Clustering

David Robinson

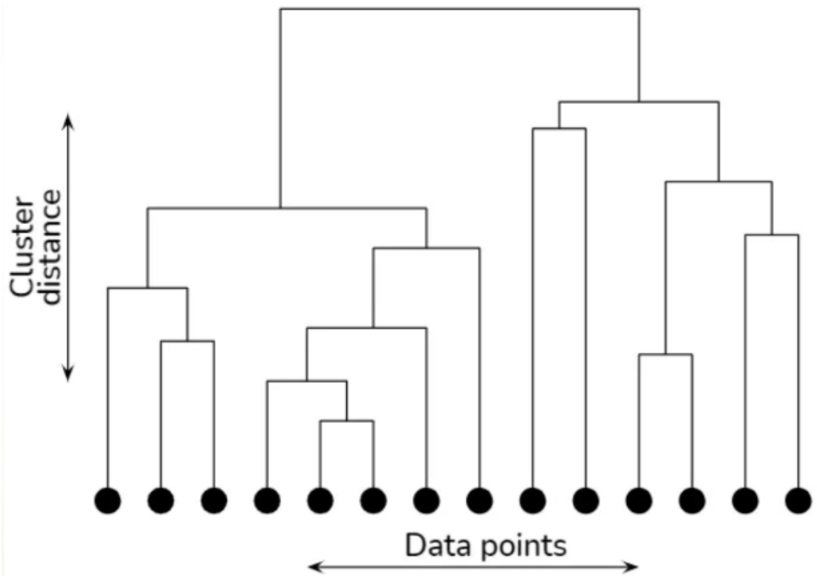
Clustering is an unsupervised learning technique used to group similar data points together based on specific criteria. The objective is to maximize similarity within a cluster and minimize similarity between clusters.

Types of Clustering

- **Hierarchical Clustering**: Builds a hierarchy of clusters.
- **K-Means Clustering**: Partitions data into a predefined number of clusters.
- **Density-based Clustering (DBSCAN)**: Clusters points based on density and handles outliers well.

Hierarchical Clustering

Hierarchical algorithms create a hierarchical decomposition of objects based on similarity. The hierarchical decomposition is represented with a **dendrogram**, which is a tree-like diagram where height measures how similar the data points are.



Bottom-Up (Agglomerative) Approach

1. Each data point is first treated as its own cluster.
2. At each step, the closest clusters are merged based on a distance metric.
3. The process continues until all data points are merged into a single cluster.

Top-Down (Divisive) Approach

1. The dataset starts a single cluster.
2. At each step, K-Means is applied to split clusters.
3. The process continues until each data point is in its own cluster.

K-Means Clustering

K-Means clustering minimizes the Euclidean distance between points and their respective cluster centroids. The cluster quality is measured with a sum of the distances from each point to the cluster centroid.

Algorithm

1. Choose the number of clusters K.
2. Initialize K cluster centroids.
3. Assign each point to the nearest centroid.
4. Update the centroids by averaging the points in each cluster.
5. Repeat steps 3–4 until the centroids do not change or the amount of change falls below a threshold.

Strengths

- $O(tKn)$ Time Complexity: K-Means has a time complexity of $O(tKn)$ where n is the number of data points, K is the number of clusters, and t is the number of iterations. K and t are much smaller than n so the runtime is relatively fast.

Limitations

- Relies on dataset mean: K-Means is only applicable when the mean of the data points can be calculated.
- Requires pre-determined K: The number of clusters must be specified at the beginning, which may not be easy to determine.
- Sensitive to noise and outliers: K-Means struggles with noisy data and outliers as it significantly affects the cluster centroids.
- Not suitable for non-convex clusters: K-Means assumes clusters are spherical.