iscte

INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Introduction to Machine Learning — 2021/2022

Final Project

> This project should be solved using Python notebooks (Jupyter) due to the ability to generate a report integrated with the code. It is assumed you are proficient with programming. All answers must be justified and the results discussed and compared to the appropriate baselines. In addition to the technical report integrated with the code, a report documenting the application of the CRISP-DM methodology should also be submitted.
> Max score of the project is 4 points. The work should be done in groups (two students) or individually. In the case of groups with two members, the report should indicate an estimate of each member's contribution to the work. For example: manuel: 60%, pedro: 40%, together with a short justification. *It is mandatory to make an oral presentation and discussion of the project.*
> **Deadline:** January 5th, 2022

The objective of this project is to apply the CRISP-DM methodology to solve a wine quality classification problem, using Machine Learning methods.

Using two datasets consisting on physicochemical data from red and white *Vinho Verde* wine samples, from the north of Portugal, and a quality classification, our client wants to build an application to automatically classify new samples of *Vinho Verde*.

The project and the report should follow the phases of the CRISP-DM methodology.

To experiment with the different machine learning models, the scikit-learn[1] toolkit [Pedregosa et al., 2011] should be used.

Dataset

The two considered datasets correspond to samples of red and white variants of the Portuguese *Vinho Verde* wine. The data consists of physicochemical (inputs) and sensory (the output) variables. For more information, see the work of Cortez et al. [2009].

[1] https://scikit-learn.org/stable/index.html

The red wine dataset has 1599 instances. The white wine dataset has 4898 intsances. Each sample has 11 features and an output classification (as mentioned, based on sensory data) that consists in a score between 0 and 10, corresponding to its quality (higher is better). The 11 features that characterize each sample (based on physicochemical tests) are the following:

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

The dataset is available on the following location:

- `https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/`

Experiments

For all experiments, you should compare the technical performance on both red and white wine datasets.

You should perform the following experiments:

- Use supervised and unsupervised methods (see following sections);
- Randomly remove 10%, 20%, and 30% of the values of the features of each dataset and explore two different strategies to handle missing values;
- Experiment with data normalization, data discretization, and data reduction. Apply these steps to the original, unchanged, dataset.

Don't forget to visually explore your data, namely presenting correlations between pairs of features.

The technical evaluation should include different metrics and means to better understand the errors of the supervised machine learning approaches. The assessment of the unsupervised machine learning approaches should compare the resulting clusters to clusters based on the quality score.

Supervised Learning Algorithms

Experiment with the following supervised learning algorithms and comment the results, based on your knowledge of how they work.

1. Decision Trees;
2. Multi-layer perceptron;
3. k-NN.

Unsupervised learning algorithms

Experiment with the following unsupervised learning algorithms and comment the results, based on your knowledge of how they work.

1. k-Means;
2. DBScan;
3. Agglomerative hierarchical clustering.

References

P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.