12 Clustering & K-Means

Clustering

- **Clustering** is a key technique in unsupervised learning, where the goal is to group a set of objects (data points) in such a way that objects in the same group (called a **cluster**) are more similar to each other than to those in other groups
- Unlike supervised learning, clustering doesn't rely on labeled data. Instead, it tries to find structure or patterns in the data by analyzing the inherent similarities or differences among data points

Key Concepts in Clustering

1. **Similarity/Dissimilarity Measure**
    - Clustering algorithms often rely on a measure of similarity or distance between data points, such as Euclidean distance, Manhattan distance, or cosine similarity
    - The closer the data points are in the feature space, the more likely they are to belong to the same cluster
2. **Centroids**
    - In some clustering algorithms like K-means, each cluster is represented by its centroid (the average of all points in the cluster)
    - The centroid acts as a representative point for the cluster
3. **Number of Clusters (K)**
    - Some algorithms, like K-means, require the user to specify the number of clusters in advance
    - Determining the right number of clusters can be challenging and often requires methods like the elbow method or silhouette analysis
4. **Cluster Assignments**
    - After clustering, each data point is assigned to a cluster
    - These assignments help to group similar data points together for further analysis or decision-making

Applications of Clustering

- **Market Segmentation** : Identifying different customer segments for targeted marketing
- **Image Segmentation** : Grouping pixels in an image to identify objects or regions
- **Anomaly Detection** : Identifying outliers in data, such as fraudulent transactions
- **Document Clustering** : Grouping similar documents together for topics or themes

K-Means

- **K-Means** is one of the most popular and widely used clustering algorithms in unsupervised learning
- The goal of K-Means is to partition a dataset into **K clusters**, where each data point belongs to the cluster with the nearest mean, also known as the cluster centroid
- The algorithm works by iteratively refining these cluster centroids to minimize the overall variance within each cluster

K-Means Steps

1. **Initialize the Centroids**
    - Choose the number of clusters, K
    - Randomly initialize K centroids. These centroids are the initial cluster centers
2. **Assign Data Points to the Nearest Centroid**
    - For each data point in the dataset, calculate the distance (usually Euclidean distance) to each of the K centroids
    - Assign the data point to the cluster whose centroid is closest to it
3. **Update Centroids**
    - Once all data points are assigned to clusters, calculate the new centroids by taking the mean of all data points in each cluster
    - These new centroids are the updated cluster centers
4. **Repeat**
    - Repeat the assignment and update steps until the centroids no longer change significantly or until a maximum number of iterations is reached
    - This means the algorithm has converged, and the clusters are stable
5. **Final Clusters**
    - The algorithm outputs the final clusters, with each data point assigned to a specific cluster

Choosing the Right Number of Clusters (K)

- **Elbow Method**
    - Plot the inertia (sum of squared distances) against different values of K
    - The point at which the decrease in inertia slows down (forming an "elbow") is often considered the optimal K
- **Silhouette Score**
    - Measures how similar a point is to its own cluster compared to other clusters
    - A higher silhouette score indicates well-defined clusters

Limitations of K-Means

- **Need to Specify K** : The number of clusters must be chosen beforehand, which may not always be obvious
- **Sensitivity to Initialization** : The final clusters can vary based on the initial random choice of centroids. Multiple runs with different initializations can help mitigate this
- **Assumption of Spherical Clusters** : K-Means assumes that clusters are spherical and of similar size. It might struggle with clusters of different shapes and densities
- **Outliers** : K-Means can be sensitive to outliers, as they can significantly affect the position of centroids