CS-411 Project Description

1. Overall Goal

Demonstrate your ability to set up a learning analytics pipe. Derive primary and secondary features from the raw clickstream. Explain/Justify the definition of features and state an a priori hypothesis of how they are related to progress and performance. Apply classification and regression methods to identify which features are most predictive for progress and which methods allow the best prediction of overall performance.

2. Data and definitions

2.1 Data structure

The data contains clickstream events for each attempt in programming assignments from the course Introduction to Object Oriented Programming in Java, Jamila Sam and Jean-Cédric Chappelier. Each submission is automatically graded and results in a score from 0-100.

1 record in the raw data corresponds to 1 submission of 1 student to 1 exercise. When students submit more than once, the events in the clickstream represent the actions they have taken in between submissions.

Submissions = grade(1), ... grade(i), ..., grade(n)
Progress = the difference in grade between two submissions = grade(i) - grade(i-1)
Performance = the difference in grade between the first and the last submission = grade(n) - grade(1)

2.2. Train and test sets

The dataset uploaded on Moodle on Nov. 7th serves as the training set for your models. The final test set would be provided to you on Nov. 29th which has the same structure as described above, but excludes the grade information.
3. Project and deliverable

In simple terms, we want to know if certain patterns and actions (or sequences of actions) are leading to better results, either in between two submissions or from the first to the last submission.

3.1. Predict progress with a classifier model

For each step below, explain and justify which observations and features you create/delete/retain from the dataset.

3.1.1. Create primary features

By using the python script provided, define at least 3 new features in addition to the ones provided by default. Run the script to augment the data with the new features. For each of the features, explain the rationale behind its formal definition. Why and how do you think this particular features will be related to progress.

**Deliverable:** (1+ pages). Describe each of the features. Why do you think it is related to progress? What is the rationale? Comment the python code to implement the feature. Report and comment descriptive statistics for each feature.

3.1.2. Train classifiers

Train a classifier to identify which features are most predictive of an increase versus decrease of the grade. To do so, recode the quantitative progress into a 2-level variable (improved, notImproved). Find out which features in combination with which machine learning method are most predictive of progress. Describe and document your experiments by contrasting at least 3 methods (e.g. SVM, Neural Networks and Random Forest). Report and interpret classification quality based on area under ROC curve (AUC).

**Deliverable:** (3 paragraphs) Compare the results from the three methods in a table. List the features for the best model and provide an interpretation of why these features work well (or do not work so well). Compare the accuracy on the training and testing subsets. Describe the quality of the fit of your model (e.g. does your model overfit the data). Submit the results for your classifier to Kaggle.
3.2. Predict performance with a regression model

3.2.1. Create secondary features

Combine the primary features across several student attempts to predict overall performance (difference in grade between the first and last submission). This step is done in R by using a function called `ddply` that combines several observations into only one (Alternatively you can use python to combine or extract secondary features). Justify the construction of your secondary features.

**Deliverable**: (1+ pages). Describe each of the secondary features. Why do you think they are related to performance? What is the rationale? Comment the R code to implement the feature. Report and comment descriptive statistics for each feature.

3.2.2. Train regression models

Train a regression model that is able to predict the performance of students. Find out which secondary features in combination with which regression model are most predictive of overall performance. Submit your results to a Kaggle competition that will be announced in the following weeks. Describe and document your experiments with at least 3 different regression methods. Report and interpret the regression accuracy based on RMSE measure.

**Deliverable**: (3 paragraphs) Compare the results from the three methods in a table. List the features for the best model and provide an interpretation of why these features work well (or do not work so well). Compare the accuracy on the training and testing subsets. Describe the quality of the fit of your model (e.g. does your model overfit the data). Submit the results for your regression model to Kaggle.
4. Grading policy

Quality of features construction and the accuracy of the predictors (classifier and regression models) would contribute in your project grade.