Activity Supervised Learning

Subject: Machine Learning - MSc. Computational Biology

Introduction

Computational biology involves the development and application of data-analytics and theoretical methods, mathematical modelling and computational simulation techniques to the study of biological systems. It also includes many aspects of bioinformatics and can be defined as the science of using biological data to develop algorithms or models in order to understand biological systems and its relationships. Within this area it is included the sub-area of computational oncology that focuses on determining the characteristics of tumors, analyzing molecules that are deterministic in causing cancer, and understanding how the human genome relates to the causation of tumors and cancer. The last is the context of this activity where you have to find the best model that relates some genetics characteristics with the good or bad response to a treatment in a cancer scenario.

Description

Colorectal cancer (CRC) is one of the principal causes of death and its early diagnosis and treatment can lead to a full recovery. It is also known that different individual response differently to the treatment due to its genetic information. The data for this activity contains information ('MM','WW','WM') of different SNPs (Single-Nucleotide polymorphism) from different individuals that have been diagnosed with rectal colon cancer and the categorization of them based on its good response (R) or bad response to the treatment (NR). The datasets are included in this activity and numbered from 1 to 33 and you have to select your assigned dataset to do the activity. You can find the list of the assigned datasets in the StudentsDatasets.pdf file.

The dataset contains 53 individuals (rows), 21 features (associated to different SNPs), and 1 target value ('Target'). The values of the data files are already transformed to 0 for 'MM', 1 for 'WW' and 2 for 'WM' values in order to be compatible with all the ML learning models format.

Objectives:

You have to develop a machine learning model that is able to classify if an individual will response well or bad to the treatment based on its SNPs information using the assigned dataset.

TO-DO

- Make a comparative study using the different machine learning classifiers that you have seen in class (Logistic Regression, Decision Trees, KNN, Random Forest, and Multilayer Perceptron) in the subject and obtain the best possible model (adjust the parameters of the models using cross-validation or any other validation method that you have seen as needed).

- Explain the results obtained (incude a table with the results for all the models), what is the best model, and interpret the results including any ideas to improve the results.
- Interpret the learned models for regression and decisión trees.

To submit

- The Jupyter Notebook source code of the study including explanations for the different steps.
- A pdf file including the study that you have done.

Evaluation rubric

The following criteria will be used for evaluating the activity:

- The accuracy of the presented best model (the higher the accuracy the better; note: remember that the best possible model doesn't have to be accountable for a perfect classification 100%.)
- The quality of the source code and explanations of the different steps taken to make the study.
- The inclusion of plots and interpretations that help to understand the presented results.

Deadline

- The deadline to submit this activity is: end of the day (23:59pm), December the 7th.