STA4026 Analytics Assignment

Unsupervised Learning

Overview

Clustering is about finding natural groupings in data. In practice, cluster analysis is an iterative process, involving the exploration of multiple strategies, because real-world data typically does not behave in an idealised manner. Such a process involves numerous decisions that can, especially when aggregated, have substantial material effects on the results. However, there is no one-size-fits-all approach, and a data scientist often relies heavily on intuition and experience to guide the workflow.

In this assignment, you will examine a synthetic dataset guided by a simplified workflow. The purpose of this assignment is two-fold:

1. Understand the Workflow: You will experience different phases of a cluster analysis workflow, and the types of decisions that need to be made at each stage.
2. Algorithm Comparison: By applying both k-means and k-medoids clustering algorithms, you will explore their relative strengths and weaknesses and understand how they perform under different data conditions.

As you will see, some aspects of this assignment are intentionally open-ended, allowing you to explore different pathways and make your own decisions based on the data. This is designed to mimic the uncertainty and decision-making processes involved in actual data science projects. This assignment isn't about finding "the right answer" but rather about exploring different strategies, making informed decisions, and learning to justify these decisions logically and coherently. Accordingly, you are expected to perform tasks, analyse the results, and justify decisions in light of these results. Your success will be measured by the clarity of your rationale and the depth of your analysis.

Marking

The total mark will be weighted as follows:

- EDA 25%
- Hyper-parameter Tuning 25%
- Cluster Analysis 25%
- Presentation 25%

Each of the three phases consists of primary and secondary tasks, denoted **P** and **S**, respectively. This is to provide an indication of the minimum level of analysis expected from you (primary tasks) and how you could deepen your analysis (secondary tasks). Primary tasks are required to pass, whilst secondary tasks are optional to improve your mark. You are not restricted to the secondary tasks indicated and may choose to do additional secondary tasks that you feel are relevant.

Nearly all tasks have a coding element and a description element. For these tasks, marks will be equally awarded for results and description. However, results produced without any description of its relevance will not be awarded any marks. The marks awarded for the code and results will depend on correctness, whereas the marks for interpretation and decisions will be based on quality and depth of analysis. Each secondary analysis will be awarded marks (i.e., more secondary tasks mean more marks). It is possible to earn more than the total marks for each section, but the maximum mark for the assignment is capped.

The write-up should be in the style of a report, rather than a list of numbered tasks, and should include a very brief introduction, the main body of work, and the conclusion. Marks for presentation include the quality of writing, overall clarity and coherence, and the quality of figures, layout, and structure. All figures must have standalone captions. The length is limited to 8 pages (excluding appendix), and going beyond this will be penalised. Keep things concise - say what you mean and mean what you say. While using ChatGPT is allowed, copying and pasting text from ChatGPT is strictly prohibited.

Notes, Instructions, Etc.

It is recommended to work in pairs. You will benefit a lot from the discussion. The submission instructions are the same as Assignment 2. **Submission instructions:**
For submission, I'd like you to give me

- A pdf containing your write-up and R code in an appendix. The write-up has an 8-page limit (excluding appendix).
  Use the naming convention `STDNUM001_STDNUM002_Analytics_2024_A3.pdf` for your file. Note the underscores.

- Your R code. Use the naming convention `STDNUM001_STDNUM002_Analytics_2024_A3.R` for your code file. Your R code should NOT contain any of the following:

  ```
  install.packages()
  rm()
  setwd()
  ```

  I want to be able to run your code on my computer without having to manually edit your code, install libraries or call external files.

ANY deviation from the above conventions WILL be penalised.

Problem Set

1. Exploratory Data Analysis (25)

   (a) (**P**) Data description: Assess and report the size of the dataset **after cleaning**, the data types, the amount of missing data, and the quartiles for each variable. Indicate if and how any data cleaning was done. (2)

   (b) (**P**) Distance metric: Based on the data description, select a distance metric and justify why it is appropriate. (2)

   (c) (**P**) Exploratory Analysis: Produce a pair plot. Comment on the shape of the univariate distributions, pointing out aspects that you think are relevant to clustering. Describe the bivariate distribution of data, highlighting aspects that you think are relevant to clustering. (Hint: clustering is about finding areas of high density. Also, how do you think different distribution shapes affect the mean?) (8)

   (d) (**P**) Exploratory Analysis of Distances: Calculate the distance matrix, and plot the distribution of pairwise distances. Comment on the shape of the distribution, pointing out aspects that you think are relevant to clustering. (Hint: how do distances affect clustering?) (8)

   (e) (**S**) Outlier Identification: Identify 10 observations that you think are outliers and could unduly influence your clustering algorithms. Describe your rationale for selecting these observations. (Hint: No need to be overly complicated about this. Keep it simple.) (4)

   (f) (**S**) Correlation Analysis: Do you think any adjustment to the data is needed to account for correlations? Produce a correlation matrix and comment on the correlations. (4)

   (g) (**S**) Scaling: Will you standardise your data? Justify your answer based on the data description and with reference to how each variable influences the total distance between observations. (4)

2. Hyper-parameter Tuning: (25)

   (a) (**P**) Selecting K: Produce an average silhouette plot for $K = 2, \ldots, 20$. Describe the plot and use it to select two possible alternatives for the number of clusters K for both k-means and k-medoids. (10)

   (b) (**P**) Initialisation Sensitivity: Analyze the sensitivity of the selected number of clusters to different initializations for both clustering algorithms. Use multiple runs to assess stability. Produce an appropriate visualisation or set of statistics to present the sensitivity analysis. Comment on the sensitivity of the results to the initialization, and choose two optimal configurations for each result. (10)

   (c) (**S**) Increasing Initialisations: Produce results for a large number of initialisations ($100s$) by parallelising the execution. (4)

   (d) (**S**) Selecting K using Gap: For robustness, produce the gap statistic plot for different K. Compare the results against the average silhouette plots. (4)

3. Cluster Analysis: (25)

(a) (**P**) Silhouette Score Analysis: From the hyper-parameter tuning results above, select two configurations for k-means, and two for k-medoids. Plot the silhouette scores and the clusters. With reference to the plots, describe and compare the cluster assignment and the cluster quality of the different configurations and the different algorithms. Identify what you think the best cluster assignment is. (12)

(b) (**S**) Outlier Analysis: Locate the previously identified outliers on the cluster and silhouette plots. Comment on whether they were or weren't as influential as you previously believed, and briefly comment on why you think this is so. Referring to the plots, identify any new points that you think are outliers. (6)

(c) (**S**) Post-Processing: For observations with negative scores, identify which clusters they may be closer to. See whether manual reassignment of these points improves the cluster quality. (5)

(d) (**P**) Conclusion: Based on your analysis, describe the overall success of the clustering algorithms. Provide some idea where each algorithm was successful, and each algorithm failed. Describe how you think you could improve the workflow, if time allowed. (6)