IE5374 Fall 2022 Project 1: Clustering

The objective of Project 1 is to a) implement different clustering methods to synthetic and real-world data and b) validate using external and internal validation techniques

Task 1

Datasets posted with the project (“Data1.csv”, “Data2.csv”, “Data3.csv”, “Data4.csv”, “Data5.csv”) contain the data points and their respective class information. For each of the 5 datasets, follow the steps below:

1. Use K-means and hierarchical clustering methods to generate clusters.
2. Evaluate the performance of the clustering algorithm using external validation metrics.
3. Plot (2D or 3D) the data points for each dataset and color them according to the original class
4. Plot (2D or 3D) the data points for each dataset and color them according to the class allocated by each clustering algorithm

Task 2

The world indicators dataset compares different countries based on selected attributes. Do the following tasks using the “WorldIndicators.csv” dataset posted with the project:

1. Use K-means and hierarchical clustering methods to group similar countries together
2. Use internal validation metrics to report the cluster quality
3. Report the best clustering solution. Give a detailed list of all the groups and the countries included within the groups
4. Generate 3 different scatter plots of your choice and color the data points according to the group. (Example: “Life expectancy vs GDP”, “Infant Mortality vs GDP”, etc.)

Submission Format

1. Submit all solutions as either an iPython notebook (.ipynb) showing all solutions OR as a PDF
2. Include text to explain the solution (don’t just show the numbers, tell what you did!)
3. Include equations for the evaluation metrics in your submission file
4. Use modules in Python that provide functions for different clustering methods and cluster validation
5. Students can also create their own custom functions if necessary
6. This is a group effort! All members of the group should actively participate and contribute to the final solution
7. Only 1 member from each group needs to submit the solution; be sure that all group member names are listed on your submission file
8. **Submit the solution by November 21 at 12 PM PT.**