Statistical & Machine Learning

Individual Project Description

The project is an individual task where the student is required to (1) select and explain a set of machine learning models; and (2) setup a benchmark experiment for those models.

Purpose

The main purpose of this project is to assess the knowledge of each student on two important subjects of statistical and machine learning: (1) the understanding of machine learning mechanism; and (2) the ability to setup a machine learning pipeline.

The student will work on the **assigned data set** and will need to fulfill the following tasks:

- Task 1: Select 5 machine learning predictive algorithms. Explain the mechanism of the algorithms using your own words (e.g. the general idea of the algorithm, the objective function, the algorithm fitting process, pros and cons of the algorithm, etc.).
- Task 2: Setup the benchmark experiment to compare the 5 selected machine learning algorithms (in the previous task). Describe your experimental setup and explain the results. The experimental setup should contain the following elements:
    - Variable selection, dimensional reduction method (Stepwise, F-score, Boruta, etc.)
    - Cross-validation method (holdout, k-fold CV, etc.)
    - Evaluation metric (AUC, Accuracy, etc.)
    - Any other methods

Submission

- Task 1 + Task 2: Written report (*.pdf, 10-15 pages)
- Task 2: Jupyter Notebook R/Python (*.ipynb) + basetable (*.csv)
- Note: Upload the project to GitHub.

Evaluation

The project will be evaluated on:

- Task 1: The correct understanding about the machine learning methods [70%].
- Task 2:
    - The solid setup of the benchmark experiment pipeline (e.g. model building, variable selection, cross-validation, model scoring, etc.) [20%].
    - The quality of the R/Python script [10%].
- Note: Using the work (text or programming scripts) of other people without citing the source is considered as plagiarism, and is strictly prohibited.

Timeline information

- Section 8, 23 March 2022