Learning Student and Content Embeddings for Personalized Lesson Sequence Recommendation

Siddharth Reddy
Dept. of Computer Science,
Cornell University
Ithaca, NY 14850, USA
sgr45@cornell.edu

Igor Labutov
Dept. of Electrical and
Computer Engineering, Cornell
University
Ithaca, NY 14850, USA
iil4@cornell.edu

Thorsten Joachims
Dept. of Computer Science,
Cornell University
Ithaca, NY 14850, USA
tj@cs.cornell.edu

L@S 2016, April 25–26, 2016, Edinburgh, UK.
ACM 978-1-4503-3726-7/16/04.
http://dx.doi.org/10.1145/2876034.2893375

Abstract
Students in online courses generate large amounts of data that can be used to personalize the learning process and improve quality of education. In this paper, we present the Latent Skill Embedding (LSE), a probabilistic model of students and educational content that can be used to recommend personalized sequences of lessons with the goal of helping students prepare for specific assessments. Akin to collaborative filtering for recommender systems, the algorithm does not require students or content to be described by features, but it learns a representation using access traces. We formulate this problem as a regularized maximum-likelihood embedding of students, lessons, and assessments from historical student-content interactions. Empirical findings on large-scale data from Knewton, an adaptive learning technology company, show that this approach predicts assessment results competitively with benchmark models and is able to discriminate between lesson sequences that lead to mastery and failure.

Author Keywords
Probabilistic Embedding; Sequence Recommendation; Adaptive Learning

Introduction
The popularity of online education platforms has soared in recent years. Companies like Coursera and EdX offer

Massive Open Online Courses (MOOCs) that improve access to high quality educational content for anyone connected to the Internet. 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 and guide learners through custom lesson plans could help learners take advantage of large and heterogeneous collections of educational content to achieve their goals.

Our aim is to build a domain-agnostic framework for modeling students and content that can be used for personalized lesson sequence recommendation. A common data source available in online learning products is *access traces* that log student interactions with modules of course content. These access traces have the form *Student A completed Lesson B* and *Student C passed assessment D*. *Lessons* are content modules that introduce or reinforce concepts, and *assessments* are content modules with pass-fail results that test student skills. 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 simple synthetic scenarios, as well as large-scale real data from Knewton, an education technology company that offers personalized recommendations and activity analytics for online courses.

Related Work Our work builds on the existing literature in psychometric user and content modeling, including the Rasch model [7], temporal and multi-dimensional item response theory [3, 8, 5], Bayesian knowledge tracing [2], sparse factor analysis [4], and deep knowledge tracing [6]. We extend this work in a multi-dimensional setting where student knowledge evolves over time in a continuous state space, and explicitly-modeled lesson prerequisites modulate knowledge gains from lesson modules.

Embedding Model

We now describe a probabilistic model that places students, lessons, and assessments together in a *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 $\vec{a} \in \mathbb{R}_+^d$. A student can be tested on an assessment module with a pass-fail result $R \in \{0, 1\}$, where the likelihood of passing is high when a student has skill levels that exceed the assessment requirements and vice-versa. A student can also 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. 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 an *input-output hidden Markov model*.

Assessment results are modeled as follows: for student $\vec{s}$, assessment $\vec{a}$, and result R,

$$R \sim \text{Bernoulli}(\phi(\Delta(\vec{s},\ \vec{a}))) \tag{1}$$

where ϕ is the logistic function, and

$$\Delta(\vec{s},\ \vec{a}) = \frac{\vec{s} \cdot \vec{a}}{||\vec{a}||} - ||\vec{a}|| + \gamma_s + \gamma_a \tag{2}$$

where γ_s is a student-specific bias term and γ_a is an assessment-specific bias term. Student learning from lessons is modeled as follows: for student $\vec{s}$ who worked on a lesson with skill gains $\vec{\ell}$ and prerequisites $\vec{q}$ at time $t+1$, the updated student state is

$$\vec{s}_{t+1} \sim \mathcal{N}\left(\vec{s}_t + \vec{\ell} \cdot \phi(\Delta(\vec{s}_t,\ \vec{q})),\ \Sigma\right) \tag{3}$$

where the covariance matrix $\Sigma = I_d \sigma^2$ is isotropic. We estimate model parameters from log data using maximum-likelihood estimation with L_2 regularization, and select hyperparameters using cross-validation. Examples of embeddings trained on synthetic log data are given in Figures 1 and 2.

Experiments on Online Course Data

To evaluate the embedding model, we use data processed by Knewton, an adaptive learning technology company whose infrastructure uses access traces to generate personalized recommendations and activity analytics for online learning products. The data describes interactions between college students and two science textbooks, and collectively consists of 2.18 million access traces from over 7,000 students, recorded in 1,939 classrooms over a combined period of 5 months. The following summarizes the main findings – details and further results can be found in [1]. We first evaluate the embedding model on the task of predicting results of held-out assessment interactions, and find that the embedding performs comparably to benchmark models from item response theory. We then introduce a novel prediction task to more directly evaluate the embedding model on its ability to distinguish between lesson sequences that lead to mastery and failure on target assessments. Using *propensity score matching* [9] to de-bias the observational data from Knewton, we find that the model is generally able to recommend successful paths.

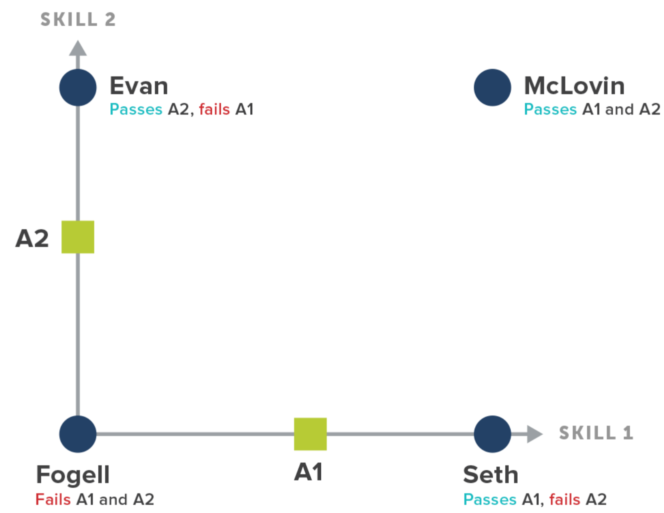

Figure 1: A two-dimensional embedding without lessons

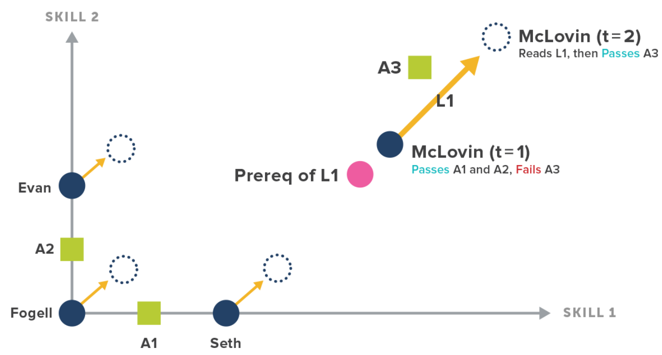

Figure 2: A two-dimensional embedding with lessons and prerequisites

Conclusions

We presented a general model that learns a representation of student knowledge and educational content that can be used for personalized instruction. The key idea lies in using a multi-dimensional 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.

An implementation of the Latent Skill Embedding and the IPython notebooks used to conduct experiments – as well as a full paper with a comprehensive description of our methods and results – are available online [1].

We thank Knewton for valuable feedback. This research was funded in part NSF Awards IIS-1247637, IIS-1217686, and IIS-1513692, a grant from the John Templeton Foundation provided through the Metaknowledge Network at the University of Chicago, and through the Cornell Presidential Research Scholars program.

References

[1] Latent skill embedding. http://siddharth.io/lentil.
[2] Corbett, A. T., and Anderson, J. R. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction 4*, 4 (1994), 253–278.
[3] Ekanadham, C., and Karklin, Y. T-skirt: Online estimation of student proficiency in an adaptive learning system. *Machine Learning for Education Workshop at ICML* (2015).
[4] Lan, A. S., Studer, C., and Baraniuk, R. 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*, ACM (2014), 452–461.
[5] Linden, W., and Hambleton, R. K. Handbook of modern item response theory. *New York* (1997).
[6] Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., and Sohl-Dickstein, J. Deep knowledge tracing. In *Advances in Neural Information Processing Systems* (2015), 505–513.
[7] Rasch, G. *Probabilistic models for some intelligence and attainment tests.* ERIC, 1993.
[8] Reckase, M. D. *Multidimensional Item Response Theory*, first ed. Springer Publishing Company, Incorporated, 2009.
[9] Rosenbaum, P. R., and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 1 (1983), 41–55.