What can machine learning do for open education?

Geoff Gordon
CMU Machine Learning
ggordon@cs.cmu.edu
Civilization advances by extending the number of important operations which we can perform without thinking about them.

—Alfred North Whitehead, 1911
CONTRIBUTION OF ML

- Machine learning can help us understand how students learn
CONTRIBUTION OF ML

- Machine learning can help us understand how students learn
  - Not just any ML, but latent variable ("hidden feature") discovery
CONTRIBUTION OF ML

- Machine learning can help us understand how students learn
  - Not just any ML, but latent variable ("hidden feature") discovery
  - Not just any latent variables, but highly structured ones
Why bother?

- **Student feedback**
  - what does the student know?
  - what are common causes for the mistake the student just made?

- **Instructor feedback**
  - what do the students know?
  - what skills does this course content address?
  - what skills *doesn’t* this course content address?

- **Evaluation**
  - help design rubrics for (peer, instructor) grading
  - cluster submissions by similar approach, skill level, …

- **Etc…**
GOAL: UNDERSTAND HOW STUDENTS LEARN SOMETHING

John took Joe for a ride on his boat. ___ boat was blue with a red stripe. A/The/[]

\[3x + 4 = x + 10\]
**Goal: Understand how students learn something**

John took Joe for a ride on his boat. **The** boat was blue with a red stripe.

\[ 3x + 4 = x + 10 \]

\[ x = 3 \]
To make metal cans, the ends for the cans are stamped out of square pieces of metal. The part of the square that is left over is then recycled as scrap. The manufacturer needs to know the area of the scrap for each end. Then the total weight of the scrap can be figured out.

1. The can end has a radius of 4 inches. If an end is punched out of a square piece of metal measuring 8 inches on a side, find the square inches of the scrap.

2. The can end has a radius of 8 inches. If an end is punched out of a square piece of metal measuring 16 inches on a side, find the square inches of the scrap.

3. The can end has a radius of 12 inches. If an end is punched out of a square piece of metal measuring 24 inches per side, find the square inches of the scrap.

NOTE: To find the area of the scrap metal remaining, you might have to first find the area of the can end, and the area of the metal square.

For this problem use an approximate value for pi, $\pi \approx 3.14$.
### Step-level Data

<table>
<thead>
<tr>
<th>Diagram Label</th>
<th>Unit</th>
<th>radius of the end of the can</th>
<th>length of the square ABCD</th>
<th>Area of the scrap metal</th>
<th>AREA OF SQUARE ABCD</th>
<th>AREA OF END OF CAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question 1</td>
<td>inches</td>
<td>4</td>
<td>8</td>
<td>13.76</td>
<td>64</td>
<td>50.24</td>
</tr>
<tr>
<td>Question 2</td>
<td>inches</td>
<td>8</td>
<td>16</td>
<td>55.04</td>
<td>256</td>
<td>200.96</td>
</tr>
<tr>
<td>Question 3</td>
<td>inches</td>
<td>12</td>
<td>24</td>
<td>123.84</td>
<td>576</td>
<td>452.16</td>
</tr>
</tbody>
</table>

- one step = fill in a box

- Record right/wrong, timing, use of hints, …
SIMPLEST MODEL: RASCH / 1-PARAMETER ITEM RESPONSE THEORY

\[ \ln \left( \frac{p_t}{1 - p_t} \right) = \theta_{i_t} + \beta_{j_t} \]

- \( p_t = P(\text{correct answer}) \)
- \( i_t = \text{student ID} \)
- \( j_t = \text{step ID} \)
- \( \Theta = \text{student mean (knowledge level)} \)
- \( \beta = \text{item mean (easy/difficult)} \)
Simplest model: Rasch / 1-parameter Item Response Theory

Steps in tutor

Each entry: does student i get step j right?

1

predict 1 if \( \theta_i + \beta_j > 0 \)

Students

\( \theta \)

\( \beta \)
STRUCTURE: SIMILARITY AMONG STEPS

- Learn a “step map”: each point = 1 step

Steps over here are more similar to each other...

... than to steps over here
STRUCTURE: SIMILARITY AMONG STEPS

- Learn a “step map”: each point = 1 step

Steps over here are more similar to each other...

... than to steps over here
# How Principal Components Analysis Got Famous

Each entry: how many stars does user $i$ give to movie $j$?
**Result of factoring**

Users

<table>
<thead>
<tr>
<th>Basis weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
</tr>
<tr>
<td>$u_2$</td>
</tr>
<tr>
<td>$u_3$</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>$u_n$</td>
</tr>
</tbody>
</table>

Movies

<table>
<thead>
<tr>
<th>Basis vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_1$</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>$v_k$</td>
</tr>
</tbody>
</table>

*Low-d basis = latent variables*

*Basis vectors represent latent properties of movies, e.g., “is a comedy”*
In our case (student-step data)

Students

\[ U_1, U_2, U_3, \ldots, U_N \]

Basis vectors are candidate “eigenskills”

Weights are students’ knowledge levels

Steps

\[ V_1, \ldots, V_K \]
Does it work?

- steps about pentagons
- steps about circles
  - other steps

Learned features let us predict held-out data better than chance ($\rho = .3, p < 0.0001$)
Does it work?

- steps about pentagons
- steps about circles
- other steps

Learned features let us predict held-out data better than chance ($\rho = .3$, $p < 0.0001$)

Yes, sort of ...
STRUCTURE: PRACTICE MAKES PERFECT

- PCA ignores step order — clearly wrong…
- Add model of student learning to PCA
  - based on “additive factor model” [Draney et al., 1995]
STRUCTURE: PRACTICE MAKES PERFECT

- PCA ignores step order — clearly wrong…
- Add model of student learning to PCA
  - based on “additive factor model” [Draney et al., 1995]
- Result: predictions of held-out data get slightly better
  - $\rho = .45$ ($p < 0.01$ vs. plain PCA)
- Step map still looks the same
- Meh…
WHAT WE REALLY WANT

- To be understandable to us humans, latents need to be sparse and binary (‘is about circles’, ‘requires subtracting areas’)
- Can’t do this fully automatically from this small data set (only 59 students, 370 steps)
- Challenge: can we discover sparse, binary, understandable latents automatically from MOOC-scale data?
“KC Hypothesis”

- Knowledge comes in atomic units (“KCs”)
- Each KC is learned independently (no transfer)
  - transfer among steps mediated by common KCs
  - or prerequisite structure (can’t learn algebra w/o knowing arithmetic)
- Each student has a (latent, scalar) proficiency level for each KC
  - learn/forget = transition to a higher/lower proficiency level
- Learning a KC happens only through exposure to that KC
  - problem, worked example, lecture, real life, …

CONSEQUENCES OF KC HYPOTHESIS

- Mistakes are at KC level: select wrong KC; apply right KC to wrong data; mistake in application of KC
  - identifying the KC at fault makes it easier to give student feedback
- If we can accurately
  - determine list of KCs
  - label instructional activities by KCs
- …then we immediately know the quality/coverage of our content
To make metal cans, the ends for the cans are stamped out of square pieces of metal. The part of the square that is left over is then recycled as scrap. The manufacturer needs to know the area of the square for each end. Then the total area...

[Stamper & Koedinger, AIED 2011]
Compose-by-addition

[Stamper & Koedinger, AIED 2011]
Why are some compose-by-addition steps harder?

Compose by addition.
Why are some compose-by-addition steps harder?

To make metal cans, the ends for the cans are stamped out of square pieces of metal. The part of the square that is left over is then recycled as scrap. The

compose by addition

hard

easy

medium
WHY ARE SOME COMPOSE-BY-ADDITION STEPS HARDER?
HYPOTHESIS: difference is in how much planning is needed

To make metal cans, the ends for the cans are stamped out of square pieces of metal. The part of the square that is left over is then recycled as scrap. The Hypothesis is that the difference in how much planning is needed is in how much scrap metal is to be composed.

Worksheet

compose by addition

plan to compose

subtract
KC discovery

Bars (shaded from the left) represent the actual error rate.

Lines show the predicted error rate for KC models.

The original KC model (square points).

A new KC model (round points) in which compose-by-addition is split into 3 KCs. The predictions of these three KCs produce better fits (with some exceptions) for the steps with high (decompose KC), medium (reduced compose-by-addition KC), and low (subtract KC) error rates.

[Stamper & Koedinger, AIED 2011]
Use data-driven model to redesign tutor

- New skill bars for planning skills
  - skill bars are a tutor interface to show students where they are in acquiring skills

- Sequence for gentle slope
  - adaptive fading of scaffolding

- New problems that focus on planning
  - next slide…
NEW PROBLEMS: ISOLATE PRACTICE ON PLANNING STEP

- Decompose complex problem into simpler ones

The given figure consists of a square and a parallelogram. The base of parallelogram QUAR is 7.5 meters and the height is 2.5 meters.

What is the area of the given figure?

A. Multiply area of SQRE by the area of QUAR: (7.5 * 7.5)(7.5 * 2.5)
B. Subtract the area of QUAR from the area of SQRE: (7.5 * 7.5) - (7.5 * 2.5)
C. Add the area of SQRE to the area of QUAR: (7.5 * 7.5) + (7.5 * 2.5)
D. Add together all sides of the figure and multiply by the height of the parallelogram: (7.5 + 7.5 + 7.5 + 7.5) * 2.5
NEW PROBLEMS: ISOLATE PRACTICE ON PLANNING STEP

- Decompose complex problem into simpler ones

The given figure consists of a square and a parallelogram. The base of parallelogram QUAR is 7.5 meters and the height is 2.5 meters.

What is the area of the given figure?

- C. Add the area of SQRE to the area of QUAR: (7.5 \times 7.5) + (7.5 \times 2.5)

- D. Add together all sides of the figure and multiply by the height of the parallelogram: (7.5 + 7.5 + 7.5 + 7.5) \times 2.5
Results

- More efficient: 25% less student time
  - instructional time by step type
- Better learning of planning skills
  - post-test %correct by item type

[Stamper & Koedinger, AIED 2011]
More structure: what’s in a KC?

- So far, each KC is just present or absent in a student or problem
- Nothing to distinguish algebra KCs from ESL KCs
- What’s going on under the hood as a student solves a problem?
RULE-BASED COGNITIVE MODEL

What does it look like inside the student’s brain?

- ... maybe a rule-based system
- ... in which case KCs might correspond to rules
  - :- president of US is Obama
  - constant C on LHS of equation E :- move C to RHS of E
Rule-based system

- Aka production system:
  - declarative knowledge held in working memory
  - production rules match declarative knowledge
  - and act on WM or external world

- Much cognitive modeling work endorses this claim explicitly or implicitly
  - ACT-R, SimStudent, Russell & Norvig, ...

- But two problems: uncertainty handling, representation learning
  - here’s where more ML research can help!
**Problem 1: Uncertainty**

*A day in the life of a rat*

<table>
<thead>
<tr>
<th>Trial</th>
<th>Bell?</th>
<th>Light?</th>
<th>Food?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>3</td>
<td>×</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>4</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>
Rat as Bayesian

- Priors over: how many trial types, sparsity of connections, reliability of connections, …

- (This is a common architecture for medical diagnosis systems)
**Quiz: Are you smarter than a rat?**

<table>
<thead>
<tr>
<th>Trial</th>
<th>Bell?</th>
<th>Light?</th>
<th>Food?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>3</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>100</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
**Quiz: Are you smarter than a rat?**

<table>
<thead>
<tr>
<th>Trial</th>
<th>Bell?</th>
<th>Light?</th>
<th>Food?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2</td>
<td>×</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>3</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>100</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>101</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
</tbody>
</table>
AND THE RAT SAYS...

<table>
<thead>
<tr>
<th>Effect name</th>
<th>2nd-order conditioning</th>
<th>Conditioned inhibition</th>
</tr>
</thead>
<tbody>
<tr>
<td>light-food trials</td>
<td>many</td>
<td>many</td>
</tr>
<tr>
<td>bell-light trials</td>
<td>few</td>
<td>many</td>
</tr>
<tr>
<td>test: bell predicts food?</td>
<td>↑</td>
<td>↓</td>
</tr>
</tbody>
</table>

- Both right! With more light-bell trials, evidence increases for a separate trial type.
**Bayesian rule learning in classical conditioning**

- Only fully Bayesian inference/learning captured both effects
  *Courville, Daw, Gordon, Touretzky, NIPS 2003*

- Few bell-light trials, 1 trial type: (bell, light, food) all associated

- More trials: (bell, light, no food) v. (light, food, no bell)

![Graph showing P(food | light) and P(food | bell) over Number of bell-light trials]
Problem 2: Representation Learning

Flaw with “KC = rule”: Many bugs come from weak features
Some student errors come from failure to correctly interpret (internally represent) a problem.

As student sees more and more examples like $3x + 5 = 8$, gets better and better “language model” to explain them (build internal representation).

—> some KCs must correspond to features of the improved language model.
**Experiment**

- Present algebra examples to a machine learning system
- As part of learning, induce a language model (an unsupervised probabilistic context free grammar) for algebra equations
- Make output of language model (grammar nonterminals, e.g., SignedNumber) available as features of each example
- Use these features in simulated problem-solving to discover KCs

[Li, Cohen, Koedinger, Matsuda, 2010]
New cognitive models are more accurate

Table 8.3: CV RMSE on SimStudent-Generated models and Human-Generated Models.

<table>
<thead>
<tr>
<th></th>
<th>Human-Generated Model</th>
<th>SimStudent-Discovered Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algebra</td>
<td>0.4024</td>
<td>0.3999</td>
</tr>
<tr>
<td>Stoichiometry</td>
<td>0.3501</td>
<td>0.3488</td>
</tr>
<tr>
<td>Fraction Addition</td>
<td>0.3232</td>
<td>0.3343</td>
</tr>
<tr>
<td>Article Selection</td>
<td>0.4044</td>
<td>0.4033</td>
</tr>
</tbody>
</table>

[Li, Cohen, Koedinger, Matsuda, 2010]
Open research question

- Can we build a new generation of rule-based system that has
  - rich uncertainty handling
  - integrated representation learning
- ... and use it to help us model student learning?
A key contribution of machine learning to education will be to help understand the educational content we’re creating and delivering.

Essential idea: ML models of structured latent variables.

Specifically, build and test hypotheses about the knowledge, procedures, and representations students use to solve problems.

- latents = KCs, rules, representations, strategies, …

Need to link uncertainty handling (traditional domain of ML) to new, harder situations encountered in understanding student knowledge.

Exciting time for research in ML and education!