Educational Data Mining

Term Project Phase I: Initial Trial

(100 points)

Assigned Date: Tuesday, April 5, 2016

Due Date:

Deadline: 4:00 PM Thursday, April 14, 2016

Educational Goal

Become familiar with supervised and unsupervised learning using real-world data.

Requirements

In this project, you will predict student performance on mathematical problems from logs of student interaction with Intelligent Tutoring Systems using the KDD 2010 Cup data. You may use any methods that have been developed for KDD Cup 2010 (in this case, you should provide citation of the methods you used), or you are welcome to design and implement your own methods.

Data Sets

For term project phase I, the data set (used for both the training and test data in this assignment) can be downloaded from
http://www.cs.umb.edu/~ding/classes/438_638/homework/term_project_data/algebra_2005_2006_train.zip

The data set is part of the Development Data Sets, which are described in the KDD Cup 2010 data web page located at https://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp:

Development Data Sets

Data sets	Students	Steps	File
Algebra I 2005-2006	575	813,661	algebra_2005_2006.zip

In this term project, you should follow the exact Data Format required by KDD Cup 2010, https://pslcdatashop.web.cmu.edu/KDDCup/rules_data_format.jsp

The actual student performance values for the prediction column, Correct First Attempt, are provided for all steps. This allows you to calculate the difference between the predicted value and the ground truth value.

Machine Learning

After you learn a machine learning model from the training data, you will use the same dataset as the test data to evaluate your approach. You will use the same rules designed by the KDD Cup 2010, which not all the attributes are used during the testing phase. You will just use the values of Student Id, Problem Hierarchy, Problem Name, Problem View, Step Name, KC(Default), Opportunity(Default) to predicate "Correct First Attempt".

Specifically, the variables that should be / should not be used during the testing phase are shown as below:

Variable	Used in Test Set	Variable	Used in Test Set	Variable	Used in Test Set
Student ID	Yes	Correct Transaction Time	No	Hints	No
Problem Hierarchy	Yes	Step End Time	No	Corrects	No
Problem Name	Yes	Step Duration	No	KC(Default)	Yes
Problem View	Yes	Correct Step Duration	No	Opportunity (Default)	Yes
Step Name	Yes	Error Step Duration	No		
Step Start Time	No	Correct First Attempt	No		
First Transaction Time	No	Incorrects	No		

During the testing phase, you will predict whether the student got the step right on the first attempt for each step in that problem. Each prediction will take the form of a **decimal between 0 and 1** for the column Correct First Attempt.

Generate the prediction result files for the data set. Each of them should contain three columns:

Row: the row number, as carried over from the original data set file.

Student ID: the Student ID, as carried over from the original data set file.

Correct First Attempt: your prediction value, **a decimal number between 0 and 1** that indicates the probability of a correct first attempt for this student-step.

Show the final **RMSE** after run your methods with the data set.

Reference:

KDD Cup 2010, Educational Data Mining Challenge, https://pslcdatashop.web.cmu.edu/KDDCup/

Grading criteria

Predication accuracy (40%), novelty of the designed methods (40%), transformality (20%) (how likely the designed methods can be used to other real-world applications).

Submission Requirements

1. One submission per team.
2. **Submit a PPT file (in PDF format) to report the RMSE on the Algebra I 2005-2006 dataset and explain the method(s) used.**
3. **Submit all the scripts (soft & hard copies) you used for this project**. Submit a single zipped file of all the files of this assignment through your Blackboard account. Submit the paper copy including the slides and scripts. Paper copy should be bound firmly together as one pack (for example, staple, but not limited to, at the left corner). 5 points will be deducted for unbounded homework.
4. Name your file with teamleadlastname_firstname_team#_ph1. For example, team 1 of lead John Smith should name their file as Smith_John_team1_ph1.zip.
5. No hard copies or soft copies results in 0 points.