Project

Machine Learning 2021-2022 - UMONS
Tanguy Bosser (tanguy.bosser@umons.ac.be)

1 Task

Given a dataset containing a target variable Y, and a set of predictors X_p, the goal of this project is to identify the most accurate classifier to assign probabilities to each of the 6 categories that the target variable Y can belong to. The competition (with related datasets) is hosted on the following Kaggle website: `https://www.kaggle.com/t/66b91ea27aae41d183076cf68c9bc009`

The training set consists of 13 predictor variables associated to 22,792 observations. The target variable Y is a categorical variable with 6 categories (C1, C2, C3, C4, C5, C6). Amongst the 13 predictors, 8 are categorical variables (X2, X4, X5, X6, X7, X8, X12, X13), and the remaining ones are continuous variables. No more information is given on the predictors. You are not allowed to use any extra datasets or information to build your classifiers.

The initial dataset has been split into training and test sets. The full training set is available to you, but only the predictors are provided for the test set. You can evaluate your predictions by submitting them to the Kaggle website. Note that only a random subset of 60% of the test set is used to compute your *public score*. Your final *private score* using the full test set will be provided at the end of the competition.

The evaluation metric for this competition is the (multi-class) log loss:

$$-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} z_{i,k}\,\log p_{i,k}$$

where:

- n is the number of data points in the training/test set.
- K is the number of classes (here, $K = 6$).
- $z_{i,k}$ is the one-hot representation of y_i, i.e. $z_{i,k} = 1$ if y_i is equal to the k^{th} class, and zero otherwise.
- $p_{i,k}$ is the predicted probability for the i^{th} observation and the k^{th} class.

1. Your first task is to form a team composed of three people.
2. Each team member should create a Kaggle account (using his/her UMONS email address).
3. Form a team on Kaggle.
4. Do some basic exploration of the dataset.
5. Build your first model and upload your predictions to Kaggle. Your predictions must follow a certain format and must be contained in a .csv file (see `https://www.kaggle.com/code/tanguybo/benchmark`).
6. Try, and try again to improve your model. You can make a maximum of five submissions per day.

2 Project Report and Code

The report can be a maximum of **10 pages**, and must abide by the section structure described below.

1. **Section 1: Exploratory Data Analysis** (max 2 pages). In this section, you must investigate the data so as to detect patterns and spot anomalies that might be present amongst the different variables, through the means of graphical representations (scatter plots, boxplots, etc...).
2. **Section 2: Methodology**. This section describes the models/methods you have used, including a justification of your choices. You should also present your model fitting, diagnostics, etc. You should discuss and compare *at least three different classifiers*.
3. **Section 3: Results and Discussion**. In this section, you must discuss and conclude on your classifiers' results, through the means of tables, curves, confusion matrices, etc... You should try to explain at best the performance that you obtained for your classifiers.

Overall, you will be graded based on clarity of writing, quality of presentation, level of machine learning content, and technical communication of main ideas. You should clearly explain what you have done, using figures to supplement your explanation. Your figures must be of proper size with labeled, readable axes. In general, you should take pride in making your report readable and clear.

Your code should be **repeatable**, properly structured and well-commented. Zip your .py file(s) before uploading them to Moodle.

3 Deadlines

- **April 22, 11:59pm**: Submit the names of each member of the team on Kaggle.
- **April 29, 11:59pm**: At least one Kaggle submission needs to have been made.
- **May 13, 11:59pm**: The Kaggle competition closes.
- **May 20, 11:59pm**: Upload to Moodle your project **report** and **code**, one per group.

Do not wait till the last minute. Late submissions will not be allowed.

4 Grading

- Total points : 20
- Accuracy of classifier on Kaggle : 6
- Report and code : 14