German International University of Applied Sciences
Informatics and Computer Science
Dr. Caroline Sabty
TA. Merna Said
TA. Sandra Samuel
TA. Amany Hussein
TA. May Magdy

Introduction to Data Science, Spring 2025
Assignment 3
Due date is May 15, 2025 at 11:59 PM

Assignment Team Composition: Each team should consist of three members.

Project Overview

In this project, we aim to explore clustering techniques using the **Mall Customers dataset** , a dataset containing information about customers such as age, gender, income, and spending score. This dataset provides valuable insights into customer behavior in a retail environment.

Our goal is to apply multiple clustering techniques to segment customers based on their demographic and spending characteristics. We will use PCA for visualization and the elbow method to identify the optimal number of clusters. We will also evaluate the clustering performance using Silhouette Score and Davies-Bouldin Index.

dataset link: https://www.kaggle.com/datasets/shwetabh123/mall-customers

Steps to Follow

a) **Data Exploration and Pre-Processing:**

1. Load the Mall Customers dataset from Kaggle. Explore its structure and understand its features (e.g., Age, Annual Income, Spending Score). Perform initial analysis to assess data quality and preprocessing needs.
2. **Feature Scaling:** Standardize or normalize the features (e.g., income and spending score) to prevent scale differences from biasing the model.

b) **Dimensionality Reduction and Data Visualization:** Use PCA to reduce the dataset to 2 dimensions for visualization purposes.

c) **Determining the Optimal Number of Clusters:** Use the elbow method to determine the optimal number of clusters for K-means. Plot the inertia for a range of k values and identify the elbow point.

d) **Model Training with K-Means:** Train a K-means clustering model on the preprocessed dataset using the optimal number of clusters obtained in step 3.

e) **Additional Clustering Techniques:**

1. **Agglomerative Hierarchical Clustering:** Apply using the same number of clusters as determined by the elbow method.
2. **Gaussian Mixture Model (GMM):** Apply with the same number of clusters and compare soft clustering results.
3. **BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies):** Apply BIRCH with the same number of clusters and compare its results to the previous methods.

f) **Evaluation and Comparison:**

- Evaluate each clustering method using:
  - **Silhouette Score**
  - **Davies-Bouldin Index**
- Discuss the differences between clustering results and how each algorithm segments the data.
- Identify which algorithm is more suitable for this dataset and explain why.

  (i.e., (c))

g) **Visualization:** Visualize the clusters using scatter plots for each method, where each cluster is represented by a different color. Use PCA-reduced data and clearly highlight the cluster assignments.

h) **Interpretation:** Interpret your findings.

Expected Outcome

By the end of the project, you should be able to:

- Preprocess data for clustering using standardization or normalization techniques.
- Visualize high-dimensional data in 2D using **Principal Component Analysis (PCA)**.
- Determine the optimal number of clusters using the **elbow method**.
- Understand and apply various clustering algorithms: **K-means, Hierarchical, GMM, and BIRCH**.
- Evaluate clustering models using **Silhouette Score** and **DaviesâBouldin Index**.
- Interpret and compare the clustering results using quantitative metrics and qualitative analysis.
- Develop customer profiles from different clustering techniques and understand the segmentation strategy.

Deliverables:

a) Your code needs to be submitted on the Google form (make sure the code runs and no errors in it).

b) Google form link: https://forms.gle/o9QRVX2DycY4xTnG8

c) Make sure the code runs without any errors.

d) Avoid writing all your code in a single cell; organize it logically into multiple cells.

e) Each team should submit only one file with the names and IDs of the other team members (file format .ipynb).

PLAGIARISM IS NOT TOLERATED AND COPIED WORK WILL BE AWARDED 0 POINTS FOR BOTH PERSONS INVOLVED!

There will be individual evaluation.