EECS 738: Machine Learning Final Project

Dropout Prediction on a MOOC Learning Platform

Background:

A massive open online course (MOOC) is an online course provided through the web so that any user can access it without limitation on time and location. Since 2012, several popular MOOC learning platforms, initiated by top universities, emerged, including edX, Coursera and Udacity. However, high dropout rate on MOOC learning platforms has been highly criticized. Thus predicting students' likelihood of dropout would be useful for maintaining and encouraging their learning activities.

For this class project, you use machine learning techniques to predict whether a user will drop a course within next 10 days based on his or her prior activities recorded in a log. We define a dropout from a course as that a user leaves no records for that course in the log during the next 10 days.

Data Source:

We will provide a training and a test data set.

Goal:

You use the training data set to build your model and maximize the performance of your model on the test data set.

Detailed Instructions:

1) Download the data from the source that is specified by the instructor.
2) You must build a model to predict the labels of the test dataset.
3) Use at least three different machine learning algorithms such as SVM, KNN, Neural Networks or any other algorithms. You need to optimize your algorithm by optimizing the parameters of the algorithms.
4) Feature selection is a big aspect of modeling. Use at least one kind of feature selection algorithm with your favorite machine learning algorithms and see how much the results improve. You also need to optimize the parameters of the feature selection algorithm
5) For model evaluation, report the performance of the above algorithms on MCC (Matthew's Correlation Coefficient) and AUC (area under curve).
   MCC = (TP*TN-FP*FN)/sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)) , where TP, TN, FP and FN are true positive, true negative, false positive and false negative respectively. When computing MCC, you need to run the experiments at least 20 times.

Report Requirement:

You will need to submit a report that follows the following guidelines in a single document.

1) Title

2) Problem statement

3) Exploratory analysis of the data set: how many features, how many samples, is the data set balanced? What types of features that you have? Do you have any missing features? Do you perform any transformation of the data set?

4) Experiment Design: how do you build your models, how to optimize your models, how do you perform feature selection, and how do you evaluate your model
   This part has to include the following information:
   How do you perform feature selection and how do you optimize feature selection?
   How do you build your model (models), how do you optimize the performance?
   How do you select the final model to work on the testing data set?
   Did you do model averaging (extra credit)?

5) Present the experimental study results
   Using appropriate tables and figures to organize your results

6) Results analysis and discussion
   Description your observations
   Using statistical analysis to compare models (extra credit)
   Did you perform error analysis of your modeling algorithms (extra credit)?

7) Construct your final model (based on (4)), with a specific and detailed description of the parameters of your model.

8) Predict results on the test data sets that were provided to you.

Your submission should have three parts:

(1) the electronic version of your report. This file should be named as EECS738_[your KU ID]_report.doc (or pdf or anything else)
(2) the electronic version of your predicting results on the test data set. This file should be named as EECS738_[your KU ID]_test.txt. This must be a text file with predictions, one prediction per line, in the same order of the samples in the test file.
(3) The electronic version of your source code for your algorithm. This file should be named as EECS738_[your KU ID]_algorithm.* (* can be java, m or anything else)

There are multiple ways that you may put extra effort into the project such as implementing a new machine learning algorithm, formalizing the problem in a slight but meaningful way, detailed discussion of the experimental results.

Grading

(1) Experimental study: accuracy, rigorousness, and comprehensiveness: 50%
(2) Test data set prediction: accuracy, completeness: 30%
(3) Report: preciseness, conciseness, clarity, and comprehensiveness: 20%

Timeline:

Oct 14, data sets provided.

Oct 30, preliminary results showing the exploratory analysis of the data sets and showing the classification accuracy for three classifiers that you selected

Nov 20, declare your intention if you wan to do a voluntary demo of your final project (with extra credit).

Dec 7/9, demo of your final project

Dec 18, your submission of the final project is due (your report, your source code, and your prediction results)