Measuring the Moment of Learning with an Information-Theoretic Approach

Brett van de Sande
Arizona State University
PO Box 878809
Tempe, AZ 85287
bvds@asu.edu

ABSTRACT

There are various methods for determining the moment at which a student has learned a given skill. Using the Akaike information criterion (AIC), we introduce an approach for determining the probability that an individual student has learned a given skill at a particular problem-solving step. We then investigate how well this approach works when applied to student log data. Using log data from students using the Andes intelligent tutor system for an entire semester, we show that our method can detect statistically significant amounts of learning, when aggregated over skills or students. In the context of intelligent tutor systems, one can use this method to detect when students may have learned a skill and, from this information, infer the relative effectiveness of any help given to the student or of any behavior in which the student has engaged.

Keywords

data mining, information theory

1. INTRODUCTION

The traditional experimental paradigm for studying student learning is to use a pre-test and post-test combined with two or more experimental conditions. Pre-test and post-test scores can indicate *whether* learning has occurred, but not *when* it may have occurred. At best, one might infer when learning has occurred if an isolated change to the instructional materials or help-giving strategy results in better post-test performance. It is more difficult to infer whether a change in student behavior at some point has resulted in greater learning, since student behavior is largely uncontrolled and must be recorded in some way. In a laboratory setting, these issues can be addressed by careful experimental design, albeit with an accompanying loss of authenticity.

Moving from the laboratory to a more realistic setting, such as a classroom study, presents a challenge since there is necessarily an extended time between any pre-test and post-test. Heckler and Sayre [4] introduce an experimental technique where they administered a test to a different subgroup of students in a large physics class each week during the quarter, cycling through the entire class over the course of the quarter (a between-students longitudinal study). With a sufficiently large number of students (1694 students over five quarters), they were able to produce plots of student mastery of various skills as a function of time, and identify exactly which week(s) students learned a particular skill. However, the shortest time scale that one could imagine for this kind of approach (administering a test in a classroom setting) can, at best, be a day or so. Can we do better?

The use of an intelligent tutor systems (ITS) provides a way forward. In this case, student activity is analyzed and logged for each user interface element change, with a granularity of typically several 10s of seconds. Instead of relying on a distant pre-test or post-test, the experimenter can examine student (or tutor system) activity in the immediate vicinity of the event of interest.

Baker, Goldstein, and Heffernan [1] construct a model that predicts the probability that a student has learned a skill at a particular time based on the Bayesian Knowledge Tracing (BKT) algorithm [3]. BKT gives the probability that the student has mastered a skill at step j using the students performance on previous opportunities to apply that skill. The authors supplement the BKT result with information on student correctness for the two subsequent steps $j+1$ and $j+2$ and infer the probability that the student learned the skill at that step. Finally, they use their model to train a second machine-learned model that does not rely on future student behavior, so it could be run in real time as the student is working

We will address the same problem using an information-theoretic approach. Starting with the Akaike information criterion and a simple model of learning, we use a multi-model strategy to predict the probability that learning has occurred at a given step, and to predict how much learning has occurred. We apply our approach to student log data from an introductory physics course. We find that, for an individual student and skill, detection of learning has large uncertainties. However, if one aggregates over skills or students, then learning can be detected at the desired level of significance.

1.1 Correct/Incorrect steps

Our stated goal is to determine student learning for an individual student as they progress through a course. What observable quantities should be used to determine student mastery? One possible observable is "correct/incorrect steps," whether the student correctly applies a given skill at a particular problem-solving step without any preceding errors or hints. There are other observables that may give us clues on mastery: for instance, how much time a student takes to complete a step that involved a given skill. However, other such observables typically need some additional theoretical interpretation. *Exempli gratia*, What is the relation between time taken and mastery? Baker, Goldstein, and Heffernan [1] develop a model of learning based on a Hidden Markov model approach. They start with a set of 25 additional observables (for instance, "time to complete a step") and construct their model and use correct/incorrect steps (as defined above) to calibrate the additional observables and determine which are statistically significant. Naturally, it is desirable to eventually include such additional observables in any determination of student learning. However, in the present investigation, we will focus on correct/incorrect steps.

What do we mean by a step? A student attempts some number of *steps* when solving a problem. Usually, a step j is associated with creating/modifying a single user interface object (writing an equation, drawing a vector, defining a quantity, *et cetera*) and is a distinct part of the problem solution (that is, help-giving dialogs are not considered to be steps). A student may attempt a particular problem solving step, delete the object, and later attempt that solution step again. A step is an *opportunity* to learn a given Knowledge Component (KC) [6] if the student must apply that KC or skill to complete the step.

For each KC and student, we select all relevant step attempts and mark each step as "correct" (or 1) if the student completes that step correctly without any preceding errors or requests for help; otherwise, we mark the step as "incorrect" (or 0). A single student's performance on a single KC can be expressed as a bit sequence, *exempli gratia* 00101011.

2. THE STEP MODEL

We need to compare the student log data to some sort of model of learning. In another paper [5], we introduced the "step model" and showed that it was competitive with other popular models of learning when applied to individual student log data. It is defined as:

$$P_{\text{step}}(j) = \begin{cases} g, & j < L \\ 1 - s, & j \geq L \end{cases} \tag{1}$$

where L is the step where the student first shows mastery of the KC, g is the "guess rate," the probability that the student gets a step correct by accident, and s is the "slip rate," the chance that the student makes an error after learning the skill. These are analogous to the guess and slip parameters of BKT [3]. This model assumes that learning occurs all at once, reminiscent of "eureka learning" discussed by [1].

2.1 Method

We examined log data from 12 students taking an intensive introductory physics course at St. Anselm College during summer 2011. The course covered the same content as a normal two-semester introductory course. Log data was recorded as students solved homework problems while using the Andes intelligent tutor homework system [7]. 231 hours of log data were recorded. Each step was assigned to one or more different KCs. The dataset contains a total of 2017 distinct student-KC sequences covering a total of 245 distinct KCs. We will refer to this dataset as student dataset $\mathcal{A}$. See Figure 1 for a histogram of the number student-KC sequences having a given number of steps.

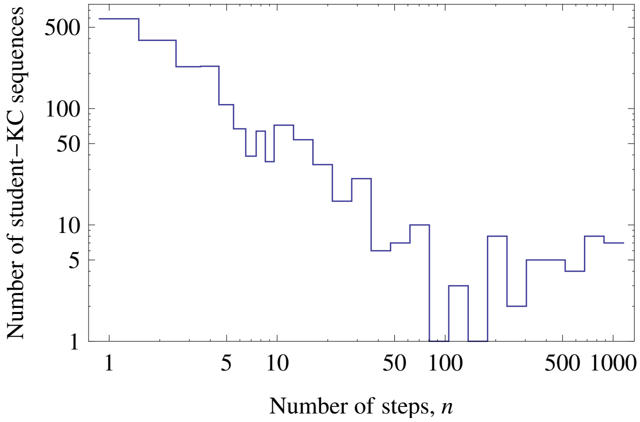

Figure 1: Histogram of number of distinct student-KC sequences in student dataset $\mathcal{A}$ having a given number of steps n.

Most KCs are associated with physics or relevant math skills while others are associated with Andes conventions or user-interface actions (such as, notation for defining a variable). The student-KC sequences with the largest number of steps are associated with user-interface related skills, since these skills are exercised throughout the entire course.

One of the most remarkable properties of the distribution in Fig. 1 is the large number of student-KC sequences containing just a few steps. The presence of many student-KC sequences with just one or two steps may indicate that the default cognitive model associated with this tutor system may be sub-optimal; there has not been any attempt, to date, to improve on the cognitive model of Andes with, say, Learning Factors Analysis [2]. Another contributing factor is the way that introductory physics is taught in most institutions, with relatively little repetition of similar problems. This is quite different than, for instance, a typical middle school math curriculum where there are a large number of similar problems in a homework assignment.

3. MULTI-MODEL APPROACH

We need to determine the step where a specific student has learned a particular skill. Our strategy is to take the step model, $P_{\text{step}}(j)$, and treat L as a constant, yielding a set of n sub-models $P_{\text{step},L}(j)$, one for each value of L. We then fit each of the n sub-models to the student data and calculate an AIC value. Finally, we find the Akaike weighs for each of the sub-models. The Akaike weights give the relative probability that learning occurred at each step.

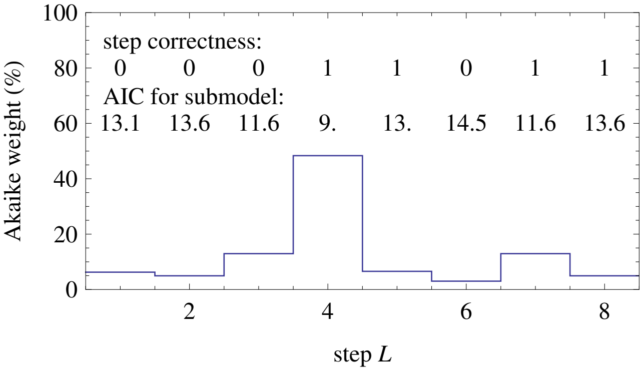

Figure 2: Akaike weights for the sub-models $P_{\mathrm{step,L}}(j)$. This gives the relative probability that the student learned the KC just before step L. The case $L = 1$ corresponds to no learning occurring during use of the ITS.

Let us illustrate this technique with a simple example. Suppose the bit sequence for a particular student-KC sequence is 00011011 (8 opportunities); see Fig. 2. We fit this bit sequence to 8 sub-models of the step model, corresponding to $L \in \{1, 2, \ldots, 8\}$, by maximizing the log likelihood, $\log \mathcal{L}_L$. The associated AIC values are given by $\mathrm{AIC}_L = 2K - \log \mathcal{L}_L$ where K is the number of fit parameters. Note that there are two parameters (s and g) when $L > 1$ and there is only one parameter (s) when $L = 1$. Not surprisingly, the best fit (lowest AIC) corresponds to the first "1" in the bit sequence at step 4. From the AICs, we calculate the Akaike weights

$$w_L = \frac{e^{-\mathrm{AIC}_L/2}}{\sum_{L'} e^{-\mathrm{AIC}_{L'}/2}} \,. \tag{2}$$

The Akaike weight w_L gives the relative probability that sub-model $P_{\mathrm{step},L}(j)$ is, of all the sub-models, the closest to the the model that actually generated the data.

Note that the case $L = 1$ corresponds to the student having "learned the skill" some time before the first step or after the last step. That is to say, the student does not acquire the skill while using the tutor system. Thus, w_1 should be interpreted as the relative probability that no learning has occurred while using the tutor system.

4. WEIGHTED GAIN

Our ultimate goal is to distinguish steps that result in learning from steps that do not. Hopefully, one can use this information to infer something about the effectiveness of the help given on a particular step, or the effectiveness of the student activity on that step.

It is not sufficient to know *when* learning has occurred but one must also determine *how much* learning has occurred. Consider the bit sequence 11011000. When fit to the step model, the best fit will occur at $L = 6$ but this would correspond to a *decrease* in student performance for that skill. In many cases seen in our log data, the change in student performance is almost zero. In order to take this into account, we propose using the Akaike weight w_L times the associated performance gain Δ_L to characterize a step. We define the performance gain $\Delta_L = 1 - \hat{g} - \hat{s}$ where $\hat{g}$ and $\hat{s}$ are the Maximum Likelihood estimators for g and s given by sub-model $P_{\mathrm{step},L}(j)$. For the "no learning" case $L = 1$, we set $\Delta_1 = 0$. We will call $w_L \Delta_L$ the "weighted gain" associated with $P_{\mathrm{step},L}(j)$. A calculation of $w_L \Delta_L$ for an example bit sequence is shown in Fig. 3. Not surprisingly, the largest gain occurs at $L = 4$, corresponding to the first 1 in the bit sequence. The remaining weighted gains are much smaller.

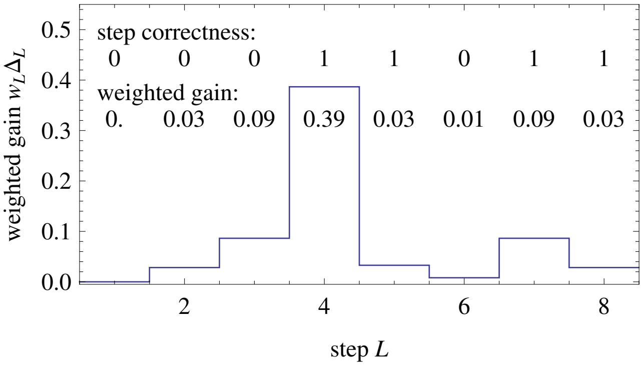

Figure 3: Weighted gain $w_L \Delta_L$ as a function of L for an example bit sequence. The associated quality factor is $Q = 0.66 \pm 0.29$; it is significantly smaller than 1 since the student makes a slip on step 6. Q is significantly greater than zero at the $p = 0.01$ level.

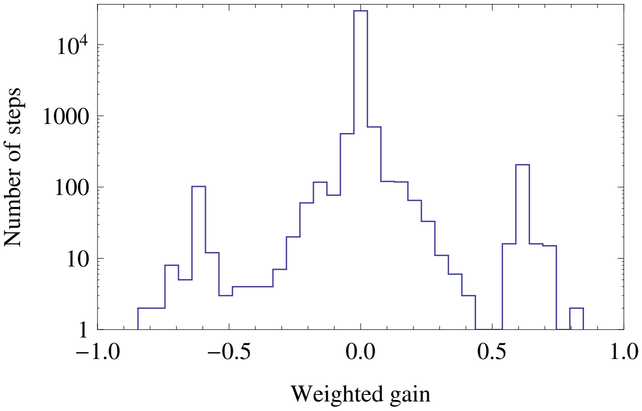

Figure 4: Histogram of weighted gains $w_L \Delta_L$ for all steps in all student-KC sequences of dataset $\mathcal{A}$.

A histogram of $w_L \Delta_L$ for student dataset $\mathcal{A}$ is shown in Fig. 4. We find that the vast majority of steps (29730) have almost zero weighted gain. We also see that there is a significant number of steps with negative gain (988), but there are somewhat more steps with positive gain (1312) .

The fact that there are so many steps with negative gain is symptomatic of bit sequences that are very noisy (a lot of randomness). Indeed, if we compare the histogram for student dataset $\mathcal{A}$ with the histogram for a randomly generated dataset $\mathcal{R}$ (we take $\mathcal{A}$ and randomly permute the steps) we find a similar distribution; see Fig. 5.

What would the distribution look like if the data weren't so noisy? To see this, we generated an artificial "ideal" dataset $\mathcal{I}$ where there were no slips or guesses, but having the same

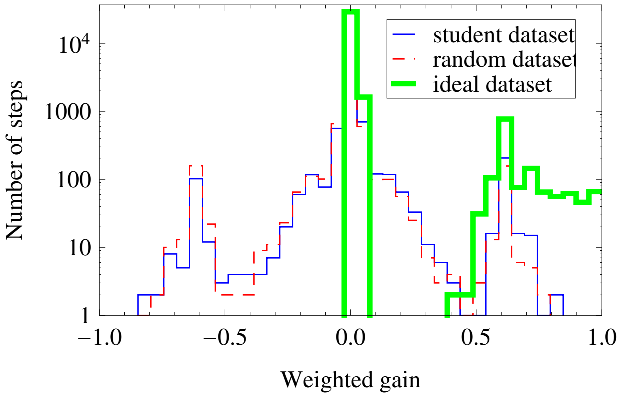

Figure 5: Histogram of weighted gains $w_L\Delta_L$ for the student dataset $\mathcal{A}$, a randomly generated dataset $\mathcal{R}$, and an artificial ideal dataset $\mathcal{I}$.

length distribution as $\mathcal{A}$ (Fig. 1). Thus, the bit sequences in $\mathcal{I}$ all have the form $00\cdots011\cdots1$. In this case, for each student-KC sequence, we expect a single large weighted gain (corresponding to the first 1 in the bit sequence) and the remaining weighted gains to be nearly zero. The resulting distribution of gains is shown in Fig. 5.

We propose to use the following average of the weighted gains as a "quality index" for determining how suitable a dataset is for determining the point of learning for an individual student-KC sequence:

$$Q = \frac{1}{N}\sum_{\alpha}\sum_{L} w_L \Delta_L \tag{3}$$

where α is an index running over all N student-KC sequences in a dataset. We use the sample standard deviation of the weighted gains $w_L\Delta_L$ to calculate the standard error associated with Q. An example calculation is shown in Fig. 3.

For the random dataset $\mathcal{R}$, the distribution of Δ_L is symmetric about zero and Q approaches zero as $N \to \infty$. For the "ideal" dataset $\mathcal{I}$, we expect that, when L coincides with the first 1 in the bit sequence, w_L will be nearly one with the associated Δ_L also nearly one so that $Q \to 1$ in the limit of many opportunities. Numerically, we obtain $Q = 0.5240 \pm 0.0003$. The fact that it is smaller than one is due to the large number of student-KC sequences having just a few steps. For the student dataset $\mathcal{A}$, we obtain $Q = 0.0467 \pm 0.0065$, which is small, but significantly larger than zero ($p < 0.001$). Thus, we conclude that one can detect statistically significant learning when applying our method to this student dataset, with the location of that learning given by the Akaike weights w_L.

5. CONCLUSION

We believe that a direct estimate of the moment when a student learns a skill could be very useful for improving instruction, improving help-giving, and understanding student learning. However, the question of whether learning has occurred at a particular step can only be answered in a probabilistic sense: unambiguous "Aha moments" seem to be relatively rare. Using the Akaike Information Criterion, we have introduced a method for determining this probability.

As can be seen in Fig. 5, there is not much difference between our student dataset and a randomly generated dataset. However, the quality index Q which can be used to quantify the size of the signal of learning as well as the size of the background. We see that the quality index $Q = 0.0467 \pm 0.0065$ for the student dataset $\mathcal{A}$ is roughly 10% the size of Q for the ideal dataset $\mathcal{I}$; we interpret this to mean that the "signal" for learning is roughly 10% as big as the "noise." However, the fact that Q for the student dataset is seven standard deviations from zero means that we have detected learning for 2000 student-KC sequences with room to spare. Using the fact that the error is proportional $1/\sqrt{N}$, where N is the number of student-KC sequences, we estimate that we could still detect learning with only 260 student-KC sequences at the $p = 0.01$ level. This gives us an initial estimate for the amount of log data needed to measure the moment of learning, at least for students using the Andes tutor system.

Finally, we see that many of the student-KC sequences are quite short, as shown in Fig. 1. We speculate that this is due to to the way that introductory physics is typically taught, with relatively little reinforcement of specific KCs, emphasizing, instead, more general problem solving meta-skills. If we were to repeat this analysis for high school or grade school math, where there is more repetition, we speculate that there would be significantly fewer KCs with less than 10 opportunities and that detecting when learning has occurred would be significantly easier.

6. ACKNOWLEDGMENTS

Funding for this research was provided by the Pittsburgh Science of Learning Center which is funded by the National Science Foundation award No. SBE-0836012.

7. REFERENCES

[1] R. S. J. D. Baker, A. B. Goldstein, and N. T. Heffernan. Detecting learning moment-by-moment. *Int. J. Artif. Intell. Ed.*, 21(1-2):5–25, Jan. 2011.

[2] H. Cen, Kenneth Koedinger, and B. Junker. Learning factors analysis - a general method for cognitive model evaluation and improvement. In *Proceedings of the 8th international conference on Intelligent Tutoring Systems*, pages 164–175, Jhongli, Taiwan, June 2006. Springer-Verlag Berlin, Heidelberg.

[3] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Model. User-Adapt. Interact.*, 4(4):253–278, 1995.

[4] A. F. Heckler and E. C. Sayre. What happens between pre- and post-tests: Multiple measurements of student understanding during an introductory physics course. *Am. J. Phys.*, 78(7):768, 2010.

[5] B. van de Sande. Applying three models of learning to individual student log data. Under review, 2013.

[6] K. VanLehn. The behavior of tutoring systems. *Int. J. Artif. Intell. Ed.*, 16(3):227–265, Jan. 2006.

[7] K. Vanlehn, C. Lynch, K. Schulze, J. A. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill. The andes physics tutoring system: Lessons learned. *Int. J. Artif. Intell. Ed.*, 15(3):147–204, Aug. 2005.