Knowledge Tracing Machines

Jill-Jênn Vie
RIKEN Center for Advanced Intelligence Project (AIP)
Tokyo, Japan
vie@jill-jenn.net

Hisashi Kashima
Kyoto University/RIKEN
Kyoto, Japan
kashima@i.kyoto-u.ac.jp

Problem: Knowledge Tracing

We want to predict the performance of students over questions.
Each student can attempt a question multiple times, and learns in-between.
Fit: Ordered triplets $(i, j, o) \in I \times J \times \{0, 1\}$
⇒ Student i attempted question j and got it correct/incorrect.
Predict: $(i, j, ?)$ for new triplets.

Existing Models

- **Prediction of sequences**: Bayesian Knowledge Tracing (BKT := HMM)
  Deep Knowledge Tracing (DKT := LSTM) [2]
- **Factor Analysis**: Item Response Theory (IRT), Performance Factor Analysis (PFA)

$$\text{BKT} < \text{PFA} \simeq^{[5]} \text{DKT} \leq^{[4]} \text{IRT} \leq^{[\text{this poster}]} \text{KTM}$$

Item Response Theory

Students $i \in I$ have unknown level θ_i
Questions $j \in J$ have unknown difficulty d_j

$$\operatorname{logit} p_{ij} = \operatorname{logit} \Pr(\text{Student } i \text{ answers correctly question } j) = \theta_i - d_j$$

⇒ **really simple, ignores skills & multiple attempts**

Multidimensional counterpart (MIRT): $\operatorname{logit} p_{ij} = \langle \boldsymbol{\theta_i}, \boldsymbol{d_j} \rangle + \delta_j$

Performance factor analysis

Students $i \in I$ have unknown level θ_i, W_{ik} prior wins and F_{ik} fails over skill k
Questions $j \in J$ have requirements $\text{KC}(j) \subseteq K$
Skills $k \in K$ have bias β_k and opportunities to be learned after win γ_k and fail δ_k

$$\operatorname{logit} p_{ij} = \theta_i + \sum_{k \in \text{KC}(j)} \beta_k + \gamma_k W_{ik} + \delta_k F_{ik}$$

⇒ **ignores item difficulty**

Additive Factor Model (AFM): only consider attempts of student i over skill k ($\gamma_k = \delta_k$)

Our proposal

Knowledge Tracing Machines

All students $i \in I$ and questions $j \in J$ and past performance are encoded into x (entities)
All entities have a bias w_k and features $\boldsymbol{v_k}$ to model the pairwise relationships between them

$$\psi(p(x)) = \mu + \sum_{k=1}^{N} w_k x_k + \sum_{1 \leq k < l \leq N} x_k x_l \langle \boldsymbol{v_k}, \boldsymbol{v_l} \rangle$$

KTM := BFM (Bayesian Factorization Machines [1]) for classification

- **KTM** take IRT, MIRT, PFA as special cases
- If ψ = probit, $w_k, v_{kf} \sim \mathcal{N}(\mu, 1/\lambda)$, $\mu \sim \mathcal{N}(0, 1)$, $\lambda \sim \Gamma(1, 1)$,
  a **Gibbs sampler** allows to train efficiently KTM [3]

Encoding of entities

Unsupervised problem becomes a supervised problem:

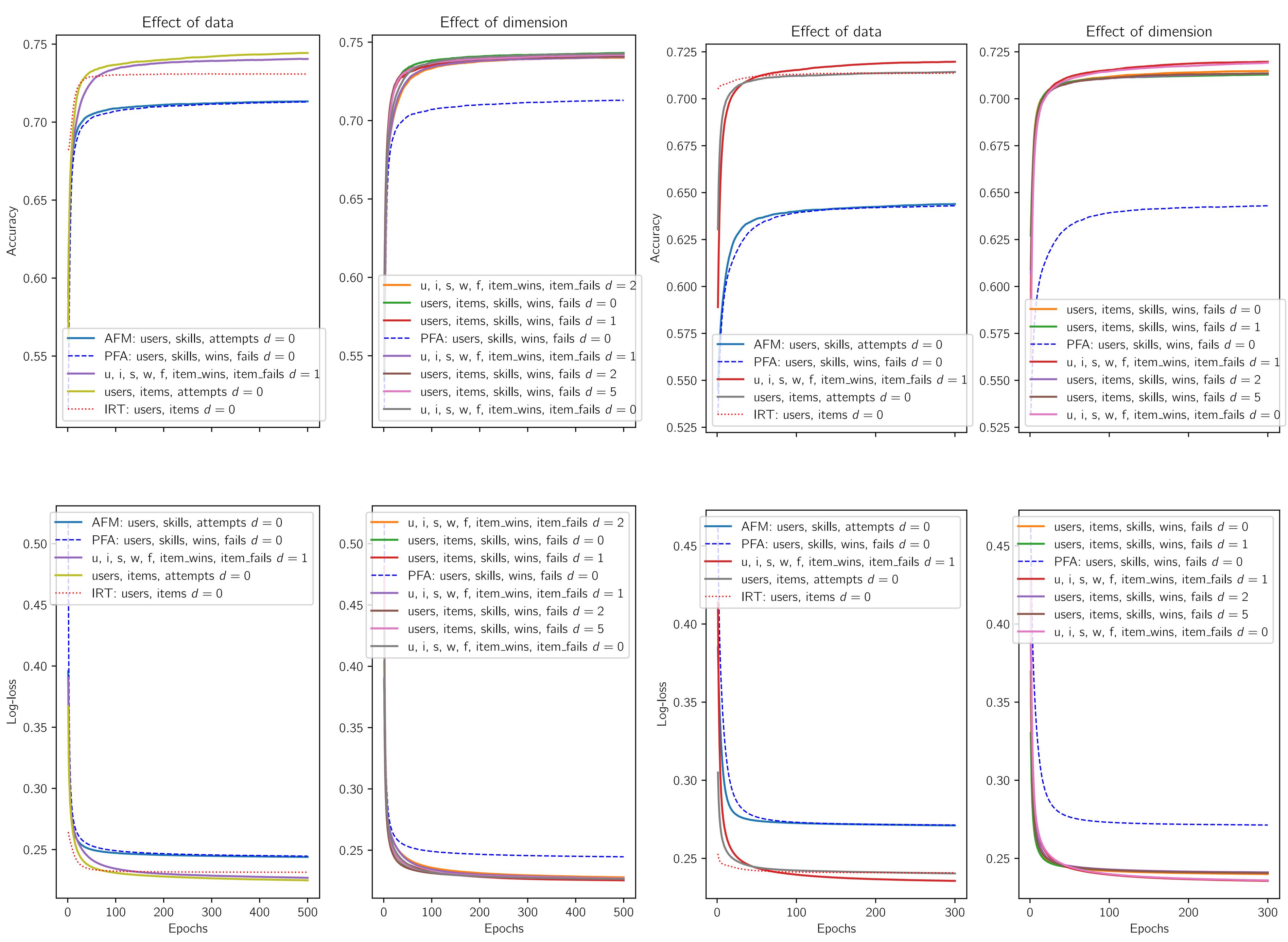
Triplet	Users		Items			Skills			Wins			Fails			Outcome
	1	2	Q_1	Q_2	Q_3	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3	
(2, 2, 1)	0	1	0	1	0	1	1	0	0	0	0	0	0	0	1
(2, 2, 0)	0	1	0	1	0	1	1	0	1	1	0	0	0	0	0
(2, 2, 1)	0	1	0	1	0	1	1	0	1	1	0	1	1	0	1
(2, 3, 0)	0	1	0	0	1	0	1	1	0	2	0	0	1	0	0
(2, 3, 1)	0	1	0	0	1	0	1	1	0	2	0	0	2	1	1
(1, 2, 1)	1	0	0	1	0	1	1	0	0	0	0	0	0	0	1
(1, 1, 0)	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Encoding users + items == IRT
Encoding users + skills + attempts == AFM
Encoding users + skills + wins + fails == PFA

What is the influence of the data over the model?

Various, large-scale educational datasets

Name	Users	Items	Skills	Skills per item	Entries	Sparsity (user-item)	Attempts per user
fraction	536	20	8	2.800	10720	1.000	1.000
timss2003	757	23	13	1.652	17411	1.000	1.000
assistments	4217	26688	123	0.796	346860	0.003	1.014
berkeley	1730	234	29	1.000	562201	0.731	1.901
castor6e	58939	17	2	1.471	1001963	1.000	1.000



AUC	PFA	AFM	IRT	MIRTb10	KTM(uia0)	KTM(uis0)	KTM(uis10)	KTM(uiswfWF1)
assistments	0.7361	0.7366	0.7748	–	**0.7932**	–	–	0.7882
berkeley	0.7	0.7005	0.7856	–	0.7873	–	–	**0.7971**
castor6e	–	–	**0.8239**	0.8235	–	**0.8240**	**0.8239**	–
fraction	–	–	0.9124	0.9126	–	0.9125	**0.9146**	–
timss2003	–	–	0.7979	0.7979	–	**0.7999**	0.7981	–

Better models found

- It is better to learn a item bias
- Bigger dimensions do not help much, sometimes harm
- IRT is simple yet competitive

model	dim	ACC	AUC	NLL
users, items, skills, wins, fails, item_wins, item_fails	1	**0.720**	**0.797**	**0.236**
users, items, skills, wins, fails, item_wins, item_fails	0	**0.719**	**0.797**	**0.236**
users, items, skills, wins, fails	0	0.715	0.788	0.24
users, items, attempts	0	0.714	0.787	0.24
users, items, skills, wins, fails	5	0.714	0.787	0.241
IRT: users, items	0	0.714	0.786	0.241
users, items, skills, wins, fails	2	0.713	0.786	0.241
users, items, skills, wins, fails	1	0.713	0.786	0.241
MIRTb: users, items	1	0.71	0.781	0.243
AFM: users, skills, attempts	0	0.644	0.701	0.271
PFA: users, skills, wins, fails	0	0.643	0.7	0.271

Optimizing Human Learning

We are organizing a workshop in Montréal on **June 12**:
CFP open on humanlearn.io

References

[1] Christoph Freudenthaler, Lars Schmidt-Thieme, and Steffen Rendle. "Bayesian factorization machines". Presented at the Workshop on Sparse Representation and Low-rank Approximation, Neural Information Processing Systems (NIPS-WS). 2011.

[2] Chris Piech et al. "Deep knowledge tracing". In: *Advances in Neural Information Processing Systems (NIPS)*. 2015, pp. 505–513.

[3] Steffen Rendle. "Factorization Machines with libFM". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 3.3 (2012), 57:1–57:22. DOI: 10.1145/2168752.2168771.

[4] Kevin H. Wilson et al. "Back to the basics: Bayesian extensions of IRT outperform neural networks for proficiency estimation". In: *Proceedings of the 9th International Conference on Educational Data Mining (EDM)*. 2016, pp. 539–544.

[5] Xiaolu Xiong et al. "Going Deeper with Deep Knowledge Tracing". In: *Proceedings of the 9th International Conference on Educational Data Mining (EDM)*. 2016, pp. 545–550.