Adaptive Testing using a General Diagnostic Model

Jill[1]-Jênn[2] Vie[3] Fabrice Popineau[4]
Yolaine Bourda[4] Éric Bruillard[2]

[1] RIKEN Center for Advanced Intelligence Project
[2] ENS Paris-Saclay
[3] Université Paris-Saclay
[4] CentraleSupélec, Gif-sur-Yvette

Context

How to predict the performance of examinees while asking as few questions as possible to them?

(AKA: I have a bunch of log files, can I use them to improve my online course?)

Outline

- Context & Adaptive Tests
- Item Response Theory & Cognitive Diagnosis
- Metrics for experiments
- Extensions

Context

We consider dichotomous data of learners over questions or tasks.

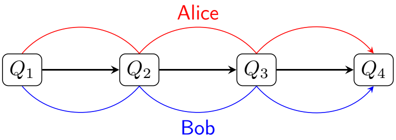
	Questions							
	1	2	3	4	5	6	7	8
Alice	0	1	1	1	0	0	0	1
Bob	1	0	1	1	0	0	0	1
Charles	1	0	1	0	0	0	0	0
Daisy	1	0	0	1	1	1	1	1
Everett	1	0	0	0	1	0	0	1
Filipe	0	1	0	1	1	1	1	1
Gwen	0	0	0	1	0	0	1	1
Henry	0	0	0	0	1	0	0	1
Ian	1	1	1	1	0	1	1	0
Jill	0	1	1	1	0	0	1	0
Ken	1	1	1	0	1	1	0	1

- Tests are too long, examinees are overtested
- Asking all questions to every examinee $\rightarrow$ boredom

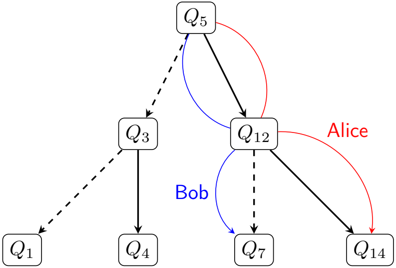
How to personalize this process?

While the test runs
  Pick the "best" next question to ask according to past

Non-Adaptive Test



Adaptive Test



Two main families in psychometrics

Do you care about explanative models or not?

Item response theory

- Answers can be explained by continuous hidden variables
- What parameters can we measure to predict performance?
- Infer them directly from student data
- Good for the examiner

Cognitive diagnosis

- Answers can be explained by the mastery or non-mastery of some knowledge components (KC)
- Expert (examiner) maps items to KCs
- Infer the KCs mastered $\Rightarrow$ predict performance
- Good for the examinee: they receive feedback

A first simple, yet reliable model: Rasch model

- $R_{ij} \in \{0, 1\}$ outcome of examinee i over item j (right/wrong)
- θ_i ability of examinee i
- d_j difficulty of item j
- $\Phi : x \mapsto 1/(1 + exp(-x))$

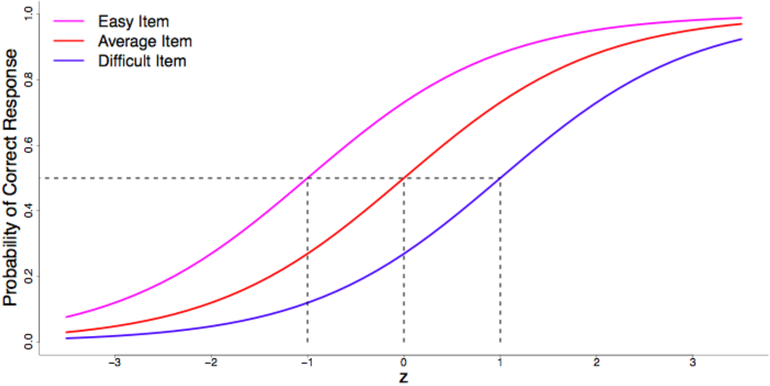
$$Pr(R_{ij} = 1) = \Phi(\theta_i - d_j).$$

Algorithm

- Learn d_j (and θ_i) for historic data (maximizing log-likelihood)
- When a new examinee arrives: initialize $\theta^{(0)} = 0$
- For each time $t = 0, \ldots, T - 1$:
  - Ask question of difficulty d_j closest to student ability $\theta^{(t)}$ (proba closest to 1/2)
  - Refine student ability $\theta^{(t+1)}$ (maximum likelihood estimate)

Response model

$$f_{d_j} : \theta_i \mapsto Pr(R_{ij} = 1) = \Phi(\theta_i - d_j).$$



Example!

Rasch model for 20 questions

	Q1	Q2	Q3	$\cdots$	Q19	Q20
Difficulty	–0.45	–0.40	–0.35	$\cdots$	0.45	0.50

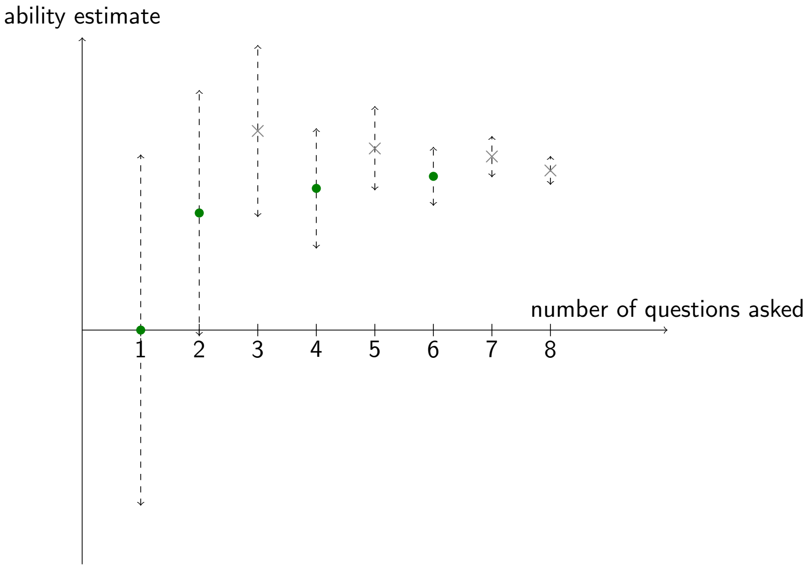
Question 10 is asked. Incorrect. $\Rightarrow$ Ability estimate $= -0.401$
Question 2 is asked. Correct! $\Rightarrow$ Ability estimate $= -0.066$
Question 9 is asked. Correct! $\Rightarrow$ Ability estimate $= 0.224$
Question 14 is asked. Correct! $\Rightarrow$ Ability estimate $= 0.478$

Feedback

Your ability estimate is 0.478.

(proba 0.7 to solve Q1, proba 0.5 to solve Q19)

Refine ability estimate over time



A cognitive diagnostic model: DINA model

- K possible skills
- $S = \{0,1\}^K$ potential latent states (subsets of mastered skills)
- Each question requires $x_j \in S$ skills.
- π: distribution of a new examinee over latent states

$$Pr(R_{ij} = 1) = \begin{cases} 1 - s_j & \text{if student } i \text{ masters all skills required } x_j \\ g_j & \text{otherwise.} \end{cases}$$

Algorithm

- Nothing to learn from historic data
- When a new examinee arrives: initialize $\pi^{(0)}$ to $Uniform(S)$
- For each time $t = 0, \ldots, T-1$:
  - Ask question that minimizes the expected entropy over $\pi^{(t+1)}$ according to the answer (using Bayes' rule)
  - Refine $\pi^{(t+1)}$ accordingly

Example of DINA-based test

Q-matrix: map between items and KCs

		Knowledge components			
		form	**mail**	**copy**	**url**
T1	Sending a mail	**form**	**mail**		
T2	Filling a form	**form**			
T3	Sharing a link			**copy**	**url**
T4	Entering a URL	**form**			**url**

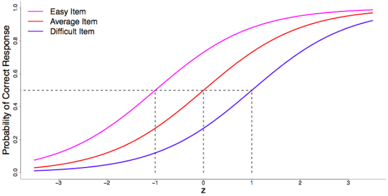
Task 1 is assigned. Correct!
$\Rightarrow$ **form** and **mail** may be mastered. No need to assign Task 2.
Task 4 is asked. Incorrect.
$\Rightarrow$ **url** may not be mastered. No need to use Task 3.

Feedback and inference

- *You master **form** and **mail** but not **url**.*

Comparison between IRT and CD

Rasch model



- Difficulty of questions
- Ability of learners
- Learners can be ranked
- No need of domain knowledge

Cognitive diagnosis

	C_1	C_2	C_3
Q_1	1	0	0
Q_2	0	1	1
Q_3	1	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$

- KCs required for each question
- Mastery or non-mastery of every KC for each learner
- Learners get feedback
- No need of prior data

GenMA: combining MIRT and a q-matrix

Rasch model

- Perf. depends on difference between learner ability and question difficulty
- Same as Elo ratings

$$\Phi(\theta_i - d_j)$$

Multidimensional Item Response Theory

- Depends on correlation between ability and question parameters
- Hard to converge

$$\Phi(\theta_i^T d_j) = \Phi\left(\sum_{k=1}^{d} \theta_{ik} d_{jk}\right)$$

$(\theta_{ik})_k$: ability of learner i

$(d_{jk})_k$: difficulty of question j

GenMA

- Depends on correlation between ability and question parameters, but only for non-zero q-matrix entries
- Easy to converge

$$\Phi\left(\sum_{k=1}^{d} \theta_{ik} q_{jk} d_{jk} + \delta_j\right)$$

$(q_{jk})_k$: q-matrix entry

δ_j: bias of question j

Recap

MIRT

- Depends on the correlation between ability and question parameters
- Hard to converge

GenMA

- Depends on the correlation between ability and question parameters, but only for non-zero q-matrix entries

Experimental protocol

		Questions							
		1	2	3	4	5	6	7	8
Train	Alice	0	1	1	1	0	0	0	1
	Bob	1	0	1	1	0	0	0	1
	Charles	1	0	1	0	0	0	0	0
	Daisy	1	0	0	1	1	1	1	1
	Everett	1	0	0	0	1	0	0	1
	Filipe	0	1	0	1	1	1	1	1
	Gwen	0	0	0	1	0	0	1	1
Test	Henry	0	0	0	0	1	0	0	1
	Ian	1	1	1	1	0	1	1	0
	Jill	0	1	1	1	0	0	1	0
	Ken	1	1	1	0	1	1	0	1

- Train student set 80% $\rightarrow$ extract features
- Test student set 20% $\rightarrow$ simulate adaptive test
- Validation question set 25% $\rightarrow$ evaluate predictions

Framework for comparing adaptive testing models

procedure $\textsc{SimulateAdaptiveTest}(\text{model } M, I_{train}, I_{test})$
  $\alpha, \kappa \leftarrow \textsc{TrainingStep}(M, D[I_{train}])$
  for every examinee s of I_{test} **do**
    $\pi_0 \leftarrow \textsc{PriorInitialization}(\alpha)$
    for $t = 0, \ldots, |Q \setminus Q_{val}| - 1$ **do**
      $q_{t+1} \leftarrow \textsc{NextItem}(\{(q_k, r_k)\}_{k=1,\ldots,t}, \kappa, \pi_t)$
      Ask question q_{t+1} to examinee s
      Receive outcome $r_{t+1} \in \{0, 1\}$
      $\pi_{t+1} \leftarrow \textsc{UpdateParams}(\{(q_k, r_k)\}_{k=1,\ldots,t+1}, \kappa)$
      $p \leftarrow \textsc{PredictPerformance}(\kappa, \pi_t, Q_{val})$
      $\sigma_{t+1} \leftarrow \textsc{EvaluatePerformance}(p, D[s][Q_{val}])$

Performance evaluation

.6	.1	.6	.7	.9	.1	.5	.5	.3	.7	.9	.4	.1	.6	.6	.7	.3	.7	.6	.3
						T													

2 correct predictions over 5 $\rightarrow$

.8	.4	.8	.6	.4
F	F	T	F	T

.6	.7	.6	.7	.9	.2	.6	.7	.4	.8	.9	.5	.6	.9	.9	.8	.4	.8	.6	.4
						T					F								

3 correct predictions over 5 $\rightarrow$

.6	.4	.8	.4	.4
F	F	T	F	T

We compute accuracy and log loss:

$$logloss(y^*, y) = \frac{1}{n} \sum_{k=1}^{n} \log(1 - |y_k^* - y_k|).$$

GenMA

Feedback

- The estimated ability $\theta_i = (\theta_{i1}, \ldots, \theta_{iK})$
- Proficiency over several KCs

Inference

- Compute the probability of success over the remaining questions

Example

- After 4 questions have been asked
- Predicted performance: $[.62, .12, .42, .13, .12]$
- True performance: $[T, F, T, F, F]$
- Computed logloss (error) is 0.350.

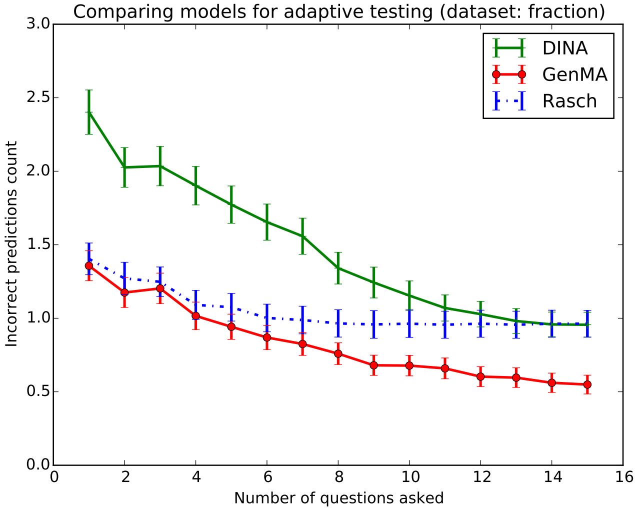
Real dataset: Fraction subtraction (DeCarlo, 2010)

- 536 middle-school students
- 20 questions of fraction subtraction
- 8 KCs

Description of the KCs

- convert a whole number to a fraction
- simplify before subtracting
- find a common denominator
- ...

Results



4 questions over 15 are enough to get a mean accuracy of $4/5$.

Summing up

Rasch model

- Really simple, competitive with other models
- But unidimensional, needs prior data, not formative

DINA model

- Formative, can work without prior data
- Needs a q-matrix

GenMA

- Multidimensional
- Formative because dimensions match KCs
- Needs a q-matrix and prior data
- Faster convergence than MIRT

Other models

Performance factor analysis

$$Pr(R_{ij} = 1) = \Phi\left(\theta_i + \sum_k q_{jk}\beta_k + \sum_k q_{jk}\gamma_k N_{ik}\right)$$

- θ_i ability of examinee i
- β_k bias of all items over KC k
- N_{ik} how many times examinee had opportunity to learn KC k
- γ_k bonus bias for each opportunity

Bandit

Ask questions so as to maximize the learning progress of the student: how well he performed recently to how well he performed before.

Further work

Consider graphs of prerequisites over KCs

Implemented in a live certification for the French MoE
(L@S 2017 poster)
Code under GPLv3 license `pix.beta.gouv.fr`

Adapting the process according to a group of answers

Method for multistage testing (ongoing work)

Train higher-dimension MIRT models

- Ongoing work (EDM 2018 submission)
- Managed to train MIRT sparse models up to 15 dimensions

Thank you for your attention!

jilljenn.github.io

Jill-Jênn Vie, Fabrice Popineau, Yolaine Bourda, and Éric Bruillard. “Adaptive Testing Using a General Diagnostic Model”. In: *European Conference on Technology Enhanced Learning*. Springer. 2016, pp. 331–339

Jill-Jênn Vie, Fabrice Popineau, Yolaine Bourda, and Éric Bruillard. “A Review of Recent Advances in Adaptive Assessment”. In: *Learning analytics: Fundaments, Applications, and Trends*. Springer, 2017, pp. 113–142

Jill-Jênn Vie, Fabrice Popineau, Françoise Tort, Benjamin Marteau, and Nathalie Denos. “A Heuristic Method for Large-Scale Cognitive-Diagnostic Computerized Adaptive Testing”. In: *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*. ACM. 2017, pp. 323–326

Do you have any questions?

jill-jenn.vie@riken.jp