CS3099 Machine Learning and Data Analysis Group Specification

As interpreted and modified by the CS class of 2019

November 7, 2017

1 General

Every group doing ML is to create a server with a RESTful API (over HTTPS). The server is to receive requests for analysis from the frontend implemented by any Data Visualization group and either retrieve the necessary data from the backend, perform the data training and return the result or, if the model has previously been trained, retrieve the model and use that to generate the result.

2 Analysis Algorithms

Depending on the type of input and request, the analysis techniques should be able to provide discrete results (eg presence of cancer tumor) or continuous results (eg estimated time until death). The following techniques for data analysis should be implemented/used by the ML:

- Naïve Bayes
- Support Vector Machines
- Random Forests
- Cox's regression
- Kaplan-Meier plots

3 Statistical Performance

K-fold cross validation is to be done on the HCI side instead of ML. The plots for the other statistical performance measures, namely:

- Receiver-operating curves
- Specificity-precision curves
- Plots of feature importance

4 API Specifics

The basic idea of the part of the API that involves the ML side is the following: HCI sends a request to train a model on particular attributes of a particular dataset (columns in a CSV file), provides the ML server with the authentication key required to access the dataset, and requests a result to be predicted from a set of variables mathching the sample data.

4.1 File Format

The basic file format that the ML server has to work with is CSV (apart from images). The backend is responsible for converting the data to CSV on request. However, the file format for the data should be specified as part of the request for the data and as an extension the ML server can support other file formats.

4.2 Model Storage

The ML server can store any model it trains on the backend server. If a request is made for a model that has already been computed, it may be reused to save time. The result of the prediction requested by the HCI server may also be stored in the backend server, if it were to request that.

4.3 Plot Generation

Plot generation should be done by both HCI and, upon request, by ML.