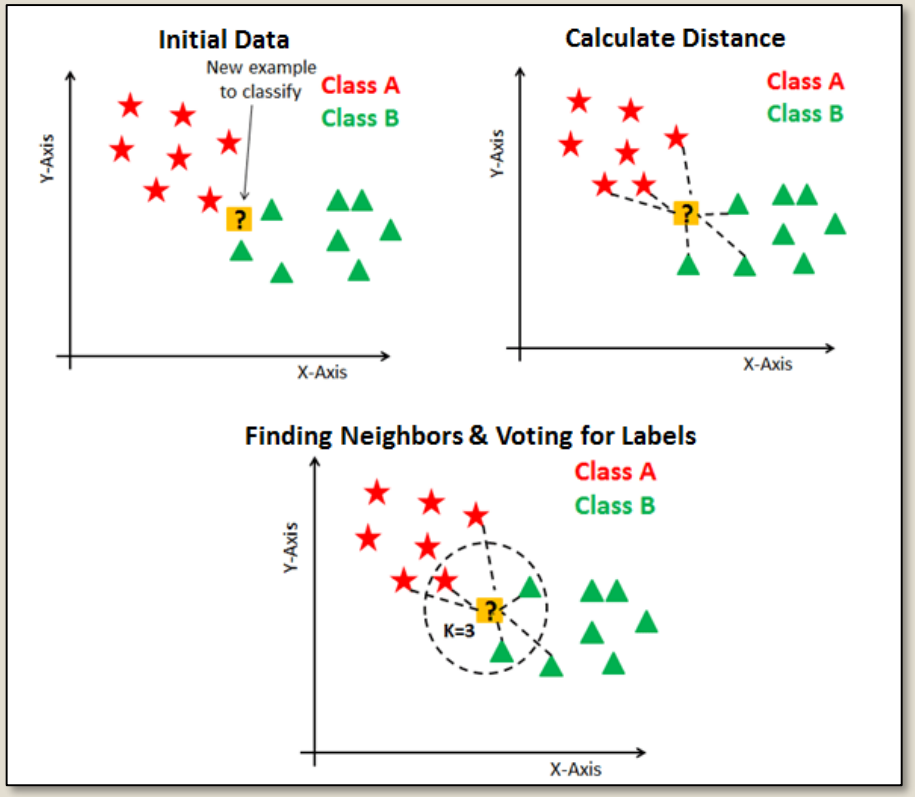
KNN Algorithm

Step1: Compute the distance metric between the test data point and all the label data points

Step2: Order the labelled data points in the increasing order of distance metric

Step3: Select the top k labelled data points and look at class labels

Step4: Find the majority of theses k-labelled data points and assign it to the test data point



Rstudio: Car service center problem

- A newly launch car service center have the capacity of 315 cars to check weather car need service or not by inspecting it.

- But on launch day they gave free service to all the cars.
- So, 450 cars came to the service center
- It is not possible to extend the working hours of the mechanics
- In this case how data scientist will help the service center
- **Solution:** we have easily measurable features of cars such as (Engine oil, tire wear, mileage etc) can be use to build model which give prediction of the remaining vehicals.

Mechanic

Data Scientist

Q) What is the primary purpose of the k-nearest neighbor (KNN) algorithm?

A. Classification
B. Regression
C. Clustering
D. Dimensionality Reduction

Solution

KNN is mainly used for classification tasks, where it assigns a data point to one of several predefined classes based on the majority class among its k-nearest neighbors.

Q) In KNN, what does "k" represent?

A. The number of clusters
B. The number of features
C. The number of neighbors to consider
D. The number of iterations

Solution

The number of neighbors to consider. "k" in KNN represents the number of nearest neighbors that the algorithm uses to make predictions

Q) How does KNN handle categorical data?

A. It cannot handle categorical data
B. It treats categorical data as continuous values
C. It uses specialized distance metrics for categorical data
D. It ignores categorical features during classification

Solution

It uses specialized distance metrics for categorical data. KNN can handle categorical data by using distance metrics like Hamming distance for categorical features.

Q) In KNN, what distance metric is commonly used for continuous numerical features?

A. Euclidean distance
B. Cosine similarity
C. Hamming distance
D. Jaccard similarity

Solution

Euclidean distance is commonly used for continuous numerical features in KNN.

Q) What is the "curse of dimensionality" in the context of KNN?

A. The model is too simple
B. The model is too complex
C. Increased computational complexity as the number of features grows
D. The model's inability to handle high-dimensional data

Solution

Increased computational complexity as the number of features grows. The "curse of dimensionality" refers to the increased computational requirements as the number of features or dimensions increases.

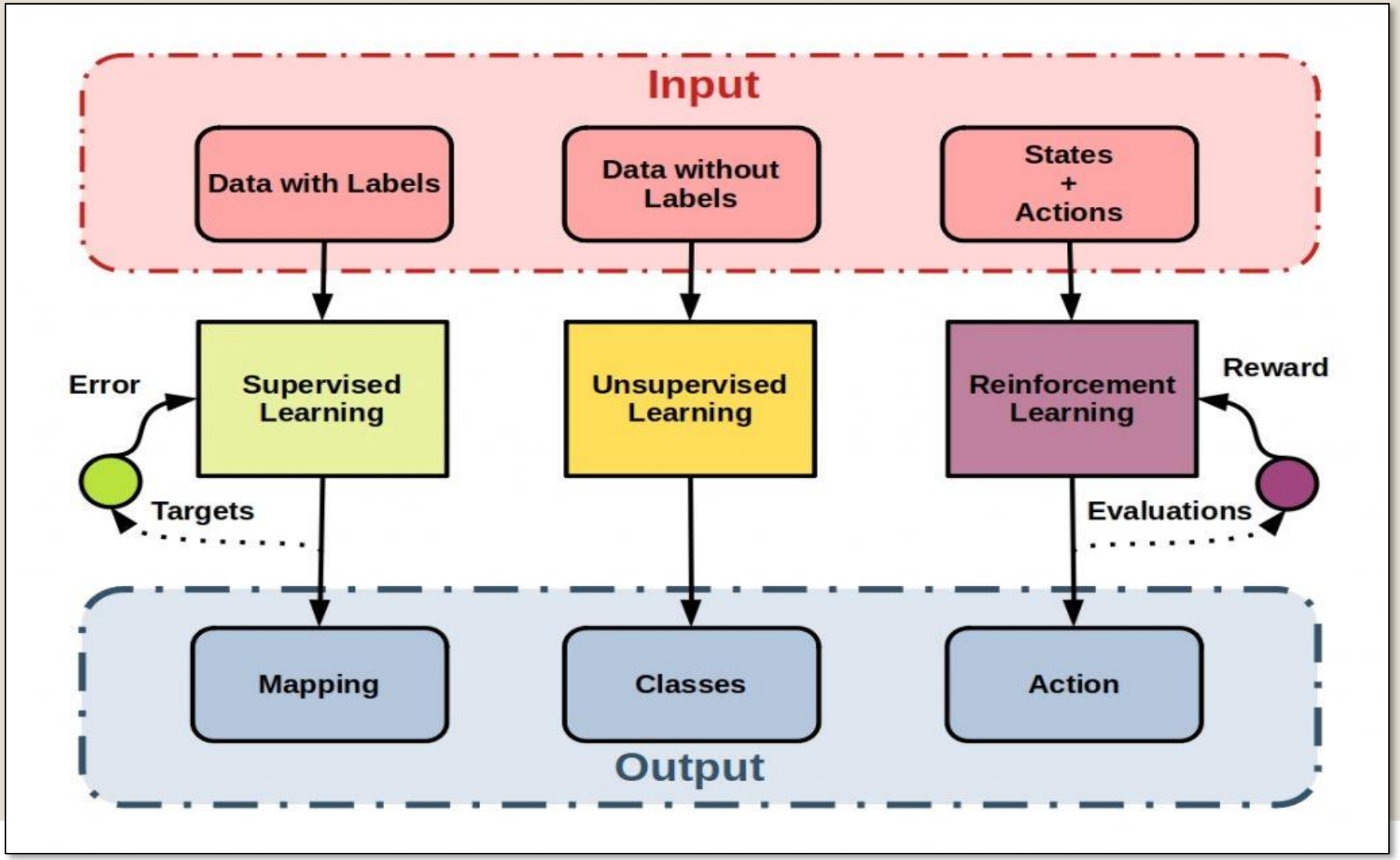
Q) You are using KNN for a regression task with "k" set to 3. The distances to the three nearest neighbors are 2.0, 3.0, and 4.0. What will be the predicted value using simple averaging?

A. 2.0
B. 3.0
C. 3.33
D. 3.67

Solution

To calculate the simple average, add up the distances and divide by "k":
(2.0 + 3.0 + 4.0) / 3 = 3

Supervised Vs Unsupervised Vs Reinforcement learning



Supervised Learning

Spam Mail Classifiers

Unsupervised Learning

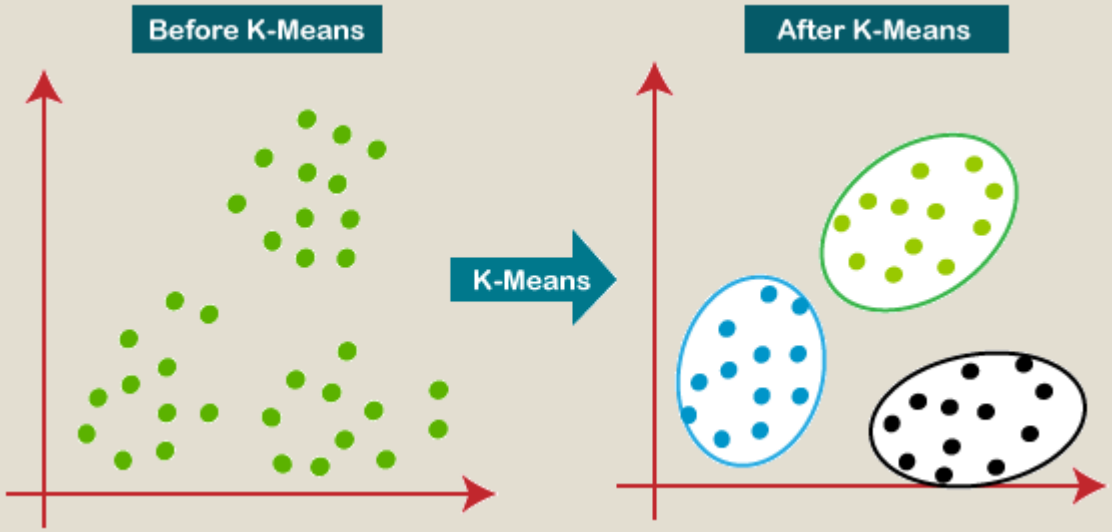
Movie Recommendation

Reinforcement Learning

Self Driving Cars

K-mean clustering

- One of the simplest **Unsupervised Algorithm**
- It is the technique to **partition N observations into k-clusters** in which each observation belongs to cluster with nearest mean



K-mean Algorithm

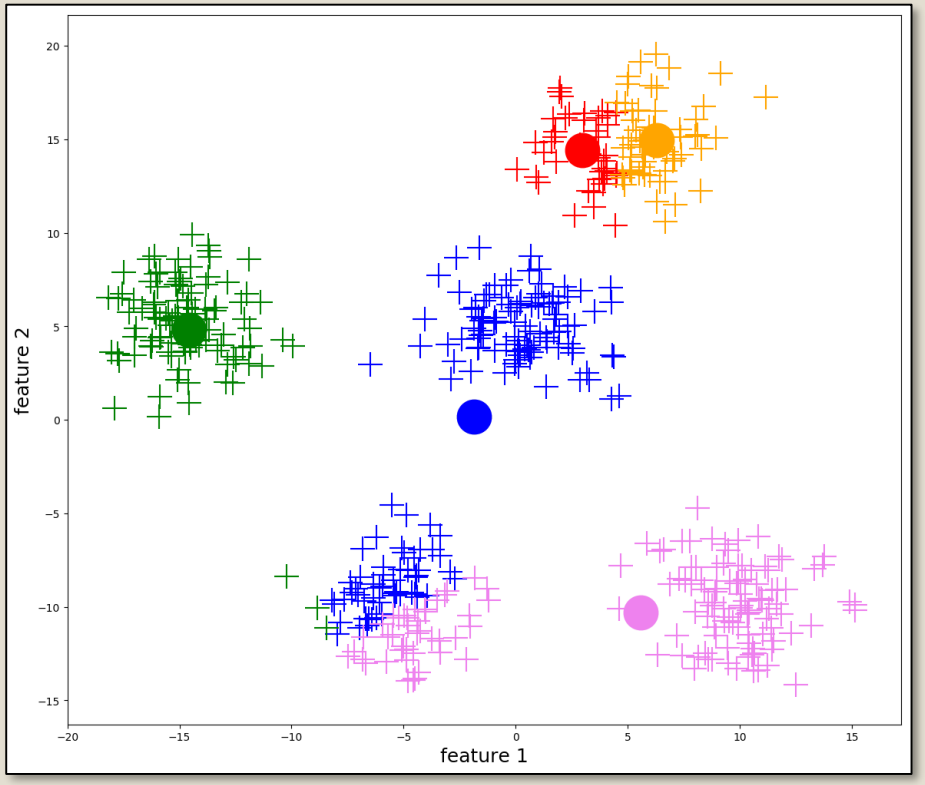
Step1: Select the number of clusters (k)

Step2: Select k points random

Step3: Make a cluster

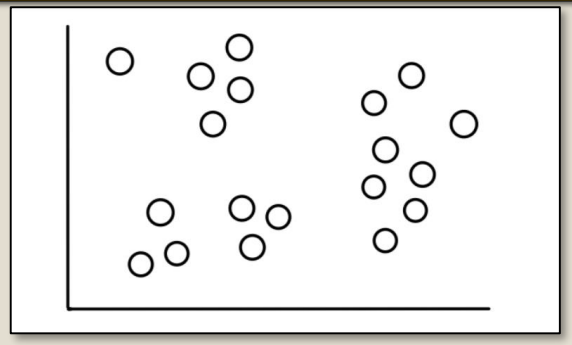
Step4: Compute new centroid of each cluster

Step5: Access the quality of each cluster



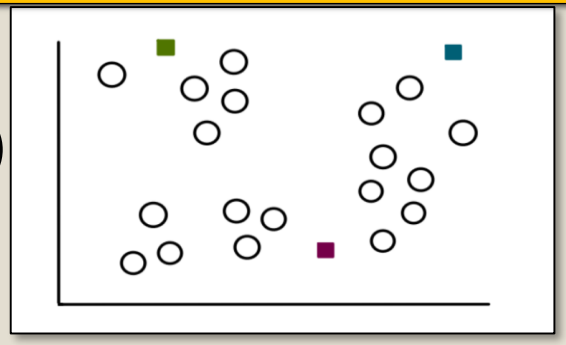
1

Select the number of clusters (k=3)



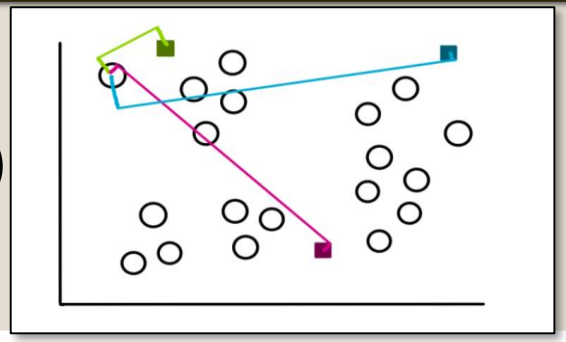
2

Select k points random



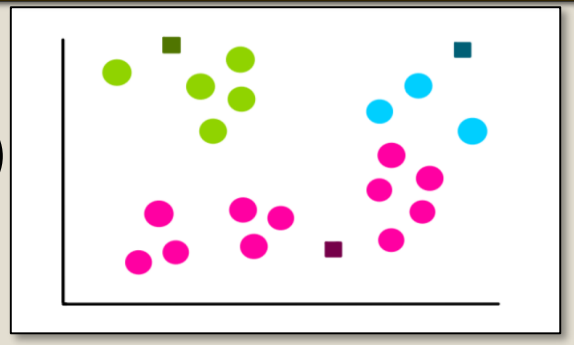
3

Make a cluster



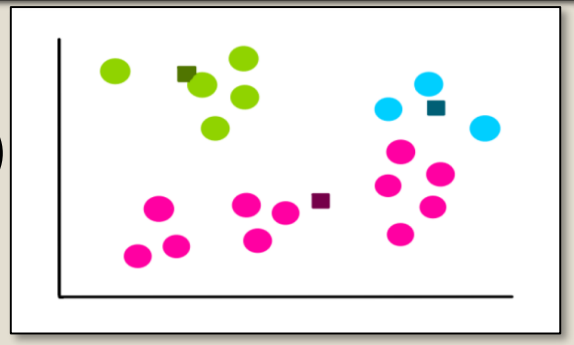
3

Make a cluster



4

Compute centroid of new cluster



5

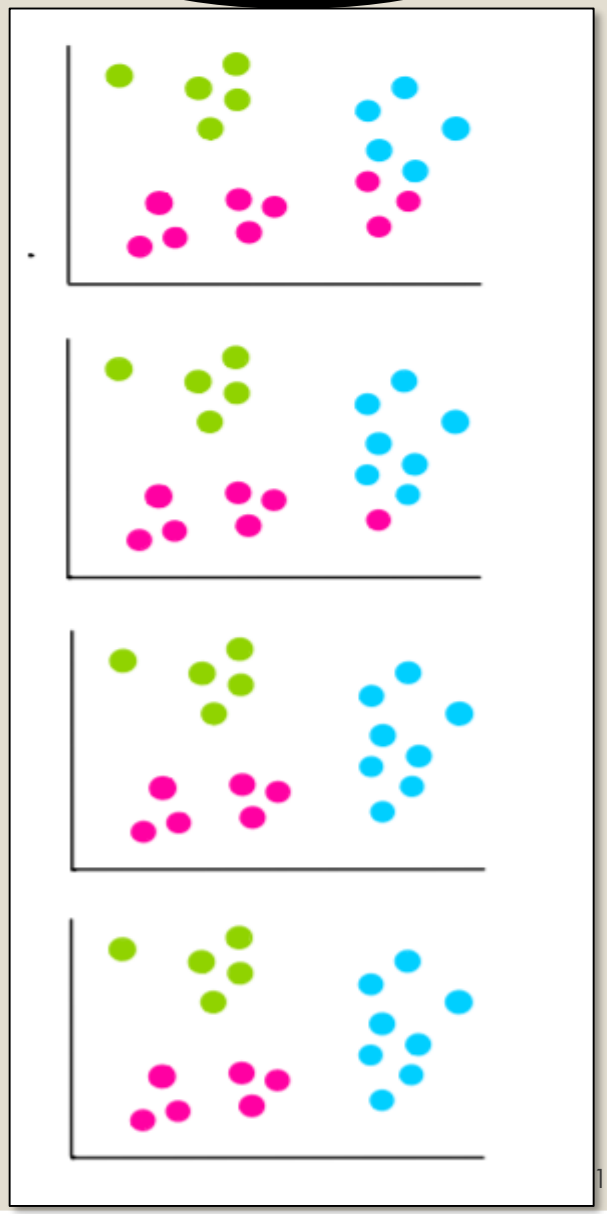
Access the quality of each cluster

$$WCSS = \sum_{C_k}^{C_n} \left(\sum_{d_i \text{ in } C_i}^{d_m} distance(d_i, C_k)^2 \right)$$

Where,

C is the cluster centroids and d is the data point in each Cluster.

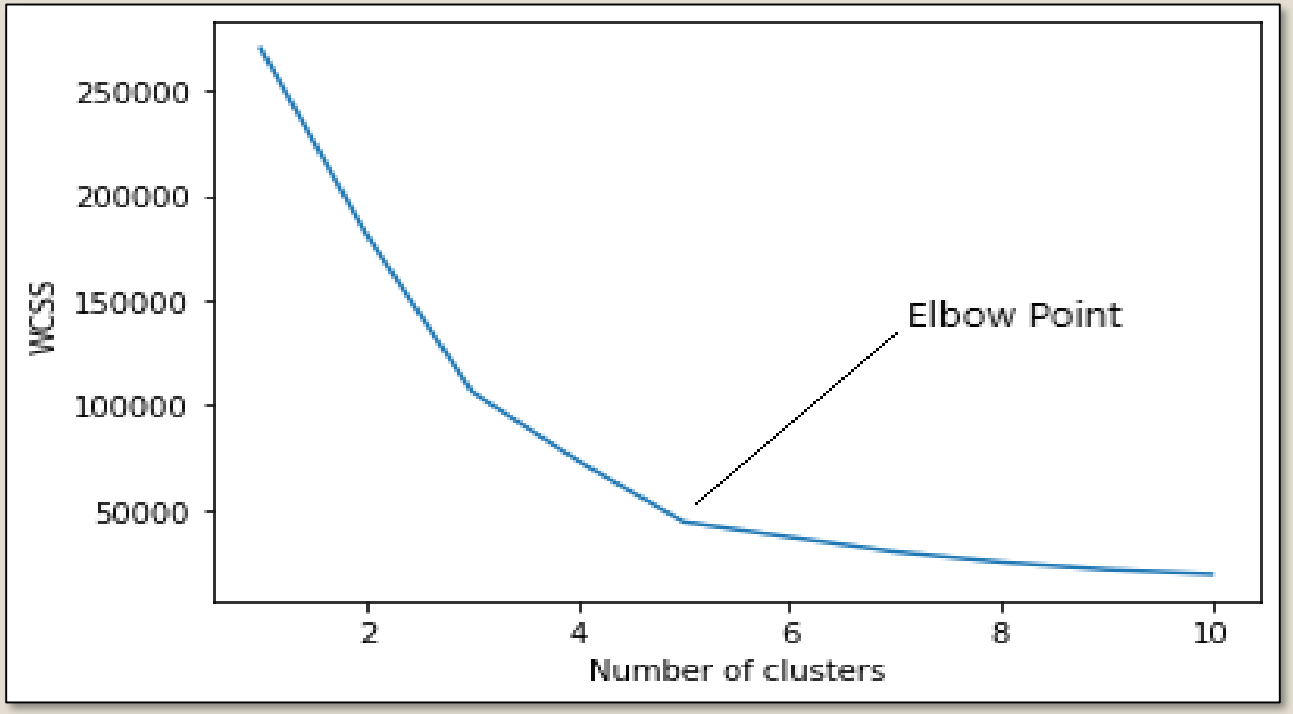
1-4 Repetition



Rstudio

How to choose k value ?

- Plot the graph between WCSS Vs k value
- The k value at which elbow shape obtained is the optimal k value



Q) What is the primary objective of the k-means clustering algorithm?

A. Classification
B. Dimensionality reduction
C. Clustering
D. Regression

Solution

The primary goal of the k-means clustering algorithm is to partition a dataset into distinct, non-overlapping groups or clusters.

Q) How does the k-means algorithm initialize cluster centroids?

A. Randomly
B. Using hierarchical clustering
C. Based on class labels
D. With gradient descent

Solution

To calculate the simple average, add up the distances and divide by "k":
(2.0 + 3.0 + 4.0) / 3 = 3

Q) What does "k" represent in k-means clustering?

A. The number of data points
B. The number of clusters
C. The number of features
D. The number of iterations

Solution

"k" in k-means represents the number of clusters you want the algorithm to create.

Q) What is the objective function that k-means tries to minimize during clustering?

A. Sum of squared differences between data points and centroids
B. Entropy
C. Mean absolute error
D. F1-score

Solution

The k-means algorithm minimizes the sum of squared distances between data points and the centroids of their respective clusters.

Q) What is the key assumption made by k-means clustering regarding cluster shapes?

A. Clusters can have any arbitrary shape
B. Clusters are spherical and equally sized
C. Clusters are linearly separable
D. Clusters have the same density

Solution

Clusters are spherical and equally sized. K-means assumes that clusters are spherical in shape and have roughly equal sizes.

Q) What is the role of the elbow method in k-means clustering?

A. To determine the number of clusters
B. To compute the final cluster assignments
C. To calculate the silhouette score
D. To measure the Davies-Bouldin index

Solution

To determine the number of clusters. The elbow method is used to find the optimal number of clusters (k) for k-means clustering by observing the change in within-cluster variance as k varies.

Q) Given a dataset with 100 data points, if you perform k-means clustering with k = 5, how many cluster centroids will be initialized initially?

A. 100
B. 10
C. 5
D. 1

Solution

In k-means clustering, "k" represents the number of clusters, and that's the number of cluster centroids that will be initialized initially.

Q) You are comparing the performance of two k-means clustering solutions. Solution A has a WCSS of 800, and Solution B has a WCSS of 700. Which solution is likely to be better?

A. Solution A
B. Solution B
C. Both solutions are equally good
D. More information is needed to determine

Solution

To determine the number of clusters. The elbow method is used to find the optimal number of clusters (k) for k-means clustering by observing the change in within-cluster variance as k varies.

Thank you