Deep Factorization Machines for Knowledge Tracing

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

Problem: Knowledge Tracing for Language Learning

We want to predict the correctness of students over words.
Each student can attempt to write a certain word multiple times, and learns in-between.
Fit: Ordered triplets $(i, j, o) \in I \times J \times \{0, 1\}$
$\Rightarrow$ Student i attempted word j and wrote it correctly/incorrectly.
Predict: $(i, j, ?)$ for new triplets.

Existing families of models

- **Prediction of sequences**: Bayesian Knowledge Tracing (BKT := HMM)
  Deep Knowledge Tracing (DKT := LSTM) [3]
- **Factor Analysis**: Item Response Theory (IRT), Performance Factor Analysis (PFA)

$$\text{BKT} < \text{PFA} \simeq^{[6]} \text{DKT} \leq^{[5]} \text{IRT} \leq^{[\text{this poster}]} \text{FM}$$

Logistic Regression (LR)

All students $i \in I$, questions $j \in J$ and metadata are encoded into sparse features x
Each feature k has a bias w_k

$$\operatorname{logit} p(x) = \operatorname{logit} \Pr(\text{event } x \text{ has positive outcome}) = \mu + w^T x$$

$\Rightarrow$ **really simple, ignores pairwise interactions** $(d = 0)$

Particular cases for user i against token j:
Item response theory (IRT): $\operatorname{logit} p_{ij} = \theta_i - d_j$
Performance Factor Analysis (PFA): $\operatorname{logit} p_{ij} = \sum_{k \in KC(j)} \beta_k + \gamma_k W_{ik} + \delta_k F_{ik}$

Factorization Machines (FM)

All students $i \in I$ and questions $j \in J$ and past performance are encoded into x
All entities have a bias w_k and features $\boldsymbol{v_k} \in \mathbf{R}^d$ to model pairwise interactions

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

$\Rightarrow$ **converting sparse features to dense embeddings**

Particular case:
Multidimensional Item Response Theory (MIRT): $\operatorname{logit} p_{ij} = \langle \boldsymbol{\theta_i}, \boldsymbol{d_j} \rangle + \delta_j$

Our proposal

Deep Factorization Machines (DeepFM)

All students $i \in I$ and questions $j \in J$ and past performance are encoded into x
FM: All entities have a bias w_k and features $\boldsymbol{v_k} \in \mathbf{R}^d$ to model pairwise interactions
Deep: Train layers $W^{(\ell)}$ and $b^{(\ell)}$ for each $\ell = 1, \ldots, L$ [2]

$$\operatorname{logit} p(x) = y_{FM}(x) + y_{DNN}(x)$$

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

$$\begin{aligned} y_{DNN}(x) &= \operatorname{ReLU}(W^{(L)} a^{(L)}(x) + b^{(L)}) \\ a^{(\ell+1)}(x) &= \operatorname{ReLU}(W^{(\ell)} a^{(\ell)}(x) + b^{(\ell)}) \\ a^0(x) &= (\boldsymbol{v}_{\texttt{user}}, \boldsymbol{v}_{\texttt{token}}, \ldots, \boldsymbol{v}_{\texttt{countries}}) \end{aligned}$$

Bayesian Factorization Machines (Bayesian FM)

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

Hyperpriors: $w_k, v_{kf} \sim \mathcal{N}(\mu_f, 1/\lambda_f)$, $\mu_f \sim \mathcal{N}(0, 1)$, $\lambda_f \sim \Gamma(1, 1)$,
Trained using **Gibbs sampling** [1, 4]

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

IRT: `user` + `token`
first: `<discrete features>`
last: `<discrete features>` + `time` + `days`
pfa: `<discrete features>` + `wins` + `fails`

Results in AUC on large-scale Duolingo dataset

data	model	d	epoch	train	first	last	pfa
fr	Bayesian FM	20	500/500	–	0.822	–	–
fr	Bayesian FM	20	500/500	–	–	0.817	–
fr	DeepFM	20	15/1000	0.872	0.814	–	–
fr	Bayesian FM	20	100/100	–	–	0.813	–
fr	DeepFM	20	21/1000	0.878	0.812	–	–
fr	FM	20	20/1000	0.874	0.811	–	–
fr	FM	20	20/1000	0.875	0.811	–	–
fr	Bayesian FM	20	500/500	–	–	–	0.806
fr	FM	20	21/1000	0.884	–	–	0.805
fr	FM	20	37/1000	0.885	–	0.8	–
fr	DeepFM	20	77/1000	0.89	–	0.792	–
fr	Deep	20	7/1000	0.826	0.791	–	–
fr	Deep	20	321/1000	0.826	–	0.79	–
fr	Deep	20	5/5	0.827	–	0.789	–
fr	LR	0	50/50	–	–	–	0.789
fr	Deep	20	127/1000	0.826	0.789	–	–
fr	LR	0	50/50	–	0.783	–	–
fr	LR	0	50/50	–	–	0.783	–

data	model	d	epoch	train	first	last	pfa
es	Bayesian FM	20	500/500	–	0.803	–	–
es	Bayesian FM	20	500/500	–	–	–	0.796
es	DeepFM	20	11/1000	0.845	0.792	–	–
es	DeepFM	20	15/1000	0.851	0.79	–	–
es	FM	20	17/1000	0.85	0.788	–	–
es	FM	20	15/1000	0.853	–	–	0.787
es	FM	20	33/1000	0.857	0.782	–	–
es	LR	0	50/50	–	–	–	0.765
es	Deep	20	94/1000	0.794	0.762	–	–
es	LR	0	50/50	–	0.759	–	–
es	Deep	20	117/1000	0.792	0.759	–	–
es	Deep	20	17/1000	0.787	–	0.756	–
es	FM	20	151/1000	0.834	–	0.748	–
es	Bayesian FM	20	500/500	–	–	0.743	–
es	DeepFM	20	323/1000	0.832	–	0.742	–
es	LR	0	50/50	–	–	0.718	–

data	model	d	epoch	train	first	last	pfa
en	Bayesian FM	20	500/500	–	0.828	–	–
en	FM	20	17/1000	0.857	0.818	–	–
en	DeepFM	20	20/1000	0.858	0.817	–	–
en	Bayesian FM	20	500/500	–	–	–	0.817
en	FM	20	20/1000	0.858	0.816	–	–
en	FM	20	15/1000	0.858	–	–	0.81
en	LR	0	50/50	–	–	–	0.792
en	Deep	20	164/1000	0.81	0.792	–	–
en	FM	20	45/1000	0.836	–	0.788	–
en	LR	0	50/50	–	0.787	–	–
en	Deep	20	32/1000	0.8	–	0.786	–
en	DeepFM	20	97/1000	0.834	–	0.784	–
en	Bayesian FM	20	500/500	–	–	0.761	–
en	LR	0	50/50	–	–	0.736	–

Take home message

- Pairwise interactions are useful
- Deep does not help much
- Time and days harm

Optimizing Human Learning

We are organizing a workshop in Montréal on **June 12**:
Proceedings 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] Huifeng Guo et al. "DeepFM: a factorization-machine based neural network for CTR prediction". In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press. 2017, pp. 1725–1731.
[3] Chris Piech et al. "Deep knowledge tracing". In: *Advances in Neural Information Processing Systems (NIPS)*. 2015, pp. 505–513.
[4] 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.
[5] 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.
[6] 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.