Project

Machine Learning 2020-2021
UMONS
Souhaib Ben Taieb

1 Task

With various predictors describing multiple aspects of residential homes, the purpose of this project is to produce the most accurate classifier to assign probabilites to five categories a home could belong to. The competition (with related datasets) is hosted on the following Kaggle website: `https://www.kaggle.com/t/1469160415a04d8b83b6727c8805999a`.

The training set consists of 58 predictors associated to 1,963 residential homes. The output y is a categorical variable with 5 categories (A, B, C, D and E), which are ranked by sales price. 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 containing 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 test set
- K is the number of classes (here, $K = 5$)
- $z_{i,k}$ is the one-hot representation of the i-th observation, i.e. $z_{i,k} = 1$ if y_i is equal to the kth class, and zero, otherwise.
- $p_{i,k}$ is the predicted probability for the i-th observation and the kth 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. Predict the test set, and upload your predictions to Kaggle.
6. Try, and try again to improve your model. You can make a maximum of five submissions per day.

2 Project report

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

1. Section 1: Introduction. The introduction will describe the data set and motivate the problem. It should be brief.
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 classification models.
3. Section 3: Results and Discussion. This includes for example graphs and tables, as well as a discussion of the results.
4. Section 4: Conclusion. This includes summary of the findings.

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.

3 Grading

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

4 Deadlines

- **May 2, 11:59pm**: Submit the names of each member of the team on Moodle.
- **May 9, 11:59pm**: At least one Kaggle submission needs to have been made.
- **May 18, 11:55pm**: The Kaggle competition closes.
- **May. 21, 11:55pm**: Upload to Moodle your project **report** and **code**, one per group.

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