Opportunities and Challenges for Education Research on Coursera

Andrew Maas
Coursera, Stanford University
Analytics @ Coursera

Turadg Aleahmad  
Thomas Barthelemy  
Zhenghao Chen  
Andrew Chou  
Peter Lofgren  
Emma Pierson

Tom Do

Andrew Maas. NIPS 2013 Workshop on Data Driven Education
Outline

1. **Demographics**: Diversity of students and courses
2. **Peer Assessments**: Developing tools for instruction and assessment at scale
3. **Item Response Theory**: Helping instructors understand student learning
Demographics
What’s different about Coursera?
Basic demographics

Gender distribution

- Female: 41.0%
- Male: 58.6%
- Other: 0.4%

Age distribution

- US: 26
- Non-US: 28, 55
Basic demographics

Geographic distribution:
- United States: 25%
- India: 10%
- Brazil: 5%
- Spain: 3%
- Canada: 2%
- United Kingdom: 2%
- Mexico: 2%
- Russian Federation: 2%
- Australia: 1%
- Germany: 1%

Educational attainment:
- Doctorate degree: 30%
- Professional school degree: 20%
- Master's degree: 15%
- Bachelor's degree: 10%
- Associate degree (academic): 5%
- Associate degree (vocational): 5%
- Some college: 4%
- High school: 3%
- Some high school: 2%
- Some primary or elementary: 1%
- No schooling: 1%

Andrew Maas. NIPS 2013 Workshop on Data Driven Education
Employment demographics

![Occupation and Employment status chart]

- Computer or mathematical education, training, or library
- Business or financial operations
- Arts, design, entertainment, sports, or media
- Architecture or engineering
- Healthcare practitioners or technical occupations
- Management
- Life, physical, or social science
- Office or administrative support
- Sales or sales-related
- Healthcare support
- Community or social service
- Legal
- Production
- Food preparation or serving related
- Transportation and materials moving
- Installation, maintenance, and repair
- Construction and extraction
- Farming, fishing, or forestry
- Personal care or service
- Protective service
- Building and grounds cleaning or maintenance

- Unable to work
- Retired
- Unemployed and not looking for work
- Unemployed and looking for work
- Homemaker or taking care of family member
- Self-employed (less than 35 hours per week)
- Self-employed (35 or more hours per week)
- Employed part-time (less than 35 hours per week)
- Employed full-time (35 or more hours per week)
Age varies by course

Among all Coursera students

[Bar chart showing age distribution by course]
Gender varies by course

Among all Coursera students

Course titles:
- Principles of Reactive Programming
- Functional Programming Principles in Scala
- Heterogeneous Parallel Programming
- Control of Mobile Robots
- Statistical Mechanics: Algorithms and Computations
- Computational Investing, Part I
- Pattern-Oriented Software Architectures for Concurrent and Networked Systems
Adjusted age distributions by course
Geographic variation in democracy interest
Peer Grading
A success in teaching tools
Peer Assessments in MOOCs

C. Kulkarni, PW Koh, H Le, D Chia, K Papadopoulos, J Cheng, D Koller, SR. Klemmer, Peer and Self Assessment in Massive Online Classes. March 2013
Peer Assessments in MOOCs

Stanford: Human-Computer Interaction

Staff Grade vs. Peer Grade

Andrew Maas. NIPS 2013 Workshop on Data Driven Education
Peer Assessments in MOOCs

![Graph showing correlation with staff grades vs. number of peer graders.](image)

- Red line: Peer Grades
- Green line: Debiased + Filtered by Time
- Blue line: Debiased + Filtered by Grader Grade
- Black line: Staff

Andrew Maas. NIPS 2013 Workshop on Data Driven Education
Peer Assessments in MOOCs

\[ b_v = \mathcal{N} \left( 0, \frac{1}{\eta_0} \right) \] (Grader bias)

\[ s_u = \mathcal{N} \left( \mu_0, \frac{1}{\gamma_0} \right) \] (Submission quality)

\[ z_u^v = \mathcal{N} \left( s_u + b_v, \frac{1}{\theta_1 s_v + \theta_0} \right) \] (Peer grade)

C Piech, J Huang, Z Chen, C Do, A Ng, D Koller, Tuned models of peer assessment in MOOCs. EDM 2013
Peer Assessments in MOOCs

Baseline
(81% within 10pp)

With debiasing model
(95% within 10pp)

C Piech, J Huang, Z Chen, C Do, A Ng, D Koller, Tuned models of peer assessment in MOOCs. EDM 2013
Item Response Theory
A challenge in instructor tools
Case study: Instructor-facing analytics

- Class composition different from typical on-campus students
- Instructors need to measure student progress to adapt or expand material
- Unclear how well existing educational tools perform in the Coursera environment:
  - Instructors struggle to generate large question banks (different from K-12 online education)
  - Tools need to be robust enough to work across a range of course content and student populations
Item Response Theory (IRT)

- Latent variable model to determine student proficiency
- 50+ years of application in education literature
- Commonly applied to check question bias (GRE/SAT)

**Question 1**

You are training a three layer neural network and would like to use backpropagation to compute the gradient of the cost function. In the backpropagation algorithm, one of the steps is to update $\Delta_{ij}^{(2)} := \Delta_{ij}^{(2)} + \delta_{j}^{(3)} \ast (a^{(2)})_i$ for every $i,j$. Which of the following is a correct vectorization of this step?

- $\Delta^{(2)} := \Delta^{(2)} + \delta^{(3)} \ast (a^{(2)})^T$
- $\Delta^{(2)} := \Delta^{(2)} + \delta^{(3)} \ast (a^{(3)})^T$
- $\Delta^{(2)} := \Delta^{(2)} + (a^{(3)})^T \ast \delta^{(2)}$
- $\Delta^{(2)} := \Delta^{(2)} + \delta^{(3)} \ast (a^{(2)})^T$

**Question 2**

Suppose $Theta1$ is a $2 \times 5$ matrix, and $Theta2$ is a $3 \times 6$ matrix. You set $thetaVec = [Theta1(:,); Theta2(:,)]$. Which of the following correctly recovers $Theta2$?

- $\text{reshape}(thetaVec(10:27), 3, 6)$
- $\text{reshape}(thetaVec(11:28), 3, 6)$
- $\text{reshape}(thetaVec(11:28), 6, 3)$
- $\text{reshape}(thetaVec(11:20), 3, 6)$
IRT Intuition

- Not all assessment total scores are equal
- Assume question grading is binary (no partial credit)
- What can we infer about questions and students?

<table>
<thead>
<tr>
<th></th>
<th>Question 1</th>
<th>Question 2</th>
<th>Question 3</th>
<th>Question 4</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhenghao</td>
<td>Correct</td>
<td>Correct</td>
<td>Correct</td>
<td>Wrong</td>
<td>75%</td>
</tr>
<tr>
<td>Emma</td>
<td>Correct</td>
<td>Wrong</td>
<td>Correct</td>
<td>Correct</td>
<td>75%</td>
</tr>
<tr>
<td>Turadg</td>
<td>Correct</td>
<td>Wrong</td>
<td>Correct</td>
<td>Wrong</td>
<td>50%</td>
</tr>
<tr>
<td>Peter</td>
<td>Correct</td>
<td>Correct</td>
<td>Wrong</td>
<td>Wrong</td>
<td>50%</td>
</tr>
<tr>
<td>Tom</td>
<td>Wrong</td>
<td>Correct</td>
<td>Wrong</td>
<td>Wrong</td>
<td>25%</td>
</tr>
</tbody>
</table>
IRT Intuition: Comparing students

- Emma correctly answers a question missed by all others
- Is Emma’s 75% is “better” than Zhenghao’s 75%?
- Turadg and Emma miss the same question. Implications?

<table>
<thead>
<tr>
<th></th>
<th>Question 1</th>
<th>Question 2</th>
<th>Question 3</th>
<th>Question 4</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhenghao</td>
<td>Correct</td>
<td>Correct</td>
<td>Correct</td>
<td>Wrong</td>
<td>75%</td>
</tr>
<tr>
<td>Emma</td>
<td>Correct</td>
<td>Wrong</td>
<td>Correct</td>
<td>Correct</td>
<td>75%</td>
</tr>
<tr>
<td>Turadg</td>
<td>Correct</td>
<td>Wrong</td>
<td>Correct</td>
<td>Wrong</td>
<td>50%</td>
</tr>
<tr>
<td>Peter</td>
<td>Correct</td>
<td>Correct</td>
<td>Wrong</td>
<td>Wrong</td>
<td>50%</td>
</tr>
<tr>
<td>Tom</td>
<td>Wrong</td>
<td>Correct</td>
<td>Wrong</td>
<td>Wrong</td>
<td>25%</td>
</tr>
</tbody>
</table>
# IRT Intuition: Comparing questions

- How does question 2 compare to question 3?
- Emma misses question 2
- Our two weakest students miss question 3

<table>
<thead>
<tr>
<th></th>
<th>Question 1</th>
<th>Question 2</th>
<th>Question 3</th>
<th>Question 4</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhenghao</td>
<td>Correct</td>
<td>Correct</td>
<td>Correct</td>
<td>Wrong</td>
<td>75%</td>
</tr>
<tr>
<td>Emma</td>
<td>Correct</td>
<td>Wrong</td>
<td>Correct</td>
<td>Correct</td>
<td>75%</td>
</tr>
<tr>
<td>Turadg</td>
<td>Correct</td>
<td>Wrong</td>
<td>Correct</td>
<td>Wrong</td>
<td>50%</td>
</tr>
<tr>
<td>Peter</td>
<td>Correct</td>
<td>Correct</td>
<td>Wrong</td>
<td>Wrong</td>
<td>50%</td>
</tr>
<tr>
<td>Tom</td>
<td>Wrong</td>
<td>Correct</td>
<td>Wrong</td>
<td>Wrong</td>
<td>25%</td>
</tr>
</tbody>
</table>
IRT Model

\[ P(\text{student s correctly answers question q}) = \sigma(\theta_q \ast \varphi_s + \alpha_q) \]

\( \theta_q \): Question discrimination

\( \alpha_q \): Question difficulty

\( \varphi_s \): Student proficiency

\[ \sigma(z) = \frac{1}{1 + \exp(-z)} \]
IRT Model

\[ P(\text{student s correctly answers question q}) = \sigma(\theta_q \ast \phi_s + \alpha_q) \]

\( \theta_q \): Question discrimination

\( \alpha_q \): Question difficulty

\( \phi_s \): Student proficiency

\[ \sigma(z) = 1/(1+\exp(-z)) \]
**IRT Example**

- Fit IRT parameters to maximize likelihood of observations

<table>
<thead>
<tr>
<th></th>
<th>Question 1</th>
<th>Question 2</th>
<th>Question 3</th>
<th>Question 4</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhenghao</td>
<td>Correct</td>
<td>Correct</td>
<td>Correct</td>
<td>Wrong</td>
<td>75%</td>
</tr>
<tr>
<td>Emma</td>
<td>Correct</td>
<td>Wrong</td>
<td>Correct</td>
<td>Correct</td>
<td>75%</td>
</tr>
<tr>
<td>Turadg</td>
<td>Correct</td>
<td>Wrong</td>
<td>Correct</td>
<td>Wrong</td>
<td>50%</td>
</tr>
<tr>
<td>Peter</td>
<td>Correct</td>
<td>Correct</td>
<td>Wrong</td>
<td>Wrong</td>
<td>50%</td>
</tr>
<tr>
<td>Tom</td>
<td>Wrong</td>
<td>Correct</td>
<td>Wrong</td>
<td>Wrong</td>
<td>25%</td>
</tr>
</tbody>
</table>

Bergner et al. (EDM 2012)
IRT Example

- Fit IRT parameters to maximize likelihood of observations
- Predict probability of correct under our model
- IRT models the observed data nearly perfectly

<table>
<thead>
<tr>
<th></th>
<th>Question 1</th>
<th>Question 2</th>
<th>Question 3</th>
<th>Question 4</th>
<th>% Correct</th>
<th>Proficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhenghao</td>
<td>97</td>
<td>86</td>
<td>86</td>
<td>2</td>
<td>75%</td>
<td>0</td>
</tr>
<tr>
<td>Emma</td>
<td>99</td>
<td>0</td>
<td>99</td>
<td>93</td>
<td>75%</td>
<td>3.7</td>
</tr>
<tr>
<td>Turadg</td>
<td>99</td>
<td>8</td>
<td>99</td>
<td>15</td>
<td>50%</td>
<td>1.2</td>
</tr>
<tr>
<td>Peter</td>
<td>84</td>
<td>99</td>
<td>8</td>
<td>0</td>
<td>50%</td>
<td>-1.2</td>
</tr>
<tr>
<td>Tom</td>
<td>6</td>
<td>99</td>
<td>0</td>
<td>0</td>
<td>25%</td>
<td>-3.7</td>
</tr>
</tbody>
</table>
IRT Example

![Graph showing item response theory example with student names and response curves]
IRT on Coursera

- Treat all quiz questions from a course as a single exam
- All students who answered at least 60% of questions
- Withhold 10% of student/question pairs for validation
- On average we have at least 100 questions and at least 5,000 students
IRT on Coursera

Graphical Models

Neural Nets

Behavioral Economics
IRT on Coursera

- Can identify broken or confusing questions
- Latent proficiencies improve final exam predictions beyond total quiz scores alone
- Basic IRT is robust enough to apply to many courses

- Unable to find multiple latent proficiencies
- Question bank size within a course limits analysis
- Questions appear easy on average which is unsurprising given the number of questions available per topic
Closing Thoughts: Opportunities

- Variety of courses and student population
- MOOCs as a technology require new pedagogy, especially considering student population variation
- Low hanging fruit! Basic models like IRT are not well understood across MOOC platforms
- Coursera is a great benchmark for whether a new tool works for students or instructors
- Instructors want to understand their students and are willing to try new approaches
- Coursera app platform is coming!
Closing Thoughts: Challenges

- Question banks are small. Young platform with specific question topics, no core concepts like K-12 or ETS
- Data is noisy. Attrition exists (though not as bad as rumors suggest)
- Student goals are mixed. Not everyone wants an “A”
- Course topics. Student learning trajectories more varied than K-12 concepts
- Analytics infrastructure is still being built. Exports for researchers are painful but quickly improving
Thank You!

Turadg Aleahmad
Thomas Barthelemy
Zhenghao Chen
Andrew Chou
Peter Lofgren
Emma Pierson

Tom Do

Andrew Maas. NIPS 2013 Workshop on Data Driven Education