ADA 2020-2021: Lab Assignments

1st ASSIGNMENT (2,5 points)

The students will use the two datasets on Portuguese wine quality in https://archive.ics.uci.edu/ml/datasets/wine+quality. These datasets are related to red and white vinho verde wine samples from the North of Portugal. The goal is to model wine quality and alcohol content based on physicochemical tests.

Variables:

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
12 - quality (score between 0 and 10)

Pre-processing & Descriptive Analysis. Data should be prepared (cleaning the data) for ulterior analysis as well as descriptive statistics should be extracted (i) to provide basic information about variables in the dataset and (ii) to identify relationships between variables and select them. For that descriptive process, students should use tabular and/or graphical layouts to provide measures about the centrality and variability of the data and measures of relationship between variables.

Regression. Regression techniques should be applied to predict a) the alcohol content (in this case the quality column will be taken out of the dataset) b) the quality score. Students are expected to identify those variables with highest influence and infer the relationship between these dependent variables and the required data. Finally, the quality of the model should be assessed.

2nd ASSIGNMENT (3,5 points)

Classification. Students will design SVM and logistic regression models for quality classification. After obtaining the models, students are expected to evaluate them according to diverse quality metrics. The design goal is, for a similar maximum achievable accuracy and precision, minimize the number of features and the number of support vectors.

As classification labels, for each dataset, we will consider three scenarios:

1) Dividing wines into three classes, low (0-4), medium (5-7) and high (8-10) qualities.
2) Low alcohol content (below average) versus high alcohol content (above average) (in this case the quality column will be taken out of the dataset)
3) Determining labelling ranges that yield high accuracies (NON COMPULSORY)

Clustering. Students are expected to apply hierarchical and k-means clustering algorithms to the datasets. For the later, the value of k will be guessed by applying heuristics. Students are expected to infer knowledge from the resulting clusters and provide measures of their quality.

SUBMISSION AND DEADLINES

Assessment will be individual during class time. For deadlines see course planning.

Each student mustupload a single compressed file with the following structure: (i) R code; (ii) data files (if any intermediate dataset has created); and (iii) a technical report (Markdown[1] or pdf). The technical report should include free text, graphics and R code in a way meaningful enough to make the analysis understandable and reproducible. More precisely, it should be structured in the following sections:

- 1st assignment: (i) Cleaning Data; (ii) Exploring Data; (iii) Prediction: Applying Linear Regression; (iv) Evaluation/Discussion
- 2nd assignment: (i) SVM Model; (ii) Clustering Models; (iv) Evaluation/Discussion.

[1] We recommend R Markdown as the authoring format for the technical report (R Markdown can be embedded in RStudio) or Jupiter Notebook