MEIC 2018/19

Ciência de Dados

Data Science Project

Project Goal

In this project the goal is that students apply their knowledge about data science techniques, for discovering information in two distinct problems (datasets).

Students are asked to create models about data, understanding and relating those models.

Additionally, students should also criticize the results achieved, and discuss the difficulties faced on mining the different datasets.

Methodology

Information discovery on both datasets has to be done using unsupervised techniques (association rules and clustering), and training classification models, including decision trees (algorithms ID3, C4.5 or CART), naïve Bayes, kNN algorithm and random forests.

Students may choose the mining tool to apply, between **R** and **python** using **scikit-learn**.

Data

The data for the two problems are available as *.cvs* files in the course webpage:

- **digital colposcopies.** Source data and information in: https://archive.ics.uci.edu/ml/datasets/Quality+Assessment+of+Digital+Colposcopies
- **APS failure at Scania trucks.** Source and aim in: https://archive.ics.uci.edu/ml/datasets/APS+Failure+at+Scania+Trucks

Non-supervised Learning

Non-supervised exploration has to be done through clustering and association rules. In both cases, *class attributes can't be used to explore the data*, unless there is a well substantiated interest of mining class-conditional data. Nevertheless, class attributes may be used to assess clustering results through a comparative evaluation. Besides this, *statistical evaluation has also to be performed*, using studied indexes.

Classification

Supervised exploration has to be done by training classification models through *kNN, Naïve Bayes, Decision Trees* and *Random Forests*. For this purpose, the use of *class attributes is mandatory*. Evaluation of the obtained models should be done as usual, through accuracy measures and evaluation charts, as studied in the classes.

Others

a) Description

The students should perform a statistical analysis of the datasets in advance and summarize relevant implications in the report, such as the underlying distributions and hypothesized forms of noise or feature dependency.

b) Preprocessing

In accordance with the properties of the input dataset and the behavior of the target learning algorithm, the students are allowed to apply preprocessing techniques when needed or under a solid conjecture of its potential impact on learning.

c) Advanced learning possibilities

In accordance with all the collected results, the students are encouraged to suggest alternative data mining algorithms, adaptations to the applied algorithms, ensembles, or/and purpose methodologies (pipeline of preprocessing, mining and postprocessing techniques) for the robust analysis of datasets with similar characteristics. Every suggestion should be clearly motivated/justified.

Deadlines

Students should register their groups and deliver the project until the **November 13th** via **Fénix**. The report should follow the template given, containing a cover, an optional index and 10 pages including any appendixes. Each additional page won't be considered.

The report should describe in a succinct form the pre-processing performed, the parameters used, and the results found, their interpretation and critical analysis for each problem and technique used. Additionally, it should include a comparison of the results achieved in both problems, and the relation among the information discovered through the different techniques. The report may be written in Portuguese or English.

Excellence

A project that applies the suggested data mining techniques over the given datasets and provides a clear and *sound analysis of the collected results is not necessarily an excelling project*.

Excelling projects have four major characteristics. First, they show an acute understanding of the data characteristics and their impact on the learning. Excelling projects formulate hypothesis behind differences in performance. Second, they have precise and succinct language: no redundancies, unnecessary or subjective statements. Third, excelling projects are often a result of a creative thinking on ways of improving the learning. Illustrating, the fundamented use of a specific preprocessing technique (whether the inclusion of new features, normalization, discretization, space transformations, feature removal, or handlers of missings, outliers or specific forms of noise) can make a difference. Fourth, robust assessments go beyond simple performance indicators. Excelling projects draw (parameter-varying) plots, test hypotheses, and establish ratios to understand less-trivial performance views such as robustness to noise, domain adequacy or overfitting propensity.

Plagiarism

Plagiarism is an act of fraud. We will apply state-of-the-art software to detect plagiarism. Students involved in projects with evidence of plagiarism will be reported to the IST scientific council in accordance with Técnico rules.

Evaluation Criteria

The project will be evaluated as a *whole*. Nevertheless, we provide below a decomposition of the total project score for the purpose of guidance and prioritization. Reference scores:

1. Succinct statistical description (**4%**)
2. **Non-supervised mining**
    a. Preprocessing (**8%**)
    b. Association Rules (**8%**)
    c. Clustering (**8%**)
    d. Comparison and critical analysis (**16%**)
3. **Classification**
    a. Preprocessing (**7%**)
    b. Instance-based Learning (**5%**)
    c. Naïve Bayes (**5%**)
    d. Decision Trees (**7%**)
    e. Random Forests (**7%**)
    f. Comparison and critical analysis (**15%**)
4. Advanced learning possibilities (**10%**)