Exercise – 5: DS203-2024-S1

Submissions due by: September 20, 2024, 3 pm

This exercise is aimed at:

- Getting introduced to and running various **Clustering** algorithms on a given data set and understanding their relative characteristics, performance, and advantages.
- Calculating, effectively documenting, and understanding various Clustering metrics and developing an approach towards effectively using them.
- Getting introduced to the relevant functions of the Python library: **sklearn**

In this Exercise you will be processing the following datasets:

- Clusters-5-v0.csv
- Clusters-5-v1.csv
- Clusters-5-v2.csv

Processing steps:

1. Enumerate and explain three measures (metrics) that you will use to assess the outcome of clustering algorithms.
2. Review the datasets using appropriate plots.
3. Use the following algorithms to cluster the observations is each of the dataset. Plot the outcomes.
    a. K-Means clustering
    b. Agglomerative clustering
    c. DBSCAN
4. Calculate and analyze the metrics you have listed in '1' above. Analyze them across the datasets and methods. Create appropriate plots and explain the trends, if any, and the outcomes of your analysis. What can you conclude (Please don't state the obvious !!)
5. List your major learnings from this exercise.
6. Your submissions should include the following:
    a. A PDF document with all the above analyses and comments. Ensure that you include the required figures and Tables (ie. metrics data) in your report, along with the explanations and analysis.
    b. Your Python source file (.py file). Please DO NOT upload Jupyter Notebooks – they get bulky! All important information (tables / plots / etc.) should be presented in the report.
    c. Name of the PDF should be **E5-your-roll-number.pdf** and the name of the Python source file should be **E5-your-roll-number.py**
    d. Upload the PDF and the source file to the assignment submission point E5.

oooOOOooo