CSC591/791: Assignment 4

October 3, 2017

Due 10/11/2017 8:30am.

In this assignment you will gain familiarity with the use of classifiers for educational data. You have been provided with a dataset in CSV format. This dataset represents a series of student observations taken in a real classroom. The students are being observed using the BROMP protocol.[1] The purpose of this protocol is to allow investigators to observe students working in real class situations and to code their affective states and on-task behaviors without direct interference in their activities.

Your task in this assignment is to:

1. Load this datafile into an appropriate tool or into your own code.
2. Train two *different* classifiers to predict **ONTASK** on a row-by-row basis using a static 70/30 train-test split with balanced random assignment.
3. Train the same two classifiers using 10-fold cross-validation with balanced random assignment.
4. Compare the performance of the four classifiers using a standard metric of the type discussed in class.

Your report should include a detailed description of the steps taken. It should also include results for the comparison packages and answers to the following questions:

1. Which classifier performed better?

[1] `http://www.columbia.edu/~rsb2162/bromp.html`

2. How did the performance change from static sampling to cross-validation?

3. What feature of the algorithm's inductive bias may explain its performance?

4. How did you control for *overfitting* in the dataset?

Dataset

The dataset has been provided as a CSV File `AssignmentData.csv`. The file contains 27,732 rows & 17 columns:

UNIQUEID Unique step id.

SCHOOL ID of the school

CLASS ID of the class

GRADE Students' grade level

CODER Coder making the observation

STUDENTID ID of the student.

GRADER Gender of the student as a binary variable.

OBSNUM Count of observations for this student.

TOTALOBS-FORSESSION Total number of observations in current session.

ACTIVITY Students' current activity.

ONTASK Is the student engaging with the task Y/N

TRANSITIONS Number of task transitions for the student.

NUMACTIVITIES Number of activities observed at this time.

FORMATCHANGES Format changes (by student) observed at this time.

NUMFORMATS Number of different formats (by student) observed.

TRANSITIONS/DURATIONS Time spent on transitions.

TOTALTIME Total observed time.

Submission

You should submit two separate files:

- Your Code (if any) code for processing the dataset. If you use a standard tool then submit a detailed list of the steps taken. (`YourName_code.zip`).
- Your written report (`YourName_report.pdf`).