Advanced Machine Learning

Project

The goal of this project is to process a dataset using machine learning methods and to make predictions on the corresponding dataset.

The project should contains at least the following 4 parts:

1. Analysis of the dataset
2. Machine Learning and Prediction
3. Visualisation of the prediction results
4. Theoretical formalism

Analysis of the dataset:

In order to analyse the dataset, you have to extract some statistical information from the given dataset, for example: the type of data, the missing values, outliers, the correlation between variables, etc. This part should contain also the analysis of the domain application and explain the goal of the prediction.

Prediction:

1. Use the Regression Line to predict values for some variables (you can choose any variable as a target. Compute the error.
2. Use a machine learning method in order to predict the class of a new set of objects. You can use the methods as K-Nearest Neighbours (K-NN), Support Vector Machine (SVM), Decision trees, Neural Networks … The obtained results should be validated using some external indexes as Prediction Error or others. Explain how the hyper-parameters are fixed and use Cros Valdiation.

The obtained results should be analysed in the report and provide **a solution to improve** the results, for example by combining multiple methods.

Visualization:

You can visualize the knowledge extracted from the classification (prediction) in order to present the results i.e. scatter plots using predicted colours, decision trees,…

Theoretical details:

Give the algorithmically (mathematical) formalism of the method which give the best results. Explain all the parameters of the used method and their impact on the results.

Some comparison should me made to conclude the project.

The report should contain all these 4 parts, and you should explain the motivation of the used methods, and discuss the obtained results.

Please note that you can use AWS, Python, R and/or Matlab and/or Knime for the experimentations.

Each group (2 students) will use one of the 10 datasets from the following web link:

http://lipn.univ-paris13.fr/~grozavu/PredA/dataProject/

Each Folder contains a file .data containing the dataset, and a file .names containing the information about the corresponding datasets. The datasets were extracted from the UCI Repository.

You can also use a different dataset from the UCI website or from other platform.