

Modèles de tests adaptatifs pour le diagnostic de connaissances dans un cadre d'apprentissage à grande échelle

Jill-Jênn Vie

5 décembre 2016

école —
normale —
supérieure —
paris-saclay —

université
PARIS-SACLAY


CentraleSupélec

Plan

1. Contexte des évaluations à grande échelle
2. État de l'art précédent et évaluation qualitative
3. Système de comparaison de modèles
4. GenMA : un modèle de test adaptatif pour le diagnostic
5. InitialD : une stratégie de choix des premières questions
6. Conclusions et perspectives

Plan

1. Contexte des évaluations à grande échelle
2. État de l'art précédent et évaluation qualitative
3. Système de comparaison de modèles
4. GenMA : un modèle de test adaptatif pour le diagnostic
5. InitialD : une stratégie de choix des premières questions
6. Conclusions et perspectives

Contexte des évaluations à grande échelle

Certifications (par ex. TOEFL)

- Besoin de classer les apprenants
- Une évaluation sommative suffit

Retour : score (+ barre d'admissibilité)

Formations (par ex. MOOC)

- Besoin de faire progresser les apprenants
- Besoin d'une évaluation formative

Retour : points maîtrisés, points à retravailler

Grande échelle $\Rightarrow$ Besoin d'automatiser les évaluations

Besoin d'évaluations personnalisées

Pour les enseignants

Pluralité des profils dans les formations sans prérequis vérifiés
⇒ **Tests de positionnement** : identifier les connaissances déjà acquises par l'apprenant

Pour les apprenants

- Lassant de répondre à trop de questions
- Frustrant d'avoir des questions peu adaptées à leur niveau

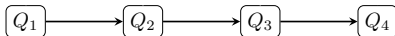
⇒ Adapter le niveau de difficulté à chaque question

Tests adaptatifs

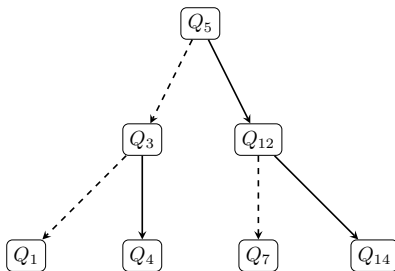
Tant que le test court

Choisir la question suivante en fonction des réponses précédentes

Test non adaptatif



Test adaptatif

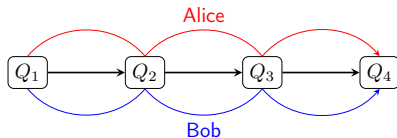


Tests adaptatifs

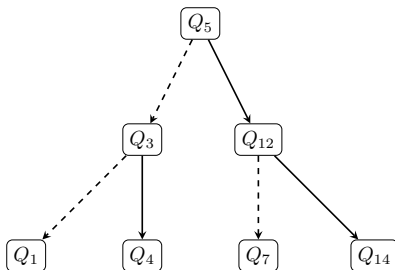
Tant que le test court

Choisir la question suivante en fonction des réponses précédentes

Test non adaptatif



Test adaptatif

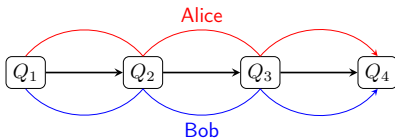


Tests adaptatifs

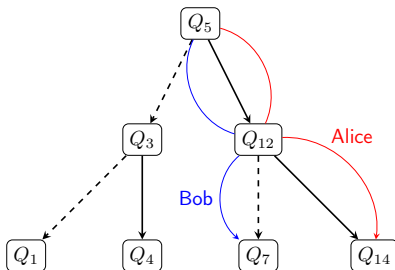
Tant que le test court

Choisir la question suivante en fonction des réponses précédentes

Test non adaptatif



Test adaptatif

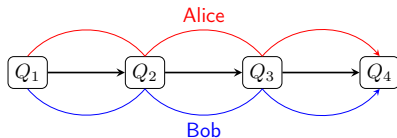


Tests adaptatifs

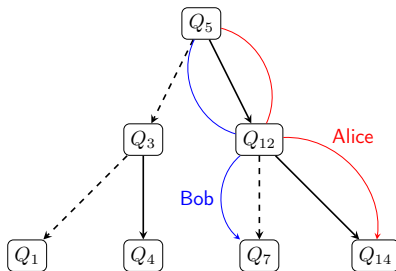
Tant que le test court

Choisir la question suivante en fonction des réponses précédentes

Test non adaptatif



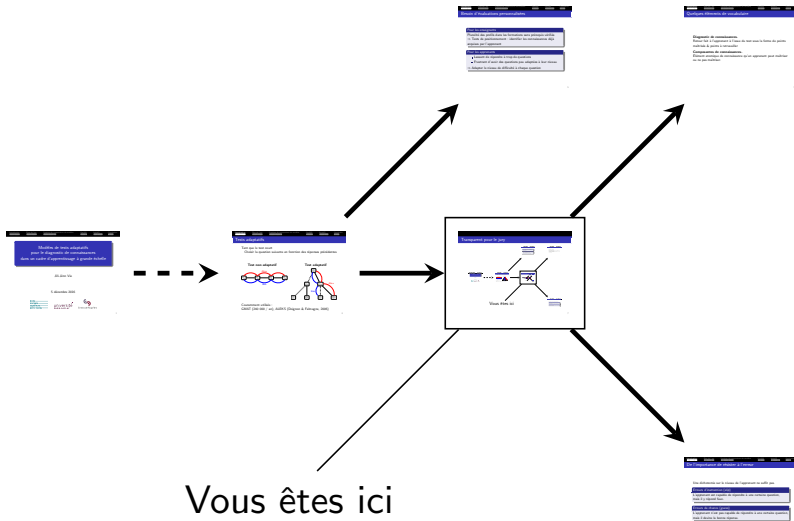
Test adaptatif



Couramment utilisés :

GMAT (240 000 / an), ALEKS (Doignon & Falmagne, 2006)

Transparent pour le jury



Quelques éléments de vocabulaire

Diagnostic de connaissances.

Retour fait à l'apprenant à l'issue du test sous la forme de points maîtrisés & points à retravailler

Composantes de connaissances.

Élément atomique de connaissance qu'un apprenant peut maîtriser ou ne pas maîtriser.

De l'importance de résister à l'erreur accidentelle

Une dichotomie sur le niveau de l'apprenant ne suffit pas.

Erreurs d'inattention (*slip*)

L'apprenant est capable de répondre à une certaine question, mais il y répond faux.

Erreurs de chance (*guess*)

L'apprenant n'est pas capable de répondre à une certaine question, mais il devine la bonne réponse.

Problèmes et contributions

- Comment comparer différents modèles de tests adaptatifs ?
 - Méthodologie de comparaison qualitative et quantitative
- Quel modèle choisir pour une situation donnée ?
 - Comparaison qualitative
- Comment améliorer un modèle de test adaptatif à partir d'un historique de réponses ?
 - Un modèle hybride adaptatif de diagnostic de connaissances
- Comment choisir les premières questions à poser ?
 - Stratégie pour choisir les premières questions à poser

Jeux de données divers utilisés dans la thèse

Soustractions de fractions (DeCarlo, 2010)

536 collégiens, 20 questions, 8 composantes de connaissance (simplifier une fraction, trouver le dénominateur commun, etc.)

Test multidisciplinaire SAT

296 apprenants, 40 questions.

Examen d'anglais ECPE

2922 apprenants, 28 questions, 3 composantes de connaissances.

TIMSS: Trends in International Mathematics and Science Study

757 collégiens de 4^e, 23 questions, 13 comp. de connaissances.

Castor informatique

58939 collégiens de 6^e–5^e, 17 tâches d'algorithmique.

Plan

1. Contexte des évaluations à grande échelle
2. État de l'art précédent et évaluation qualitative
3. Système de comparaison de modèles
4. GenMA : un modèle de test adaptatif pour le diagnostic
5. InitialD : une stratégie de choix des premières questions
6. Conclusions et perspectives

Modèles de théorie de la réponse à l'item

Caractéristiques des modèles

- K est le nombre de paramètres mesurés par apprenant
- Caractéristiques d'un apprenant $\theta_1, \dots, \theta_K \in \mathbf{R}$
- Caractéristiques d'une question $d_1, \dots, d_K \in \mathbf{R}$

Modèle de Rasch (Hambleton, 1985)

Performance dépend de la **différence** entre niveau θ de l'apprenant et difficulté d de la question

Φ logistique : $x \mapsto 1/(1 + e^{-x})$

$$\Phi(\theta - d)$$

Théorie de la réponse à l'item multidim.

Performance dépend de la **corrélacion** entre caractéristiques de l'apprenant et caractéristiques de la question

$$\Phi(\vec{\theta} \cdot \vec{d} + \delta) = \Phi \left(\sum_{k=1}^K \theta_k d_k + \delta \right)$$

δ : facilité de la question

Modèle 1 : Rasch

Calibrage automatique des difficultés de 20 questions

	Q1	Q2	Q3	...	Q19	Q20
Difficulté	-0,45	-0,40	-0,35	...	0,45	0,50

On pose la question 10. **Incorrect.**

⇒ Niveau estimé = -0,40

On pose la question 2. **Correct !**

⇒ Niveau estimé = -0,07

On pose la question 14. **Correct !**

⇒ Niveau estimé = 0,48

Retour

Votre niveau estimé est 0,48.

Inférence

- Les Q1-7 peuvent être résolues avec probabilité 0,7
- Les Q8-15 peuvent être résolues avec probabilité 0,6
- Les Q16-20 peuvent être résolues avec probabilité 0,5

Modèles basés sur les composantes de connaissances

Caractéristiques des questions & apprenants

- K composantes de connaissances
- États des apprenants $c_1, \dots, c_K \in \{0, 1\}^K$
- Pour les questions, une q-matrice :

	C_1	C_2	C_3	guess	slip
Q_1	1	1	0	0,12	0,08
Q_2	0	1	0	0,05	0,10

Modèle DINA (Cheng, 2009)

Toute composante impliquée dans la résolution d'une question doit être maîtrisée par l'apprenant.

→ Complexité $O(K2^K)$

$$\begin{cases} 1 - s & \text{si } \forall k | q_k = 1, k \\ & \text{est maîtrisée} \\ g & \text{sinon} \end{cases}$$

$(q_k)_k$: entrées de la q-matrice

g : guess de la question

s : slip de la question

Modèle 2 : DINA (Deterministic Input, Noisy And)

Spécification manuelle des composantes de connaissances

		Composantes de connaissances			
		form	mail	copy	url
T1	Envoyer un mail	form	mail		
T2	Remplir un formulaire	form			
T3	Partager un lien			copy	url
T4	Entrer une URL	form			url

On affecte la tâche 1. **Correct !**

⇒ **form** et **mail** : maîtrisées. Tâche 2 peu informative.

On affecte la tâche 4. **Incorrect.**

⇒ **url** semble non maîtrisée. Tâche 3 peu informative.

Retour

*Vous semblez maîtriser **form** et **mail** mais pas **url**.*

D'autres modèles

Modèle de hiérarchie sur les attributs (Leighton, 2003)

Considère un **graphe de prérequis** sur les composantes de connaissances pour réduire la complexité du modèle DINA.

Théorie des espaces de connaissances (Doignon et Falmagne, 2006)

- Presque identique au modèle de hiérarchie sur les attributs
- Ne considère pas de paramètres de *slip* et *guess*

SPARFA : Sparse Factor Analysis (Lan et al., 2012)

- Presque identique au modèle de théorie de réponse à l'item multidimensionnelle
- Tente d'extraire peu de paramètres de difficulté par question (vecteurs creux)

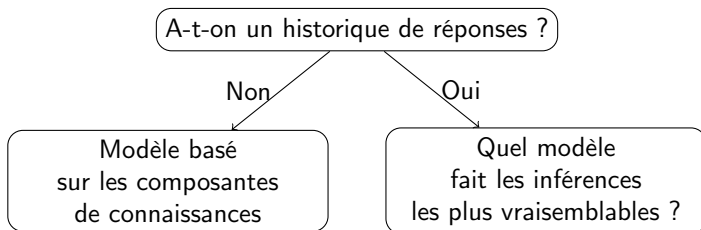
Comparaison qualitative

- Combien de paramètres de l'apprenant chaque modèle permet-il de mesurer ?
- Nécessite-t-il de spécifier manuellement des paramètres ?
- Nécessite-t-il un historique de réponses ou non ?

	Nb paramètres	Calibrage	Zéro donnée ?
Rasch	1	Auto	Non
TRI multidimensionnelle	$K \leq 4$	Auto	Non
Sparse Factor Analysis	$K \leq 16$	Auto	Non
DINA	$K \leq 15$	Manuel	Oui
Hierarchie sur les attributs	$K \leq 90$	Manuel	Oui
Espaces de connaissances	$K \leq 90$	Manuel	Oui

Publié comme chapitre d'un livre sur les *learning analytics*
(Vie, Bourda, Popineau et Bruillard, 2016)

Aide au choix d'un modèle



Rasch (théorie de la réponse à l'item)

- Simple à calculer
- Ne calcule qu'un score

DINA (basé sur les composantes de connaissances)

- Renvoie un diagnostic
- Coûteux : $O(K^2 K)$

Plan

1. Contexte des évaluations à grande échelle
2. État de l'art précédent et évaluation qualitative
3. **Système de comparaison de modèles**
4. GenMA : un modèle de test adaptatif pour le diagnostic
5. InitialD : une stratégie de choix des premières questions
6. Conclusions et perspectives

Comparer la capacité prédictive des modèles

Étant donné un modèle :

Existe-t-il un groupe de questions permettant à ce modèle d'inférer le comportement de l'apprenant sur des questions qui ne lui ont pas été posées ?

→ Questions qui ont un fort **pouvoir prédictif** sur les autres questions du test

Composants communs aux modèles de tests adaptatifs

- Un modèle de réponse de l'apprenant
- Phase de calibrage des caractéristiques des questions & apprenants à partir de l'historique du test (s'il existe)
- Initialisation des caractéristiques d'un nouvel apprenant
- Mise à jour des paramètres de l'apprenant
- Critère de choix de la question suivante

Méthode de validation bicroisée

		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

- Ensemble d'entraînement I_{train} : 80 % des apprenants
- Ensemble de validation Q_{val} : 25 % des questions

Protocole expérimental

procédure SIMULATEADAPTIVETEST(modèle M , I_{train} , I_{test})
 $\alpha, \kappa \leftarrow \text{TRAININGSTEP}(M, D[I_{train}])$
 pour tout apprenant s de l'ensemble I_{test} **faire**
 $\pi_0 \leftarrow \text{PRIORINITIALIZATION}(\alpha)$
 pour t de 0 à $|Q \setminus Q_{val}| - 1$ **faire**
 $q_{t+1} \leftarrow \text{NEXTITEM}(\{(q_k, r_k)\}_{k=1, \dots, t}, \kappa, \pi_t)$
 Poser la question q_{t+1} à l'apprenant s
 Récupérer la valeur de succès ou échec r_{t+1}
 $\pi_{t+1} \leftarrow \text{UPDATEPARAMS}(\{(q_k, r_k)\}_{k=1, \dots, t+1}, \kappa)$
 $p \leftarrow \text{PREDICTPERFORMANCE}(\kappa, \pi_t, Q_{val})$
 $\sigma_{t+1} \leftarrow \text{EVALUATEPERFORMANCE}(p, D[s][Q_{val}])$

Protocole expérimental

procédure SIMULATEADAPTIVETEST(modèle M , I_{train} , I_{test})
 $\alpha, \kappa \leftarrow \text{TRAININGSTEP}(M, D[I_{train}])$
pour tout apprenant s de l'ensemble I_{test} **faire**
 $\pi_0 \leftarrow \text{PRIORINITIALIZATION}(\alpha)$
 pour t de 0 à $|Q \setminus Q_{val}| - 1$ **faire**
 $q_{t+1} \leftarrow \text{NEXTITEM}(\{(q_k, r_k)\}_{k=1, \dots, t}, \kappa, \pi_t)$
 Poser la question q_{t+1} à l'apprenant s
 Récupérer la valeur de succès ou échec r_{t+1}
 $\pi_{t+1} \leftarrow \text{UPDATEPARAMS}(\{(q_k, r_k)\}_{k=1, \dots, t+1}, \kappa)$
 $p \leftarrow \text{PREDICTPERFORMANCE}(\kappa, \pi_t, Q_{val})$
 $\sigma_{t+1} \leftarrow \text{EVALUATEPERFORMANCE}(p, D[s][Q_{val}])$

Protocole expérimental

procédure SIMULATEADAPTIVETEST(modèle M , I_{train} , I_{test})
 $\alpha, \kappa \leftarrow \text{TRAININGSTEP}(M, D[I_{train}])$
 pour tout apprenant s de l'ensemble I_{test} **faire**
 $\pi_0 \leftarrow \text{PRIORINITIALIZATION}(\alpha)$
 pour t de 0 à $|Q \setminus Q_{val}| - 1$ **faire**
 $q_{t+1} \leftarrow \text{NEXTITEM}(\{(q_k, r_k)\}_{k=1, \dots, t}, \kappa, \pi_t)$
 Poser la question q_{t+1} à l'apprenant s
 Récupérer la valeur de succès ou échec r_{t+1}
 $\pi_{t+1} \leftarrow \text{UPDATEPARAMS}(\{(q_k, r_k)\}_{k=1, \dots, t+1}, \kappa)$
 $p \leftarrow \text{PREDICTPERFORMANCE}(\kappa, \pi_t, Q_{val})$
 $\sigma_{t+1} \leftarrow \text{EVALUATEPERFORMANCE}(p, D[s][Q_{val}])$

Protocole expérimental

procédure SIMULATEADAPTIVETEST(modèle M , I_{train} , I_{test})
 $\alpha, \kappa \leftarrow \text{TRAININGSTEP}(M, D[I_{train}])$
 pour tout apprenant s de l'ensemble I_{test} **faire**
 $\pi_0 \leftarrow \text{PRIORINITIALIZATION}(\alpha)$
 pour t de 0 à $|Q \setminus Q_{val}| - 1$ **faire**
 $q_{t+1} \leftarrow \text{NEXTITEM}(\{(q_k, r_k)\}_{k=1, \dots, t}, \kappa, \pi_t)$
 Poser la question q_{t+1} à l'apprenant s
 Récupérer la valeur de succès ou échec r_{t+1}
 $\pi_{t+1} \leftarrow \text{UPDATEPARAMS}(\{(q_k, r_k)\}_{k=1, \dots, t+1}, \kappa)$
 $p \leftarrow \text{PREDICTPERFORMANCE}(\kappa, \pi_t, Q_{val})$
 $\sigma_{t+1} \leftarrow \text{EVALUATEPERFORMANCE}(p, D[s][Q_{val}])$

Protocole expérimental

procédure SIMULATEADAPTIVETEST(modèle M , I_{train} , I_{test})
 $\alpha, \kappa \leftarrow \text{TRAININGSTEP}(M, D[I_{train}])$
 pour tout apprenant s de l'ensemble I_{test} **faire**
 $\pi_0 \leftarrow \text{PRIORINITIALIZATION}(\alpha)$
 pour t de 0 à $|Q \setminus Q_{val}| - 1$ **faire**
 $q_{t+1} \leftarrow \text{NEXTITEM}(\{(q_k, r_k)\}_{k=1, \dots, t}, \kappa, \pi_t)$
 Poser la question q_{t+1} à l'apprenant s
 Récupérer la valeur de succès ou échec r_{t+1}
 $\pi_{t+1} \leftarrow \text{UPDATEPARAMS}(\{(q_k, r_k)\}_{k=1, \dots, t+1}, \kappa)$
 $p \leftarrow \text{PREDICTPERFORMANCE}(\kappa, \pi_t, Q_{val})$
 $\sigma_{t+1} \leftarrow \text{EVALUATEPERFORMANCE}(p, D[s][Q_{val}])$

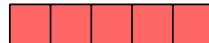
Protocole expérimental

procédure SIMULATEADAPTIVETEST(modèle M , I_{train} , I_{test})
 $\alpha, \kappa \leftarrow \text{TRAININGSTEP}(M, D[I_{train}])$
 pour tout apprenant s de l'ensemble I_{test} **faire**
 $\pi_0 \leftarrow \text{PRIORINITIALIZATION}(\alpha)$
 pour t de 0 à $|Q \setminus Q_{val}| - 1$ **faire**
 $q_{t+1} \leftarrow \text{NEXTITEM}(\{(q_k, r_k)\}_{k=1, \dots, t}, \kappa, \pi_t)$
 Poser la question q_{t+1} à l'apprenant s
 Récupérer la valeur de succès ou échec r_{t+1}
 $\pi_{t+1} \leftarrow \text{UPDATEPARAMS}(\{(q_k, r_k)\}_{k=1, \dots, t+1}, \kappa)$
 $p \leftarrow \text{PREDICTPERFORMANCE}(\kappa, \pi_t, Q_{val})$
 $\sigma_{t+1} \leftarrow \text{EVALUATEPERFORMANCE}(p, D[s][Q_{val}])$

Protocole expérimental

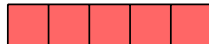
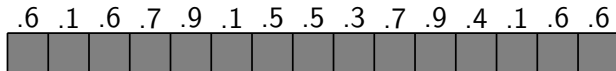
procédure SIMULATEADAPTIVETEST(modèle M , I_{train} , I_{test})
 $\alpha, \kappa \leftarrow \text{TRAININGSTEP}(M, D[I_{train}])$
 pour tout apprenant s de l'ensemble I_{test} **faire**
 $\pi_0 \leftarrow \text{PRIORINITIALIZATION}(\alpha)$
 pour t de 0 à $|Q \setminus Q_{val}| - 1$ **faire**
 $q_{t+1} \leftarrow \text{NEXTITEM}(\{(q_k, r_k)\}_{k=1, \dots, t}, \kappa, \pi_t)$
 Poser la question q_{t+1} à l'apprenant s
 Récupérer la valeur de succès ou échec r_{t+1}
 $\pi_{t+1} \leftarrow \text{UPDATEPARAMS}(\{(q_k, r_k)\}_{k=1, \dots, t+1}, \kappa)$
 $p \leftarrow \text{PREDICTPERFORMANCE}(\kappa, \pi_t, Q_{val})$
 $\sigma_{t+1} \leftarrow \text{EVALUATEPERFORMANCE}(p, D[s][Q_{val}])$

Évaluation des performances



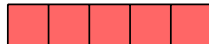
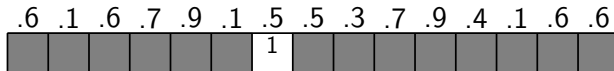
Début du test

Évaluation des performances



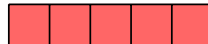
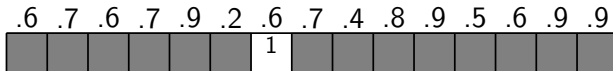
On initialise les paramètres de l'apprenant

Évaluation des performances



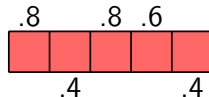
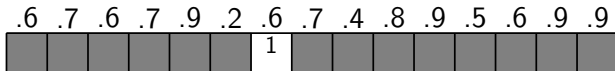
On choisit la question suivante, l'apprenant répond correctement

Évaluation des performances



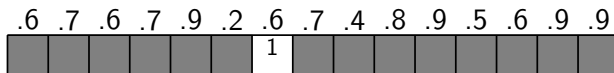
On met à jour les paramètres de l'apprenant

Évaluation des performances

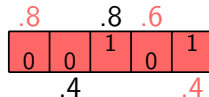


On calcule les prédictions du modèle sur les questions de validation

Évaluation des performances

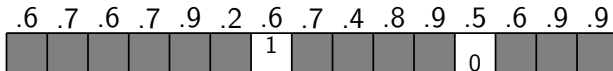


3 prédictions incorrectes sur 5 →

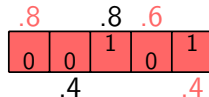


On évalue les prédictions du modèle en fonction des vraies données

Évaluation des performances

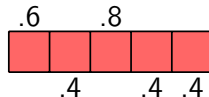
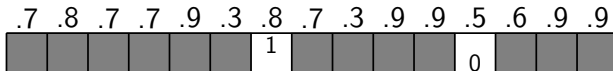


3 prédictions incorrectes sur 5 →



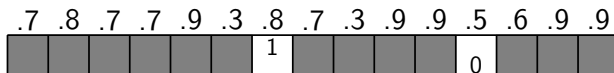
On choisit la question suivante, l'apprenant répond faux

Évaluation des performances

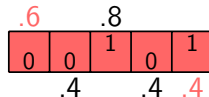


On met à jour les paramètres de l'apprenant

Évaluation des performances



2 prédictions incorrectes sur 5 →



On calcule les prédictions du modèle sur les questions de validation

Validation

Protocole implémenté en Python, testé sur :

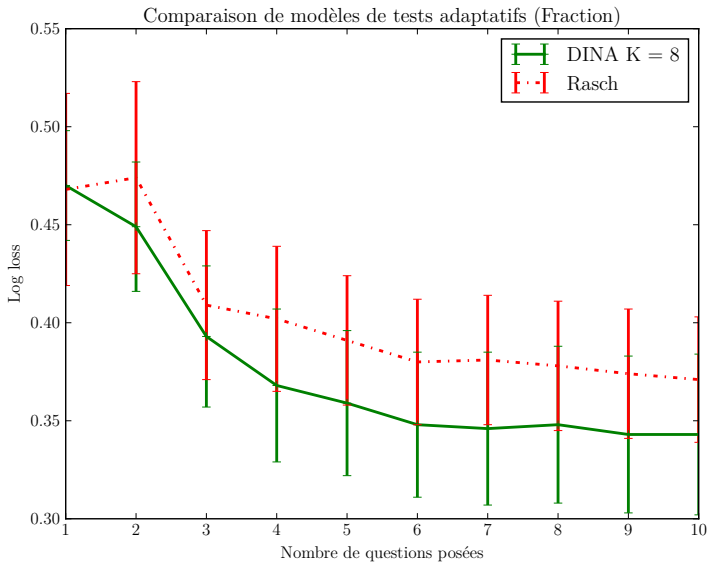
- Modèle de **Rasch** (niveau, difficulté)
- Modèle **DINA**
(composantes maîtrisées / non maîtrisées, slip, guess)

Jeux de données :

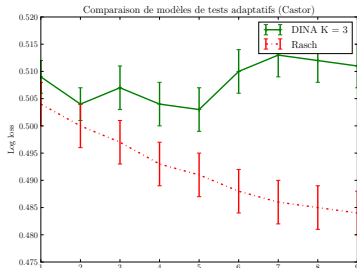
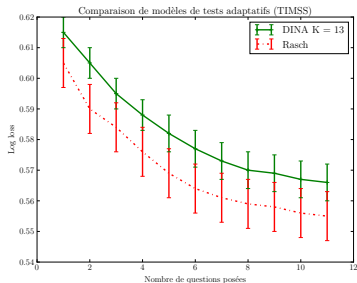
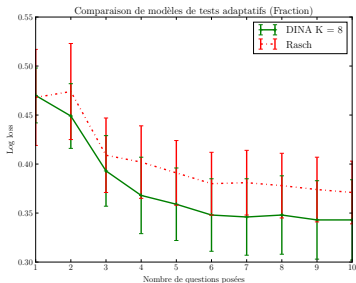
- SAT, ECPE, Fraction, TIMSS, Castor

Calcul de l'erreur **moyenne** de prédiction sur l'ensemble d'apprenants de test.

Résultats



Résultats



Aide à la comparaison de modèles

Pour comparer un nouveau modèle aux autres, il suffit d'implémenter :

- TRAININGSTEP
- PRIORINITIALIZATION
- NEXTITEM
- UPDATEPARAMS
- PREDICTPERFORMANCE

Plan

1. Contexte des évaluations à grande échelle
2. État de l'art précédent et évaluation qualitative
3. Système de comparaison de modèles
4. GenMA : un modèle de test adaptatif pour le diagnostic
5. InitialD : une stratégie de choix des premières questions
6. Conclusions et perspectives

Intuition pour la conception d'un modèle hybride

- Faire intervenir des composantes de connaissances avec des poids différents dans une question

	C_1	C_2	C_3	C_4	C_5		C_1	C_2	C_3	C_4	C_5
Q_1	1	1	0	1	1	$\Rightarrow$	0.47	0.18		0.15	0.63
Q_2	1	0	1	1	0		0.05		0.27	0.07	
Q_3	1	0	1	0	1		0.03		0.05		0.36
Q_4	0	0	0	1	1					0.41	0.10
Q_5	1	1	0	0	0		0.25	0.54			
Q_6	1	0	1	1	1		0.63		0.71	1.00	0.00
Q-matrix							GenMA				

GenMA : General Multidimensional Adaptive

Caractéristiques des questions & apprenants

- Niveau des apprenants selon chaque CC $\vec{\theta} = (\theta_1, \dots, \theta_K)$
- Discrimination selon chaque CC $\vec{d} = (d_1, \dots, d_K) + \text{facilité } \delta$
- Pour toute CC k , $q_k = 0 \Rightarrow d_k = 0$

Théorie de la réponse à l'item multidim.

→ Difficile à calibrer sur mes jeux de données (Desmarais, 2012)

$$\Phi(\vec{\theta} \cdot \vec{d} + \delta) = \Phi\left(\sum_{k=1}^K \theta_k d_k + \delta\right)$$

Modèle GenMA

Dépend de la corrélation, mais seulement pour les valeurs **non nulles** de la q-matrice

→ Plus facile à calibrer

$$\Phi\left(\sum_{k=1}^K \theta_k q_k d_k + \delta\right)$$

$(q_k)_k$: entrées de la q-matrice

GenMA: General Multidimensional Adaptive

Phase d'apprentissage

Minimisation de la fonction d'erreur sur l'historique de réponses avec un algorithme de Metropolis-Hastings Robbins-Monro (Cai, 2010)
Implémenté dans le package R `mirt`

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

Choix de la question suivante

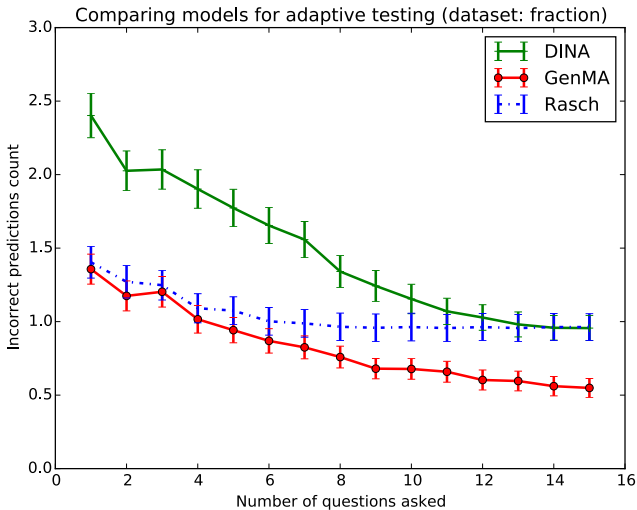
La question dont la probabilité de réponse est la plus proche de 0,5

Mise à jour des paramètres

Régression logistique

- Publié à EC-TEL 2016 (Vie, Bourda, Popineau et Bruillard)
- Disponible sur GitHub : <http://github.com/jilljenn/qna>

Résultats



4 questions sur 15 suffisent → précision moyenne de 80 %.

Plan

1. Contexte des évaluations à grande échelle
2. État de l'art précédent et évaluation qualitative
3. Système de comparaison de modèles
4. GenMA : un modèle de test adaptatif pour le diagnostic
5. InitialD : une stratégie de choix des premières questions
6. Conclusions et perspectives

Démarrage à froid du modèle

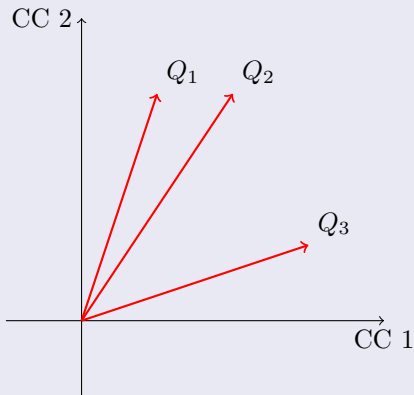
Avec le modèle GenMA :

- Quelles premières questions poser, lorsqu'on ne sait rien de l'apprenant ?
- Faut-il les poser une par une ou poser un groupe de questions avant de faire le premier diagnostic ?

Intuition géométrique sur le choix des premières questions

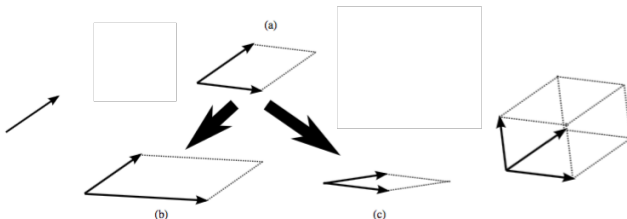
Comment choisir un bon premier ensemble de questions ?

Exemple : questions selon deux composantes de connaissances



Caractérisation de la diversité d'un ensemble de questions

Ensembles de questions qui mesurent des dimensions différentes



- Diversité d'un ensemble de questions
= **volume du paralléloptope formé par les vecteurs**
- Or, déterminant de la matrice de Gram VV^T
= (volume formé par les vecteurs)²

Processus à point déterminantal

- $K = VV^T$: $n \times n$ matrice de similarité entre questions
- Pour tout $A \subset \{1, \dots, n\}$, K_A désigne la sous-matrice carrée indexée par A en ligne et colonne

Un **processus à point déterminantal** (PPD) est une distribution de probabilité vérifiant que si l'ensemble Y est tiré, alors :

$$\forall A \subset Y, \quad Pr(A \subseteq Y) \text{ est proportionnel à } \det(K_A)$$

Avantages

- Algorithmes efficaces pour échantillonner selon un PPD
- Tirer k questions parmi n en $O(nk^3)$ (+ précalcul $O(n^3)$)
- Ex. tirer 10 exercices parmi une banque de 10000 questions
- Probabiliste : ne pose pas les mêmes questions à tout le monde

Différentes stratégies comparées

Random

Choisir un groupe de questions au hasard

Uncertainty

Choisir le groupe de questions dont la probabilité que l'apprenant y réponde correctement est le plus proche de 0,5

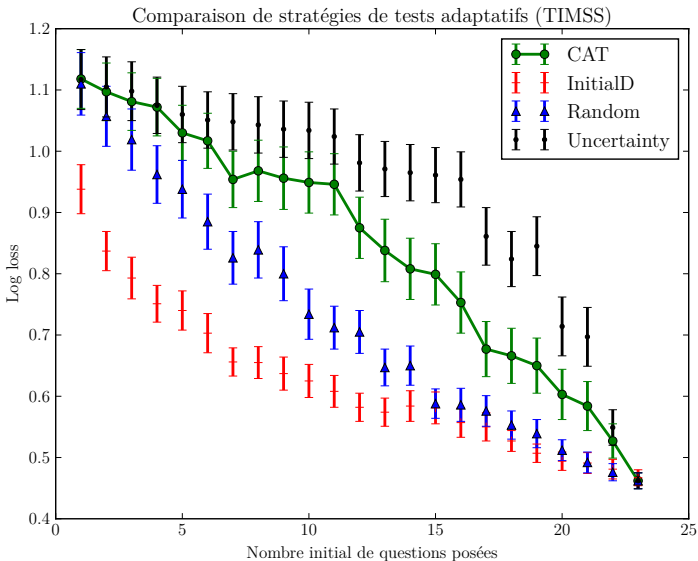
InitialID (Initial Determinant)

Échantillonner un groupe de questions selon un PPD
(= proportionnellement au volume du parallélotope formé)

CAT (Computer Adaptive Test)

Test adaptatif classique GenMA

Résultats



Plan

1. Contexte des évaluations à grande échelle
2. État de l'art précédent et évaluation qualitative
3. Système de comparaison de modèles
4. GenMA : un modèle de test adaptatif pour le diagnostic
5. InitialID : une stratégie de choix des premières questions
6. Conclusions et perspectives

Conclusion et perspectives

Système de comparaison générique

⇒ Comparer d'autres modèles

GenMA : un nouveau modèle formatif plus précis que l'existant

⇒ À tester dans les MOOC

InitialD : une nouvelle stratégie pour poser les premières questions

⇒ Que se passe-t-il si on alterne non adaptatif puis adaptatif ?

Perspectives

Couplages à des travaux existants

- Systèmes de recommandation de ressources pédagogiques
- Génération automatique d'exercices
- Interfaces plus riches pour l'évaluation (jeux sérieux)

Perspectives

Couplages à des travaux existants

- Systèmes de recommandation de ressources pédagogiques
- Génération automatique d'exercices
- Interfaces plus riches pour l'évaluation (jeux sérieux)

Prendre en compte la progression de l'apprenant

- Bandits à plusieurs bras (Clement et al., JEDM 2015)
- Compromis entre mesurer les connaissances de l'apprenant (exploration) et le faire progresser (exploitation)

Perspectives

Couplages à des travaux existants

- Systèmes de recommandation de ressources pédagogiques
- Génération automatique d'exercices
- Interfaces plus riches pour l'évaluation (jeux sérieux)

Prendre en compte la progression de l'apprenant

- Bandits à plusieurs bras (Clement et al., JEDM 2015)
- Compromis entre mesurer les connaissances de l'apprenant (exploration) et le faire progresser (exploitation)

Certification des compétences numériques

- Projet PIX porté par le ministère de l'Éducation nationale
pix.beta.gouv.fr
- Présenté au congrès de la Société informatique de France (2–3 février 2017)

Merci !

Jill-Jênn Vie, Fabrice Popineau, Éric Bruillard et Yolaine Bourda. "A review of recent advances in adaptive assessment". In *Learning analytics: Fundamentals, applications, and trends: A view of the current state of the art*. Springer, 2016, sous presse.

Jill-Jênn Vie, Fabrice Popineau, Yolaine Bourda, and Éric Bruillard. "Adaptive Testing Using a General Diagnostic Model." In *European Conference on Technology Enhanced Learning (EC-TEL 2016)*, 331–339. Springer.

Jill-Jênn Vie, Fabrice Popineau, Jean-Bastien Grill, Éric Bruillard, and Yolaine Bourda. "Predicting Performance over Dichotomous Questions: Comparing Models for Large-Scale Adaptive Testing." In *8th International Conference on Educational Data Mining (EDM 2015)*.

Jill-Jênn Vie, Fabrice Popineau, Jean-Bastien Grill, Éric Bruillard, and Yolaine Bourda. "Prédiction de performance sur des questions dichotomiques : comparaison de modèles pour des tests adaptatifs à grande échelle." In *Atelier Évaluation des Apprentissages et Environnements Informatiques, EIAH 2015*.