Lab 4: Clustering

August 2023

1 In This Lab

1.1 Topics to Cover

- Different types of clustering technique

1.2 Requirements

- Python installed on your computer.
- Basic Python programming, Jupyter Notebook.
- Libraries: scikit-learn, numpy, matplotlib, seaborn

2 Clustering

Clustering is an unsupervised machine learning technique that involves grouping similar data points together into clusters based on their inherent characteristics. It is used to discover patterns, similarities, and structures within data without any predefined labels.

2.1 Key Concepts

- **Cluster:** A cluster is a collection of data points that are more similar to each other than to those in other clusters.
- **Distance Metric:** Clustering often relies on a distance metric (e.g., Euclidean distance) to measure the similarity or dissimilarity between data points.
- **Centroid:** In centroid-based clustering, each cluster is represented by a central point, known as the centroid, which is the mean or median of the data points in the cluster.
- **Hierarchical Clustering:** This approach builds a hierarchy of clusters by iteratively merging or splitting clusters based on similarity measures.

- **Density-Based Clustering:** Density-based methods identify clusters as regions of high data point density separated by areas of lower density.
- **Partitioning Clustering:** Partitioning methods aim to divide the dataset into non-overlapping clusters where each data point belongs to exactly one cluster.
- **K-Means Clustering:** K-means is a popular partitioning-based clustering algorithm that aims to minimize the sum of squared distances between data points and their cluster centroids.
- **DBSCAN:** Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based algorithm that groups data points based on their density and identifies outliers as noise.
- **Evaluation:** The quality of clustering can be evaluated using metrics like silhouette score, Davies-Bouldin index, or visual inspection.

2.2 Applications

- Data Analysis: Clustering helps discover underlying patterns and structures in datasets for exploratory analysis.
- Customer Segmentation: Businesses use clustering to group customers with similar behaviors or preferences for targeted marketing.
- Image Segmentation: Clustering is used to segment images into regions with similar characteristics, aiding object recognition and computer vision tasks.
- Anomaly Detection: Unusual data points can be identified as anomalies by considering them as separate clusters or as distant points from clusters.
- Document Clustering: Text documents can be clustered to group similar content for information retrieval and topic modeling.

2.3 Task 1: sample clustering using kmeans

1. Follow kmeans.ipynb for data visualization and use of different clustering techniques

2.4 Report

1. Use mall customer dataset (provided)
2. Apply Kmeans, DBSCAN, and Agglomerative Clustering techniques (try using a different number of cluster centers) to find different clusters and cluster centers. Agglomerative is a type of hierarchical clustering.
3. Consider age and spending, and income and spending for 2D clustering.

4. Consider age, income, and spending for 3D clustering.