DS605 Fundamentals of Machine Learning
Academic Year 2025-26 (Autumn)

LAB - 6

Date: September 26, 2025

1. Objective

The objective of this lab is to gain practical experience with unsupervised learning by implementing and evaluating three different clustering algorithms: K-Means, Agglomerative Hierarchical Clustering, and DBSCAN. By the end of this assignment, you will be able to:

- Preprocess data for clustering analysis.
- Determine the optimal number of clusters for a given dataset using methods like the Elbow Method and Dendrograms.
- Implement K-Means, Hierarchical, and DBSCAN clustering algorithms using Python's scikit-learn library.
- Visualize and interpret the results of each clustering method.
- Compare and contrast the performance and characteristics of different clustering algorithms.

3. Dataset

For this assignment, we will use the **"Mall Customers"** dataset. This dataset contains information about customers of a mall, including their gender, age, annual income, and a calculated "spending score."
The dataset has the following columns:

- CustomerID: A unique identifier for each customer.
- Gender: The customer's gender.
- Age: The customer's age.
- Annual Income (k$): The customer's annual income in thousands of dollars.
- Spending Score (1-100): A score assigned by the mall based on customer behavior and spending habits.

You can download the dataset from this Kaggle link or use a direct CSV link in your code. For this lab, we will focus on clustering customers based on their **Annual Income** and **Spending Score**.

4. Tasks & Procedure

Follow the steps below to complete the assignment. Be sure to comment on your code and provide explanations for your findings in your final notebook.

Part 1: Data Exploration and Preprocessing

1. **Load the Data:** Load the Mall_Customers.csv file into a pandas DataFrame.
2. **Explore the Dataset:**
    - Display the first few rows using .head().
    - Get a summary of the data, including data types and non-null values, using .info().
    - Generate descriptive statistics using .describe().
3. **Data Selection:** For this lab, we are only interested in the Annual Income (k$) and Spending Score (1-100) columns. Create a new DataFrame or a NumPy array containing only these two features.
4. **Initial Visualization:** Create a scatter plot of Annual Income vs. Spending Score to visualize the distribution of the data points. This will give you an initial intuition about the potential clusters.

Part 2: K-Means Clustering

1. **Finding the Optimal Number of Clusters (k):**
    - Use the **Elbow Method** to determine the optimal value of 'k'.
    - Iterate through a range of k values (e.g., 1 to 10).
    - For each k, fit a KMeans model and calculate the Within-Cluster Sum of Squares (WCSS), which is the inertia_ attribute of the fitted model.
    - Plot the WCSS values against the number of clusters (k). The "elbow" point on the graph indicates the optimal k.
2. **Applying K-Means:**
    - Based on the Elbow Method, choose the optimal k.
    - Initialize and fit the KMeans model to your data with the chosen number of clusters.
    - Get the cluster labels for each data point from the labels_ attribute.
3. **Visualize K-Means Results:**
    - Create a scatter plot of the data points, coloring each point according to its assigned cluster label.
    - Plot the centroids of the clusters on the same graph to clearly mark the center of each group.
    - Add a title, axis labels, and a legend to your plot.

Part 3: Agglomerative Hierarchical Clustering

1. **Creating a Dendrogram:**
    - Use the scipy.cluster.hierarchy library to generate a dendrogram for your data. A ward linkage method is recommended.
    - Visualize the dendrogram. The y-axis represents the distance between clusters. The optimal number of clusters can be determined by finding the tallest vertical line that doesn't cross any extended horizontal lines and counting the number of vertical lines it intersects.
2. **Applying Hierarchical Clustering:**
    - Based on your analysis of the dendrogram, choose the optimal number of clusters.
    - Initialize and fit an AgglomerativeClustering model from scikit-learn with your chosen number of clusters and ward affinity.
    - Get the cluster labels for each data point.
3. **Visualize Hierarchical Clustering Results:**
    - Create a scatter plot similar to the one for K-Means, coloring each data point according to its assigned cluster from the hierarchical model.

Part 4: DBSCAN Clustering

1. **Applying DBSCAN:**
    - DBSCAN does not require you to specify the number of clusters beforehand. Instead, it requires two parameters: eps (the maximum distance between two samples for one to be considered as in the neighborhood of the other) and min_samples (the number of samples in a neighborhood for a point to be considered as a core point).
    - Experiment with different values for eps and min_samples. A good starting point for this dataset could be eps=5 and min_samples=5.
    - Initialize and fit the DBSCAN model to your data.
    - Get the cluster labels. Note that DBSCAN will label noise points as -1.
2. **Visualize DBSCAN Results:**
    - Create a scatter plot of the data, coloring points by their cluster label.
    - Make sure to handle the noise points (label -1). It is common practice to plot them in a distinct color, like black, to differentiate them from the actual clusters.

5. Analysis and Questions

After implementing all three algorithms, answer the following questions

1. **Optimal Clusters:** What was the optimal number of clusters you identified for K-Means and Hierarchical Clustering? Justify your choices using the Elbow Method plot and the Dendrogram.
2. **Cluster Comparison:** Visually compare the results of the three algorithms. Did they produce similar clusters? Describe any notable differences.
3. **DBSCAN Performance:** How did DBSCAN perform on this dataset? Did it identify any noise points? How did its results compare to the other methods which force every point into a cluster?
4. **Algorithm Suitability:** Based on your results, which algorithm do you think was most suitable for this specific dataset and why? Consider the shape and density of the clusters.
5. **Real-World Application:** Describe a hypothetical real-world business scenario where the customer segments you identified could be used by the mall's marketing team. For example, how would you target the group with high income but low spending score?

6. Submission Guidelines

- Submit a single Jupyter Notebook (.ipynb) file containing all your code, visualizations, and answers to the analysis questions.
- Ensure your code is clean, well-commented, and runs without errors.
- Use markdown cells in your notebook to structure your work and provide explanations for each part of the assignment.