K-Means Clustering

Key Points

- **Definition**: K-Means is an unsupervised machine learning algorithm used to partition a dataset into K distinct, non-overlapping subsets (clusters). Each cluster is defined by its centroid, which is the mean of all points in the cluster.
- **Purpose**: To group similar data points together based on feature similarity, making it easier to analyze and interpret the data.
- **Assumptions**:
  - The number of clusters (K) is predefined and must be specified before running the algorithm.
  - Data points are assigned to the cluster whose centroid is nearest to the point based on distance (usually Euclidean distance).
- **Evaluation Metrics**:
  - **Within-Cluster Sum of Squares (WCSS)**: Measures the total distance between each point and its cluster centroid. Lower WCSS indicates better clustering.
  - **Silhouette Score**: Measures how similar a data point is to its own cluster compared to other clusters. Ranges from -1 (poor) to +1 (excellent).
  - **Elbow Method**: Used to determine the optimal number of clusters by plotting WCSS against the number of clusters and looking for an "elbow" point.

Algorithm

1. **Initialize Centroids**:
   - Randomly select K data points from the dataset as the initial centroids of the clusters.
2. **Assign Clusters**:
   - For each data point, assign it to the nearest centroid based on the distance metric (usually Euclidean distance).
3. **Update Centroids**:
   - Recalculate the centroids of each cluster by computing the mean of all data points assigned to that cluster.
4. **Repeat**:
   - Repeat the assign-and-update steps until the centroids no longer change significantly or a maximum number of iterations is reached. This indicates convergence.
5. **Final Clustering**:
   - Once converged, the algorithm outputs K clusters with their respective centroids.
6. **Evaluate**:
   - Assess the quality of the clustering using evaluation metrics like WCSS, Silhouette Score, or the Elbow Method to determine the optimal number of clusters.