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

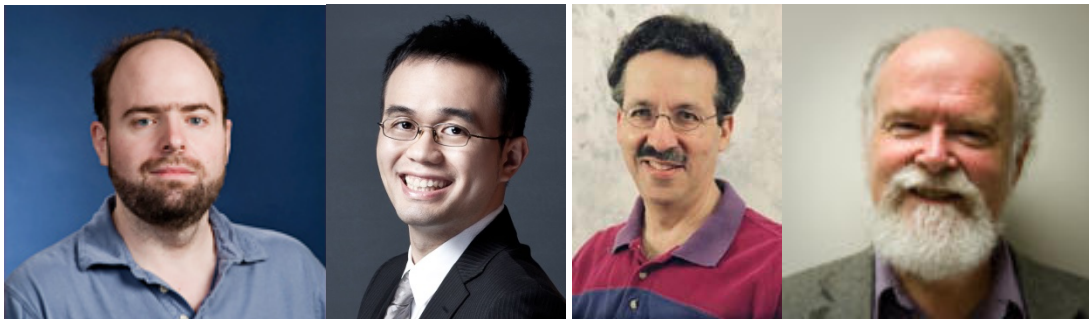
- Largely take the form of relaxing the assumption that parameters vary by skill, but are constant for all other factors

Advanced BKT

- Beck's Help Model
- Individualization of L_o
- Moment by Moment Learning
- Contextual Guess and Slip

Beck, Chang, Mostow, & Corbett 2008

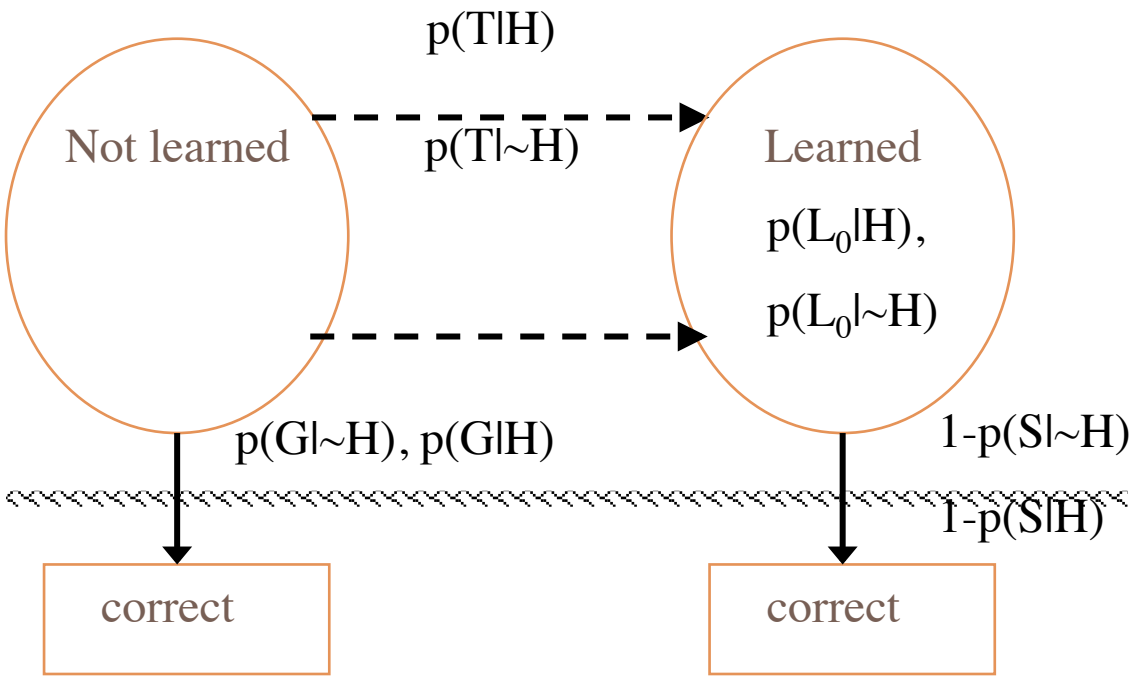
- Beck, J.E., Chang, K-m., Mostow, J., Corbett, A. (2008) Does Help Help? Introducing the Bayesian Evaluation and Assessment Methodology. *Proceedings of the International Conference on Intelligent Tutoring Systems.*



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

Beck et al.'s (2008) Help Model



Beck et al.'s (2008) Help Model

- Parameters per skill: 8
- Fit using Expectation Maximization
  - Takes too long to fit using Brute Force

Beck et al.'s (2008) Help Model

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

	KT model	Help model	
		No Help Given	Help Given
Already know	0.618	0.660	0.278
Learn	0.077	0.083	0.088
Guess	0.689	0.655	0.944
Slip	0.056	0.058	0.009

Beck et al.'s (2008) Help Model

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

	KT model	Help model	
		No Help Given	Help Given
Already know	0.618	0.660	0.278
Learn	0.077	0.083	0.088
Guess	0.689	0.655	0.944
Slip	0.056	0.058	0.009

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

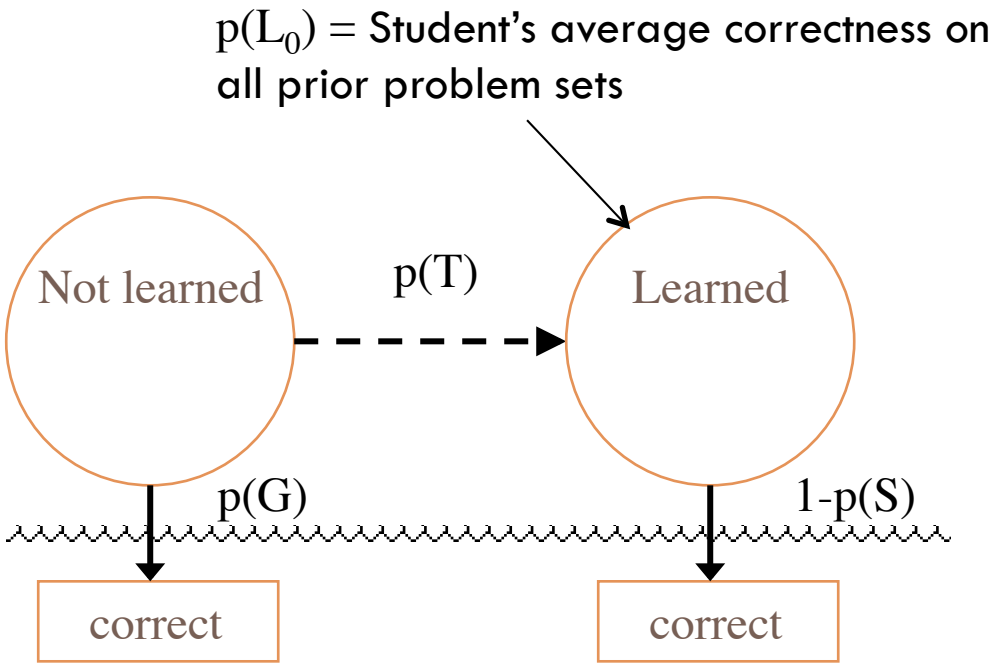
Advanced BKT

- Beck's Help Model
- Individualization of L_o
- Moment by Moment Learning
- Contextual Guess and Slip

Pardos & Heffernan (2010) BKT-Prior Per Student Model

- Pardos, Z.A., Heffernan, N.T. (2010) Modeling individualization in a bayesian networks implementation of knowledge tracing. *Proceedings of User Modeling and Adaptive Personalization.*

BKT-Prior Per Student



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)

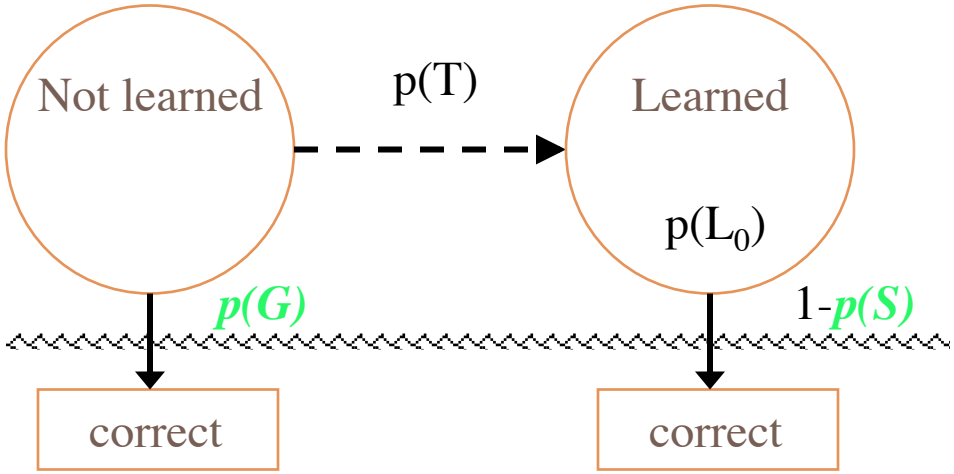
Advanced BKT

- Beck's Help Model
- Individualization of L_o
- Contextual Guess and Slip
- Moment by Moment Learning

Contextual Guess-and-Slip

- Baker, R.S.J.d., Corbett, A.T., Aleven, V. (2008) More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, 406-415.

Contextual Guess and Slip model



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 + (P (S) model size)/skills + (P (G) model size)/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

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

- Within Bayesian Knowledge Tracing
- Exact same formulas
- Just substitute a contextual prediction about guessing and slipping for the prediction-for-each-skill

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

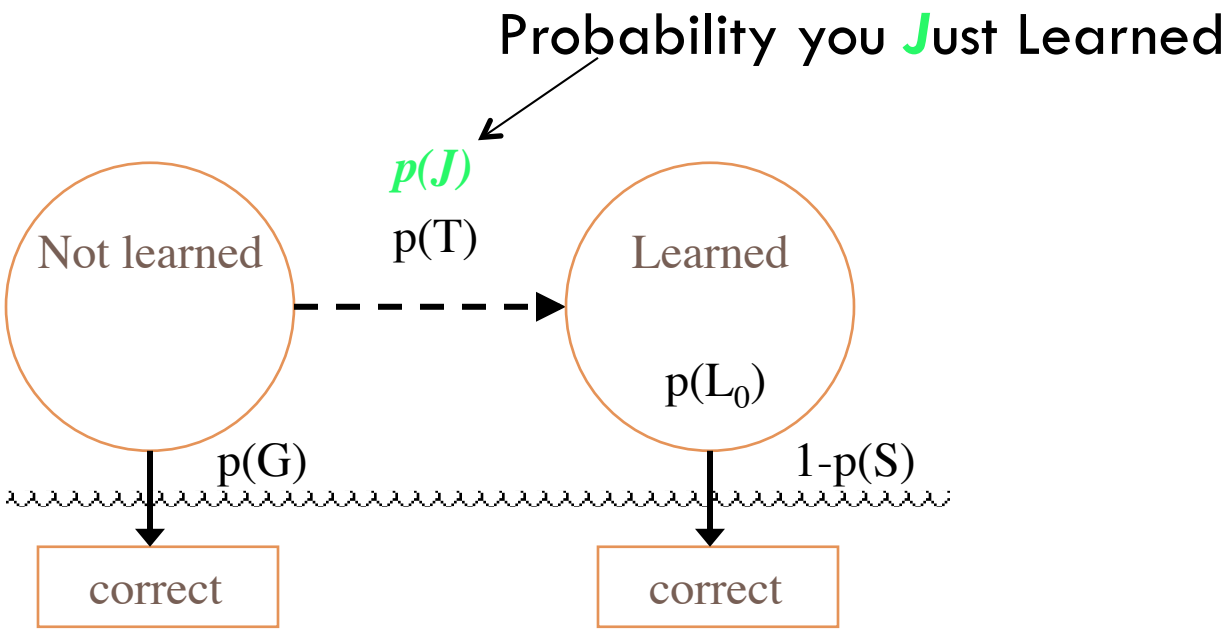
Advanced BKT

- Beck's Help Model
- Individualization of L_o
- Contextual Guess and Slip
- Moment by Moment Learning

Moment-By-Moment Learning Model

- Baker, R.S.J.d., Goldstein, A.B., Heffernan, N.T. (2011) Detecting Learning Moment-by-Moment. *International Journal of Artificial Intelligence in Education*, 21 (1-2), 5-25.

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



P(J)

- P(T) = chance you will learn if you didn't know it

- P(J) = probability you JustLearned
  - P(J) = P(~L_n ^ T)

P(J) is distinct from P(T)

- For example:

$P(L_n) = 0.1$	$P(L_n) = 0.96$
$P(T) = 0.6$	$P(T) = 0.6$
$P(J) = 0.54$	$P(J) = 0.02$
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."

$$P(J) = P(\sim L_n \text{ ^ } T \mid A_{+1+2})$$

*For full list of equations, see
Baker, Goldstein, & Heffernan (2011)

Breaking down $P(\sim L_n \wedge T \mid A_{+1+2})$

- We can calculate $P(\sim L_n \wedge T \mid A_{+1+2})$ with an application of Bayes' theorem

- $$P(\sim L_n \wedge T \mid A_{+1+2}) = \frac{P(A_{+1+2} \mid \sim L_n \wedge T) * P(\sim L_n \wedge T)}{P(A_{+1+2})}$$

Bayes' Theorem: $$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

Breaking down $P(A_{+1+2})$

- $P(\sim L_n \wedge T)$ is computed with BKT building blocks $\{P(\sim L_n), P(T)\}$
- $P(A_{+1+2})$ is a function of the only three relevant scenarios, $\{L_n, \sim L_n \wedge T, \sim L_n \wedge \sim T\}$, and their contingent probabilities

- $P(A_{+1+2}) =$

$$P(A_{+1+2} \mid L_n)\, P(L_n)$$
$$+ P(A_{+1+2} \mid \sim L_n \wedge T)\, P(\sim L_n \wedge T)$$
$$+ P(A_{+1+2} \mid \sim L_n \wedge \sim T)\, P(\sim L_n \wedge \sim T)$$

Breaking down $P(A_{+1+2} \mid L_n)\, P(L_n)$: One Example

- $P(A_{+1+2} = C, C \mid L_n) = P(\sim S)P(\sim S)$
- $P(A_{+1+2} = C, \sim C \mid L_n) = P(\sim S)P(S)$
- $P(A_{+1+2} = \sim C, C \mid L_n) = P(S)P(\sim S)$
- $P(A_{+1+2} = \sim C, \sim C \mid L_n) = P(S)P(S)$

skill	problemID	userID	correct	L_{n-1}	L_n	G	S	T	P(J)
similar-figures	71241	52128	0	.56	.21036516	.299	.1	.067	.002799
similar-figures	71242	52128	0	.21036516	.10115955	.299	.1	.067	.00362673
similar-figures	71243	52128	1	.10115955	.30308785	.299	.1	.067	.00218025
similar-figures	71244	52128	0	.30308785	.12150209	.299	.1	.067	.00346442
similar-figures	71245	52128	0	.12150209	.08505184	.299	.1	.067	.00375788

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 $P(\sim L_n \text{ ^ } T \mid A_{+1+2})$ 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

Learn More

- 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)

Learn More

- 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