Learning Representations of Student Knowledge and Educational Content

Siddharth Reddy SGR45@CORNELL.EDU
Department of Computer Science, Cornell University, Ithaca, NY USA
Knewton, Inc., 100 5th Avenue., New York, NY 10011 USA

Igor Labutov IIL4@CORNELL.EDU
Department of Electrical and Computer Engineering, Cornell University, Ithaca, NY USA

Thorsten Joachims TJ@CS.CORNELL.EDU
Department of Computer Science, Cornell University, Ithaca, NY USA

Abstract

Students in online courses generate large amounts of data that can be used to personalize the learning process and improve quality of education. The goal of this work is to develop a statistical model of students and educational content that can be used for a variety of tasks, such as adaptive lesson sequencing and learning analytics. We formulate this problem as a regularized maximum-likelihood embedding of students, lessons, and assessments from historical student-module interactions. Akin to collaborative filtering for recommender systems, the algorithm does not require students or modules to be described by features, but it learns a representation using access traces. An empirical evaluation on large-scale data from Knewton, an adaptive learning technology company, shows that this approach predicts assessment results more accurately than baseline models and is able to discriminate between lesson sequences that lead to mastery and failure.

1. Introduction

The popularity of online education platforms has soared in recent years. Companies like Coursera and EdX offer Massive Open Online Courses (MOOCs) that attract millions of students and high-calibre instructors. Khan Academy has become a hugely popular repository of videos and interactive materials on a wide range of subjects. E-learning products offered by universities and textbook publishers are also gaining traction. These platforms improve access to high quality educational content for anyone connected to the Internet. As a result, people who would otherwise lack the opportunity are able to consume materials like video lectures and problem sets from courses offered at top universities. However, in these online environments learners often lack the personalized instruction and coaching that can potentially lead to significant improvements in educational outcomes. Furthermore, the educational content may be contributed by many authors without a formal underlying structure. Intelligent systems that learn about the educational properties of the content, guide learners through custom lesson plans, and quickly adapt through feedback could help learners take advantage of large and heterogeneous collections of educational content to achieve their goals.

The extensive literature on intelligent tutoring systems (ITS) and computer-assisted instruction (CAI) dates back to the 1960s. Early efforts focused on approximating the behavior of a human tutor through rule-based systems that taught students South American geography (Carbonell, 1970), electronics troubleshooting (Lesgold et al., 1988), and programming in Lisp (Corbett & Anderson, 1994). Today's online education platforms differ from early ITSes in their ability to gather data at scale, which facilitates the use of machine learning techniques to improve the educational experience. Relatively little academic work has been done to design systems that use the massive amounts of data generated by students in online courses to provide personalized learning tools. Learning and content analytics (Lan et al., 2014c), instructional scaffolding in educational games (ORourke et al., 2015), hint generation (Piech et al., 2015b), and feedback propagation (Piech et al., 2015a) are a few topics currently being explored in the personalized learning space.

Our aim is to build a domain-agnostic framework for mod-

Proceedings of the 31^{st} *International Conference on Machine Learning*, Lille, France, 2015. JMLR: W&CP volume 37.

eling students and content that can be used in many online learning products for applications such as adaptive lesson sequencing, learning analytics, and content analytics. A common data source available in products is a stream of interaction data, or *access traces* that log student interactions with modules of course content. These access traces come in the form *Student A completed Lesson B* and *Student C passed assessment D*. *Lessons* are content modules that introduce or reinforce concepts; for example, an animation of cellular respiration or a paragraph of text on Newton's first law of motion. *Assessments* are content modules with pass-fail results that test student skills; for example, a true-or-false question halfway through a video lecture. By relying on a coarse-grained, binary assessment result, we are able to gracefully handle many types of assessments (e.g., free response and multiple choice) so long as a student response can be labelled as correct or incorrect.

We use access traces to embed students, lessons, and assessments together in a joint semantic space, yielding a representation that can be used to reason about the relationship between students and content (e.g., the likelihood of passing an assessment, or the skill gains achieved by completing a lesson). The model is evaluated on synthetic data, as well as large-scale real data from Knewton, an education technology company that offers personalized recommendations and activity analytics for online courses (Knewton, 2015). The data set consists of 2.18 million access traces from over 7,000 students, recorded in 1,939 classrooms over a combined period of 5 months.

2. Related Work

Our work builds on the existing literature in psychometric user modeling. The Rasch model estimates the probability of a student passing an assessment using latent concept proficiency and assessment difficulty parameters (Rasch, 1993). The two-parameter logistic item response theory (2PL IRT) model adds an assessment discriminability parameter to the result likelihood (Linden & Hambleton, 1997). Both models assume that a map from assessments to a small number of underlying concepts is known a priori. We propose a data-driven method of learning content representation that does not require a priori knowledge of content-to-concept mapping. Though this approach sacrifices the interpretability of expert ratings, it has two advantages: 1) it does not require labor-intensive expert annotation of content and 2) it can evolve the representation over time as existing content is modified or new content is introduced.

Lan et al. propose a sparse factor analysis (SPARFA) approach to modeling graded learner responses that uses assessment-concept associations, concept proficiencies, and assessment difficulty (Lan et al., 2014c). The algorithm does not rely on an expert concept map, but instead learns assessment-concept associations from the data. Multi-dimensional item response theory (Reckase, 2009) also learns these assocations from the data. We extend the ideas behind SPARFA and multi-dimensional item response theory to include a model of student learning from lesson modules, which is a key prerequisite for recommending personalized lesson sequences.

Bayesian Knowledge Tracing (BKT) uses a Hidden Markov Model to model the evolution of student knowledge (which is discretized into a finite number of states) over time (Corbett & Anderson, 1994). Further work (González-Brenes et al., 2014) has modified the BKT framework to include the effects of lessons (described by features) through an input-output Hidden Markov Model. Similarly, SPARFA has been extended to model time-varying student knowledge and the effects of lesson modules (Lan et al., 2014a). Item response theory has also been extended to capture temporal changes in student knowledge (Ekanadham & Karklin, 2015; Sohl-Dickenstein, 2014). Recurrent neural networks have been used to trace student knowledge over time and model lesson effects (Piech et al., 2015c). Similar ideas for estimating temporal student knowledge from binary-valued responses have appeared in the cognitive modeling literature (Prerau et al., 2008; Smith et al., 2004). We extend this work in a multi-dimensional setting where student knowledge lies in a continuous state space and lesson prerequisites modulate knowledge gains from lesson modules.

Our model also builds on previous work that uses temporal embeddings to predict music playlists (Moore et al., 2013). While Moore et al. focused on embedding objects (songs) in a metric space, we propose a non-metric embedding where the distances between objects (students, assessments, and lessons) are not symmetric, capturing the natural progression in difficulty of assessments and the positive growth of student knowledge.

3. Embedding Model

We now describe the probabilistic embedding model that places students, lessons, and assessments in a joint semantic space that we call the *latent skill space*. Students have trajectories through the latent skill space, while assessments and lessons are placed at fixed locations. Formally, a student is represented as a set of d latent skill levels $\vec{s} \in \mathbb{R}_+^d$; a lesson module is represented as a vector of skill gains $\vec{\ell} \in \mathbb{R}_+^d$ and a set of prerequisite skill requirements $\vec{q} \in \mathbb{R}_+^d$; an assessment module is represented as a set of skill requirements $\vec{a} \in \mathbb{R}_+^d$.

Students interact with lessons and assessments in the following way. First, a student can be tested on an assessment

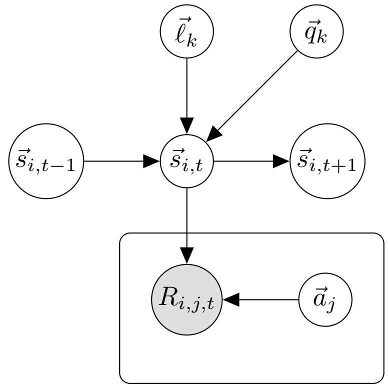

Figure 1. A graphical model of student learning and testing, i.e. a continuous state space Hidden Markov Model with inputs and outputs. $\vec{s}$ = student knowledge state, $\vec{\ell}$ = lesson skill gains, $\vec{q}$ = lesson prerequisites, $\vec{a}$ = assessment requirements, and R = result. Assessments can be completed by different students multiple times, each resulting in a separate result. Students can complete different assessments multiple times, each resulting in a separate result. Lessons can be completed by different students multiple times. Students depend on their knowledge state at the previous timestep.

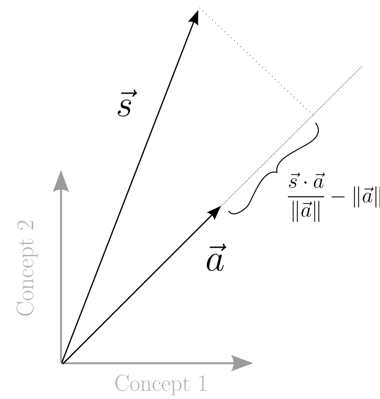

Figure 2. Geometric intuition underlying the parametrization of the assessment result likelihood (Equation 1). Only the length of the projection of the student's skills $\vec{s}$ onto the assessment vector $\vec{a}$ affects the pass likelihood of that assessment, meaning only the "relevant" skills (with respect to the assessment) should determine the result.

module with a pass-fail result $R \in \{\pm 1\}$, where the likelihood of passing is high when a student has skill levels that exceed the assessment requirements and vice-versa. Second, a student can work on lesson modules to improve skill levels over time. To fully realize the skill gains associated with completing a lesson module, a student must satisfy prerequisites (fulfilling some of the prerequisites to some extent will result in relatively small gains, see Equation 3 for details). Time is discretized such that at every timestep $t \in \mathbb{N}$, a student completes a lesson and may complete zero or many assessments. The evolution of student knowledge can be formalized as the graphical model in Figure 1, and the following subsections elaborate on the details of this model.

3.1. Modeling Assessment Results

For student $\vec{s}$, assessment $\vec{a}$, and result R,

$$Pr(R = r) = \frac{1}{1 + \exp\left(-r \cdot \Delta(\vec{s},\ \vec{a})\right)} \qquad (1)$$

where $r \in \{\pm 1\}$ and $\Delta(\vec{s},\ \vec{a}) = \frac{\vec{s}\cdot\vec{a}}{\|\vec{a}\|} - \|\vec{a}\| + \gamma_s + \gamma_a$. $\vec{s}$ and $\vec{a}$ are constrained to be non-negative (for details see Section 4). A pass result is indicated by $r = 1$, and a fail by $r = -1$. The term $\frac{\vec{s}\cdot\vec{a}}{\|\vec{a}\|}$ can be rewritten as $\|\vec{s}\|\cos(\theta)$, where θ is the angle between $\vec{s}$ and $\vec{a}$; it can be interpreted as "relevant skill". The term $\|\vec{a}\|$ can be interpreted as general (i.e. not concept-specific) assessment difficulty. The expression $\frac{\vec{s}\cdot\vec{a}}{\|\vec{a}\|} - \|\vec{a}\|$ is visualized in Figure 2. The bias term γ_s is a student-specific term that captures a student's general (assessment-invariant and time-invariant) ability to pass; it can be interpreted as a measure of how well the student guesses correct answers. The bias term γ_a is a module-specific term that captures an assessment's general (student-invariant and time-invariant) difficulty. γ_a differs from the $\|\vec{a}\|$ difficulty term in that it is not bounded; see Section 4 for details. These bias terms are analogous to the bias terms used for modeling song popularity in (Chen et al., 2012).

Our choice of Δ differs from classic multi-dimensional item response theory, which uses $\Delta(\vec{s}, \vec{a}) = \vec{s}\cdot\vec{a} + \gamma_a$ where s and a are not bounded (although in practice, suitable priors are imposed on these parameters). We have also chosen to use the logistic link function instead of the normal ogive.

3.2. Modeling Student Learning from Lessons

For student $\vec{s}$ who worked on a lesson with skill gains $\vec{\ell}$ and no prerequisites at time $t+1$, the updated student state is

$$\vec{s}_{t+1} \sim \mathcal{N}\left(\vec{s}_t + \vec{\ell},\ \mathbf{\Sigma}\right) \qquad (2)$$

where the covariance matrix $\mathbf{\Sigma} = I_d\sigma^2$ is diagonal. For a lesson with prerequisites $\vec{q}$,

$$\vec{s}_{t+1} \sim \mathcal{N}\left(\vec{s}_t + \vec{\ell} \cdot \frac{1}{1 + \exp\left(-\Delta(\vec{s}_t,\ \vec{q})\right)},\ \mathbf{\Sigma}\right) \qquad (3)$$

where $\Delta(\vec{s}_t,\ \vec{q}) = \frac{\vec{s}_t\cdot\vec{q}}{\|\vec{q}\|} - \|\vec{q}\|$. The intuition behind this equation is that the skill gain from a lesson should be

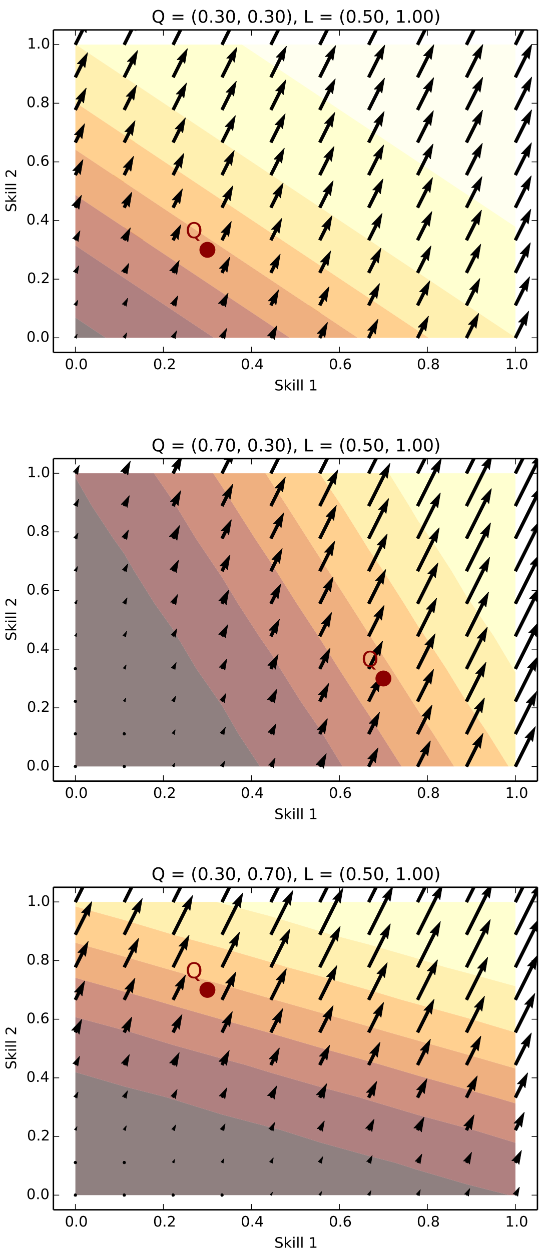

Figure 3. $\vec{\ell}$ = skill gains, $\vec{q}$ = prerequisites. Here, we illustrate the vector field of skill gains possible for different students under different lesson prerequisites. A student can compensate for lack of prerequisites in one skill through excess strength in another skill, but the extent to which this trade-off is possible depends on the lesson prerequisites.

weighted according to how well a student satisfies the lesson prerequisites. A student can compensate for lack of prerequisites in one skill through excess strength in another skill, but the extent to which this trade-off is possible depends on the lesson prerequisites. The same principle applies to satisfying assessment skill requirements. Figure 3 illustrates the vector field of skill gains possible for different students under different lesson prerequisites. Without prerequisites, the vector field is uniform.

Our model differs from (Lan et al., 2014a) in that we explicitly model the effects of prerequisite knowledge on gains from lessons. Lan et al. model gains from a lesson as an affine transformation of the student's knowledge state plus an additive term similar to $\vec{\ell}$.

4. Parameter Estimation

We compute maximum-likelihood estimates of model parameters Θ by maximizing the following objective function:

$$\begin{aligned} L(\Theta) = \sum_{\mathcal{A}} \log\left(Pr(R \mid \vec{s}_t, \vec{a}, \gamma_s, \gamma_a)\right) \\ + \sum_{\mathcal{L}} \log\left(Pr(\vec{s}_{t+1} \mid \vec{s}_t, \vec{\ell}, \vec{q})\right) - \beta \cdot \lambda(\Theta) \end{aligned} \tag{4}$$

where $\mathcal{A}$ is the set of assessment interactions, $\mathcal{L}$ is the set of lesson interactions, $\lambda(\Theta)$ is a regularization term that penalizes the L_2 norms of embedding parameters (not bias terms), and β is a regularization parameter. Non-negativity constraints on embedding parameters (not bias terms) are enforced.

L_2 regularization is used to penalize the size of embedding parameters to prevent overfitting. The bias terms are not bounded or regularized. This allows $-\|\vec{a}\| + \gamma_a$ to be positive for assessment modules that are especially easy, and $\frac{\vec{s}\cdot\vec{a}}{\|\vec{a}\|} + \gamma_s$ to be negative for students who fail especially often. We solve the optimization problem with box constraints using the L-BFGS-B (Zhu et al., 1997) algorithm. We randomly initialize parameters and run the iterative optimization until the relative difference between consecutive objective function evaluations is less than 10^{-3}. Averaging validation accuracy over multiple runs during cross-validation reduces sensitivity to the random initializations (since the objective function is non-convex).

5. Experiments on Synthetic Data

To verify the correctness of our model and to illustrate the properties of the embedding geometry that the model captures, we conducted a series of experiments on small, synthetically-generated interaction histories. Each scenario is intended to demonstrate a different feature of the model (e.g., recovering student knowledge and assessment requirements in the absence of lessons, or recovering sensible skill gain vectors for different lessons). For the sake of simplicity, the embeddings are made without bias terms. The figures shown are annotated versions of plots made by our embedding software.

5.1. Recovering Assessments

The embedding can recover intuitive assessment requirements and student skill levels from small histories without lessons. See Figures 7 and 8 for examples. These examples also illustrate the fact that a multidimensional embedding is more expressive than a one-dimensional embedding, i.e. increasing the embedding dimension improves the model's ability to capture the dynamics of complicated scenarios.

5.2. Recovering Lessons

The embedding can recover orthogonal skill gain vectors for lessons that deal with two different skills. See Figure 9 for an example in two dimensions.

The embedding can recover sensible lesson prerequisites. In Figure 10, we recover prerequisites that explain a scenario where a strong student realizes significant gains from a lesson while weaker students realize nearly zero gains from the same lesson.

6. Experiments on Online Course Data

We use data processed by Knewton, an adaptive learning technology company. Knewton's infrastructure uses student-module access traces to generate personalized recommendations and activity analytics for partner organizations with online learning products. The data describes interactions between college students and two science textbooks. Both data sets are filtered to eliminate students with fewer than five lesson interactions and content modules with fewer than five student interactions. To avoid spam interactions and focus on the outcomes of initial student attempts, we only consider the first interaction between a student and an assessment (subsequent interactions between student and assessment are ignored). Each content module in our data has both a lesson and assessment component. To simulate a more general setting where the sets of lesson and assessment modules are disjoint, each module is randomly assigned to the set of lessons or the set of assessments. We performed our experiments several times using different random assignments, and observed that the results did not change significantly.

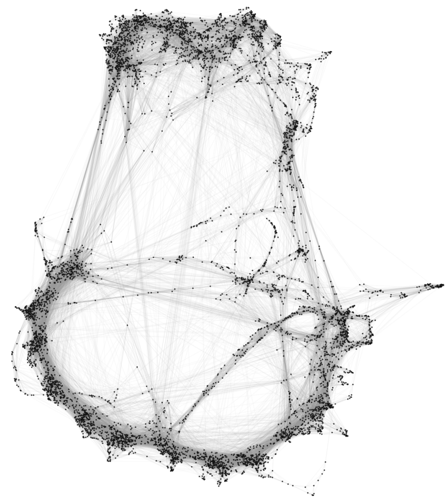

Figure 4. The directed graph of module sequences for all students in *Book A*. V = set of assessment and lesson modules, $(n_1, n_2) \in E$ with edge weight k if k students had a transition from module n_1 to n_2 in their access traces. This graph shows that module sequences are heavily directed by the order imposed by the textbook, teachers, or a recommender system.

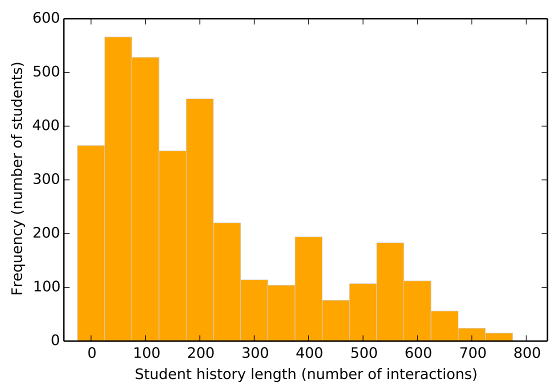

Figure 5. Distribution of the lengths of student histories in the *Book A* data set

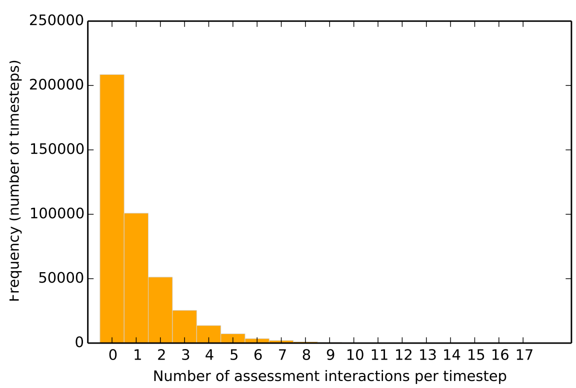

Figure 6. Distribution of the number of assessment interactions at each timestep in the *Book A* data set. Long chains of assessment interactions are rare, implying that students complete lesson modules with consistent frequency.

							Book A				*Book B*			
							RANDOM		LAST		RANDOM		LAST	
	d	s	a	ℓ	q	γ	TRAINING	TEST	TRAINING	TEST	TRAINING	TEST	TRAINING	TEST
1	1	Y	N	N	N	N	0.659	0.605	0.659	0.626	0.723	0.696	0.723	0.686
2	1	N	Y	N	N	N	0.743	0.734	0.743	0.737	0.716	0.712	0.715	0.661
3	2	Y	Y	N	N	N	0.994	0.687	0.994	0.687	0.992	0.659	0.991	0.614
4	2	Y	Y	N	N	Y	0.995	0.673	0.994	0.705	0.993	0.745	0.992	0.726
5	2	Y	Y	Y	N	N	0.900	0.713	0.897	0.730	0.896	0.733	0.897	0.696
6	2	Y	Y	Y	N	Y	0.873	0.741	0.872	0.756	0.869	0.787	0.871	0.748
7	2	Y	Y	Y	Y	N	0.898	0.709	0.898	0.727	0.902	0.742	0.898	0.694
8	2	Y	Y	Y	Y	Y	0.890	0.732	0.889	0.753	0.872	0.787	0.882	0.749

Table 1. We conduct a lesion analysis to gain insight into which components of an embedding contribute most to prediction accuracy (AUC)

6.1. Data

Here, we give summary statistics of the data sets we use to evaluate our model.

6.1.1. *Book A*

This data set was collected from 869 classrooms from January 1, 2014 through June 1, 2014. It contains 834,811 interactions, 3,471 students, 3,374 lessons, 3,480 assessments, and an average assessment pass rate of 0.712. See Figures 4, 5, and 6 for additional data exploration.

6.1.2. *Book B*

This data set was collected from 1,070 classrooms from January 1, 2014 through June 1, 2014. It contains 1,349,541 interactions, 3,563 students, 3,843 lessons, 3,807 assessments, and an average assessment pass rate of 0.693.

6.2. Assessment Result Prediction

We evaluate the embedding model on the task of predicting results of held-out assessment interactions using the following scheme:

- We use ten-fold cross-validation to compute the validation accuracies of variations of the embedding model
- On each fold, we train on the full histories of 90% of students and the truncated histories of 10% of students, and validate on the assessment interactions immediately following the truncated histories
- Truncations are made at random locations in student histories, or just before the last batch of assessment interactions for a given student (maximizing the size of the training set)

We start by constructing two baseline models that predict the student pass rate and the assessment pass rate respectively. The student baseline is equivalent to a one-dimensional embedding of only students that is not regularized or bounded (see row 1 of Table 1). Analogously, the assessment baseline is equivalent to a one-dimensional embedding of only assessments that is not regularized or bounded (see row 2 of Table 1). We then gradually add components to the embedding model to examine their effects on prediction accuracy. Initially, we omit lessons and bias terms from the embedding, and only consider students and assessments. We progressively include lesson parameters $\vec{\ell}$ without prerequisites, prerequisite parameters $\vec{q}$ for lessons, and bias terms γ. Each variant of the model corresponds to a row in Table 1. Our performance metric is area under the ROC curve (AUC), which measures the discriminative ability of a binary classifier that assigns probabilities to class membership.

Table 1 shows AUC performance on the assessment result prediction task for different data sets (*Book A* and *Book B*), student history truncation styles (random and last), and models. All embeddings on both data sets use the default parameter values $\sigma^2 = 0.5$ and $\beta = 10^{-6}$, which were selected in exploratory experiments. All p-values are computed using Welch's t-test (Welch, 1947) for the results in column (*Book A*, Last, Test). From these results, we observe the following:

- Including lessons in the embedding improves performance significantly ($p = 0.003$ for rows 3 vs. 5, $p = 0.0005$ for 4 vs. 6)
- Including prerequisites gives a modest performance gain on *Book B* and a statistically insignificant decrease in performance on *Book A* ($p = 0.913$ for rows 5 vs. 7, $p = 0.744$ for 6 vs. 8)
- Including bias terms gives a large performance gain ($p = 0.02$ for 5 vs. 6, $p = 0.028$ for 7 vs. 8)
- The assessment baseline performs surprisingly well on *Book A* (see row 2)

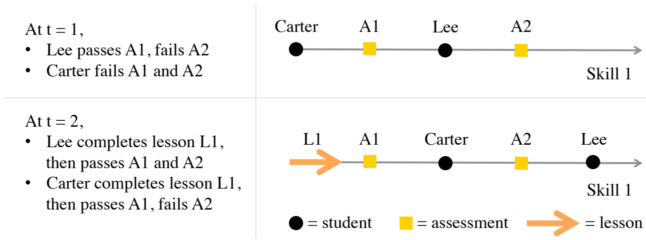

Figure 7. A one-dimensional embedding, where a single latent skill is enough to explain the data. The key observation here is that the model recovered positive skill gains for L_1, and "correctly" arranged students and assessments in the latent space. Initially, Carter fails both assessments, so his skill level is behind the requirements of both assessments. Lee passes A_1 but fails A_2, so his skill level is beyond the requirement for A_1, but behind the requirement for A_2. In an effort to improve their results, Lee and Carter complete lesson L_1 and retake both assessments. Now Carter passes A_1, but still fails A_2, so his new skill level is ahead of the requirements for A_1 but behind the requirements for A_2. Lee passes both assessments, so his new skill level exceeds the requirements for A_1 and A_2. This clear difference in results before completing lesson L_1 and after completing the lesson implies that L_1 had a positive effect on Lee and Carter's skill levels, hence the non-zero skill gain vector recovered for L_1.

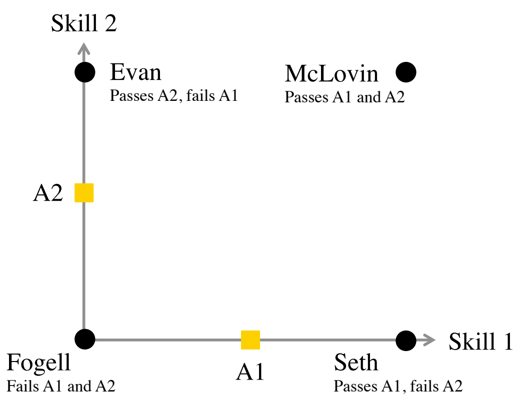

Figure 8. A two-dimensional embedding, where an intransitivity in assessment results requires more than one latent skill to explain. The key observation here is that the assessments are embedded on two different axes, meaning they require two completely independent skills. This makes sense, since student results on A_1 are uncorrelated with results on A_2. Fogell fails both assessments, so his skill levels are behind the requirements for A_1 and A_2. McLovin passes both assessments, so his skill levels are beyond the requirements for A_1 and A_2. Evan and Seth are each able to pass one assessment but not the other. Since the assessments have independent requirements, this implies that Evan and Seth have independent skill sets (i.e. Evan has enough of skill 2 to pass A_2 but not enough of skill 1 to pass A_1, and Seth has enough of skill 1 to pass A_1 but not enough of skill 2 to pass A_2).

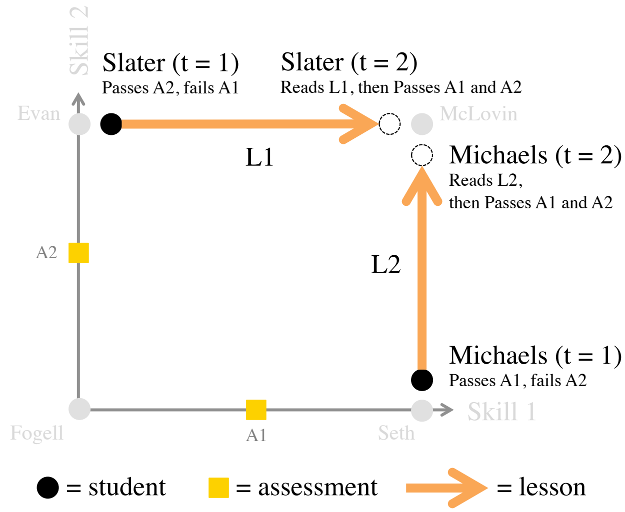

Figure 9. We replicate the setting in Figure 8, then add two new students Slater and Michaels, and two new lesson modules L_1 and L_2. Slater is initially identical to Evan, while Michaels is initially identical to Seth. Slater reads lesson L_1, then passes assessments A_1 and A_2. Michaels reads lesson L_2, then passes assessments A_1 and A_2. The key observation here is that the skill gain vectors recovered for the two lesson modules are orthogonal, meaning they help students satisfy completely independent skill requirements. This makes sense, since initially Slater was lacking in Skill 1 while Michaels was lacking in Skill 2, but after completing their lessons they passed their assessments, showing that they gained from their respective lessons what they were lacking initially.

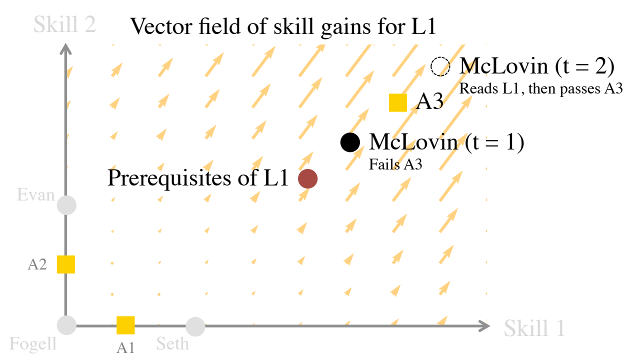

Figure 10. We replicate the setting in Figure 8, then add a new assessment module A_3 and a new lesson module L_1. All students initially fail assessment A_3, then read lesson L_1, after which McLovin passes A_3 while everyone else still fails A_3. The key observation here is that McLovin is the only student who initially satisfies the prerequisites for L_1, so he is the only student who realizes significant gains.

One issue that may have affected the findings is the biased nature of student paths, which has been discussed by (González-Brenes et al., 2014). In the data, we observe that student paths are heavily directed along common routes through modules. We conjecture that this bias dulls the effect of including prerequisites in the embedding, allowing the assessment baseline to perform well in certain settings, and causing the inclusion of bias terms to give a large performance boost. Most students attack an assessment with the same background knowledge, so an embedding that captures the variation in students who work on the same assessment is not as valuable. In a regime where students who work on an assessment come from a variety of skill backgrounds, the embedding may outperform the baselines by a wider margin.

Here, we explore the parameter space of the embedding model by varying the regularization constant β, learning update variance σ^2, and embedding dimension d:

- From Figure 11, we find that regularizing via β is not necessary and that any small value of β gives good performance. Our conjecture is that the Gaussian lesson model already acts as a regularizer. Results for changing d are shown in Figure 12. In summary, we find that increasing embedding dimension d substantially improved performance for embedding models without bias terms, but that it has little effect on performance for embeddings with bias terms. The former is expected, since the embedding itself must be used to model general student passing ability and general assessment difficulty.
- From Figure 13, very small and very large σ perform poorly. For very large σ, the Gaussian steps caused by lesson interactions have so much variance that they might as well not occur, causing the embedding with lessons to effectively degenerate into an embedding without lessons. We see this occur when performance for *d=2, with bias* approaches that of *d=2, without lessons* as σ is increased.

Here, we measure the sensitivity of model performance to the size of the training set:

- Performance is mostly affected by a student's recent history. We see this in the drastic plateau of validation accuracy after a student's history is extended more than fifty interactions into the past in Figure 19.
- The number of full student histories in the training set has a strong effect on performance (via the quality of module embeddings). We see this in the positive relationship between validation accuracy and the number of full histories in Figure 20.

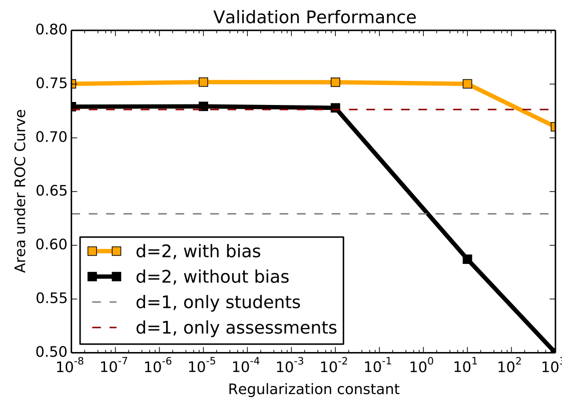

Figure 11. Book A, fixed student history truncations, $\sigma^2 = 0.5$. Note that a high regularization constant for *d=2, with bias* simulates a one-parameter logistic item response theory (1PL IRT) model.

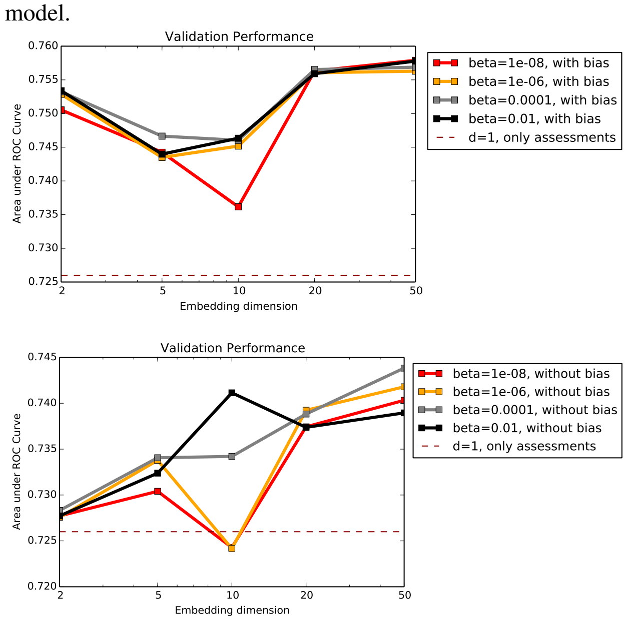

Figure 12. Book A, fixed student history truncations, $\sigma^2 = 0.5$

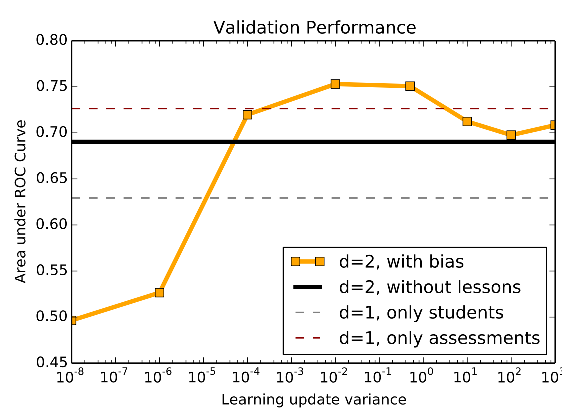

Figure 13. Book A, fixed student history truncations, $\beta = 10^{-6}$

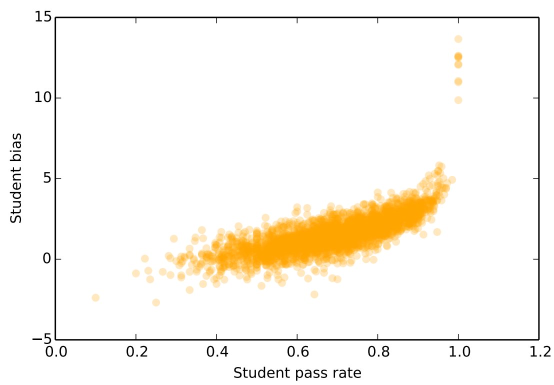

Figure 14. Student pass rate seems to be the result of passing γ_s through a logistic function, which corresponds to our model of assessment results (see Equation 1)

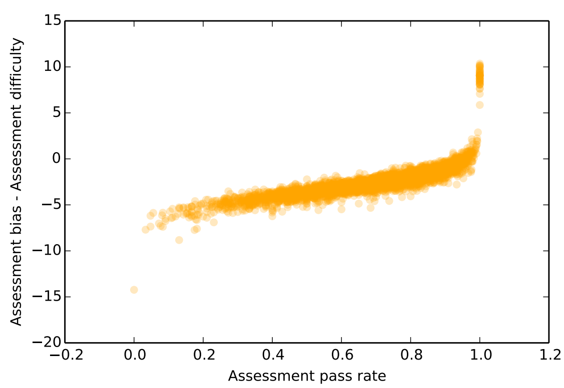

Figure 15. Assessment pass rate seems to be the result of passing $-\|\vec{a}\| + \gamma_a$ through a logistic function, which corresponds to our model of assessment results (see Equation 1)

6.3. Qualitative Evaluation

Here, we explore the recovered model parameters of a two-dimensional embedding of the *Book A* data set. In Figures 14 and 15, we see that the recovered student and assessment parameters are tied to student and assessment pass rates in the training data. In Figure 16, we see that the learning signals aggregated across students grow over time (i.e., students are generally learning and improving their skills over time); Figure 18 gives an example of one of these learning signals, or "student trajectories". Together, these figures demonstrate the ability of the embedding to recover sensible model parameters.

6.4. Lesson Sequence Discrimination

The ability to predict future performance of students on assessments, while a useful metric for evaluating the learned

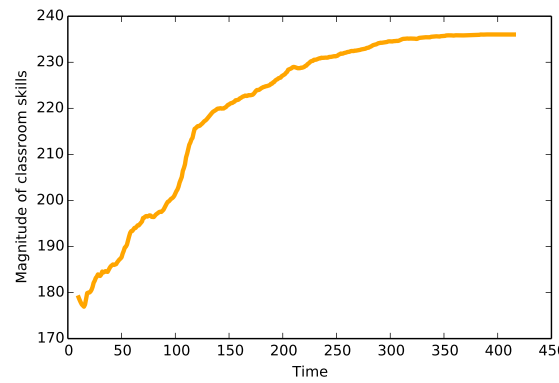

Figure 16. Students' skill levels, i.e. the Frobenius norm of the matrix of embeddings of all students for a given t, improve over time

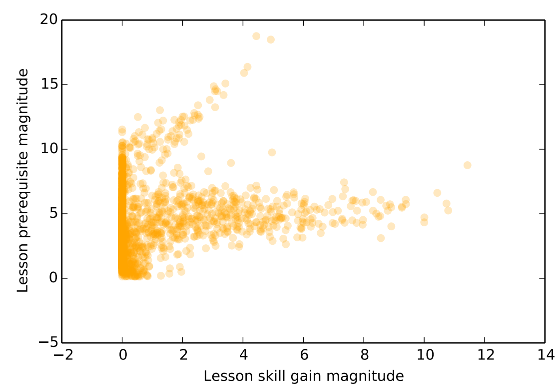

Figure 17. $\|\vec{\ell}\|$ vs. $\|\vec{q}\|$ (see Equation 3). The reason for the shape of this plot is unclear.

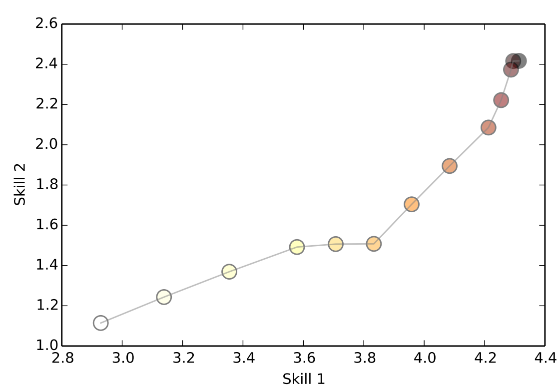

Figure 18. A two-dimensional embedding of a student over time. White indicates where the student starts, and black indicates where the student ends. The variance of the learning update σ^2 has an effect on the tortuosity of student trajectories. Smaller σ leads to paths that show less student forgetting (i.e. do not double back on themselves), while larger σ allows for more freedom.

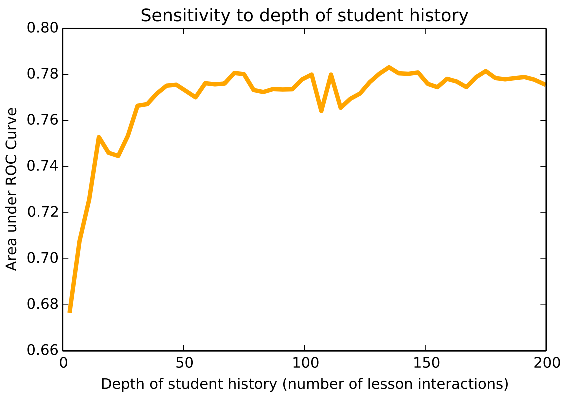

Figure 19. Area under ROC curve is computed for assessment interactions at $T \approx 201$ after training from $T - depth$ to T. For students who do not have assessment interactions at $T = 201$, we select the soonest $T > 201$ such that the student has assessment interactions. Performance is mostly affected by a student's recent history. We see this in the drastic plateau of validation accuracy after a student's history is extended more than 50 interactions into the past.

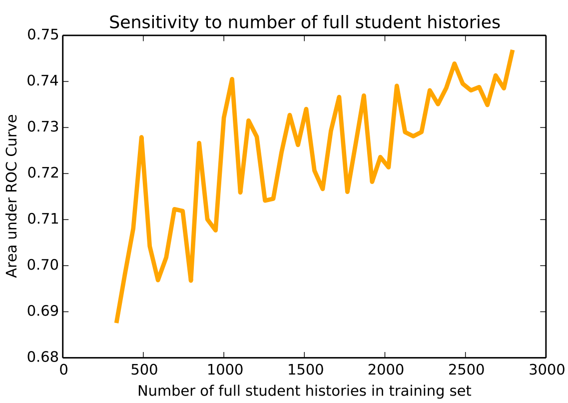

Figure 20. Area under ROC curve is computed for 10% of students, using fixed history truncations. The number of full student histories in the training set has a strong effect on performance (via the quality of module embeddings). We see this in the positive relationship between validation accuracy and the number of full histories.

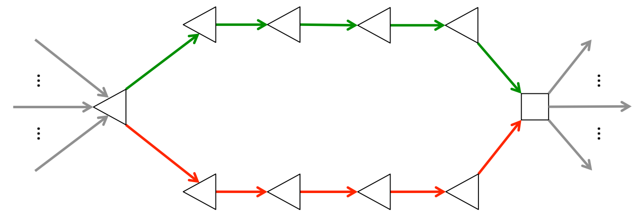

Figure 21. A schematic diagram of a *bubble*, where triangles are lessons and squares are assessments. The green path is the recommended path (note that recommendations are personalized to each student).

embedding, does not address the more important task of adaptive tutoring via customized lesson sequence recommendation. We introduce a surrogate task for evaluating the sequence recommendation performance of the model based entirely on the observational data of student interactions, by assessing the model's ability to recommend "productive" paths amongst several alternatives.

The size of the data set creates a unique opportunity to leverage the variability in learning paths to simulate the setting of a controlled experiment. For this evaluation, we use a larger version of the *Book A* data set examined in Section 6.2, containing 14,707 students and 14,327 content modules. We find that the data contains many instances of student paths that share the same lesson module at the beginning and the same assessment module at the end, but contain different lessons along the way. We call these instances *bubbles*, for example see Figure 21, which present themselves as a sort of natural experiment on the relative merits of two different learning progressions. We can thus use these *bubbles* to evaluate the ability of an embedding to recommend a learning sequence that leads to success, as measured by the relative performance of students who take the recommended vs. the not recommended path to the assessment module at the end of the *bubble*.

We use the full histories of 70% of students to embed lesson and assessment modules, then train on the histories of held-out students up to the beginning of a *bubble*. The lesson sequence for a student is then played over the initial student embedding, using the learning update (Equation 3) to compute an expected student embedding at the end of the *bubble* (which can be used to predict the passing likelihood for the final assessment using Equation 1). The path that leads the student to a higher pass likelihood on the final assessment is the "recommended" path. Our performance measure is $\mathbb{E}\left[\frac{\mathbb{E}[R']-\mathbb{E}[R]}{\mathbb{E}[R]}\right]$, where $R' \in \{0,1\}$ is the outcome at the end of the recommended path and $R \in \{0,1\}$ is the outcome at the end of the other path (0 is failing and 1 is passing). This measure can be interpreted as "expected gain" (averaged over many *bubbles*) from taking recommended paths, or how "successful" the paths recommended

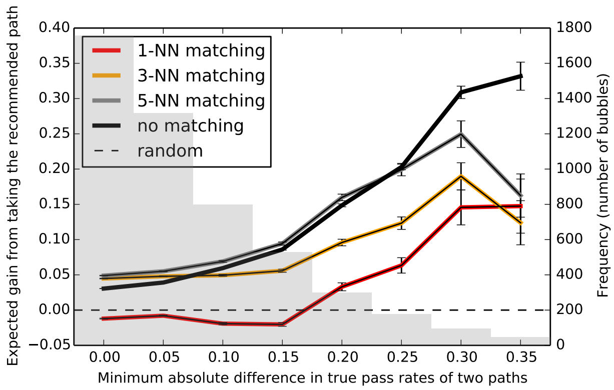

Figure 22. The *x*-axis represents a threshold on absolute difference between pass rates of the two *bubble* paths. The *y*-axis ticks on the left correspond to the gain curves, and the y-axis ticks on the right correspond to the gray bins that illustrate the number of *bubbles* that meet the threshold given on the x-axis. *Bubbles* are filtered to meet the following criteria: at least ten students take each branch, each branch must contain at least two lessons, and both branches must contain the same number of lessons.

by the model are when compared to the alternative.

This observational study is potentially confounded by many hidden variables. For example, it may be that one group of students systematically takes recommended paths while another group of students does not, leading to results at the end of a *bubble* that are mostly dictated by the teachers directing the groups, or other student-specific hidden factors, rather than path quality. To best approximate the settings of a randomized controlled trial in our observational study, we use the standard *propensity score matching* approach for de-biasing observational data (Rosenbaum & Rubin, 1983; Caliendo & Kopeinig, 2008). The key idea behind *propensity score matching* is to subset the observed data in a way that balances the distribution of the features ("hidden variables") describing subjects in the two conditions, as it would be expected in a randomized experiment. The validity of any conclusion drawn from the observational data de-biased in this way hinges on the assumption that all confounding variables that determine self-selection have been accounted for in the features prior to matching. In this study, we hypothesize that the set of all lesson modules and assessment modules (with outcomes) that the learner attempted throughout his or her duration in the online system is sufficient to compensate for any self-selection in the taken learning paths. Formally, we represent learners in a feature space X such that $X_{ij} \in \{-1, 0, 1\}$, where $X_{ij} = 1$ if student i passed module j (lessons are always "passed"), $X_{ij} = 0$ if student i has not completed module j, and $X_{ij} = -1$ if student i failed module j.

We use PCA to map X to a low-dimensional feature space where students are described by 1,000 features, which capture 80% of the variance in the original 14,327 features. A logistic regression model with L_2 regularization is used to estimate the probability of a student following the recommended branch of a *bubble*, i.e. the *propensity score*, given the student features (the regularization constant is selected using cross-validation to maximize average log-likelihood on held-out students). Within each *bubble*, students who took their recommended branch are matched with their nearest neighbors (by absolute difference in *propensity scores*) from the group of students who did not take their recommended branch. Matching is done with replacement (so the same student can be selected as a nearest neighbor multiple times) to improve matching quality, trading off bias for variance. Multiple nearest neighbors can be matched (we examine the effect of varying the number of nearest neighbors), trading off variance for bias.

Naturally, our evaluation metric of gain in the pass rate from following a recommended path would depend strongly on the relative merits of the recommended and alternative paths. From Figure 22, the two-dimensional embedding model with lessons, prerequisites, and bias terms (the same configuration as row 8 in Table 1) is able to recommend more successful paths when there is a significant difference between the quality of the two paths, as measured by the absolute difference in pass rates between the two paths (regardless of choice of *propensity score matching*).

7. Outlook and Future Work

7.1. Model Extensions

An issue with our model is that we cannot embed new modules that do not have any existing access traces. One way to solve this cold start problem is to jointly embed modules with semantic tags, i.e. impose a prior distribution on the locations of modules with certain tags. This approach has been previously explored in the context of music playlist prediction (Moore et al., 2012). Embedding with tags has the added benefit of making the dimensions of the latent skill space more interpretable, as demonstrated in (Lan et al., 2014b).

Another approach to solving the cold start problem is to use an expert content-to-concept map to impose a prior on module embeddings that captures 1) the grouping of modules by underlying concept and 2) the prerequisite relationships between concepts (which translate into prerequisite relationships between modules). The objective function (recall Section 4) would penalize the cosine distance between modules governed by the same concept, and reward arrangements of modules that reflect prerequisite edges in the concept graph.

The variance parameter σ^2 in the learning update (Equation 2) can be used to model offline learning and forgetting. Though we treat it as a constant across students in earlier experiments, it can also be estimated as a student-specific parameter and as a function of the real time elapsed between lesson interactions.

The *forgetting effect*, where student knowledge diminishes during the gap between interactions, may be modeled as a penalty in the skill gains from a lesson that increases with the time elapsed since the previous interaction. A substantially similar approach has been explored in (Lan et al., 2014a).

7.2. Adaptive Lesson Sequencing

Given a student's current skill levels and a set of assessments the student wants to pass, what is the optimal sequence of lessons for the student? We can formulate the problem two ways: specify a minimum pass likelihood for the goal assessments and find the shortest path to mastery, or specify a time constraint and find the path that leads to the highest pass likelihood for the goal assessments while not exceeding the prescribed length. Both problems can be tackled by using a course embedding to specify a Markov Decision Process (Bellman, 1957), where state is given by the student embedding, the set of possible actions corresponds to the set of lesson modules that the student can work on, the transition function is the learning update (Equation 3), and the reward function is the likelihood of passing all goal assessments (Equation 1).

An issue we have not considered is the spacing effect (Dempster, 1988). If a student learns on a known time frame (e.g., 1-3pm on weekdays), it is possible for the adaptive tutor to adjust the learning schedule to present review modules in a timely manner or help a student cram. For theoretical work on the timing of modules see (Novikoff et al., 2012).

7.3. Learning and Content Analytics

We speculate that a course embedding can be used to measure the following characteristics of educational content: lesson quality, through magnitude of the skill gain vector $||\vec{\ell}||$; the diversity of skills and concept knowledge relevant for a course, through the embedding dimension that optimizes validation accuracy on the assessment result prediction task; the availability of content that teaches and assesses different skills, through the density of modules embedded in different locations in the latent skill space. The following characteristics of student learning may also be of interest: the rate of knowledge acquisition, through the velocity of a student moving through the latent skill space; the rate of knowledge loss (i.e. forgetting), through the amount of backtracking in a student trajectory.

8. Conclusions

We presented a general model that learns a representation of student knowledge and educational content that can be used for personalized instruction, learning analytics, and content analytics. The key idea lies in using a multidimensional embedding to capture the dynamics of learning and testing. Using a large-scale data set collected in real-world classrooms, we (1) demonstrate the ability of the model to successfully predict learning outcomes and (2) introduce an offline methodology as a proxy for assessing the ability of the model to recommend personalized learning paths. We show that our model is able to successfully discriminate between personalized learning paths that lead to mastery and failure.

Acknowledgements

We would like to thank members of Knewton for valuable feedback on this work. This research was funded in part by the National Science Foundation under Awards IIS-1247637 and IIS-1217686, and through the Cornell Presidential Research Scholars program.

References

Bellman, Richard. A markovian decision process. *Indiana Univ. Math. J.*, 6:679–684, 1957. ISSN 0022-2518.

Caliendo, Marco and Kopeinig, Sabine. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72, 2008.

Carbonell, Jaime R. Ai in cai: An artificial-intelligence approach to computer-assisted instruction. *Man-Machine Systems, IEEE Transactions on*, 11(4):190–202, 1970.

Chen, Shuo, Moore, Josh L, Turnbull, Douglas, and Joachims, Thorsten. Playlist prediction via metric embedding. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 714–722. ACM, 2012.

Corbett, AlbertT. and Anderson, JohnR. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1994. ISSN 0924-1868. doi: 10.1007/BF01099821. URL `http://dx.doi.org/10.1007/BF01099821`.

Dempster, Frank N. The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, 43(8):627, 1988.

Ekanadham, Chaitanya and Karklin, Yan. T-skirt: Online estimation of student proficiency in an adaptive learning system. *Machine Learning for Education Workshop at ICML*, 2015.

González-Brenes, JP, Huang, Yun, and Brusilovsky, Peter. General features in knowledge tracing: Applications to multiple subskills, temporal item response theory, and expert knowledge. In *Proceedings of the 7th International Conference on Educational Data Mining (accepted, 2014)*, 2014.

Knewton. The knewton platform: A general-purpose adaptive learning infrastructure. Technical report, 2015. URL http://www.knewton.com/wp-content/uploads/knewton-technical-white-paper-201501.pdf.

Lan, Andrew S, Studer, Christoph, and Baraniuk, Richard G. Time-varying learning and content analytics via sparse factor analysis. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 452–461. ACM, 2014a.

Lan, Andrew S, Studer, Christoph, Waters, Andrew E, and Baraniuk, Richard G. Tag-aware ordinal sparse factor analysis for learning and content analytics. *arXiv preprint arXiv:1412.5967*, 2014b.

Lan, Andrew S., Waters, Andrew E., Studer, Christoph, and Baraniuk, Richard G. Sparse factor analysis for learning and content analytics. *J. Mach. Learn. Res.*, 15(1):1959–2008, January 2014c. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2627435.2670314.

Lesgold, Alan et al. Sherlock: A coached practice environment for an electronics troubleshooting job. 1988.

Linden, WJ and Hambleton, Ronald K. Handbook of modern item response theory. *New York*, 1997.

Moore, J., Chen, Shuo, Joachims, T., and Turnbull, D. Learning to embed songs and tags for playlist prediction. In *Conference of the International Society for Music Information Retrieval (ISMIR)*, pp. 349–354, 2012.

Moore, J., Chen, Shuo, Turnbull, D., and Joachims, T. Taste over time: The temporal dynamics of user preferences. In *Conference of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 401–406, 2013.

Novikoff, Timothy P., Kleinberg, Jon M., and Strogatz, Steven H. Education of a model student. *Proceedings of the National Academy of Sciences*, 109(6):1868–1873, 2012. doi: 10.1073/pnas.1109863109. URL http://www.pnas.org/content/109/6/1868.abstract.

ORourke, Eleanor, Andersen, Erik, Gulwani, Sumit, and Popovic, Zoran. A framework for automatically generating interactive instructional scaffolding. 2015.

Piech, Chris, Huang, Jonathan, Nguyen, Andy, Phulsuksombati, Mike, Sahami, Mehran, and Guibas, Leonidas. Learning program embeddings to propagate feedback on student code. *arXiv preprint arXiv:1505.05969*, 2015a.

Piech, Chris, Sahami, Mehran, Huang, Jonathan, and Guibas, Leonidas. Autonomously generating hints by inferring problem solving policies. 2015b.

Piech, Chris, Spencer, Jonathan, Huang, Jonathan, Ganguli, Surya, Sahami, Mehran, Guibas, Leonidas J., and Sohl-Dickstein, Jascha. Deep knowledge tracing. *CoRR*, abs/1506.05908, 2015c. URL http://arxiv.org/abs/1506.05908.

Prerau, Michael J, Smith, Anne C, Eden, Uri T, Yanike, Marianna, Suzuki, Wendy A, and Brown, Emery N. A mixed filter algorithm for cognitive state estimation from simultaneously recorded continuous and binary measures of performance. *Biological cybernetics*, 99(1):1–14, 2008.

Rasch, Georg. *Probabilistic models for some intelligence and attainment tests.* ERIC, 1993.

Reckase, M. D. *Multidimensional Item Response Theory*. Springer Publishing Company, Incorporated, first edition, 2009.

Rosenbaum, Paul R and Rubin, Donald B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Smith, Anne C, Frank, Loren M, Wirth, Sylvia, Yanike, Marianna, Hu, Dan, Kubota, Yasuo, Graybiel, Ann M, Suzuki, Wendy A, and Brown, Emery N. Dynamic analysis of learning in behavioral experiments. *The journal of neuroscience*, 24(2):447–461, 2004.

Sohl-Dickenstein, Jascha. Temporal multi-dimensional item response theory, 2014. URL http://lytics.stanford.edu/datadriveneducation/slides/sohldickstein.pdf.

Welch, B. L. The generalization of student's problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947. doi: 10.1093/biomet/34.1-2.28. URL http://biomet.oxfordjournals.org/content/34/1-2/28.short.

Zhu, Ciyou, Byrd, Richard H, Lu, Peihuang, and Nocedal, Jorge. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM*

Transactions on Mathematical Software (TOMS), 23(4): 550–560, 1997.