Machine Learning 2022 - Assignement 1 - Unsupervised Learning

Francisco P. Romero - University of Castilla La Mancha 17/10/2022

The overall objective is to use unsupervised learning techniques to make a preliminary exploration of the data and to extract conclusions from relevant variables, discarded elements, etc.

The specific objectives are as follows:

1. Identification of outliers elements (teams) in the dataset
2. Use clustering algorithms to identify groups and characterize them.
3. (optional) Feature Selection using clustering algorithms.

Deliverable

Student submissions must include the .ipynb file(s) generated by Google Colab.

The notebook must include the name of the members of the team.

Only one member of the team must upload the file to moodle.

Deadline: 2nd Nov

Tasks

The following steps could be executed in different order.

Note: It is recommendable to select a subset of the existing features previously to carry out the unsupervised learning tasks.

Dimensionality Reduction

1. Extract the correlation among features and obtain conclusions.
2. Execute PCA and plot the results. Some conclusions are welcomed.

Outlier Identification

1. Find outliers in your data. DBSCAN uses distance and a minimum number of points per cluster to classify a point as an outlier.
2. Analyze why these elements are outliers and decide whether or not to consider them for further analysis.

Clustering by K-means

Apply K-means algorithm to your data.

1. Don't forget to normalize your dataset (excluding the primary keys, of course) or use the PCA data.
2. Specify the chosen number of clusters (k) and a brief explanation about why you have chosen this value of k.
3. Execute k-means, test with different options of initialization (random, k-means++),
4. Try to assign a label to each group. Try to interpret the meaning of each cluster through its centroid.
5. The graphical result of your clustering (only one chart - PCA results- with elements represented by in different colours) would be welcomed.

Hierarchical Clustering Algorithm

- Compute the similarity matrix. Execute the hierarchical clustering algorithm. Test several cluster-distances-measures and choose the best solution in your opinion.
- Cut the dendrogram and characterize the obtained groups. Try to assign a label to each group. (Don't forget to read the feature descriptions in the competition page)
- Your best dendrogram and a brief explanation of your choices must be included in your report.
- The graphical result of your clustering (only one chart - PCA results- with elements represented by in different colours) would be welcomed.