CMSC 422 “MACHINE LEARNING”

Fall 2016

PROJECT DESCRIPTION

You are given a training dataset represented as a spreadsheet in XLS format. The training data has just over 2750 rows of data.

- Columns A through BE (including both) of the spreadsheet represent a host of independent variables.
- Column BF represents the dependent variable (0 or 1).

You are required to learn different types of accurate predictors for this project using this training data.

We will test your predictor by giving you test data sets for which column BF is missing (i.e. where you do not know the true class to which rows in the test data belong). You will submit both your code and results in a format that we will specify. The grade you receive for your project will depend upon the accuracy of your predictors and the quality/readability of your code. Submission instructions will be provided shortly.

Task 1 (8 points: due Sep 20 2016) Use a Naïve Bayes Classifier to predict the class of rows in a sample test data set.

Task 2 (8 points: due Sep 30 2016) Use Decision Trees to predict the class of rows in a sample test data set.

Task 3 (8 points: due Oct 18 2016) Use Support Vector Machines to predict the class of rows in a sample test data set.

Task 4 (8 points: due Nov 3 2016) Use Logistic Regression to predict the class of rows in a sample test data set.

Task 5 (8 points: due Nov 17 2016) You are required to submit a PowerPoint presentation consisting of a maximum of 20 slides. This presentation should explain what you learned about the relationship between the attributes (columns) in the data and whether a row should be classified as a 0 or a 1. For this task, you are welcome to use anything else you have learned in class, any other classifiers out there in the literature and/or anything else you choose to do.

All projects will be due by 1159pm on the deadline date. Late submissions will not be accepted without a doctor’s note.

NOTES:

- Student submissions for tasks 1-4 should contain functions that abstract the training and testing phases:
    - **train(data)**, where input data is the xls filename of the train dataset, and output is a classifier C
    - **predict(C, row)**, where input C is a classifier from train(...), input row is an array corresponding to some row of IVs in the xls test dataset (a single row only, to account for variable dataset sizes), output is 0 or 1
- Projects must be coded in Python using the SciKit Learn libraries ((downloadable from http://scikit-learn.org/stable/install.html). If you want to code your project in another language or using another library, please secure the TA's permission before doing so.
- Code must be well-documented.

More detailed code and output submission information will be provided.