Diagnosing University Student Subject Proficiency and Predicting Degree Completion in Vector Space

Yuetian Luo  
UW-Madison, USA  
Statistics  
yluo86@wisc.edu

Zach A. Pardos  
UC Berkeley, USA  
Information / Education  
zp@berkeley.edu

PDF: tiny.cc/ylzpEAAI
Why should this be studied?

• On-time graduation remains a problem nation wide
  • 38.9% among 2004 freshman (DeAngelo et al., 2011)
  • 37.5% among 2010 freshman (Shapiro et al., 2016)
    • Tracking 3,600 post-secondary institutions

• High adviser to student ratio of 1:400 (2014)
  • Approaches needed to help with advising at scale
Related work

Dropout on campus
• Academic failure vs voluntary (Tinto, 1975)
• Institutional loyalty & routinization factors (Bean, 1980)
• Prediction approaches using student demographics and subject GPA features (Dekker et al., 2009; Aulck et al., 2016)
• Early warning system (Jayaparakash et al., 2014)

Dropout in MOOCs
• Features engineered from clickstream (Whitehill et al., 2017)
• Features derived from forum activity (Yang et al., 2013)
• Evaluation (Gardner & Brooks, in-press)

Time to degree
• Pathway modeling (Lin, 2009; Imbrie et al., 2008)
• Prediction using FFNN & logistic regression (Herzog, 2006)

Representation learning
• Language (Mikolov et al., 2013)
• E-commerce (Grbovic et al., 2015)
• Tutoring systems (Pardos & Dadu, 2017)
• Courses (Pardos & Nam, 2017)
Methodology: vector space embedding

“All happy families are alike; each unhappy family is unhappy in its own way.”  

(Mikolov, Chen, Corrado, & Dean (2013))

Each unhappy is unhappy

Family

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Example 1</th>
<th>Example 2</th>
<th>Example 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>France - Paris</td>
<td>Italy: Rome</td>
<td>Japan: Tokyo</td>
<td>Florida: Tallahassee</td>
</tr>
<tr>
<td>big - bigger</td>
<td>small: larger</td>
<td>cold: colder</td>
<td>quick: quicker</td>
</tr>
<tr>
<td>Miami - Florida</td>
<td>Baltimore:</td>
<td>Dallas: Texas</td>
<td>Kona: Hawaii</td>
</tr>
<tr>
<td>Einstein - scientist</td>
<td>Maryland</td>
<td>Mozart:</td>
<td>Picasso: painter</td>
</tr>
<tr>
<td>Sarkozy - France</td>
<td>Messi:</td>
<td>violinist</td>
<td>Koizumi: Japan</td>
</tr>
<tr>
<td>copper - Cu</td>
<td>midfieldier</td>
<td>Merkel:</td>
<td>uranium: plutonium</td>
</tr>
<tr>
<td>Berlusconi - Silvio</td>
<td>Berlusconi:</td>
<td>Germany:</td>
<td>Obama: Barack</td>
</tr>
<tr>
<td>Microsoft - Windows</td>
<td>Italy</td>
<td>gold: Au</td>
<td>Apple: iPhone</td>
</tr>
<tr>
<td>Microsoft - Ballmer</td>
<td>zinc: Zn</td>
<td>Putin: Medvedev</td>
<td>IBM: Linux</td>
</tr>
<tr>
<td>Japan - sushi</td>
<td>Sarkozy: Nicolas</td>
<td>IBM: McNealy</td>
<td>Apple: Jobs</td>
</tr>
<tr>
<td></td>
<td>Google:</td>
<td>IBM:</td>
<td>USA: pizza</td>
</tr>
<tr>
<td></td>
<td>Yahoo</td>
<td>McNealy</td>
<td></td>
</tr>
</tbody>
</table>
“All happy families are alike; each unhappy family is unhappy in its own way.” (natural language)

“CS61A MATH1B SPA12 STAT200B CUE100A CS188 CS C267 CS268 ENN1B.” (course enrollments)

Methodology: vector space embedding

(Pardos & Name, 2017 & in-preparation)
Methodology: vector space embedding

“How to get a student embedding?”

Document: “All happy families are alike; each unhappy family is unhappy in its own way.” (natural language)

Student: “CS61A MATH1B SPA12 STAT200B CUE100A CS188 CS C267 CS268 ENN1B.” (course enrollments)
“All happy families are alike; each unhappy family is unhappy in its own way.” (natural language)

“CS61A MATH1B SPA12 STAT200B CUE100A CS188 CS C267 CS268 ENN1B.” (course enrollments)

“stu1 stu2 stu5 stu9 stu3 stu51 stu19 stu8…”

stu1 stu2 stu9 stu3

stu5 (student embedding)
Methodology: vector space embedding

Document: “All happy families are alike; each unhappy family is unhappy in its own way.” (natural language)

Student: “CS61A MATH1B SPA12 STAT200B CUE100A CS188 CS C267 CS268 ENN1B.” (course enrollments)

Course: “stu1 stu2 stu5 stu9 stu3 stu51 stu19 stu8…”

stu1    stu2    stu9    stu3

 stu5
(student embedding)
Methodology: vector space embedding

“All happy families are alike; each unhappy family is unhappy in its own way.” (natural language)

“CS61A MATH1B SPA12 STAT200B CUE100A CS188 CS C267 CS268 ENN1B.” (course enrollments)

“stu1 stu2 stu5 stu9 stu3 stu51 stu19 stu8…”

stu1    stu2    stu9    stu3

stu5
(student embedding)
Methodology: vector space embedding

"All happy families are alike; each unhappy family is unhappy in its own way." (natural language)

“CS61A MATH1B SPA12 STAT200B CUE100A CS188 CS C267 CS268 ENN1B.” (course enrollments)

SPA12_A  “stu1 stu2 stu5 stu9 stu3 stu51 stu19 stu8…”
SPA12_B  “stu4 stu91 stu22 stu31 stu46 stu21 stu67 stu39…”
SPA12_C  “stu33 stu72 stu41 stu28 stu71 stu89 stu99 stu11…”
SPA12_F  “stu37 stu27 stu47 stu20 stu30 stu55 stu12 stu63…”

Fall 2009  Fall 2010
Methodology: skip-gram model selection

- Hyperparameters: vector size, window size, min_count, negative samples
- Model selected based on how well majors are clustered together
- No student validation set available
- Course validation set accuracy correlates with subject clustering (Hungarian metric) with $p = 0.499$
Dataset

- 3.6M enrollments at UCB from Fall 2008 through Fall 2015
- 110,335 undergraduates (anon)
- 38,147 graduates (anon)
- 9,038 unique lectures courses
  - across 17 colleges
  - 124 departments

![Line Plot of Total Unique Students Enrolled in Classes per Semester](image)

Data obtained with permission from the UC Office of the Registrar & CPHS
Dataset

- Overtime graduation less of an issue in the UC system than at the California State and CC levels
- Using the data available to us
- UCB Chancellor has called for improvement

Top overtime majors among freshman entering in 2008-2010

<table>
<thead>
<tr>
<th>Major</th>
<th>Overtime graduate number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integrative Biology</td>
<td>87 (12%)</td>
</tr>
<tr>
<td>Interdisciplinary Studies</td>
<td>61 (14%)</td>
</tr>
<tr>
<td>American Studies</td>
<td>55 (11%)</td>
</tr>
<tr>
<td>Political Economy</td>
<td>49 (9%)</td>
</tr>
<tr>
<td>Sociology</td>
<td>44 (9%)</td>
</tr>
</tbody>
</table>

On-time/overtime IB majors by year

<table>
<thead>
<tr>
<th>Integrative biology</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-time graduate (&lt;= 4 years)</td>
<td>277</td>
<td>258</td>
<td>250</td>
</tr>
<tr>
<td>Overtime graduate (&gt; 4 years)</td>
<td>36</td>
<td>27</td>
<td>24</td>
</tr>
</tbody>
</table>
Case Study

• Integrative Biology majors will serve as case study
  • Physics, Chemistry, Math course requirements
  • Incoming freshman in 2008 & 2009
Methodology: evaluation

• Visualization of the student embedding
  • t-SNE visualization of all students in the dataset

• Proficiency in Chemistry, Physics, Math
  • Defined as the average vector of on-time graduated students in those majors (who entered in 2008)
  • Semester vector representations of IB students entering in 2009 compared (using cosine) to 2008 major averages

• Prediction of on-time graduation
  • Student level leave-on-out CV of 2009 IB cohort
  • Logistic regression and 200 node FFNN using features: student vector; cosine, Euclidian, and dot product of the student vector to each of the major averages
Visualization

- Each data point is a student
- Students from 2008-2015
- Color corresponds to major

2D projection of the student embedding using Barnes-Hutt t-SNE
Misc. notable observations
Majors associated with our case study of IB
Proficiency Results

Correlation of 2009 IB student-semester vectors to 2008 major requirement vectors

- Chemistry $R^2 = 0.10$
- Mathematics $R^2 = 0.05$
- Physics $R^2 = 0.13$
- Integrative Biology $R^2 = 0.23$
Proficiency Results

Correlation of 2009 IB student-semester vectors to an average of themselves
### On-time Prediction Results

Predicting if a student will graduate on-time based on their student-semester vector features

<table>
<thead>
<tr>
<th>Method</th>
<th>Input</th>
<th>Vector</th>
<th>Euclidian</th>
<th>Cosine similarity</th>
<th>Dot projection</th>
<th>Ave. Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>0.880</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>0.873</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>0.873</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>0.895</td>
</tr>
<tr>
<td>Neural Network 1 layer 200 nodes</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>0.873</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>0.874</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>0.873</td>
</tr>
<tr>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>0.956</td>
</tr>
</tbody>
</table>
On-time Prediction Results

Predicting if a student will graduate on-time based on their student-semester vector features.
Limitations

• We used on-time graduating 2008 freshman to correlate to 2009 students (on-time wouldn’t be known)
• Student-semester approach requires re-training every semester (subject to variability)
• In need of a better validation for model selection
Conclusion

• Robust on-time graduation prediction with
  • No demographic information
  • No manual feature engineering
  • Using only enrollment data
Future Work

• Additional semantic annotation of the vector space
  • Career information (from exist survey)
  • Financial aid burden

• Proficiency interventions involving intelligent pairing (i.e. Stu1 + Stu? = improved proficiency metric)

• RNN based student modeling work (askoski.berkeley.edu)
Thank You

Questions?

Yuetian Luo
UW-Madison, USA
Statistics
yloo86@wisc.edu

Zach A. Pardos
UC Berkeley, USA
Information / Education
zp@berkeley.edu

PDF: tiny.cc/ylzpEAAI