Data Science Lab: Process and methods
Politecnico di Torino

Project Assignment
Summer Call, A.Y. 2020/2021

Last update: June 9, 2021

1 Project dates

Start date: June 9, 2021 at 23:59 AM CET
Due date: June 23, 2021 at 23:59 AM CET

Due date is a **strict deadline**.

2 Problem description

This project consists in the prediction of a quality score associated with a beer review. Practically, you are required to build a regression model capable of inferring the quality score assigned to the beer by the reviewer.

Warning: For this competition, you will not be allowed to use external datasets other than that provided for the competition. Adoption of external resources will result in failure of the exam.

2.1 Dataset

The dataset for this project contains beer reviews in tabular format. It counts 100,000 entries, each of which corresponds to a review expressed by a user on a website for beer benchmarks.

Each review is characterized by both numerical and categorical attributes. Each reviewer evaluates the beer on four different categories, namely appearance, aroma, palate and taste. A textual description is provided as well. Additionally, several information on the user is included. The quality score is reported on the feature named `review/overall` and is expressed as a number between 1 and 5. Half scores, such as 1.5, 2.5, etc., are allowed[1]. The dataset is located at:

`https://dbdmg.polito.it/wordpress/wp-content/uploads/2021/06/DSL_project_2021.zip`

Within the archive, you will find the following files:

- **development.tsv** (development set): a tab-separated values file containing the reviews from the development set. This portion does have the `review/overall` feature, which you should use to train and validate your models.
- **evaluation.tsv** (evaluation set): a tab-separated values file containing the reviews from the evaluation set. This portion does not have the `review/overall` feature.
- **sample_submission.csv**: a sample submission file.

[1] Execute `numpy.arange(1, 5.5, .5)` to get the values used by reviewers.

2.2 Task

You are required to build a regression pipeline to assign an overall quality score to each record in the Evaluation set.

2.3 Evaluation metric

Your submissions will be evaluated through r2_score.

3 Submit your result

Submission file To get your results evaluated, you have to upload a result file on our submission competition platform, Data Science Lab Environment (DSLE). The submission file has to be a `.csv` file formatted as follows:

```
Id,Predicted
10,2.5
123,1.5
21,1
345,5
42,3.5
...
```

The submission file must contains a header line and a row for each record in the Evaluation collection. Each row must have two fields:

- the `Id` of the corresponding record in the Evaluation set, as an integer number.

  Info: The `Ids` in the submission file must correspond to the positions of the records in the Evaluation set. **The first record in the Evaluation set has Id=0, the second has Id=1 ans so on.**

- the `Predicted` label for the corresponding record.

You can find a sample submission file on the project material.

Submission platform The submission platform is the same one you used during the course laboratories. Therefore, you have to use your personal key. If you do not have the key or have problems using it, please send an email to `giuseppe.attanasio@polito.it`. Please refer to the guide on the course website, to go through the submission procedure.

You can find the DSLE platform at http://trinidad.polito.it:8888

4 Upload the report and the software

Warning: The report and the software have to be submitted by the due date reported in Section 1. This is a **strict deadline**.

Submission All the required files (i.e. for the report and the software) must be included in a **single *.zip*** file[2]. The archive must be uploaded to the "Portale della Didattica", under the *Homework* section.
Please use as description: **report_exam_june_2021**.

Formatting rules The formatting rules for both the report and the software are described in the exam rules document. You can find it on the course website.

[2] A ZIP archive is a ZIP archive, not a RAR, a 7z or, a tarball archive, nor any of those renamed with a trailing `.zip` extension.