ENGR 421 DASC 521
Homework 07: Modeling Late Payments for Credit Card Bills

Deadline: January 2, 2020, 11:59 PM

In this homework, you will develop a machine learning solution in R, Matlab, or Python for three real-life classification problems from finance industry. Your machine learning algorithm needs to predict whether a customer will delay his/her credit card bill payment more than 1 day (named as target1), more than 31 days (named as target2), or more than 61 days (named as target3) using the information given about each customer. Here are the steps you need to follow:

1. For each binary classification problem, you are given three input data files.
    a. For the first problem, the files are named as hw07_target1_training_data.csv, hw07_target1_training_label.csv, and hw07_target1_test_data.csv. The training and test sets contain 11,000 and 5,813 data instances, respectively, where each data instance has 162 features.
    b. For the second problem, the files are named as hw07_target2_training_data.csv, hw07_target2_training_label.csv, and hw07_target2_test_data.csv. The training and test sets contain 9,000 and 4,752 data instances, respectively, where each data instance has 211 features.
    c. For the third problem, the files are named as hw07_target3_training_data.csv, hw07_target3_training_label.csv, and hw07_target3_test_data.csv. The training and test sets contain 5,000 and 2,951 data instances, respectively, where each data instance has 202 features.

    You are also given a very simple solution strategy using a boosting classifier in the file named hw07_quick_and_dirty_solution.R.

2. Develop your own machine learning solution for these three problems. You are free to use any publicly available packages in R, Matlab, or Python. The predictive quality of your solutions will be evaluated in terms of AUROC (area under the receiver operating characteristics curve) values on the test sets.

3. Use the trained algorithms from the previous step to perform predictions for the test data sets, which contain 5,813, 4,752, and 2,951 customers for three problems. You are not given the correct labels for test instances. You need to predict the scores or posterior probabilities for positive class in each problem and to write these estimates into three files. For example, the strategy implemented in hw07_quick_and_dirty_solution.R file generates the estimates for the test sets and writes these values into three different files named as hw07_target1_test_predictions.csv, hw07_target2_test_predictions.csv and hw07_target3_test_predictions.csv.

What to submit: You need to submit your source code in a single file (.R file if you are using R, .m file if you are using Matlab, or .py file if you are using Python), the estimated scores or posterior probabilities for positive class on the test sets (hw07_target1_test_predictions.csv, hw07_target2_test_predictions.csv, and hw07_target3_test_predictions.csv), and a detailed report explaining your approach (.doc, .docx, or .pdf file). You will put these five files in a single zip file named as ***STUDENTID.zip***, where ***STUDENTID*** should be replaced with your 7-digit student number.

How to submit: Submit the zip file you created to Blackboard. Please follow the exact style mentioned and do not send a zip file named as ***STUDENTID.zip***. Submissions that do not follow these guidelines will not be graded.

Late submission policy: Late submissions will not be graded.

Cheating policy: Very similar submissions will not be graded.