Learning Representations of Student Knowledge and Educational Content

Siddharth Reddy\textsuperscript{1,2}, Igor Labutov\textsuperscript{3}, Thorsten Joachims\textsuperscript{1}

\textsuperscript{1} Department of Computer Science
Cornell University
Ithaca, NY, USA

\textsuperscript{2} Knewton, Inc.
New York, NY, USA

\textsuperscript{3} Department of Electrical and Computer Engineering
Cornell University
Ithaca, NY, USA

Machine Learning for Education Workshop, ICML 2015
Motivation

How can we guide students through large, heterogeneous collections of educational content to help them achieve their goals?

- **Content** = lessons and assessments
- **Goal** = passing an assessment
- **Guide** = recommending personalized lesson sequences
How can we guide students through large, heterogeneous collections of educational content to help them achieve their goals?

- Content = lessons and assessments
- Goal = passing an assessment
- Guide = recommending personalized lesson sequences
General Approach

1. Alice completed Lesson B
2. Alice passed Assessment C
3. Alice failed Assessment D
4. Alice completed Lesson E
5. Alice passed Assessment D
6. ...

1. Bob completed Lesson B
2. Bob failed Assessment C
3. Bob failed Assessment D
4. Bob completed Lesson E
5. Bob passed Assessment C
6. ...

---

Reddy, Labutov, Joachims
Learning Representations of Student Knowledge and Educational Content
### Related Work

Existing frameworks:
- Bayesian Knowledge Tracing (BKT)
- Item Response Theory (IRT)
- Sparse Factor Analysis (SPARFA)

<table>
<thead>
<tr>
<th>Framework</th>
<th>Knowledge Tracing</th>
<th>Lesson Effects</th>
<th>Evaluations</th>
</tr>
</thead>
<tbody>
<tr>
<td>DKT</td>
<td>RNN</td>
<td>-</td>
<td>outcome prediction</td>
</tr>
<tr>
<td>T-SKIRT</td>
<td>input-output HMM</td>
<td>-</td>
<td>outcome prediction</td>
</tr>
<tr>
<td>SPARFA-Trace</td>
<td>input-output HMM</td>
<td>skill gains</td>
<td>outcome prediction</td>
</tr>
<tr>
<td>Embedding</td>
<td>input-output HMM</td>
<td>skill gains, prereqs</td>
<td>outcome prediction, lesson recommendation</td>
</tr>
</tbody>
</table>
Student and Content Representations

- **Student** = a set of latent skill levels that vary over time
- **Lesson module** = a vector of skill gains, and a set of prerequisite skill requirements
- **Assessment module** = a set of skill requirements

A student can be tested on an assessment module, which has a pass-fail result. The likelihood of passing should be high when a student has skill levels that exceed the assessment requirements, and vice-versa.

A student can complete lesson modules to learn over time, though the skill gains from a lesson module are modulated by prerequisite knowledge.
Student and Content Representations

- **Student** = a set of latent skill levels that vary over time
- **Lesson module** = a vector of skill gains, and a set of prerequisite skill requirements
- **Assessment module** = a set of skill requirements
- A student can be tested on an assessment module, which has a pass-fail result. The likelihood of passing should be high when a student has skill levels that exceed the assessment requirements, and vice-versa.
- A student can complete lesson modules to learn over time, though the skill gains from a lesson module are modulated by prerequisite knowledge
Modeling Assessment Outcomes

For student $\vec{s}_t \in \mathbb{R}^d$, assessment $\vec{a} \in \mathbb{R}^d$, and result $R \in \{0, 1\}$,

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

where $\phi$ is the logistic function and

$$\Delta(\vec{s}_t, \vec{a}) = \frac{\vec{s}_t \cdot \vec{a}}{||\vec{a}||} - ||\vec{a}|| + \gamma_s + \gamma_a$$
Modeling Student Learning from Lessons

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

$$\vec{s}_{t+1} \sim \mathcal{N}(\vec{s}_t + \vec{\ell}, \Sigma)$$

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

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

where $\Delta(\vec{s}_t, \vec{q}) = \frac{\vec{s}_t \cdot \vec{q}}{||\vec{q}||} - ||\vec{q}||$
One-Dimensional Embedding

Alice (t = 1) Fails A1

Alice (t = 2) Reads L1, then Passes A1

A1 = student

= lesson

= assessment

Skill 1

Skill 2

L1

A1

Reddy, Labutov, Joachims

Learning Representations of Student Knowledge and Educational Content
Two-Dimensional Embedding

- **Evan**: Passes A2, fails A1
- **Fogell**: Fails A1 and A2
- **McLovin**: Passes A1 and A2
- **Seth**: Passes A1, fails A2

**Skills**

- **Skill 1**
- **Skill 2**
Embedding with Lessons

- **Fogell**
  - Skill 1: Passes A1, fails A2

- **McLovin**
  - Skill 2: Reads L1, then Passes A1 and A2

- **Slater**
  - (t = 1): Passes A2, fails A1
  - (t = 2): Reads L1, then Passes A1 and A2

- **Seth**
  - (t = 1): Passes A1, fails A2
  - (t = 2): Reads L2, then Passes A1 and A2

- **Evan**
  - Skill 2: Passes A2, fails A1

- **A1**

- **A2**

- **L1**

- **L2**

- **Skill 1**

- **Skill 2**

- **= student**
- **= assessment**
- **= lesson**

---

Reddy, Labutov, Joachims  
Learning Representations of Student Knowledge and Educational Content 10 / 21
Embedding Model

Embedding with Lesson Prerequisites

Skill 2

Fogell

A2

A1

Seth

Evan

Skill 1

McLovin (t = 1)

Reads L1, then Passes A3

Prereq of L1

A3

L1

Fogell

A2

A1

Seth

Evan

McLovin (t = 2)

Passes A1 and A2, Fails A3

L1

Prereq of L1
The two data sets from Knewton collectively contain:

- 2,184,352 interaction logs
- 1,939 classrooms
- 6 months
- 7,034 students
- 7,217 lessons
- 7,287 assessments
- Average assessment pass rates of 0.712 and 0.693
Parameter Estimation

- We compute MAP estimates of model parameters $\Theta$ by maximizing

$$L(\Theta) = \sum_{A} \log (Pr(R | \vec{s}_t, \vec{a}, \gamma_s, \gamma_a))$$

$$+ \sum_{L} \log (Pr(\vec{s}_{t+1} | \vec{s}_t, \vec{\ell}, \vec{q})) - \beta \cdot \lambda(\Theta)$$

where $A$ is the set of assessment interactions, $L$ is the set of lesson interactions, $\lambda(\Theta)$ is a regularization term that penalizes the $L_2$ norms of embedding parameters (not bias terms $\gamma$), and $\beta$ is a regularization parameter.

- Solved with L-BFGS-B and random parameter initializations
Assessment Result Prediction

- Hold out assessment interactions at the end of student histories
- Area under ROC Curve (AUC)
- Ten-fold cross-validation
## Assessment Result Prediction

<table>
<thead>
<tr>
<th></th>
<th>d</th>
<th>s</th>
<th>a</th>
<th>ℓ</th>
<th>q</th>
<th>γ</th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>0.723</td>
<td>0.686</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>0.715</td>
<td>0.661</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>0.991</td>
<td>0.614</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>0.992</td>
<td>0.726</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>0.897</td>
<td>0.696</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>0.871</td>
<td>0.748</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>0.898</td>
<td>0.694</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>0.882</td>
<td>0.749</td>
</tr>
</tbody>
</table>

- Joint embedding beats baselines
  - Including **lesson embeddings** improves performance significantly
  - Including **prerequisite embeddings** has a statistically insignificant effect on performance
  - Including **bias terms** improves performance significantly
Assessment Result Prediction

<table>
<thead>
<tr>
<th>d</th>
<th>s</th>
<th>a</th>
<th>l</th>
<th>q</th>
<th>γ</th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>0.723</td>
<td>0.686</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>0.715</td>
<td>0.661</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>0.991</td>
<td>0.614</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>0.992</td>
<td>0.726</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>0.897</td>
<td>0.696</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>0.871</td>
<td>0.748</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>0.898</td>
<td>0.694</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>0.882</td>
<td>0.749</td>
</tr>
</tbody>
</table>

- Joint embedding beats baselines
- Including **lesson embeddings** improves performance significantly
- Including **prerequisite embeddings** has a statistically insignificant effect on performance
- Including **bias terms** improves performance significantly
## Assessment Result Prediction

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>0.723</td>
<td>0.686</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>0.715</td>
<td>0.661</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>0.991</td>
<td>0.614</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>0.992</td>
<td>0.726</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>0.897</td>
<td>0.696</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>0.871</td>
<td>0.748</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>0.898</td>
<td>0.694</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>0.882</td>
<td>0.749</td>
</tr>
</tbody>
</table>

- Joint embedding beats baselines
- Including **lesson embeddings** improves performance significantly
- Including **prerequisite embeddings** has a statistically insignificant effect on performance
- Including **bias terms** improves performance significantly
## Assessment Result Prediction

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>0.723</td>
<td>0.686</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>0.715</td>
<td>0.661</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>0.991</td>
<td>0.614</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>0.992</td>
<td>0.726</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>0.897</td>
<td>0.696</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>0.871</td>
<td>0.748</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>0.898</td>
<td>0.694</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>0.882</td>
<td>0.749</td>
</tr>
</tbody>
</table>

- Joint embedding beats baselines
- Including **lesson embeddings** improves performance significantly
- Including **prerequisite embeddings** has a statistically insignificant effect on performance
- Including **bias terms** improves performance significantly
Lesson Sequence Discrimination

Predicted worse

\[ \triangle = \text{lesson} \]
\[ \square = \text{assessment} \]

Predicted better

...
Lesson Sequence Discrimination

- Expected relative gain from taking recommended path

\[ E \left[ \frac{\mathbb{E}[R'] - \mathbb{E}[R]}{\mathbb{E}[R]} \right] \]

where \( R' \) is the outcome at the end of the recommended path and \( R \) is the outcome at the end of the other path

- Propensity score matching
  - Student features = past outcomes
  - Logistic regression for propensity score estimation
  - Nearest neighbor matching
Lesson Sequence Discrimination

- Expected relative gain from taking recommended path

\[ \mathbb{E} \left[ \frac{\mathbb{E}[R'] - \mathbb{E}[R]}{\mathbb{E}[R]} \right] \]

where \( R' \) is the outcome at the end of the recommended path and \( R \) is the outcome at the end of the other path

- Propensity score matching
  - Student features = past outcomes
  - Logistic regression for propensity score estimation
  - Nearest neighbor matching
Lesson Sequence Discrimination

![Graph showing the expected gain from taking the recommended path vs. minimum difference in path quality. The graph compares 'no matching' and 'random' scenarios. The x-axis represents the minimum difference in path quality, ranging from 0.00 to 0.35, and the y-axis represents the expected gain from taking the recommended path, ranging from -0.2 to 1.4. The graph includes error bars for the 'no matching' scenario.]
Lesson Sequence Discrimination

![Graph showing expected gain from taking recommended path against minimum difference in path quality. Three lines represent 2-NN matching, no matching, and random.](image)

- **2-NN matching**
- **no matching**
- **random**

Reddy, Labutov, Joachims
Learning Representations of Student Knowledge and Educational Content 18 / 21
Lesson Sequence Discrimination

The graph illustrates the expected gain from taking the recommended path for different matching strategies. The x-axis represents the minimum difference in path quality, while the y-axis shows the expected gain. Four strategies are compared:

- **2-NN matching** (red line)
- **3-NN matching** (yellow line)
- **no matching** (black line)
- **random** (dashed line)

The graph shows that as the minimum difference in path quality increases, the expected gain also increases, with the 2-NN and 3-NN matching strategies generally outperforming the other two.
Lesson Sequence Discrimination

![Graph showing expected gain from taking recommended path vs. minimum difference in path quality for different matching techniques. The graph includes lines for 2-NN matching, 3-NN matching, 4-NN matching, no matching, and random matching. The x-axis represents the minimum difference in path quality, and the y-axis represents the expected gain from taking the recommended path. The graph illustrates the performance of each matching technique under varying path quality differences.]
Summary

- Demonstrated the ability of an embedding model to successfully predict assessment results
  - Modeling skill gains from lessons is helpful
  - Modeling prerequisites for lessons is not helpful
- Introduced an offline methodology as a proxy for assessing the ability of a model to recommend personalized lesson sequences
  - Embedding model can distinguish between “good” and “bad” paths in bubble scenarios
Ongoing Work

- **Cold Start Problem**
  - Impose prior distribution on content embeddings based on *content-to-concept map* and *concept dependency graph* made by experts

- **Personalized Scheduling**
  - Jointly model *assessment results*, *response times*, and *number of attempts*
  - Model the *forgetting effect* and *offline learning* between interactions
  - Handle wall-clock time constraints on recommended sequences

- **Content Analytics**
  - Small-scale Mechanical Turk experiments for teaching basic programming
  - Measure lesson quality
Contact

- **Email**: sgr45@cornell.edu
- **Paper, Slides, Poster, Software**: http://siddharth.io