K-means clustering

is a popular unsupervised machine learning algorithm used for clustering data points into groups or clusters based on their similarity. The algorithm aims to minimize the sum of squared distances between the data points and their corresponding cluster centroids.

Here's a step-by-step overview of the K-means clustering algorithm:

1. Choose the number of clusters (K) you want to create.

2. Initialize K centroids randomly or by selecting K data points as centroids.

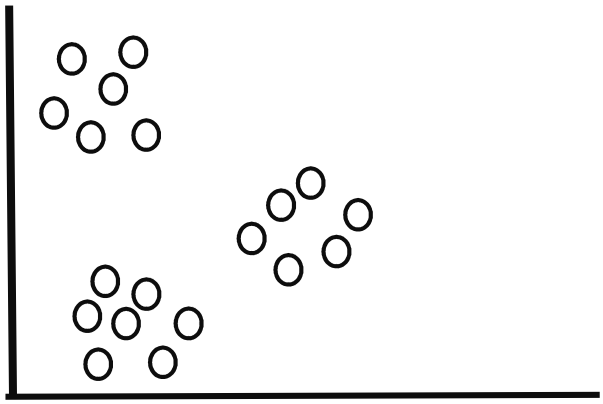
3. Assign each data point to the nearest centroid based on the Euclidean distance or other distance metrics.

4. Recalculate the centroids of each cluster by taking the mean of all the data points assigned to that cluster.

5. Repeat steps 3 and 4 until convergence or until a maximum number of iterations is reached.

6. Once convergence is reached, the algorithm has successfully formed K clusters. Each data point belongs to the cluster whose centroid it is closest to.

K-means clustering has several applications, such as customer segmentation, image compression, document clustering, and anomaly detection. However, it is worth noting that the algorithm's effectiveness can depend on the nature of the data and the appropriate choice of K.



Elbow curve

The elbow curve, also known as the "elbow method" or "scree plot," is a visual technique used to determine the optimal number of clusters (K) in a K-means clustering algorithm. The name "elbow curve" comes from the shape of the plot, which often resembles an elbow or knee.

Here's how you can create an elbow curve:

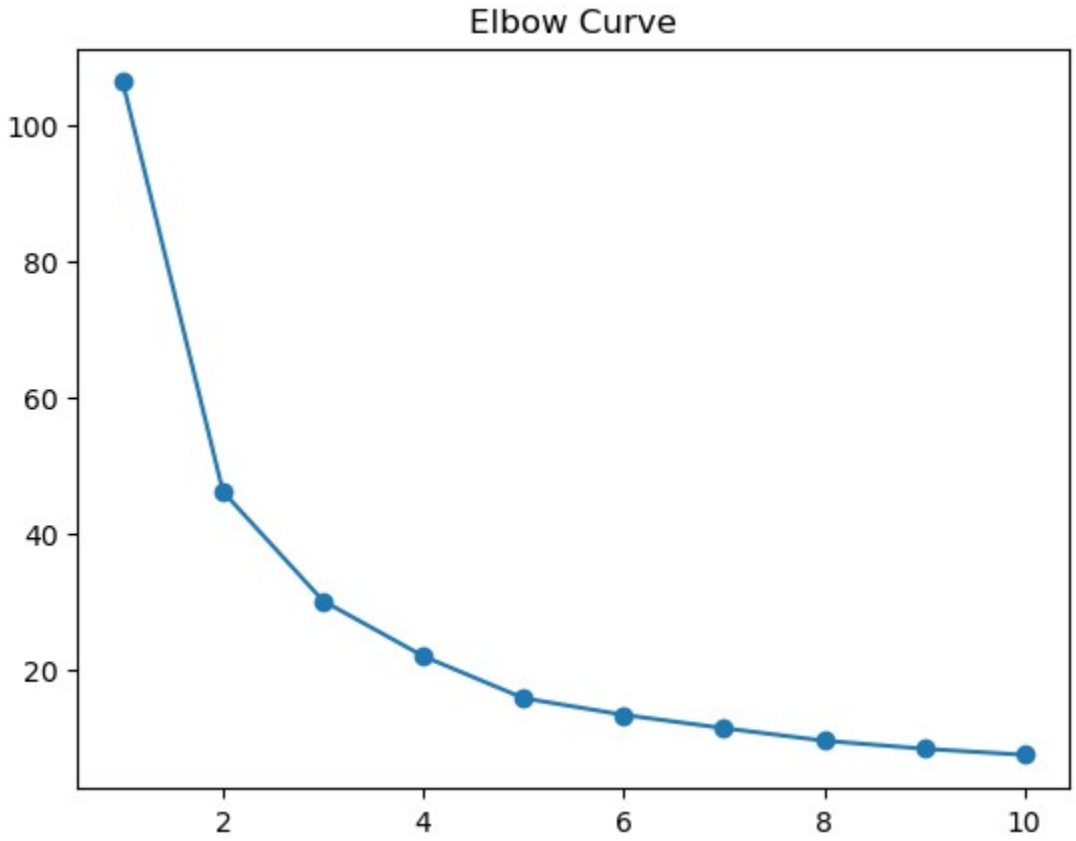
1. Run the K-means algorithm on your dataset for a range of values of K, typically starting from 1 and increasing incrementally. For each value of K, calculate the sum of squared distances (SSE) between the data points and their assigned centroids.

2. Plot the SSE values against the corresponding K values on a line graph.

3. Examine the plot and look for the "elbow point," which is the point of inflection in the curve. The elbow point is where the SSE decreases significantly at a decreasing rate. It indicates the value of K at which adding more clusters does not provide a significant improvement in the clustering performance.

4. Choose the value of K at the elbow point as the optimal number of clusters for your dataset.

The intuition behind the elbow curve is that as you increase the number of clusters, the SSE tends to decrease. However, beyond a certain point, the rate of decrease slows down, resulting in a less significant improvement in clustering quality. The elbow point represents the trade-off between the SSE and the number of clusters.



Total within cluster Sum of Squares (WCSS)

$$\sum_{i=}^{k_i} \sum_{i=1}^{X_i} (X_i - \overline{X_k})^2$$

Silhouette method

The silhouette method is a popular technique for evaluating the quality of clustering results, including K-means clustering. It provides a measure of how well each data point fits into its assigned cluster and can help determine the optimal number of clusters.

The silhouette method calculates a silhouette coefficient for each data point, which ranges from -1 to 1. The coefficient quantifies the cohesion within the cluster (how close the data point is to other points in its cluster) and the separation between clusters (how far the data

point is from points in other clusters). The higher the silhouette coefficient, the better the clustering result.

Here's how you can use the silhouette method to evaluate clustering results:

1. Run the K-means algorithm on your dataset for a range of values of K.

2. For each value of K, calculate the silhouette coefficient for each data point. The formula for the silhouette coefficient of a data point "i" is as follows:

$$\textbf{silhouette\_coefficient(i)} = \frac{(\boldsymbol{b_i} - \boldsymbol{a_i})}{\max(\boldsymbol{a_i}, \boldsymbol{b_i})}$$

- "a(i)" represents the average distance between data point "i" and other points within the same cluster.

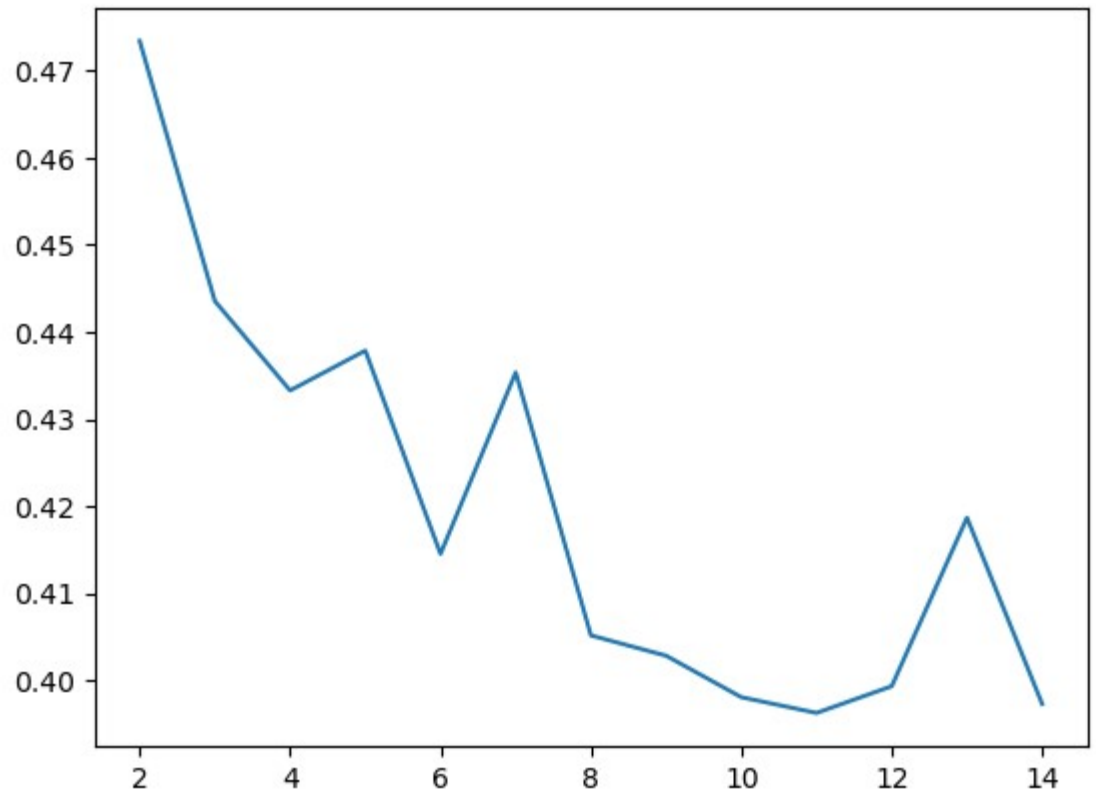
- "b(i)" represents the average distance between data point "i" and points in the nearest neighboring cluster.

3. Calculate the average silhouette coefficient across all data points for each value of K.

4. Plot the average silhouette coefficients against the corresponding K values on a line graph.

5. Examine the plot and look for the highest average silhouette coefficient. This indicates the number of clusters that provides the best separation and cohesion among the data points.

The silhouette method helps in identifying the optimal number of clusters by selecting the value of K that maximizes the average silhouette coefficient. A high average silhouette coefficient suggests that the clustering result is well-defined and distinct, while a low value indicates that the clusters are overlapping or poorly separated.


0.47
0.46
0.45
0.44
0.43
0.42
0.41
0.40
2
4
6
8
10
12
14