

Data Analytics 1

Assignment 5

Release: November 2nd, 2024
Deadline: November 14th, 2024

Introduction

This assignment is on clustering and data visualization. The goal is to gain experience with common clustering techniques and data preprocessing/visualization tools. There are three parts in this assignment; the first part involves data visualization and preprocessing using PowerBI, the second part requires the implementation of K-Means from scratch, and the third part requires different clustering algorithms. In the third part, `sklearn` can be used for implementations, including for K-Means.

Part 1: Data Visualization and Preprocessing Using PowerBI (20 marks)

Use Dataset1/ for this part.

In this part, you are required to use PowerBI to perform data visualization and preprocessing on the provided dataset. This dataset encompasses data for approximately 2,240 customers. Each row within the dataset includes details about an individual customer, including their personal information, purchasing behavior, and interactions with marketing campaigns.

- **Data Exploration and Visualization:**

- Import the dataset into PowerBI.
- Create various visualizations to explore the dataset, such as:
 - * Bar charts to show the distribution of categorical variables (e.g., Education levels, Marital Status).
 - * Histograms to show the distribution of numerical variables (e.g., Age, Income).
 - * Scatter plots to identify relationships between variables (e.g., Income vs. Total Spending).
 - * Heatmaps or correlation matrices to identify correlations between variables.
- Provide insights based on the visualizations. For example, identify any trends, patterns, or anomalies in the data.
- Handle the basics of data preprocessing using the tool.

Part 2: K-Means Clustering Implementation (35 marks)

Use the preprocessed dataset from Part 1 for this part.

This part requires you to implement the K-Means clustering algorithm from scratch.

- Implement the K-Means clustering algorithm without using any built-in clustering functions.
- Choose $k = 2, 5, 7, 9$.
- Use the elbow method and silhouette score to find the optimal number of clusters (from 2, 5, 7, and 9).
- Analyze the clusters formed and interpret the results in the context of customer segmentation.

Part 3: Clustering with Different Algorithms (35 marks)

Use datasets in Dataset2/ folder for this part.

- For the given 4 datasets in Part 3, use the following clustering algorithms:
 - K-Means
 - Agglomerative Clustering (you can try any linkage strategies: single, complete, etc.)
 - DBSCAN

Perform the following tasks:

1. Apply each clustering algorithm to each dataset and visualize the clusters using a scatter plot (different color for each cluster). Since there are only 2 features, the data points can be plotted directly.
 2. Analyze which algorithm performs the best for each dataset and mention the reason if an algorithm fails for a particular dataset.
 3. Tabulate silhouette scores for each dataset and each algorithm.
 - Feel free to visualize the data before applying the algorithms and set the parameters accordingly (or you can use the elbow method).
 - For DBSCAN, try `eps` values in the range of 1 to 5 and adjust other parameters (`min_samples` from 1 to 50) based on data visualization.
 4. Try using another metric apart from silhouette score (e.g., sum of all intra-cluster distances) and analyze if there is any difference in the results.
 5. Ignore the noise points predicted by DBSCAN while computing the silhouette score (mention the number of noise points in your table).
- Plot the visualizations in your notebook and write the reasoning/inference. Tabulate the scores in the notebook markdown.

Report and Analysis (10 marks)

Write a comprehensive report detailing the results and conclusions derived from all parts of the assignment. Your report should include the various plots and insights as well as reflection on the use of PowerBI for data visualization and preprocessing. You should take into account both insights from visualization and from clustering to reach conclusion.

Submission Format

Submit a zip file named `teamId_assignment5.zip` consisting of the following files:

- `Part1.pbix` (PowerBI file)
- `Part2.ipynb` (Jupyter Notebook for Part 2)
- `Part3.ipynb` (Jupyter Notebook for Part 3)
- Report document (in PDF or Word format)
- Any intermediate files or datasets used