Week 4 Video 5

Knowledge Inference: Advanced BKT
Friendly Warning

- This lecture is going to get mathematically intense by the end

- You officially have my permission to stop this lecture mid-way
Extensions to BKT

- Largely take the form of relaxing the assumption that parameters vary by skill, but are constant for all other factors
Advanced BKT

- Beck’s Help Model
- Individualization of L
- Moment by Moment Learning
- Contextual Guess and Slip
Note

- In this model, help use is not treated as direct evidence of not knowing the skill.

- Instead, it is used to choose between parameters.

- Makes two variants of each parameter:
  - One assuming help was requested.
  - One assuming that help was not requested.

Not learned

\[ p(T|H), \quad p(T|\sim H), \quad p(G|\sim H), \quad p(G|H) \]

Learned

\[ p(L_0|H), \quad p(L_0|\sim H), \quad 1-p(S|H), \quad 1-p(S|\sim H) \]

correct

- Parameters per skill: 8
- Fit using Expectation Maximization
  - Takes too long to fit using Brute Force
Table 1. Comparing the parameters estimated by the KT model and the Help model

<table>
<thead>
<tr>
<th></th>
<th>KT model</th>
<th>Help model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No Help Given</td>
</tr>
<tr>
<td>Already know</td>
<td>0.618</td>
<td>0.660</td>
</tr>
<tr>
<td>Learn</td>
<td>0.077</td>
<td>0.083</td>
</tr>
<tr>
<td>Guess</td>
<td>0.689</td>
<td>0.655</td>
</tr>
<tr>
<td>Slip</td>
<td>0.056</td>
<td>0.058</td>
</tr>
</tbody>
</table>


Table 1. Comparing the parameters estimated by the KT model and the Help model

<table>
<thead>
<tr>
<th></th>
<th>KT model</th>
<th>Help model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No Help Given</td>
<td>Help Given</td>
</tr>
<tr>
<td>Already know</td>
<td>0.618</td>
<td>0.660</td>
<td>0.278</td>
</tr>
<tr>
<td>Learn</td>
<td>0.077</td>
<td>0.083</td>
<td>0.088</td>
</tr>
<tr>
<td>Guess</td>
<td>0.689</td>
<td>0.655</td>
<td>0.944</td>
</tr>
<tr>
<td>Slip</td>
<td>0.056</td>
<td>0.058</td>
<td>0.009</td>
</tr>
</tbody>
</table>
Note

- This model did not lead to better prediction of student performance

- But useful for understanding effects of help
  - We’ll discuss this more in week 8, on discovery with models
Advanced BKT

- Beck’s Help Model
- Individualization of L
- Moment by Moment Learning
- Contextual Guess and Slip
BKT-Prior Per Student

\[ p(L_0) = \text{Student’s average correctness on all prior problem sets} \]

- Not learned
  - \( p(G) \) to correct
  - \( p(T) \) to Learned

- Learned
  - 1\(-p(S)\) to correct
BKT-Prior Per Student

- Much better on
  - ASSISTments (Pardos & Heffernan, 2010)
  - Cognitive Tutor for genetics (Baker et al., 2011)

- Much worse on
  - ASSISTments (Pardos et al., 2011)
Advanced BKT

- Beck’s Help Model
- Individualization of $L_0$
- Contextual Guess and Slip
- Moment by Moment Learning
Contextual Guess-and-Slip

Contextual Guess and Slip model

Not learned

\[ p(G) \]

correct

\[ p(T) \]

Learned

\[ p(L_0) \]

\[ 1 - p(S) \]

correct
Contextual Slip: The Big Idea

- Why one parameter for slip
  - For all situations
  - For each skill

- When we can have a different prediction for slip
  - For each situation
  - Across all skills
In other words

- $P(S)$ varies according to context

- For example
  - Perhaps very quick actions are more likely to be slips
  - Perhaps errors on actions which you’ve gotten right several times in a row are more likely to be slips
Contextual Guess and Slip model

- Guess and slip fit using contextual models across all skills

- Parameters per skill: $2 + \frac{(P(S) \text{ model size})}{\text{skills}} + \frac{(P(G) \text{ model size})}{\text{skills}}$
How are these models developed?

1. Take an existing skill model
2. Label a set of actions with the probability that each action is a guess or slip, using data about the future
3. Use these labels to machine-learn models that can predict the probability that an action is a guess or slip, without using data about the future
4. Use these machine-learned models to compute the probability that an action is a guess or slip, in knowledge tracing
2. Label a set of actions with the probability that each action is a guess or slip, using data about the future.

- Predict whether action at time $N$ is guess/slip.
- Using data about actions at time $N+1, N+2$.
- This is only for labeling data!
- Not for use in the guess/slip models.
2. Label a set of actions with the probability that each action is a guess or slip, using data about the future

- The intuition:
  - If action $N$ is right
  - And actions $N+1, N+2$ are also right
    - It’s unlikely that action $N$ was a guess
  - If actions $N+1, N+2$ were wrong
    - It becomes more likely that action $N$ was a guess

- I’ll give an example of this math in few minutes...
3. Use these labels to machine-learn models that can predict the probability that an action is a guess or slip.

- Features distilled from logs of student interactions with tutor software.

- Broadly capture behavior indicative of learning.
  - Selected from same initial set of features previously used in detectors of:
    - gaming the system (Baker, Corbett, Roll, & Koedinger, 2008)
    - off-task behavior (Baker, 2007)
3. Use these labels to machine-learn models that can predict the probability that an action is a guess or slip

- Linear regression
  - Did better on cross-validation than fancier algorithms
- One guess model
- One slip model
Within Bayesian Knowledge Tracing

Exact same formulas

Just substitute a contextual prediction about guessing and slipping for the prediction-for-each-skill

4. Use these machine-learned models to compute the probability that an action is a guess or slip, in knowledge tracing
Contextual Guess and Slip model

- Effect on future prediction: very inconsistent

- Much better on Cognitive Tutors for middle school, algebra, geometry (Baker, Corbett, & Aleven, 2008a, 2008b)

- Much worse on Cognitive Tutor for genetics (Baker et al., 2010, 2011) and ASSISTments (Gowda et al., 2011)
But predictive of longer-term outcomes

- Average contextual P(S) predicts
  - post-test (Baker et al., 2010)
  - shallow learners (Baker, Gowda, Corbett, & Ocumpaugh, 2012)
  - college attendance several years later (San Pedro et al., 2013)
    - Higher P(S) means lower college attendance, once you control for student knowledge
  - STEM major several years later (San Pedro et al., 2013)
    - Higher P(S) means lower probability of STEM major, once you control for student knowledge
What does $P(S)$ mean?
What does P(S) mean?

- Carelessness? (San Pedro, Rodrigo, & Baker, 2011)
  - Maps very cleanly to theory of carelessness in Clements (1982)

- Shallow learning? (Baker, Gowda, Corbett, & Ocumpaugh, 2012)
  - Student’s knowledge is imperfect and works on some problems and not others, so it appears that the student is slipping
Advanced BKT

- Beck’s Help Model
- Individualization of L₀
- Contextual Guess and Slip
- Moment by Moment Learning
Moment-By-Moment Learning Model

Moment-By-Moment Learning Model (Baker, Goldstein, & Heffernan, 2010)

Not learned

\[ p(G) \]
\[ p(T) \]

Probability you just learned

Learned

\[ p(L_0) \]
\[ 1-p(S) \]

correct

\[ p(J) \]
\( \text{P}(\text{J}) \)

- \( \text{P}(\text{T}) = \) chance you will learn if you didn’t know it

- \( \text{P}(\text{J}) = \) probability you \text{JustLearned}
  - \( \text{P}(\text{J}) = \text{P}(\sim \text{L}_n \wedge \text{T}) \)
P(J) is distinct from P(T)

- For example:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P(L_n) = 0.1</td>
<td>P(L_n) = 0.96</td>
</tr>
<tr>
<td>P(T) = 0.6</td>
<td>P(T) = 0.6</td>
</tr>
<tr>
<td>P(J) = 0.54</td>
<td>P(J) = 0.02</td>
</tr>
</tbody>
</table>

Learning!      Little Learning
Labeling $P(J)$

- Based on this concept:
  - “The probability a student did not know a skill but then learns it by doing the current problem, given their performance on the next two.”

$$P(J) = P(\sim L_n \land T | A_{+1+2})$$

*For full list of equations, see Baker, Goldstein, & Heffernan (2011)*
We can calculate $P(\neg L_n \wedge T \mid A_{+1+2})$ with an application of Bayes’ theorem.

\[
P(\neg L_n \wedge T \mid A_{+1+2}) = \frac{P(A_{+1+2} \mid \neg L_n \wedge T) \cdot P(\neg L_n \wedge T)}{P(A_{+1+2})}
\]

Bayes’ Theorem:
\[
P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}
\]
Breaking down $P(A_{+1+2})$

- $P(\neg L_n \land T)$ is computed with BKT building blocks \{\(P(\neg L_n), P(T)\}\}
- $P(A_{+1+2})$ is a function of the only three relevant scenarios, \{\(L_n, \neg L_n \land T, \neg L_n \land \neg T\}\, and their contingent probabilities
- \[
    P(A_{+1+2}) = \\
    \quad P(A_{+1+2} \mid L_n) \, P(L_n) \\
    + P(A_{+1+2} \mid \neg L_n \land T) \, P(\neg L_n \land T) \\
    + P(A_{+1+2} \mid \neg L_n \land \neg T) \, P(\neg L_n \land \neg T)
\]
Breaking down \( P(A_{n+1} + 2 \mid L_n) \ P(L_n) \): One Example

\[
\begin{align*}
\mathbf{P(A_{n+1} + 2 = C, C \mid L_n)} & = P(\sim S)P(\sim S) \\
\mathbf{P(A_{n+1} + 2 = C, \sim C \mid L_n)} & = P(\sim S)P(S) \\
\mathbf{P(A_{n+1} + 2 = \sim C, C \mid L_n)} & = P(S)P(\sim S) \\
\mathbf{P(A_{n+1} + 2 = \sim C, \sim C \mid L_n)} & = P(S)P(S)
\end{align*}
\]

<table>
<thead>
<tr>
<th>skill</th>
<th>problemID</th>
<th>userID</th>
<th>correct</th>
<th>( L_{n-1} )</th>
<th>( L_n )</th>
<th>G</th>
<th>S</th>
<th>T</th>
<th>P(J)</th>
</tr>
</thead>
<tbody>
<tr>
<td>similar-figures</td>
<td>71241</td>
<td>52128</td>
<td>0</td>
<td>.56</td>
<td>.21036516</td>
<td>.299</td>
<td>.1</td>
<td>.067</td>
<td>.002799</td>
</tr>
<tr>
<td>similar-figures</td>
<td>71242</td>
<td>52128</td>
<td>0</td>
<td>.21036516</td>
<td>.10115955</td>
<td>.299</td>
<td>.1</td>
<td>.067</td>
<td>.00362673</td>
</tr>
<tr>
<td>similar-figures</td>
<td>71243</td>
<td>52128</td>
<td>1</td>
<td>.10115955</td>
<td>.30308785</td>
<td>.299</td>
<td>.1</td>
<td>.067</td>
<td>.00218025</td>
</tr>
<tr>
<td>similar-figures</td>
<td>71244</td>
<td>52128</td>
<td>0</td>
<td>.30308785</td>
<td>.12150209</td>
<td>.299</td>
<td>.1</td>
<td>.067</td>
<td>.00346442</td>
</tr>
<tr>
<td>similar-figures</td>
<td>71245</td>
<td>52128</td>
<td>0</td>
<td>.12150209</td>
<td>.08505184</td>
<td>.299</td>
<td>.1</td>
<td>.067</td>
<td>.00375788</td>
</tr>
</tbody>
</table>
Features of P(J)

- Distilled from logs of student interactions with tutor software

- Broadly capture behavior indicative of learning
  - Selected from same initial set of features previously used in detectors of
    - gaming the system (Baker, Corbett, Roll, & Koedinger, 2008)
    - off-task behavior (Baker, 2007)
    - carelessness (Baker, Corbett, & Aleven, 2008)
Features of $P(J)$

- All features use only **first response data**

- Later extension to include subsequent responses only increased model correlation very slightly – not significantly
Uses

- Patterns in P(J) over time can be used to predict whether a student will be prepared for future learning (Hershkovitz et al., 2013; Baker et al., 2013) and standardized exam scores (Jiang et al., 2015)

- P(J) can be used as a proxy for Eureka moments in Cognitive Science research (Moore et al., 2015)
Alternate Method

- Assume at most one moment of learning
- Try to infer when that single moment occurred, across entire sequence of student behavior

- (Van de Sande, 2013; Pardos & Yudelson, 2013)

- Some good theoretical arguments for this – more closely matches assumptions of BKT

- Has not yet been studied whether this approach has same predictive power as $P(\neg L_n \wedge T | A_{+1+2})$ method
Key point

- Contextualization approaches do not appear to lead to overall improvement on predicting within-tutor performance

- But they can be useful for other purposes
  - Predicting robust learning
  - Understanding learning better
Another type of extension to BKT is modifications to address multiple skills

Addresses some of the same goals as PFA

(Pardos et al., 2008; Koedinger et al., 2011)
Another type of extension to BKT is modifications to include item difficulty.

 Addresses some of the same goals as IRT.

(Pardos & Heffernan, 2011; Khajah, Wing, Lindsey, & Mozer, 2013)
Next Up

- Knowledge Structure Inference: Q-Matrices