A Comparison of Error Metrics for Learning Model Parameters in Bayesian Knowledge Tracing [*]

Asif Dhanani[†] Seung Yeon Lee[†] Phitchaya Mangpo Phothilimthana[†] Zachary Pardos
University of California, Berkeley
{asifdhanani, sy.lee, mangpo, pardos}@berkeley.edu

ABSTRACT

In the knowledge-tracing model, error metrics are used to guide parameter estimation towards values that accurately represent students' dynamic cognitive state. We compare several metrics, including log-likelihood (LL), RMSE, and AUC, to evaluate which metric is most suited for this purpose. In order to examine the effectiveness of using each metric, we measure the correlations between the values calculated by each and the distances from the corresponding points to the ground truth. Additionally, we examine how each metric compares to the others. Our findings show that RMSE is significantly better than LL and AUC. With more knowledge of effective error metrics for learning parameters in the knowledge-tracing model, we hope that better parameter searching algorithms can be created.

1. INTRODUCTION

In Bayesian Knowledge Tracing (BKT), one of the essential elements is the error metric that is used for learning model parameters: prior, learn, guess, and slip. Choice of a type of error metric is crucial because the error metric takes a role of guiding the search to the best parameters. The BKT model can be fit to student performance data by using a method which finds a best value calculated from the error metric such as log-likelihood (LL), root-mean-squared error (RMSE), or area under the ROC curve (AUC).

As a modeling method, grid search/brute force [1] is often used to find the set of parameters with optimal values of the error metric, and Expectation Maximization (EM) algorithm [5] is also commonly used to choose parameters maximizing the LL fit to the data. Many studies have compared different modeling approaches [1, 4]. However, the findings are varied across the studies, and it has still been unclear which method is the best at predicting student performance [2].

Pardos and Yudelson compares different error metrics to investigate which one has the most accuracy of estimating the moment of learning [6]. Our work extends this comparison by looking closer into the relationship between three popular error metrics: LL, RMSE, and AUC, and particularly elucidating the relationship to one another closer to the ground truth point.

2. METHODOLOGY

To assess whether LL, RMSE, or AUC is the best error metric to use in parameter searching for the BKT model, we needed datasets with known parameter values in order to compare these with the parameter values predicted by using different error metrics. Therefore, we synthesized 26 datasets by simulating student responses based on diverse known ground truth parameter values.

Correlations to the ground truth. For each dataset, we evaluated LL, RMSE, and AUC values on all points over the entire prior/learn/guess/slip parameter space with a 0.05 interval. On each point, we calculated students' predicted responses (probability that students will answer questions correctly). We then used these predicted responses with the actual responses to calculate LL, RMSE, and AUC for all points. To determine which error metric is the best for this purpose, we looked at the correlations between values calculated from error metrics (i.e. LL, RMSE, and AUC) and the euclidean distances from the points to the ground truth. We applied logarithm to all error metrics other than LL in order to compare everything on the same scale. Finally, we tested whether the correlation between the values calculated by any particular error metric and the distances is significantly stronger than the others' by running one-tailed paired t-tests comparing all three metrics against one another.

Distributions of values. We visualized the values of LL and -RMSE of all points over the 2 dimensional guess/slip space with a 0.02 interval while fixing prior and learn parameter values to the actual ground truth values. Using the guess and slip parameters as the axes, we visualize LL and -RMSE values by color. The colors range from dark red to dark blue corresponding to the values ranging from low to high.

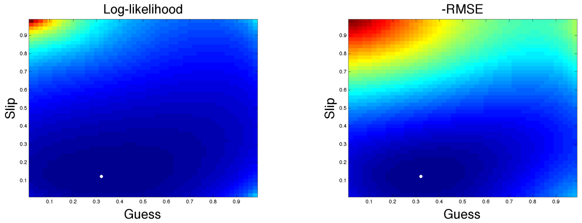
Direct comparison: LL and RMSE. We plotted LL values and RMSE values of all points against each other in order to observe the behavior of the two metrics in detail. We then labeled each data point by its distance to the ground truth with a color. The range of colors is the same as used in the previous method.

[*] For more details of this work, please refer to the full technical report [3].

[†] Asif Dhanani, Seung Yeon Lee, and Phitchaya Mangpo Phothilimthana contributed equally to this work and are listed alphabetically.

Comparision	Δ of correlations	t	p-value
RMSE > LL	0.0408	8.9900	<< 0.0001
RMSE > AUC	0.0844	2.7583	0.0054
LL > AUC	0.0436	1.4511	0.0796

Figure 1: T-test statistics



(a) LL Heatmap

(b) -RMSE Heatmap

Figure 2: LL and -RMSE values when fixing prior and learn parameter values and varying guess and slip parameter values. Red represents low values, while blue represents high values. The white dots represent the ground truth.

3. RESULTS

Correlations to the ground truth. The average LL, RMSE, and AUC correlations were 0.4419, 0.4827, and 0.3983 respectively. We define that an error metric A is *better* than B if the correlation between values calculated by an error metric A and the distances to the ground truth is higher than that of B. By this definition, RMSE was better than LL on all 26 datasets and better than AUC on 18 of 26 datasets. This is validated by the one-tailed paired t-test shown in Figure 1 revealing RMSE as statistically significantly better than both LL and AUC.

Distributions of values. Figure 2 shows the heat maps of LL and RMSE on a representative dataset. If we follow the gradient from the lowest value to the highest value in the LL heat map, we see that it is very high at the beginning (far from the ground truth) and is very low at the end (close to the ground truth). Conversely, in the -RMSE heat map, the change in the gradient is low. Additionally, notice that the darkest blue region in -RMSE heat map is smaller than that in LL heat map. This suggests that we may be able to refine the proximity of the ground truth better with RMSE.

Direct comparison: LL and RMSE. Figure 3 shows a LL vs -RMSE graph from the most representative dataset. As expected, LL values and RMSE values correlate logarithmically. Additionally, a secondary curve, which we will refer to as the *hook*, is observed in varying sizes among datasets. The hook converges with the main curve when the -RMSE and LL values are both sufficiently high and the points are very close to the ground truth.

Before this point, when we look at a fixed LL value with varied RMSE values, most points in the hook have higher -RMSE values and are closer to the ground truth than do the points in the main curve. However, this same pattern is not seen for a fixed RMSE value with varied LL values. After the curve and hook converge, we can infer that both RMSE and LL will give similar estimates of the ground truth. However, for a portion of the graph before this point, RMSE is a better predictor of ground truth values.

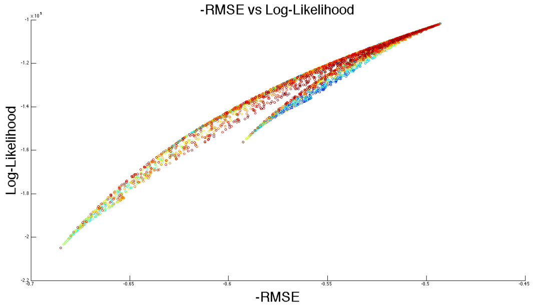


Figure 3: LL vs -RMSE of dataset 25 when prior = 0.564, learn = 0.8, guess = 0.35 , and slip = 0.4

4. CONCLUSION

In our comparison of LL, RMSE, and AUC as metrics for evaluating the closeness of estimated parameters to the true parameters in the knowledge tracing model, we discovered that RMSE serves as the strongest indicator. RMSE has a significantly higher correlation to the distance from the ground truth on average than both LL and AUC, and RMSE is notably better when the estimated parameter value is not very close to the ground truth. The effectiveness of teaching systems without human supervision relies on the ability of the systems to predict the implicit knowledge states of students. We hope that our work can help advance the parameter learning algorithms used in the knowledge tracing model, which in turn can make these teaching systems more effective.

5. REFERENCES

[1] R. Baker, A. Corbett, S. Gowda, A. Wagner, B. MacLaren, L. Kauffman, A. Mitchell, and S. Giguere. Contextual slip and prediction of student performance after use of an intelligent tutor. In *User Modeling, Adaptation, and Personalization*, volume 6075 of *Lecture Notes in Computer Science*. 2010.

[2] R. S. Baker, Z. A. Pardos, S. M. Gowda, B. B. Nooraei, and N. T. Heffernan. Ensembling predictions of student knowledge within intelligent tutoring systems. In *Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization*, 2011.

[3] A. Dhanani, S. Y. Lee, P. Phothilimthana, and Z. Pardos. A comparison of error metrics for learning model parameters in bayesian knowledge tracing. Technical Report UCB/EECS-2014-131, EECS Department, University of California, Berkeley, May 2014.

[4] Y. Gong, J. Beck, and N. Heffernan. Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures. In *Intelligent Tutoring Systems*, volume 6094 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2010.

[5] Z. Pardos and N. Heffernan. Modeling individualization in a bayesian networks implementation of knowledge tracing. In *User Modeling, Adaptation, and Personalization*. 2010.

[6] Z. A. Pardos and M. V. Yudelson. Towards moment of learning accuracy. In *Proceedings of the 1st AIED Workshop on Simulated Learners*, 2013.