COSC74/174: Machine Learning and Statistical Data Analysis

Homework 1 (due: 10AM on Thursday, 17 Jan 2019)

Instructor: Prof. V.S. Subrahmanian (vs@dartmouth.edu)

You are given a training dataset in CSV format. The training data has 5,600 rows:

- Columns 1 through 6 of the given CSV file represent independent variables
- The last column ("Label") represents the dependent variable (0 or 1)

You are required to learn an accurate Naïve Bayes predictor for this project using this training data. You may use any form of Naïve Bayes and you may choose to add or drop features. We will test your predictor by giving you a test set with 2,400 rows for which the last column is blank (i.e. you do not know the true class to which rows in the test data belong). You will submit both your results and code. The grade you receive for your project will depend upon the accuracy of your predictors relative to the highest accuracy of the class, and the quality/readability of your code. Please include the following in your submissions:

1. A new CSV file of the test set with an added column, "Label", showing the dependent variable (0 or 1) that you predicted
2. Your code (e.g. Jupyter Notebook or python script).

TASK: Use a Naïve Bayes classifier to predict the class of rows in the test set. You may use any flavor of Naïve Bayes and you may add or drop features. You are encouraged to use cross validation to find the best method.

NOTES:

- Your code should contain functions that abstract the training and prediction phases:
    - ***train(data)***, where input data is the CSV filename of the training dataset, and output is a classifier *C*
    - ***predict(C, row)***, where input *C* is a classifier object on *train(...)* was called, input row is an array corresponding to some row of IVs in the CSV 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 library (downloadable from http://scikit-learn.org/stable/install.html). You are responsible for making sure that your project is properly submitted and your code can be properly run.
- Please be sure to submit all parts of your homework in the required format.
- All work must be your own. Academic Honor Principle applies to all parts of the project. Please refer to http://student-affairs.dartmouth.edu/policy/academic-honor-principle for more detail.
- All projects will be due by the deadline posted on Canvas. If your submission is up to 1 day late, you will only get 80% of your original score (e.g. if your scored 18/20, you will get 0.8*18=14.4). If your submission is up to 2 days late, you will get 60% of the points you scored on that part. If your submission is 2 or more days late, your homework will not be accepted.
- Your grade will be the accuracy of your classifier on the test set divided by the max accuracy obtained by any one in the class pool (consisting of all students in the class plus the TAs).