DAT-330 Data Mining

Assignment 2: Cluster Analysis

Due: 11/07/2024 Points: 100

Instructions

Read all instructions in this section thoroughly. You will lose 50% on the assignment if the instructions are not followed.

Collaboration: You must complete this assignment individually; you cannot collaborate with anyone else. You may discuss the homework to understand the problems and the mathematics behind it. Still, you are not allowed to share problem solutions or your code with any other students.

Citing Your Sources: Any sources of help that you consult while completing this assignment (other students, textbooks, websites, etc.) *MUST* be noted in your write-up. This includes anyone you briefly discussed the homework with. If you received help from the following sources, you do not need to cite it: course instructor, course lecture notes, course textbooks, or other course materials.

Formatting and Submission: This assignment consists of two parts: programming and write-up. Both parts must follow the same order.

1. Programming:
   a. Submission: <your_name>_PA02.ipnyb
   b. Add each question that you are solving in the ***Markdown*** cell on your Juypter notebook.
   c. Use descriptive identifiers.
   d. Comment your code.
   e. All plots must be complete.
   f. Your code must be optimized. (If you are repeating any section of the code, it must be in a loop or make it a function)
2. Write-up:
   a. Submission: <your_name>_PA02.pdf
   b. Add each question to your write-up before answering them.
   c. Each answer must reflect and be correlated with the code.
   d. Use Times New Roman font, a font size of 14 for the questions, and a font size of 12 for the answers. Justify the document.
   e. All the plots and tables must be complete, be of the same size, and be centered with a figure/table number and a figure/table name.
   f. Screen shots are not allowed.
   g. Cite the sources (if you are using them) for every question.

Reading: This assignment is on the application of Cluster analysis which includes:

a. K-means
b. Agglomerative
c. DBSCAN

Supporting materials:

a. In class lecture notes.
b. In class coding session (WB_Clustering_10_17.ipynb).
c. Sample codes for plotting bar graph and line graph (Plots.ipynb).

Question 1:

Dataset: *Iris_PA02.csv*

Part A (15 Points)

Task: Identify two attributes from the dataset that are most suitable for performing cluster analysis.

Hint: use `seaborn.pairplot()`

Explanation:

> **Process:** Describe the criteria and process used to evaluate and select these attributes for clustering. Justify why these two attributes are expected to yield meaningful cluster separation.
>
> **Results:** Name the chosen attributes and discuss any preliminary insights gained from exploring these attributes using visualization.

Part B (20 Points)

Task: Perform cluster analysis using the k-Means, Agglomerative, and DBSCAN clustering methods on the dataset.

Explanation:

> **Process**: Explain how each clustering method was applied, and specify the chosen parameters (e.g., number of clusters for k-Means, linkage criteria for Agglomerative, epsilon and minimum points for DBSCAN). Summarize and organize the parameters used for each clustering method. Include any rationale behind parameter choices.
>
> **Results**: Report and visualize the clusters produced by each method with their respective parameters and label each cluster appropriately. Discuss any observable differences in cluster shapes, sizes, or distributions.

Part C (15 Points)

Task: Identify which clustering method and parameter combination produced the best cluster separation.

Explanation:

> **Process**: Evaluate each clustering result and detail the criteria used to judge cluster quality (e.g., compactness, separation).
>
> **Results**: Describe the clustering method and parameters that yielded the most distinct and well-separated clusters, explaining why it was deemed the best approach.

Part D (10 Points)

Task: Evaluate if any clustering method failed to produce meaningful clusters and explain why.

Explanation:

> **Process**: Analyze each clustering result and identify any methods where clusters were poorly defined or overlapping. Discuss any limitations or assumptions of the method that could have led to these results.
>
> **Results**: Identify any method(s) that did not perform well and explain the potential causes, including dataset characteristics or method-specific limitations.

Question 2: (40 Points)

Dataset: refer PA_02.ipynb

Prove that the DBSCAN clustering method works best on the above data over k-Means and Agglomerative clustering methods.