Fundamentals of Machine Learning - 2022

Report 1 - Classification task
02 September 2022 - Due on 13 September 2022

1. Instructions

This is the course first deliverable. You will analyze the problem below and describe in detail the steps taken to solve it. You are expected to deliver:

1. a written report in a Jupyter notebook (or equivalently in pdf, in that case with the complete associated code in a separate file),
2. the trained model in a pickle file,
3. an *env* file with the necessary modules for the code to run smoothly. Your code should execute with no error with the associated *env* file, reproducing your reported results. Additionally, the model should run on a test file.

The report may be in Spanish or English. The logistics regarding how to deliver the report will be explained a few days prior to the deadline. The report is expected to reflect all stages of an end-end machine learning workflow, as seen in class (for instance, in Practice 1). In particular, it is mandatory that you include in your report your model's best prediction error estimate.

2. Problem

You are provided with a dataset which contains observations on 30m by 30m spacial cells containing different kinds of trees. The target variable reflect the most frequent tree class in a given cell. The predictor variables provide different geographic information regarding that cell, such as floor slope, distance to water bodies or possible fire spots, etc.

Your task is to build a machine learning model to predict the vegetation classes as best as you can. You are expected to describe as much as you can the modeling process, use figures liberally to convey information and, most important of all, provide an estimate for the prediction error.

3. Data

The dataset is provided as a zipped csv file. Consider this dataset as your training/validation partition (you may partition it as you like). After the report deadline (i.e. after 13 September) you will be given a **test dataset** to compare with you validation score, which will serve as an external assesment of your model's performance.

All columns are measurements in meters, except columns 2 and 3 which are in degrees and 7, 8 and 9, which are in 256 values (integers in the interval [0, 255]). Columns 11, 12 and 13 are categorical (i.e., integers represent classes).

4. Models allowed

You are allowed to experiment with two specific models: `random forests` and `xgboost`. You can play around with whichever combination of feature engineering, fine tuning, or any modeling step that you see appropriate, as long as it is well documented.

5. Grades

The report will not be graded by the model's final performance, but by how much of the course content you were able to correctly apply (and clearly communicate) in it. Try to reflect as much as possible the stages described until this point (data analysis, feature engineering, correct error estimation, fine tuning, etc.). The more explicit you are about your modeling decisions and findings, the better, as you will give us the means to improve your grade. Good model performance is cool, but will not warrant a good report grade. Good writing, polished figures and correct modeling logic will. Finally, if your code does not run correctly, this will certainly impact your final grade.