Machine Learning Engineer Nanodegree

Capstone Proposal

Sai Raj Reddy

Proposal

Domain Background

Educational Technology and Learning Analytics (Also known as Educational Data Mining) has been my keen interest since past few years. Identifying the Knowledge Components and interaction data from Intelligent Tutoring Systems (ITS) used by students help us gather structured (or semi-structured) data. Picked up interest naturally with amalgamation of deep curiosity about the education data. The insights from educational datasets are intented to help students, teachers or school administrators to help them grow.

One of the EdTech products I am aware of -- Carnegie Learning's Math Tutor (Mathia Software ITS) is intented to help high school students learn Math better. PSLC Datashop platform is meant for Education Data Scientists to look for learning analytics' data. Founded by noted researcher Ryan Baker, PSLC hosted an open challenge in 2010 called KDD EDM Challenge which was sponsored by Facebook and IBM Research. They released the Mathia ITS data which consisted of both Training and Testing set. Related research fields that are in spotlight are 'Cognitive Tutors'. Some of the paper publications that talk about the software can be found here :

- http://pact.cs.cmu.edu/pubs/Koedinger,%20Corbett,%20Ritter,%20Shapiro%2000.pdf
- https://link.springer.com/article/10.3758/BF03194060
- https://arxiv.org/pdf/1802.08616.pdf
- https://arxiv.org/pdf/1707.09308.pdf

Problem Statement

Students who use the Mathia Tutor to solve problems -- their interactions are recorded as logs. This log data can be mined to get insights. My aim is to *predict students' correct first attempt (CFA)* while solving the math. CFA is a feature in the dataset which is recorded in terms of bit or boolean (0 or 1). Therefore, precise problem domain involved is classification.

Probability can be calculated by sklearn. We can find some models or methods at Sklearn Home and sklearn.svm.libsvm.predict_proba() So, we can get model's performance metric like log loss on unseen data for measurement.

Datasets and Inputs

Download the data from http://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp

Data set	Students	Steps	File
Algebra I 2008-2009	3,310	9,426,966	algebra_2008_2009.zip

Stamper, J., Niculescu-Mizil, A., Ritter, S., Gordon, G.J., & Koedinger, K.R. (2010). Algebra I 2008-2009. Challenge data set from KDD Cup 2010 Educational Data Mining Challenge.

Data featrues :

|----------------------------|--|

Row	Row number for record.
Anon Student Id	Unique ID for a student
Problem Hierarchy	The hierarchy of curriculum levels containing the problem.
Problem Name	Unique identifier for one problem.
Problem View	The total number of times the student encountered the problem so far.
Step Name	Each problem consists of one or more steps (e.g., "find the area of rectangle ABCD" or "divide both sides of the equation by x"). The step name is unique within each problem, but there may be collisions between different problems, so the only unique identifier for a step is the pair of problem_name and step_name.
Step Start Time	The starting time of the step. Can be null.
First Transaction Time	The time of the first transaction toward the step.
Correct Transaction Time	The time of the correct attempt toward the step, if there was one.
Step End Time	The time of the last transaction toward the step.
Step Duration (sec)	The elapsed time of the step in seconds, calculated by adding all the duration for transactions that were attributed to the step. Can be null (if step start time is null).
Correct Step Duration (sec)	The step duration if the first attempt for the step was correct.
Error Step Duration (sec)	The step duration if the first attempt for the step was an error (incorrect attempt or hint request).
Correct First Attempt	The student’s first attempt on a step — 1 if correct, 0 if an error.
Incorrects	Total number of incorrect attempts by the student on the step.
Hints	Total number of hints requested by the student for the step.
Corrects	Total correct attempts by the student for the step. (Only increases if the step is encountered more than once.)
KC (KC Model Name)	The identified skills that are used in a problem, where available.
)Opportunity (KC Model Name)	Steps (contains KC) count for student that are needed to solve a problem. So, student can have multi chances to solve a problem. Steps with multiple KCs will have multiple opportunity numbers separated by ~~.

It has 8918055 records (algebra_2008_2009_train.txt)

For more info about Data Format -- Have a look at http://pslcdatashop.web.cmu.edu/KDDCup/rules_data_format.jsp

Solution Statement

Based on existing data given by the competition, we only have target feature Correct First Attempt. Since the CFA only contains value 1 and 0. So we can use classifier to predict the target result. We will train a classifier on training data set with validation data set together. Since original competition KDD CUP 2010 ask for probability of students' correct first attempt on problems. We will transform the problem from classification to numeric measurable problem. My approach involves replacing each categorical feature with a numerical one by using the "correct first attempt rate" (CFAR). The CFAR can be expressed by: - CFA: Student's correct first attempt - N: Total number of one student's all records(CFA = 1) - T: Total number of one student's all records(both CFA = 0 and CFA = 1) This CFAR directly connects a feature and CFA, which is now the target for prediction. CFAR is numeric between 0 and 1. So the classification problem can be transformed to numeric measurable problem. The model on test data can be measure with log loss. Our goal is to minimize the log loss on test data set which coming from portion of training data set. The test data set in original data is used to predict and submit to competition's leader board so that attendants can achieve their ranks. But in our experiments, there is no leader board. So we only use portion of training data as test data. That's important to know.

Since it's a classification problem, and a lot of data records involved in, sklearn's SGD classifier would be a good choice rather than linear SVM. Cause SVM takes a lot of time in large data set. LightGBM is a gradient boosting framework that uses tree based learning algorithms.I would consider LightGBM as my best choice. LightGBM has a very good performance not only in time dimension but also in accuracy. LightGBM is 10 times faster than xgboost, and 20 ~ 50 times faster than sklearn's alogorithm.

Benchmark Model

KNN is one of the most popular algorithms for find the nearest neighbors in data. For our KDD CUP 2010 competition problem, we suppose to find the nearest K students for one student. So these neighbors' average probability of first correct attempts on problems is thought to be the student's probability on that problem. Their average probability of first correct attempt will be calculated by the number(K) of stu- dents' whose first correct attempt is 1 divide by the total number of students in K.

• N: Number of Students(CFA = 1). • K: Number of Students in K. • Py: Probability of student's correct first attempt on one problem. P = N • CFARy': find this term and represent formula in Solution Statement section.

Log loss on test portion of training data could be represented as follows: -log P(y'|y) = -(y' log(y) + (1 - y') log(1 - y)) The probability calculated by benchmark model will also calculate log loss. To compare the result to our design solution result. We should optimize our solution result in better performance than Benchmark Model result. So, we can achieve a good performance.

Evaluation Metrics

we will only train on the training portion of each data set, meanwhile, use part of training portion data as validation set, and will then be evaluated on our performance at providing correct first attempt values for the test portion(part of training data).

We will compare the predictions we provided against test portion(part of training data) true values and calculate the difference as log loss.

• y’: Predicted target probability of Correct First Attempt(CFA=1) • y: Target’s original CFAR

$-\log P(y'|y) = -(y'\log(y) + (1 - y') \log(1 - y))$

The use of log loss is very common and it makes an excellent general purpose error metric for numerical predictions. We will dedicate our best to acquire the lowest log loss as possible.

Project Design

1. First step to load data and prepare data.
2. Secondly, One-hot encoding and regularization on numeric features for train data.
3. Thirdly, Visualize data into chart, so that we can see trends and scatters. Remove scatters and add potential feature to data. We are gonna use GMM to cluster data into different clusters so that we can add additional features.
4. Finally, we will use ensemble algorithms model to train data(train set and validation set). For e.g., XGBoost, LightGBM, GBDT or etc. And tuning hyper parameters. Then predict result on test data set. Compute log loss.