Data Mining Project: Part 2
Regression & Clustering tasks

Regression : Decision Trees + Random Forest

Regression is a form of data analysis that extracts models describing relationships between variables to deduce a given output. These models enable the estimation or prediction of new data based on previously observed examples. Regression has numerous applications, particularly in finance (for stock price forecasting), marketing (for estimating consumer demand), and social sciences (for analyzing relationships between demographic and behavioral variables) …etc. Several methods are available, each with its advantages and disadvantages depending on the data type, ranging from simple techniques like linear regression and k-NN to advanced techniques like decision trees, random forests, and neural networks for more complex relationships.

Clustering : CLARANS + DBSCAN

Clustering is a data partitioning method that groups objects into clusters, where objects within the same cluster are similar to each other but different from those in other clusters. Performed automatically by algorithms, it allows for the discovery of previously unknown groups within the data. Cluster analysis is widely used in various fields, such as business intelligence, image recognition, web search, biology, and security. Numerous algorithms have been developed to effectively explore large datasets and identify relevant groupings.

Thus, in this second part of the project, you are required to implement, test, and compare regression algorithms and clustering algorithms on the **project dataset** obtained **after preprocessing (you may use different preprocessings for the 2 distincts tasks)**. You are then asked to:

Application of Regression Algorithms:

1. Consider the "Near-surface specific humidity" as the output to predict.
2. Split the dataset into training and test data.
3. Program the two classification algorithms, "**Decision Trees**" and "**Random Forest**".
4. Apply both algorithms to the dataset and test different parameters.
5. Evaluate and compare the two regression models by calculating the adequate metrics as well as the **average execution time**.
6. Compare the two algorithms with the pre-defined DT and RF of the library of your choice.

Application of Clustering Algorithms:

7. Program the clustering algorithms "**CLARANS**" and "**DBSCAN**".
8. Evaluate, compare, and analyze the results of CLARANS with various parameters.
9. Evaluate, compare, and analyze the results of DBSCAN with various parameters.
10. Compare the two algorithms, CLARANS and DBSCAN, using appropriate metrics.

HOUACINE Naila Aziza : n.a.houacine@gmail.com

11. Illustrate with graphs and interpret the results.

Advanced UI Options:

- Select the Data Mining method to execute.
- Enter a new data instance and deduce the output of the regression.
- Display the PCA result of a clustering.

Deadline: Sunday, January 05, 2025.

Have fun !