Exoplanet Classification

Practical I - Machine Learning

1 Objectives

The goal of this practical is to practice the concepts studied during the module and to acquire experience in the use of some of the main classification methods, model evaluation and interpretation and presentation of experimental results. To achieve this you will use and compare multiple classification methods based on different principles in a binary classification problem regarding exoplanet candidates.

2 Task

In this practical you are tasked with comparing six classification methods: Naive Bayes, Decision Tree, k-Nearest Neighbors, Support Vector Machines, Random Forest and Gradient Tree Boosting. You also have to perform the experiments listed below which are specific to each method. It might be necessary to normalise the data and test different values for the methods' hyperparameters to achieve good results (it is not necessary to submit all tested combinations of hyperparameters, only the one with the best results, except for the ones requested below). The models have to be evaluated using k-fold cross-validation with 5 folds.

- **Naive Bayes**: Only a single experiment to serve as a baseline
- **Decision Tree**: Test multiple values for the maximum tree height (including unlimited height) and present the results using plots
- **SVM**: Test the linear, sigmoid, polynomial and RBF kernels
- **k-NN**: Test multiple values for the number k of neighbours and present the results using plots
- **Random Forest**: Test multiple values for the number of trees and present the results using plots
- **Gradient Tree Boosting**: Test multiple values for the number of boosting rounds and present the results using plots

You do not need to implement the methods listed above. All methods are available in the library *scikit-learn* for the Python programming language. You can also use auxiliary libraries for tasks such as plotting graphs and mathematical operations, as only long as they do not implement the experiments themselves.

For each of the experiments performed you are required to explain what is the goal of the experiment (what is the meaning of the hyperparameter being tested for instance) and also include an interpretation of the results based on the theoretical concepts studied during the module. At the end you are required to present a comparison of the performance of all the methods tested, including the ROC curve and the precision and recall metrics.

3 Dataset

The methods will be tested on the task of binary classification of exoplanet candidates found by NASA's Kepler spacecraft[1]. An exoplanet is a planet outside of the solar system (i.e. does not orbit the sun). The spacecraft first identifies possible signs of exoplanets, referred to as *Kepler Object of Interest* (KOI). Not all KOIs are actual exoplanets however, some are false positives of different natures. The task is then to classify KOIs between confirmed exoplanets and false positives. Each observation corresponds to a KOI and the features are estimates of the physical properties of the (possible) exoplanet (radius, temperature, features of the host star, etc).

The dataset will be ready for use and will be available at Moodle in the file koi_data.csv. The file will be in CSV format separated by commas. The first column identifies the KOI, the second contains the ground-truth classification (FALSE POSITIVE or CONFIRMED) and the remaining columns are KOI features of various sources. For this practical you will not be required to understand the meaning of each feature.

4 Submission

The practical needs to submitted as a Jupyter (Ipython) notebook. The notebook has to include all the code (with appropriate comments) necessary to run all the experiments, present the results using text, plots and tables, the explanations of what is being done and the interpretation of the results. Only the notebook is to be submitted. The clarity and organisation of the notebook will also be evaluated. The teaching assistant has to be able to reproduce all the experiments simply by running all the notebook cells in order.

The notebook has be submitted to Moodle by 22/05 at 23:59 (only the .ipynb file has to be submitted). **Practicals submitted after the deadline will not evaluated**. **There will be no deadline extensions**. **In case of plagiarism, all students involved will be graded zero and reported to the CS department**.

5 Assessment Criteria

- Correct implementation of the use of the methods, experiments and model evaluation (50%)
- Clear, concise and non-ambiguous presentation of the experiments and results. (20%)
- Correctness of the explanations and interpretation of the experiments. (20%)
- Organisation of the information presented in the notebook. (10%)

Note that you will not be evaluated based on the accuracy obtained in the experiments, only the process itself.

[1]Data was retrieved from the NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu/)