Modeling Student Knowledge as a Latent Variable in Intelligent Tutoring Systems: A Comparison of Multiple Approaches

Richard Davis, Alex Kolchinski, Qandeel Tariq
Fall 2016, CS229

Background

- Students who receive mastery-oriented one-on-one tutoring have been shown to outperform their traditionally-schooled classmates by two standard deviations (Bloom, 1984).
- Computer-based tutors called Intelligent Tutoring Systems (ITS) aim to replicate Bloom's findings on a large scale.
- Intelligent tutors need to model student understanding as students progress through problems on a computer.
- ITS's only see student responses, so constructing a model of student understanding is a latent-variable problem.
- A traditional approach to this is known as Bayesian Knowledge Tracing (BKT).
- BKT uses Hidden Markov Models (HMM's) to model latent student knowledge using observations of student performance (e.g., answers to test questions) as observed variables.

Research Questions

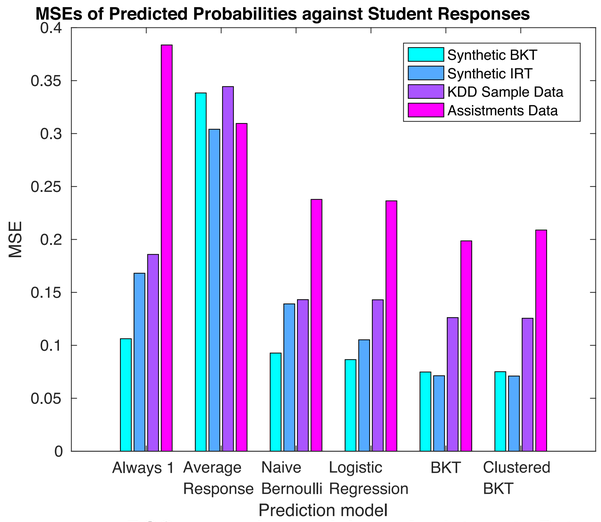
- How much better (if at all) does an HMM-based knowledge tracing algorithm compare to basic statistical approaches like independent Bernoulli and logistic regression?
- How much of a performance boost can knowledge tracing algorithms reap from modeling students as different rather than identical?

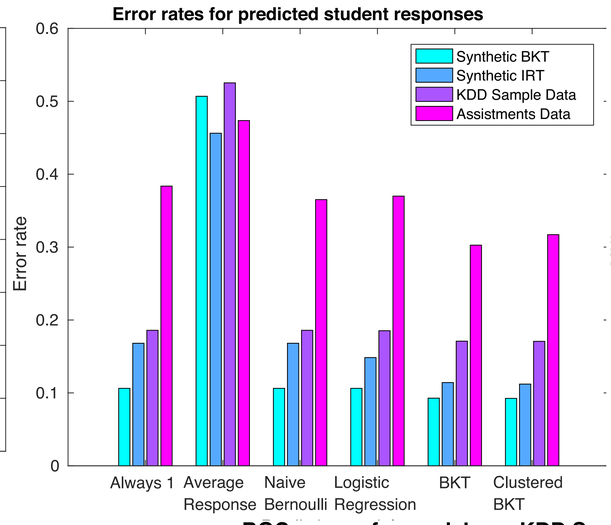
Data

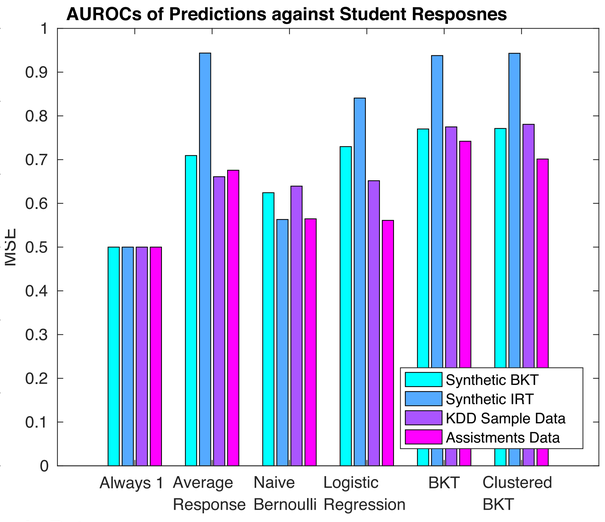
- **Synthetic data generated according to IRT assumptions** (500 synthetic students and 10 concepts). P(correct) depends on student skill on question's concept and concept difficulty (logit link).
- **Synthetic data generated according to BKT assumptions** (500 synthetic students and 10 concepts). P(correct) modeled using an HMM with fixed transition and emission probabilities.
- **Bridge to Algebra 2006–2007** data set (out of 808 concepts and 5986 students, we used a subset of 500 students and 10 concepts)
- **Assistments** (out of 125 concepts and 4218 students, we used a subset of 1678 students and 10 concepts)

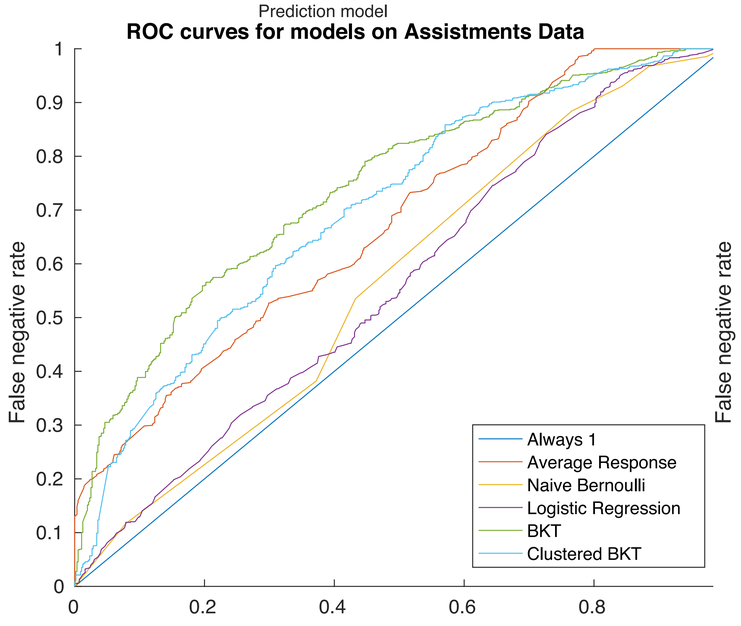
Discussion

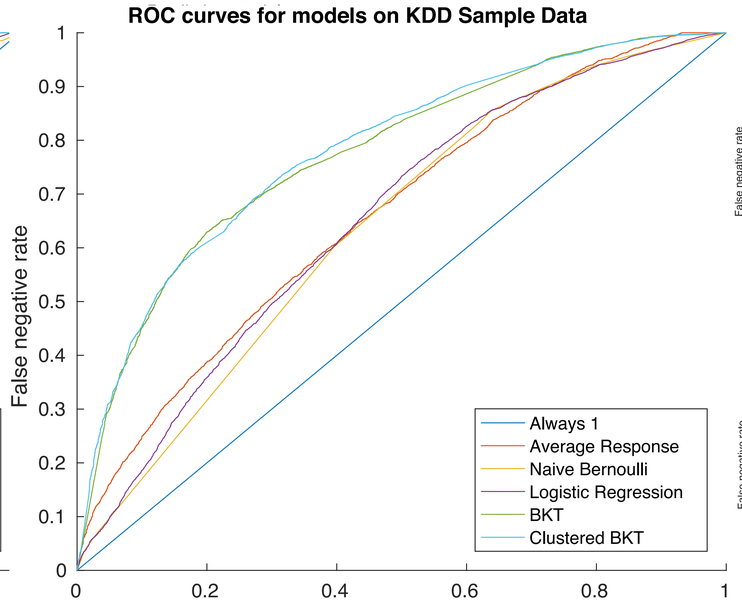
- Only BKT and Clustered BKT reliably outperform baseline on all datasets. The average response model performed nearly as well as BKT and Clustered BKT on the Assistments and Synthetic data, but performed far worse than BKT on the KDD data. This shows that BKT is the most reliable model to use when making predictions about student knowledge across multiple datasets.
- Clustered BKT and BKT achieved nearly identical performance across all datasets. This was unexpected. As splitting HMMs into one for high-scoring students and one for low-scoring students did not improve performance, evidently fitting two sets of transition and emission parameters did not bring the mixture model closer in line with the true distribution.

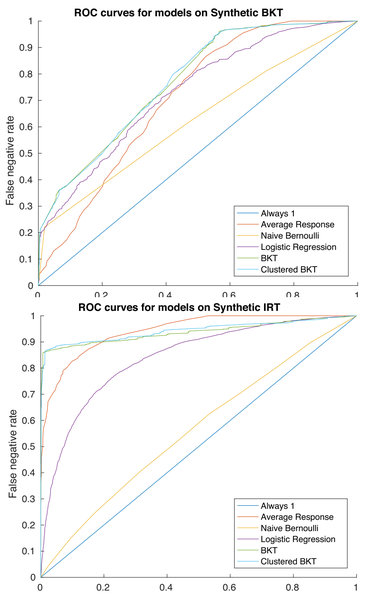












Models

- **Always 1 (Baseline)**: Predict that all students answer all questions correctly.
- **Naive Bernoulli**: For each concept, there is a constant probability that a question is answered correctly (learned from training data). Students assumed identical.
- **Logistic regression**: For each concept, model P(correct) with a logit link on (# of questions that the student saw for that concept) as the only independent variable. Students asssumed identical.
- **Average Response**: P(correct) for a given question is modeled per student as the percent of questions the student got correct on that concept up to that point
- **Bayesian Knowledge Tracing**: Model students as independent and identical, concepts as independent. For each concept, train an HMM (states: know/don't know; emissions: correct/incorrect) and then make predictions by using the HMM model to predict state sequence probabilities for new students.
- **Clustered Bayesian Knowledge Tracing**: Group students and train separate HMMs for each group on each concept. We divided students in the training data using median split on percent of total correct answers. We made predictions for a new answer sequence such that at each index, the proportion of correct answers up to that index determines which HMM's predicted state sequence is used to predict the next emission.

Findings

- Both BKT models consistently outperformed all other models across all datasets
- Clustered BKT did not provide additional gains over BKT on any model
- The Average Response model performed surprisingly well on the synthetic and KDD data, but performed near baseline on the KDD data
- Naive Bernoulli performed passably well, especially on synthetic data. Even "Always 1" achieved <20% error rates on the synthetic data sets and KDD data, where students got most questions right!
- Logistic regression performed almost as well as the BKT models on the synthetic data and the KDD data, but substantially worse on the Assistments data.