K-means Clustering: Theory and Case Study

August 30, 2024

Introduction to K-means Clustering

- K-means clustering is an unsupervised learning algorithm used to partition data into K clusters.
- The algorithm aims to minimize the within-cluster sum of squares (WCSS) or variance.
- Common applications include customer segmentation, market research, image compression, and pattern recognition.

K-means Clustering Algorithm

Algorithm Steps:

1. Initialize K cluster centroids randomly.
2. Assign each data point to the nearest cluster centroid.
3. Update the centroids by calculating the mean of all data points in each cluster.
4. Repeat steps 2 and 3 until convergence (i.e., centroids do not change significantly).

Objective Function

The objective of K-means is to minimize the within-cluster sum of squares (WCSS):

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu_k\|^2 \tag{1}$$

where:

- K: Number of clusters
- C_k: Set of points assigned to cluster k
- x_i: Data point
- μ_k: Centroid of cluster k
- $\|x_i - \mu_k\|^2$: Squared Euclidean distance between data point x_i and centroid μ_k

Cluster Assignment and Centroid Update

Cluster Assignment:

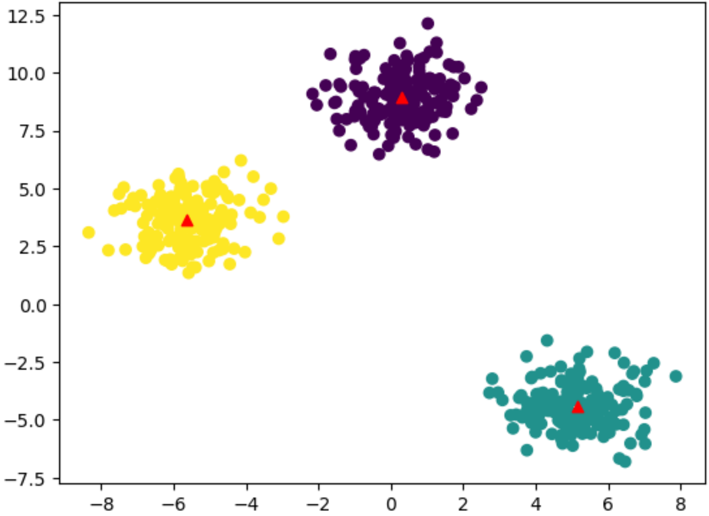
$$C_k = \{x_i : \|x_i - \mu_k\|^2 \leq \|x_i - \mu_j\|^2 \ \forall j,\ 1 \leq j \leq K\} \tag{2}$$

Centroid Update:

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i \tag{3}$$

- Reassign each data point to the cluster with the closest centroid.
- Recalculate the centroids based on the current cluster assignments.

K-means Illustration



Case Study: Customer Segmentation

Use Case: Segmenting customers based on purchasing behavior.
Dataset: Customer purchase history with features such as age, annual income, and spending score.
Objective: Group customers into clusters to identify distinct segments for targeted marketing strategies.
Steps:

1. Select the number of clusters K using the elbow method.
2. Apply K-means clustering to partition customers into K clusters.
3. Analyze the characteristics of each cluster to understand different customer segments.

Choosing the Number of Clusters

Elbow Method:

- Plot the within-cluster sum of squares (WCSS) against the number of clusters K.
- The "elbow" point on the graph indicates the optimal number of clusters where adding more clusters yields diminishing returns.

Silhouette Score:

- Measures the quality of clustering by calculating the mean silhouette coefficient over all samples.
- Values close to $+1$ indicate that the sample is far from the neighboring clusters, whereas values close to 0 indicate that the sample is on or very close to the decision boundary between two neighboring clusters.

Evaluation of Clustering Results

- **Inertia (WCSS):** Measures the sum of squared distances between each point and its assigned centroid.
- **Silhouette Score:** Measures how similar a point is to its cluster compared to others.
- **Dunn Index:** Ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance.

Interpreting Results:

- Clusters should be compact, well-separated, and interpretable.
- Validate results with domain knowledge or external benchmarks when available.

Conclusion

- K-means clustering is an effective technique for partitioning data into distinct groups based on feature similarity.
- Selecting the appropriate number of clusters is crucial for meaningful segmentation.
- Case study demonstrated the application of K-means in customer segmentation for targeted marketing.
- Future work can explore advanced clustering techniques, such as hierarchical clustering or DBSCAN, for more complex datasets.