Inferring History-Dependent Memory Strength via Collaborative Filtering

Robert Lindsey, Jeff Shroyer, Michael Mozer
University of Colorado
December 8, 2012
Retrieval probability is strongly influenced by the temporal spacing of study opportunities.

Diagram:
- Study 1
- Interstudy interval
- Study 2
- Retention interval
- Test
- Time

Saturday, December 8, 2012
Retrieval probability is strongly influenced by the temporal spacing of study opportunities.
Retrieval probability is strongly influenced by the temporal spacing of study opportunities.
Effective teaching requires predicting the dynamic properties of memory. Item response theory is focused on measuring a fixed memory state at an instant in time.
Study A
Study Item X

Study A
Study Item Y

Study A
Study Item Z

Study B
Study Item X

Inference Problem

Study 1

Study 2

Study N

Training Data

Student A
Study Item Y

0 0 1

Student A
Study Item Z

0 1 1

Student B
Study Item X

1 0
Study 1
Student A
Study Item X

Study 2
Student A
Study Item Y

Study N
Student A
Study Item Z

Study N
Student B
Study Item X

Same student, other items
Same item, other students

Saturday, December 8, 2012
Naive Approach: Ignore Study History

**Likelihood**

\[ \pi_n = \sigma(\alpha_s - \delta_i) \]

- **Ability**: \(\alpha_s\)
- **Difficulty**: \(\delta_i\)

**Hierarchical Priors**

- \(\alpha_s \sim \text{Normal}(\mu_\alpha, \tau_\alpha)\)
- \((\mu_\alpha, \tau_\alpha) \sim \text{NG}(\mu_0^{(\alpha)}, \kappa_0^{(\alpha)}, a_0^{(\alpha)}, b_0^{(\alpha)})\)
- \(\delta_i \sim \text{Normal}(\mu_\delta, \tau_\delta)\)
- \((\mu_\delta, \tau_\delta) \sim \text{NG}(\mu_0^{(\delta)}, \kappa_0^{(\delta)}, a_0^{(\delta)}, b_0^{(\delta)})\)

---

Rasch Model

- Item: \(\delta_i\)
- Student: \(\alpha_s\)
- Trial: \(R\)
- Parameters: \(\mu_\delta, \tau_\delta\)
Additive Factor Model: Binning Approach

Likelihood

\[ \pi_n = \sigma(\alpha_s - \delta_i + h_{si}) \]

Hierarchical Priors

\[ \alpha_s \sim \text{Normal}(\mu_\alpha, \tau_\alpha) \]
\[ (\mu_\alpha, \tau_\alpha) \sim \text{NG}(\mu_0^{(\alpha)}, \kappa_0^{(\alpha)}, a_0^{(\alpha)}, b_0^{(\alpha)}) \]
\[ \delta_i \sim \text{Normal}(\mu_\delta, \tau_\delta) \]
\[ (\mu_\delta, \tau_\delta) \sim \text{NG}(\mu_0^{(\delta)}, \kappa_0^{(\delta)}, a_0^{(\delta)}, b_0^{(\delta)}) \]

\[ h_{si} = \sum_k w_k f_{si k} \]

\[ f_{si k} = ? \]
Simple Idea:
count the number of correct responses made and the total number of trials

\[ f_{sik} = ? \]
Contribution of the previous trials to the logit recall probability decays according to a power law with exponent $d_k$. This power law exponent depends on the retrieval probability at the time of that trial $\beta_{si}$ is a student-item specific fudge factor and $h$ is a time-scaling parameter. Pavlik and Anderson (2008) do not fully specify how they constrained the model. What little the paper does specify is nonsense. After each trial, adjusting $\alpha_s$ by 60 fl?1 is done. In my simulations of this model, I chose to fit it via Monte Carlo Maximum Likelihood. Because it's a non-Bayesian model, there's no principled way to make predictions for new items or new students. In our project, this is a major problem because new material is introduced weekly. Note that I find that this model typically does worse at making predictions than a baseline model, which assumes that retrieval probability for a student on an item is the fraction of times it has been correctly recalled in the past by that student.

$$\pi_n = \sigma \left( \alpha_s - \delta_i + \sum w \phi(c) \log(1 + n_{si}) - \phi \cdot \log(1 + n_{si}) \right)$$

$$p(\phi) \propto \text{constant}$$

There are some sound assumptions in the Pavlik and Anderson model, for example, the power law decaying influence of previous trials and the fact that $d_k$ decreases over time. I therefore cleaned up the model further—beyond the already cleaned up Equation 9—and added hierarchical Bayesian priors.

I assume the likelihood

$$\logit(p_{sik}) = \alpha_s - \delta_i + \sum w \phi(c) \log(1 + n_{si}) - \phi \cdot \log(1 + n_{si})$$

with priors

$$\alpha_s \sim \text{Normal}(\mu_{\alpha}, \sigma^2_{\alpha})$$

$$\delta_i \sim \text{Normal}(\mu_{\delta}, \sigma^2_{\delta})$$

and improper priors over $h$. Note that I split the ACT-R time-scaling parameter into two— one for when the student responded correctly and one for when the student responded incorrectly. I also dropped the log operator (I find empirically that doing so greatly helped the model.)

6 Time-Augmented Item Response Theory

$π_n = \sigma (\alpha_s - \delta_i + \sum w \phi(c) \log(1 + n_{si}) - \phi \cdot \log(1 + n_{si}))$
proportion recalled
depends on the retrieval probability at the time of that trial. This power-law exponent depends on the retrieval probability at the time of that trial. Alpha is a student-specific fudge factor and \( h \) is a time-scaling parameter. What little the paper does specify is nonsense. After each trial, Pavlik and Anderson do not fully specify how they constrained the model. What little the paper does specify is nonsense. In my simulations of this model, I chose to fit it via Monte Carlo Maximum Likelihood. Because it's a non-Bayesian model, there's no principled way to make predictions for new items or new students. In our project, this is a major problem because new material is introduced weekly. Note that I find that this model typically does worse at making predictions than a baseline model, which assumes that retrieval probability for a student on an item is the fraction of times it has been correctly recalled in the past by that student.

BACTR: Hierarchical Bayesian ACT-R

There are some sound assumptions in the Pavlik and Anderson model, such as the power-law decaying influence of previous trials and the fact that \( d_k \) decreases over time. I therefore cleaned up the model further – beyond the already cleaned up Equation 9 – and added hierarchical Bayesian priors. I assume the likelihood

\[
\text{logit}(p_{sik}) = \alpha_s - \delta_i + \left( \sum_{k < i} \phi(c) \right)\log(1 + n_{si}) - \left( \sum_{k > i} \phi(c) \right)\log(1 + n_{ci})
\]

where

\[
d_k \equiv m \exp(\text{logit}(p_{sik})) + b
\]

with priors

\[
\text{Normal}(\mu_{\alpha}, \sigma_{\alpha}^2) \quad \text{Normal}(\mu_{\delta}, \sigma_{\delta}^2) \quad \text{Normal}(\mu_{\phi}, \kappa_{\phi})
\]

and improper priors over \( h \) and \( m \), and \( \phi \). Note that I split the ACT-R time-scaling parameter into two—one for when the student responded correctly and one for when the student responded incorrectly. I also dropped the log operator, I find empirically that doing so greatly helped the model.
\[
\pi_n = \sigma(\alpha_s - \delta_i + h_{si})
\]

counts in a particular time window

\[
h_{si} = \sum_k \left[ w_k(c) \log(1 + n_{sik}^{(c)}) - w_k2 \log(1 + n_{sik}) \right]
\]

summation over time windows \(k\)

\[n_{s1}, n_{s1}^{(c)}\]

\[n_{s2}, n_{s2}^{(c)}\]

\[n_{s3}, n_{s3}^{(c)}\]

\[n_{s4}, n_{s4}^{(c)}\]

\[
\text{window 1}
\]

\[
\text{window 2}
\]

\[
\text{window 3}
\]

\[
\text{window 4}
\]

\[
\ldots
\]
\[ \pi_n = \sigma(\alpha_s - \delta_i + h_{si}) \]

summation over previous presentations \( j \)

\[ h_{si} = c \log(1 + \sum_j m_j t_{sij}^{-d}) \]

power law decay of memory trace \( j \)
Dataset

- Online vocabulary tutor
- 8th grade Spanish classes
- 180 students
- 409 study items
- 14 week study
- 600,000+ study trials

Simulations

Predict recall on day X given data from days 1 through X-1
Average Surprisal

Running Average

(good)

(bad)
Average Surprisal

(Pavlik & Anderson, 2008)
Average Surprisal

(Pavlik & Anderson, 2008)
To what extent can this scheme improve long term retention?
individualized predictions

Predictive Model

Control Policy

Intelligently chosen material reviewed by student

Experiment Timeline

14 weeks studying

Immediate test

Delayed test
5 weeks later

Delayed test
9 weeks later

Saturday, December 8, 2012
Regardless of what’s being learned, forgetting occurs. Accounting for it is important.

Collaborative filtering enables us to make strong inferences despite weak behavioral data.

We can use the models to intelligently choose individual items to present to individual students.