Data Analytics 1

Assignment 5

Clustering

Release : 6th November 2023
Deadline : 15th November 2023 (11:55 pm)

The objective of the assignment is to gain practical experience in implementing and evaluating different clustering algorithms

Note : Only KMeans needed to be implemented from scratch for PART A , you can use library functions for all the algorithms in PART B(including KMeans)

Dataset 1:
This dataset encompasses data for approximately 18,000 football players. Each row within the dataset includes details about an individual player, including their personal information, current club affiliation, financial data, and football-related characteristics like Shot Power, Stamina, Reflexes, and so on. When a player is operating as a goalkeeper, attributes are prefixed with GK (e.g., GKReflexes). Generally, when performing clustering, it is essential to take into account these football-specific attributes. Nonetheless, you also have the option to consider personal information for clustering purposes, such as comparing attributes like height and jumping ability.

Part A (Datasets to be used is in Dataset1/ folder) (50 marks):

- Implement k-means clustering algorithm from scratch.
- Choose k = 3, 5, 7. You can use both numerical and categorical attributes. For categorical attributes you may need to convert them into numerical before clustering
- Use elbow method and silhouette score for finding optimal number of clusters (from 3 ,5 and 7)

Part B (Datasets to be used are in Dataset2/ folder) (50 Marks):

- Now for the given 4 datasets in Part B use KMeans , Aggloromative clustering (You can try any linkage strategies single,complete etc) , DBSCAN and do the following tasks
    - Apply the clustering algorithm for each of the data and visualize the clusters using a scatter plot (Different color for each cluster) (there are only 2 features so the datapoi)

    - Analyze which algorithm performs the best for each and every data and also mention the reason if an algorithm fails for a particular data
    - Tabulate silhouette scores for each data , for each algorithm
        - Feel free to visualize the data before applying the algorithm and set the parameters accordingly for the algorithm (or ideally you can use elbow method for that)
        - For DBSCAN you can try eps values(if you are using sklearn) in the range of 1 to 5 and check for other parameters similarly (min_smaples From 1 to 50) by visualizing the data.
    - You can try using any other metric apart from silhouette score (eg: Sum of of all intra cluster distances) and analyze if there is any difference in results
    - You can ignore the noise points predicted by DBSCAN while computing silhouette score (mention the number of noise points in table)
- Plot the visualizations in notebook and write the reasoning/inference and tabulate the scores in notebook markdown .

Submission details:

- Submit a zip file <teamId_assignment5.ipynb> consisting of two files:
    - PartA.ipynb
    - PartB.ipynb
    - Any other implementation files (if required)