Towards Moment of Learning Accuracy

Zachary A. Pardos† and Michael V. Yudelson‡

†Massachusetts Institute of Technology
77 Massachusetts Ave., Cambridge, MA 02139
‡Carnegie Learning, Inc.
437 Grant St., Pittsburgh, PA 15219, USA
zp@csail.mit.edu,yudelson@carnegielearning.com

1 Introduction

Models of student knowledge have occupied a significant portion of the literature in the area of Educational Data Mining[1]. In the context of Intelligent Tutoring Systems, these models are designed for the purpose of improving prediction of student knowledge and improving prediction of skill mastery. New models or model modifications need to be justified by marked improvement in evaluation results compared to prior-art. The standard evaluation has been to forecast student responses with an N-fold student level cross-validation and compare the results of prediction to the prior-art model using a chosen error or accuracy metric. The hypothesis of this often employed methodology is that improved performance prediction, given a chosen evaluation metric, translates to improved knowledge and mastery prediction. Since knowledge is a latent, the estimation of knowledge cannot be validated directly. If knowledge were directly observable, would we find that models with better prediction of performance also estimate knowledge more accurately? Which evaluation metrics of performance would best correlate with improvements in knowledge estimation? In this paper we investigate the relationship between performance prediction and knowledge estimation with a series of simulation studies. The studies allow for observation of the ground truth knowledge states of simulated students. With this information we correlate the accuracy of estimating the moment of learning (mastery) with a host of error metrics calculated based on performance.

2 Bayesian Knowledge Tracing

Among the various models of knowledge, a model called Bayesian Knowledge Tracing [2] has been a central focus among many investigators. The focus on this model has been in part motivated by its use in practice in the Cognitive Tutors [4], used by over 600,000 students, and by its grounding in widely adopted cognitive science frameworks for knowledge acquisition. For our experiments we will be employing the most frequently used basic Bayesian Knowledge Tracing

[1] A session during the main proceedings of EDM 2012 was dedicated to papers on Knowledge Tracing, a frequently used approach to modeling student knowledge.

model for both simulation and evaluation; however, there are implications beyond BKT models. Knowledge Tracing is a simple Hidden Markov Model of Knowledge defined by four parameters; two performance parameters and two knowledge parameters. The performance parameters, guess and slip, are the emission parameters in an HMM which respectively correspond to the probability that a student answers correct even if she is in the negative knowledge state (guess) and the probability that she answers incorrectly even if she is in the positive knowledge state (slip). The knowledge parameters, prior and learn rate, are the probability that a student knows the skill before answering any questions and the probability that, if the student is in the negative knowledge state, she will transition to the positive state at any given opportunity.

3 Related Work

There has been a limited amount of prior work focusing on detecting the moment of learning. We were able to track one relevant publication by Baker and colleagues [1]. They investigated detection of moment of learning in student data by modifying BKT structure. Another relevant result was published by [5]. They looked at scoring student model fits on simulated data and found a disparity between rankings of two frequently used metrics: root mean squared error and area under ROC curve. In this work we would like to address the question of the quality of detecting the moment of learning and investigate the problem of choosing a goodness-of-fit metric for that purpose.

4 Data

Our simulation dataset consisted of 1,000 simulated students and 100 skills with 30 questions per skill. Every student answered all 30 questions for each of the 100 skills. In the BKT simulation model we included no dependencies between skills and also no student specific parameters; therefore, the data can be thought of as either being produced by 1,000 students total or a new 1,000 students for every skill. Programmatically, data for each skill is stored in a separate file. Each row in each file represents one students data for that skill. The data stored from the simulation contains the students ground truth binary state of knowledge (mastered or not) at each of the 30 opportunities to answer (first 30 columns) and also the students correctness of responses to the 30 questions (stored in the second set of 30 columns).

In addition to the simulated data files containing student knowledge states and observed responses, we had corresponding files containing inferences of knowledge states and predictions of responses made with 16 different parameter sets resulting in 1,600 prediction files. Details of the parameter selection for simulation and prediction are discussed in the next section.

5 Methodology

5.1 Simulation

We generated 1,000 students knowledge and performance for 100 skills. Skills are defined by a set of four knowledge tracing parameters which the skill data is generated from. The 100 sets of four parameters were selected at random, uniformly sampling from the following constrained ranges for the parameters; prior between 0.01-0.80, learn rate between 0.01-0.60, and guess and slip between 0.05-0.40. After the 100 sets of parameters were selected, simulated data was produced by specifying a Dynamic Bayesian Network representation of Knowledge Tracing with a time slice length of 30. This representation, defined in Kevin Murphys Bayes Net Toolbox, with a particular parameter set fixed in the conditional probability tables, was then sampled 1,000 times, representing each simulated student. The sample Dynamic Belief Network function in BNT for simulation is a simple one; a random number between 0 and 1 is generated, if the number is equal to or lower than the prior parameter, the simulated student begins in the negative (not learned) state at time slice 1. To generate the observed response at this time slice, another random number is generated, if that number is greater than the guess parameter, the observed response is incorrect. To determine if the students knowledge state is positive (learned) at the next time slice; a random number is generated, if that number is less than or equal to the learning rate, then the students state is positive. With a positive state, the new random number needs to be greater than the slip parameter in order to produce a correct response. This is repeated for 30 times to simulate 30 knowledge states and observed responses per student.

5.2 Prediction

Typically, to predict student data, a hold-out strategy is used whereby a fraction of the students and their data is used to find a good fitting set of parameters. That good fitting set is then used to predict the fraction of students not used in training. The research question of this paper did not involve parameter fitting but rather required us to evaluate various models and observe how the models prediction of performance corresponded to its inference of knowledge. To do this we needed variation in models which we accomplished by choosing 16 candidate parameter sets with which to predict student data from each of the 100 skills. Since no training was involved, all data served as the test set. The top five sets of parameters used in the Cognitive Tutors was used, as well as 10 randomly generated parameters sets using the the same parameter constraints as the simulation, and, lastly, the ground truth parameter set for the skill was used to predict. The the same 15 parameter sets were used to predict the 100 skills, only the ground truth parameter set changed.

The prediction procedure is the same one used in all papers that use Knowledge Tracing; the prior, guess and slip parameters dictate the probability of correct on the first question. After the prediction is made, the correctness of

Table 1: Confusion Table

		Actual	
		Correct	Incorrect
Predicted	Correct	True Topisitve(TP)	False Positive (FP)
	Incorrect	False Negative (FN)	True Negative (TN)

the first question is revealed to the KT algorithm, which incorporates this observation using Bayes Theorem to infer the likelihood that the knowledge was known at that time. A learning rate transition function is applied and the processes is repeated 30 times in total to create 30 predictions of knowledge and 30 predictions of correctness per student for a skill.

6 Metrics

The most common metrics used to evaluate prediction performance in the EDM literature has been Area Under the Receiver Operator Curve (AUC) and Root Mean Squared Error (RMSE). One of the goals of our experiment is to reveal how indicative these measures are of the models accuracy in inferring knowledge. While these are the most common metrics, many others have been used in machine learning to evaluate predictions. We utilize a suite of metrics to investigate which metric is best at forecasting knowledge inference accuracy.

6.1 Model Performance

We selected a set of metrics in wide use today to score models when predicting student performance and knowledge state. Below is a short description of them.

Confusion Table Metrics Confusion table (rf. Table 1) is a table widely used in information retrieval and is a basis for a set of metrics capturing correctness of a retrieval or classification algorithm. Rows and columns of the confusion table denote the predicted and actual classes respectively and the cells in the intersection contain the counts of cases. Refer to Table 1 for an illustration. Here we illustrate a case for binary classification akin to the problem of binary classification of student performance (correct or incorrect) and state of knowledge (known or unknown).

If prediction is not categorical, say a probability from [0, 1], it is customary to round it: probabilities of 0.5 and greater become 1. For example, the cases when prediction matches the reality are captured in True Positive cell and the cases when the actually incorrect responses are marked as correct are captured in False Positive cell. We will use the confusion table metrics below.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1a}$$

$$precision = \frac{TP}{TP + FP} \tag{1b}$$

$$recall = \frac{TP}{TP + FN} \tag{1c}$$

$$F - measure = 2\frac{precision*recall}{precision + recall} \tag{1d}$$

As opposed to the so-called point measures described above, there is also a frequently used Area Under Receiver Operating Characteristic curve (AUROC), which is a curve measure. The curve is produced by varying the rounding threshold (0.5 for point measures) from 0 to 1 and computing and plotting False Positive Rate (FPR) vs. True Positive Rate (TPR) (see below).

$$TPR = \frac{TP}{TP + FN} \tag{2a}$$

$$FPR = \frac{FP}{FP + FN} \tag{2b}$$

An area under resulting curve is the sought metric. An area of 0.5 is equivalent to random chance for a binary classifier. An area greater than 0.5 is, thus, better than chance. An exact AUC calculation can also be derived by enumerating all possible pairs of predictions. The percentage of the pairs in which the true positive prediction is higher is the AUC. This is the ability of the predictor to discriminate between true and false.

Pseudo R^2 R^2 or percent variance explained is often used as a goodness of fit metric in linear regression analysis. For with binary classification, there exist several versions of R^2 called pseudo R^2. Applicable to our situation is Efrons pseudo R2 (refer to Equation below).

$$R^2 = 1 - \frac{\sum_{i=1}^{N} y_i - \hat{y}_i}{\sum_{i=1}^{N} y_i - \bar{y}} \tag{3}$$

Where N is the number of data points, y_i is the i-th component of the observed variable, $\bar{y_i}$ is the mean observed value, and $\hat{y_i}$ the prediction of i-th component of the observed variable.

Metrics Based on Log-Likelihood Likelihood functions are widely used in machine learning and classification. Likelihood captures the probability of the observing data given parameters of the model. In binary classification a natural log transformation of the likelihood function is often used (see below). Here

N is the total number of datapoints, y_i is the i-th component of the dependent variable, $\hat{y_i}$ is the predicted value of the i-th component of the dependent variable.

$$loglikelihood = \sum_{i=1}^{N} y_i \ln(\hat{y_i}) + (1 - y_i) \ln(1 - \hat{y_i}) \tag{4}$$

In addition to log-likelihood itself, there are several metrics that use log-likelihood as kernel component. For example, Akaike Information Criterion (AIC), Akaike Information Criterion with correction for finite sample size (AICc), Bayesian Information Criterion (BIC), and several others. These metrics introduce various forms of penalty for the size of the model (number of parameters) and number of datapoints in the sample in order to put overfitting models at disadvantage when performing model selection. Here k is the number of model parameters, N is the number of datapoints.

$$AIC = -2loglikelihood + 2k \tag{5a}$$

$$AICc = AIC + \frac{2k(k+1)}{N-k-1} \tag{5b}$$

$$BIC = -2loglikelihood + k \ln(N) \tag{5c}$$

Since we are comparing models that are only different in the parameter values and are doing so on the same dataset, we will not see difference in ranks assigned by log-likelihood, AIC, AICc, and BIC metrics.

Capped Binomial Deviance In addition to log-likelihood and log-likelihood-based metrics, we include the Capped Binomial Deviance (CBD). Capped binomial deviance is a version of the log-likelihood where prediction values are mandated to be at least away from 0 and 1 values and uses a logarithm with base 10 instead of natural logarithm. The is usually set to a small value of 0.001.

6.2 Moment of Learning

To capture the quality of detecting the moment of learning we devised a metric based on mean absolute deviation (MAD). Namely, moment of learning MAD is the average absolute difference of number of skill application opportunities between the moment when the internal state of the generating skill model switched to learned state and the moment when the probability of the skill being in a learned state reaches 0.95 (a traditionally used threshold in the area of intelligent tutoring systems). A perfect model would have a moment of learning MAD of 0. The larger the moment of learning MAD is the worse the model prediction of model of learning is.

7 Experiments and Results

7.1 Experiment 1

Research question: Among accuracy metrics used for ranking various parameter sets (models), which ones correlate best with accuracy of moment of learning prediction?

7.2 Results

The Table 2 below contains the correlations of performance prediction value, knowledge prediction value for all metrics, and moment of learning mean absolute error. Since prediction of performance is most widely adopted as a standard approach and the fact that we are trying to contrast it to the moment of learning mean absolute error, we sorted the rows corresponding to various statistical metrics by the respective column. The first column lists the metric used to evaluate the goodness of performance and knowledge prediction. The second column is the correlation between knowledge and performance prediction using the particular metric on both (this is the column the table is sorted by). The third column is the correlation between the particular metric used to evaluate performance and Mean Absolute Deviation (MAD) of Moment of Learning prediction. This is the column which tells us if the metrics used to evaluate performance are correlated with error in mastery / Moment of Learning prediction. The fourth column gives correlations of Moment of Learning MAD and metric values for predicting internal knowledge state. This correlation captures agreement between identifying the moment student learned a skill (this happens once per student-skill tuple) and the correctness of identifying the skills knowledge state for the student across all skill attempts.

7.3 Experiment 2

Hypothetically, the ground truth parameter sets should be the best at both making predictions of performance and estimating knowledge. A good metric should favor the ground truth parameters, therefore we ask: How often is the ground truth model the best at prediction performance according to the various metrics?

7.4 Results

The correlations of the performance and knowledge state prediction metrics from prior section targeted the 15 model parameter combinations that were different from the generating ground truth model parameters. Now, let us look at how the ground truth model compares to the other 15 we tested with respect to the statistical metrics we chose. Table 3, for each metric, gives the number of times a ground truth model parameter set is the best with respect to a given metric, and an average rank of the ground model parameter set as compared to the

Table 2: Metric correlations

Metric	Correlation of performance and knowledge metric	Correlation of performance metric and Moment of Learning MAD	Correlation of knowledge metric and Moment of Learning MAD
recall	0.878 ***	-0.954 ***	-0.819 ***
F-measure	0.561 ***	-0.839 ***	-0.792 ***
accuracy	0.522 ***	-0.802 ***	-0.822 ***
precision	0.334 ***	-0.797 ***	-0.628 ***
RMSE	0.470 ***	0.754 ***	0.828 ***
AIC	0.375 ***	0.751 ***	0.702 ***
AICc	0.375 ***	0.751 ***	0.702 ***
BIC	0.375 ***	0.751 ***	0.702 ***
CBD	0.409 ***	0.751 ***	0.762 ***
log-likelihood	0.375 ***	0.751 ***	0.702 ***
pseudo R^2	0.592 ***	-0.236 *	-0.296 **
AU ROC	0.335 ***	-0.119	-0.652 ***

Note: with respect to correlations with moment of learning MAD, in some cases a negative correlation is desirable (e.g., for accuracy), and for some cases a positive correlation is desirable (e.g., for RMSE). This is due to the fact that the smaller the moment of learning MAD the better, which is true for some metrics and the inverse is true for others. The table is sorted while observing this phenomenon (effectively sorting by the absolute value of the correlation coefficient).

Table 3: ground truth model rank vs. the other 15 models

Metric	Ground truth model has rank of 1	Mean rank of ground truth model
AIC	88/100	1.82/16
AICc	88/100	1.82/16
BIC	88/100	1.82/16
CBD	88/100	1.82/16
log-likelihood	88/100	1.82/16
RMSE	88/100	1.82/16
pseudo R^2	88/100	1.83/16
accuracy	33/100	2.52/16
F-measure	12/100	4.27/16
AU ROC	26/100	4.35/16
recall	0/100	6.65/16
precision	5/100	9.71/16

other 15 model. In each case we are aggregating across 100 different sets of 15 models plus one ground truth model. As we can see log-likelihood based models and RMSE form a group of metrics that gives ground truth models a large edge over the 15 reference models. Confusion table metrics, Area under ROC curve and the pseudo R2 gibe a drastically smaller support for it.

7.5 Experiment 3

Ground truth parameters do not always predict the data the best, but often do when using metrics like RMSE or log-likelihood. Do the parameter sets that are not predicted well by ground truth share a common pattern? Does the relative performance of ground truth correlate with high or low values of prior, learn, guess or slip in the generating parameters?

7.6 Results

Seeing log-likelihood based and RMSE metrics score the ground truth model at the same level of mean rank, we are wondering whether, across all 100 of generating parameter sets, the data produced by the same sets of parameters is equally hard to predict with ground truth model. For that we looked at whether the BKT parameter values correlate with ranks ground truth model receives on the moment of learning MAD metric.

First of all, moment of learning MAD metric ranked ground truth as best only 33/100 times with an average rank of 2.53/16. Correlations of moment of learning MAD ranks for ground truth models showed that theres a small marginally significant effect of pInit probability on the moment of learning MAD score ($r = 0.18$, $p - val = 0.07$). Guessing probability does not correlates with moment of learning MAD ($r = -.06$, $p - val = 0.55$).

Probability of learning and slip probability, however, are very strongly related to the moment of learning metric. The larger the learning rate of a simulated skill is, the higher the rank of the ground truth model is ($r = 0.68$, $p - val < 0.001$). Namely, the faster the skill is learned, the worse job ground truth model is doing. In the case of pSlip, the relation is the opposite: the higher the guess rate is, the higher rank moment of learning MAD assigns to the ground truth model ($r = -0.52$, $p - val < 0.001$).

Both the pLearn and pSlip parameters are controlling the process of skills transitioning into the learned state. Strong negative correlation of moment of learning MAD and pSlip is quite logical. Higher pSlip results in more errors even when the skill is mastered, as a result the transition to the learned state becomes more blurred. In this situation the ground truth model has an edge over other models. However, it is high to explain a high positive correlation of moment of learning MAD and pLearn. Higher pLearn means more correct responses overall, this should put ground truth model at an advantage. Additional investigation is necessary to address this phenomenon.

8 Discussion

In our first experiment we found that three less commonly used accuracy metrics showed the best correspondence to accuracy of moment of learning estimation. These metrics were: recall, F-measure, and accuracy, with recall giving a very high correlation of 0.954. Also noteworthy was the poor performance of AUC

with a correlation of -0.119. This was the worst correlation and suggests that AUC should not be used to determine the relative goodness of models based on prediction performance if the underlying goal is to rank models based on knowledge estimation goodness. Metrics like recall and F-measure ought to be adopted in place of AUC for these purposes.

We also found that ground truth model parameters did not always perform the best and that RMSE and log-likelihood based metrics tended to predicted ground truth being the best parameter set more than the others. AUC, recall, F-measure, and precision, however, were among the worst. Therefore, if the underlying goal of an analysis is to recover ground truth parameters (such as with inferring pedagogical efficacy), RMSE and log-likelihood measures should be used and the aforementioned accuracy metrics should be avoided. The experiments 2 raised the question of why ground truth may not always predict the best experiment 3 indicated that high learning rate and low slip in the generating parameters can prove difficult for mastery prediction.

Overall detecting the moment of learning in the generated data by observing a switch from a string of all 0s (unknown state) to the string of all 1s (known state) is often not easy even when ground truth parameters are used. Especially if guess and slip parameters are larger, several back-and-forths between known and unknown state are common. In the area of ITS it is customary to wait till three correct attempts in a row to be sure student has mastered the underlying skill. In our case, when we assumed the moment of learning is the first time when probability of knowing the skill crosses the 0.95 threshold. Following from recent results on the lag with detecting the moment of learning that occurs in the Bayesian Knowledge Tracing [3], in future, we will experiment with adjustments to our computation of the moment of learning to compensate for this.

References

1. Baker, R.S.J.d., Goldstein, A.B., Heffernan, N.T. (2010) Detecting the Moment of Learning. Proceedings of the 10th Annual Conference on Intelligent Tutoring Systems, 25-34.
2. Corbett, A. T. and Anderson, J. R.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4(4), 253-278. (1995)
3. Fancsali, S.E., Nixon, T., Ritter, S. (2013) Optimal and Worst-Case Performance of Mastery Learning Assessment with Bayesian Knowledge Tracing. In: Proceedings of the 6th International Conference on Educational Data Mining.
4. Koedinger, K. R., Anderson, J. R., Hadley, W. H., and Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. International Journal of Artificial Intelligence in Education, 8, 3043.
5. Pardos, Z. A., Wang, Q. Y., Trivedi, S. (2012) The real world significance of performance prediction. In Proceedings of the 5th International Conference on Educational Data Mining. Crete, Greece. pp 192-195.