Assignment 5

Data Analytics - I

Deadline: 11 Nov. 2022

Dataset

This dataset contains features of around 18000 football players. Each row contains the following information related to one player: Personal information, current Club information, financial information, football-specific attributes such as Shot Power, Stamina, Reflexes etc. A prefix of GK (eg. GKReflexes) represents the attributes when the player is playing as a goalkeeper. Typically you have to consider these football-specific attributes while clustering. However, personal information can also be used (eg. clustering on height vs jumping ability etc.)

You are expected to complete the following tasks. Please attach a clear report along with analysis (pdf). Please feel free to use library functions, except for question 1.1. Finally, don't forget about data preprocessing/cleaning !

Total: 100 points

1. K-means (45)

1. Implement k-means clustering algorithm from *scratch*. (25)
2. 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. (5)
3. Use elbow method and Silhouette Score to get the optimal number of clusters. (5)
4. Analyze the results in every case and try to mark each cluster.(A few helpful pointers for analysis have been provided below. Feel free to add your own insights as well.) (10)

2. Hierarchical Clustering (40)

1. Cluster the data using the following Agglomerative (bottom-up strategy) distance measures. (30)
    - Minimum distance (Single-Linkage)
    - Maximum distance (Complete-Linkage)
    - Average distance (Average-Linkage)
    - Mean distance
2. Plot a dendrogram and then analyze the clusters formed. (10)

3. Analysis (15)
Finally, compare the clusters formed by each of the above techniques. Which method is the
best according to you for clustering the given dataset? (If K-means, why? Else if Agglomerative, then also mention the distance measure along with reasoning.) Which clustering technique made the most meaningful clusters?

Analysis of clusters

This is a naive example of how you can approach analyzing clusters. Suppose you got 3 clusters.

- How good are the clusters? Use intra-class similarity and inter-class similarity to measure the goodness of clusters.
- Which attributes are the most similar in a cluster?
- Can these clusters be named according to mean of attributes present in each cluster like Forwards, Midfielders and Defenders. (Note: Some domain-specific knowledge is required for this task. Thus, we will accept both technical terms as well as non-technical cluster names simply based on the parameters on which clustering was done. For example: With high shot power and high finishing, it is likely that these players are strikers/forwards. However, simply mentioning these high/low comparisons is also fine.)
- Are there any outliers? If any, then what do you interpret from those outliers? Which attributes were different in them?

Submission details

- Naming convention:

```
teamName_A5.zip
|
|--teamName_KMeans.ipynb
|--teamName_Agglomerative.ipynb
|--teamName_Report.pdf
```

- Only one team member needs to submit.
- Any kind of plagiarism will be severely punished.