CS229 Spring 2018-2019
Modeling Student Learning in Mobile App with Machine Learning

Zhaolei (Henry) Shi
zshi2@stanford.edu

Problem Statement

This study applies machine learning to this new data source to understand how investing in tutoring products change students' learning trajectories. Machine learning techniques are used to model student's progress with student-question interactions as inputs and predicted achievement (operationalized as probability of a correct answer across all questions) as output.

Motivation

Online education promise to revolutionize education but companies struggle to show that their products improve student learning. To tackle this problem, I seek to leverage machine learning to estimate student learning.

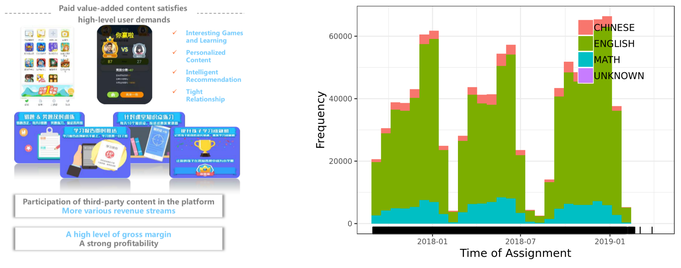

Figure 1: In-App Purchases and Homework Assignments Over Time

In the context of my study, the space of possible questions that a student can be exposed to is large. **Each student is only exposed to a small subset of problems** in any given session. Hence, I will leverage machine learning techniques to **predict the students' performance on the questions that she did not see**.

Dataset

Data for this project comes from a mobile education company in China. The data contains logs of student questions responses on homework assigned by their school teacher. For the estimation, I use $138,001$ records of student question responses from one school in one month in 2018.

Models

- Naive Bayes

$$\hat{\Pr}(y_i = 1; t) = \frac{\sum_j \gamma(t_{ij} - t) \cdot \mathbb{1}\{y_{ij} = 1\}}{\sum_j \gamma(t_{ij} - t) \cdot \mathbb{1}\{i \text{ attempted } j\}}$$

where $\gamma(.)$ is a Gaussian kernel.

- Difficulty-Modulated Naive Bayes (with Laplace smoothing)

$$\hat{\theta}_j = \frac{\sum_i \mathbb{1}\{y_{ij} = 1\}}{\sum_i \mathbb{1}\{i \text{ attempted } j\}}$$
$$\hat{\Pr}(y_i = 1) = \frac{\sum_j \gamma(t_{ij} - t)\mathbb{1}\{y_{ij} = 1\}/\hat{\theta}_j}{\sum_j \gamma(t_{ij} - t)\mathbb{1}\{i \text{ attempted } j\}}$$

- Siamese-Like Neural Network

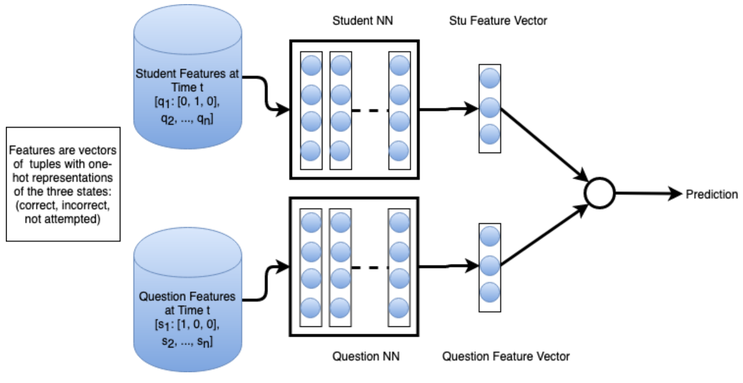

Figure 2: Siamese-Like Neural Network

Results

Model Class	Accuracy	Precision	Recall	F1 Score
Siamese	0.873	0.907	0.937	0.922

Batch Size	Learning Rate	Epochs	Network Structure	Feature Vector Size
128	0.001	15	Student: [400-80] Question: [200-80]	10

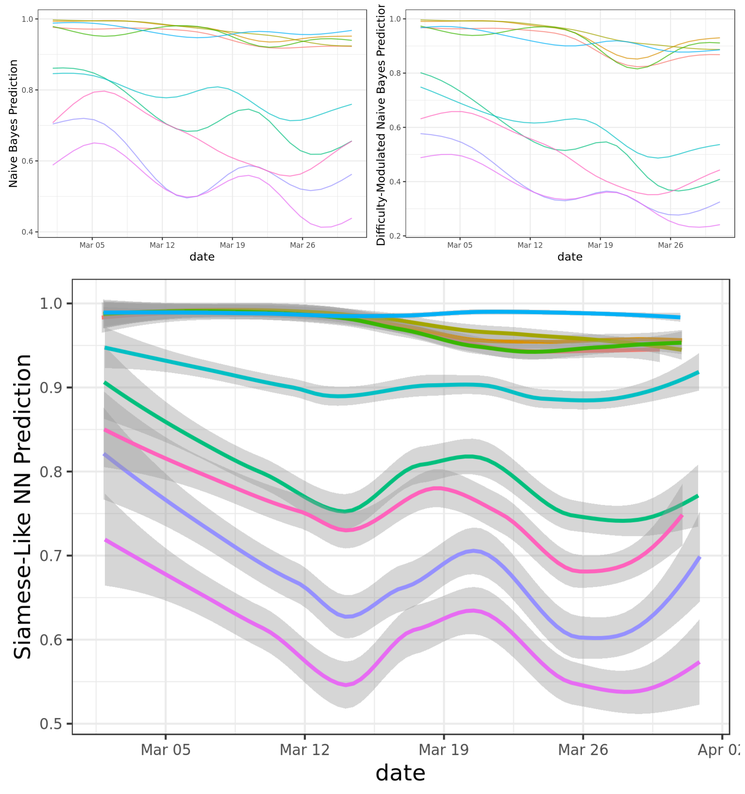

Figure 3: Prediction of Learning of the Three Models

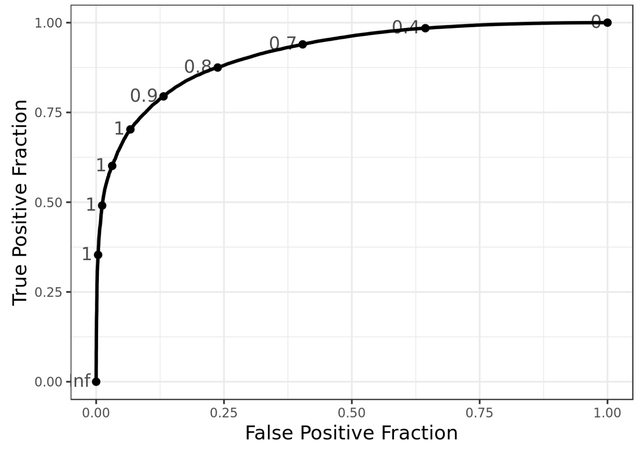

Figure 4: ROC Curves of Siamese-Like NN