Understanding Dropouts in MOOCs

AAAI

American Association for Artificial Intelligence

Author : Wenzheng Feng, Jie Tang and Tracy Xiao Liu

Reporter : gbl555

MOOCs

Massive open online courses (MOOCs)

By the end of 2017,9400 courses, 81,000,000 registered students

MOOCs

MOOCs are really beneficial to the learners who complete courses,

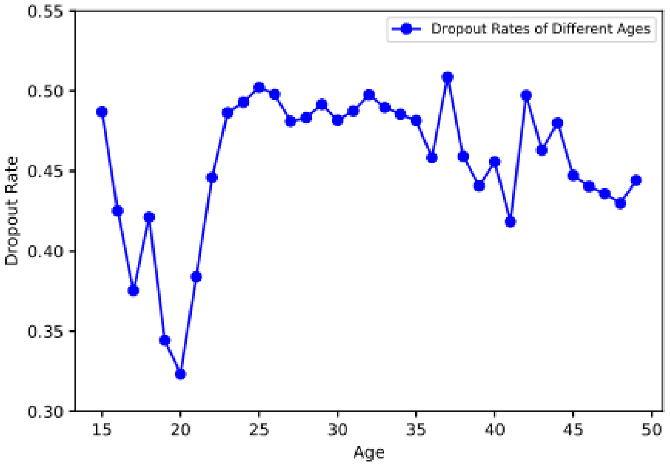
61% of survey respondents report MOOCs’ education benefits

72% of those report career benefits

But…

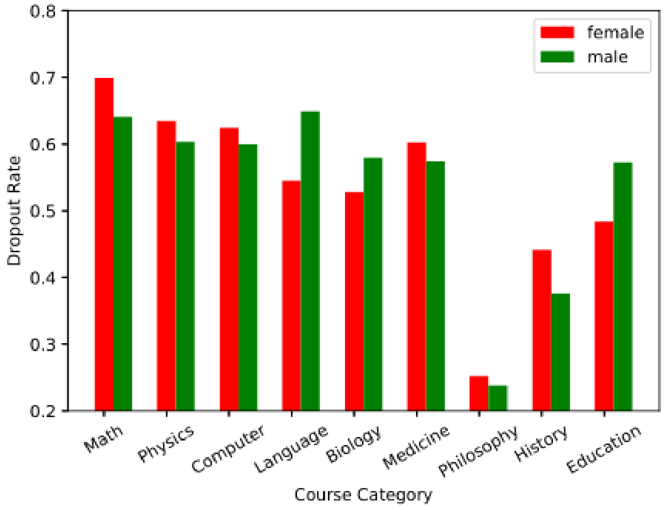
The biggest threat to MOOCs

Observational analyses



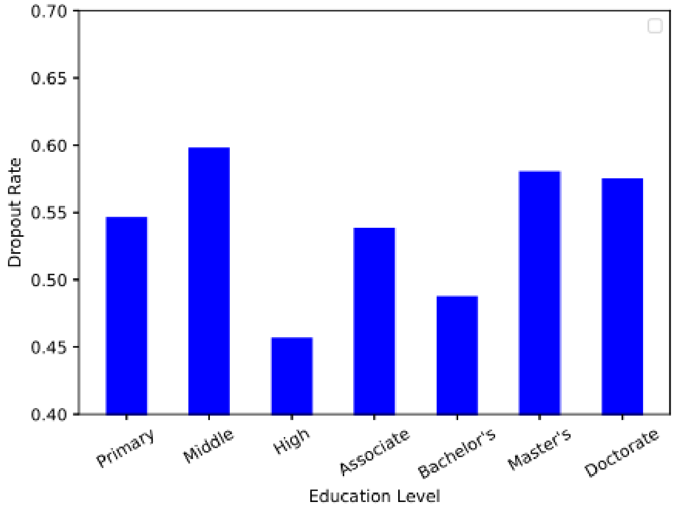
(a) Age

young people are more inclined to drop out



(b) Course Category

female users are more likely to drop science courses and male users are more likely to give up non-science courses



(c) Education Level

Educational background is also important

Interesting questions

Q1: What are the major motivations that drive the users to study in MOOCs?

Q2: What are the major dropout reasons?

Q3: Is that possible to predict users' dropout behavior in advance, so that the MOOCs platform could deliver some kind of useful interventions?

XuetangX

Launched in 2013

2 learning mode,
IPM and SPM

1.IPM(Instructor-paced mode): same course schedule,16weeks
2.SPM(Self-paced mode):flexible Schedule,maybe a longer period

1000 courses,
12 categories

学堂在线
xuetangx.com

10,000,000
registered users

System records multiple types of activities

1.Video watching(watch,stop,jump)
2.Forum discussion(ask questions and replies)
3.Assignment completion(with
4.correct/incorrect answers,and reset)
Web page clicking(click and close)

Datasets

Table 1: Statistics of the KDDCUP dataset.

Category	Type	Number
log	# video activities	1,319,032
	# forum activities	10,763,225
	# assignment activities	2,089,933
	# web page activities	738,0344
enrollment	# total	200,904
	# dropouts	159,223
	# completions	41,681
	# users	112,448
	# courses	39

Compare with other methods

Table 2: Statistics of the XuetangX dataset.

Category	Type	#IPM*	#SPM*
log	# video activities	50,678,849	38,225,417
	# forum activities	443,554	90,815
	# assignment activities	7,773,245	3,139,558
	# web page activities	9,231,061	5,496,287
enrollment	# total	467,113	218,274
	# dropouts	372,088	205,988
	# completions	95,025	12,286
	# users	254,518	123,719
	# courses	698	515

* #IPM and #SPM respectively stands for the number for the corresponding IPM courses and SPM courses.

Test robustness and generalization

Users' learning behavior

For Q1 : Clustering analysis on users' learning activities

Definition 1:Temporal Code

For each user,the temporal code is $\mathbf{S}^u = [\mathbf{s}^u_{c_1}, \mathbf{s}^u_{c_2}, \ldots, \mathbf{s}^u_{c_M}]$ (very sparse!)

M is the number of courses

$$\mathbf{s}^u_c = [s^u_{c,1}, s^u_{c,2}, \ldots, s^u_{c,K}]$$

$s^u_{c,k} \in \{0, 1\}$ indicates whether user u visits course c in the k-th week

Users' learning behavior

K-means (k=5)

Table 3: Results of clustering analysis. C1-C5 — Cluster 1 to 5; CAR — average correct answer ratio.

Category	Type	C1	C2	C3	C4	C5
video	#watch	21.83	46.78	12.03	19.57	**112.1**
	#stop	28.45	68.96	20.21	37.19	**84.15**
	#jump	16.30	16.58	11.44	14.54	**21.39**
forum	#question	0.04	**0.38**	0.02	0.03	0.03
	#answer	0.13	**3.46**	0.13	0.12	0.17
assignment	CAR	0.22	**0.76**	0.19	0.20	0.59
	#revise	0.17	0.02	0.04	**0.78**	0.01
session	seconds	1,715	714	**1,802**	1,764	885
	count	3.61	**8.13**	2.18	4.01	7.78
enrollment	#enrollment	21,048	9,063	**401,123**	25,042	10,837
	total #users	2,735	4,131	**239,302**	4,229	4,121
	dropout rate	0.78	0.29	**0.83**	0.66	0.28

Cluster 2, may simply want to meet friends with similar interest.

Cluster 4 ,probably with difficulties to learn the corresponding courses

Cluster 5, use MOOC to seriously study knowledge(hard workers)

Users' learning behavior

For Q2 : Correlation Between Courses & Influence From Dropout Friends

Conclusion

- **High** correlation between dropouts of different courses
- **Strong** influence between friends' dropout behaviors.

Users’ learning behavior

Correlation Between Courses : regression analysis

Qustion : Will someone's dropout for one course increase or decrease the probability that she drops out from another course?

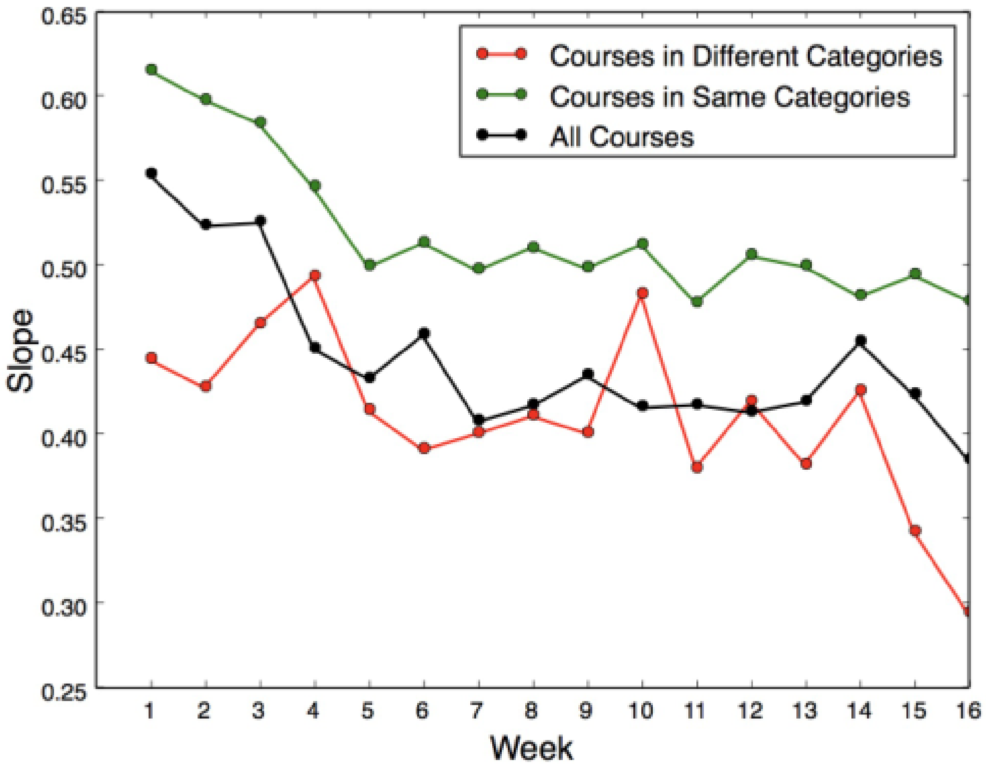
Technology :

$$course_1 = a \times course_2 + \mathrm{b}$$

$where\ course_1\ and\ course_2\ indicate\ a\ user's\ dropout\ behavior\ for\ two\ different\ courses\ in\ the\ same\ semester$

$course_i$ is a 16-dim dummy vector, with each element representing whether the user has visited the course in the corresponding week (thus 16 corresponds to the 16 weeks for studying the course).

Users' learning behavior



A significantly positive correlation between users' dropout probabilities of different enrolled courses.

The correlation between courses of the same category is higher than courses from different categories.

Users' learning behavior

Influence From Dropout Friends

Friend relationship is implicitly defined using co-learning relationships.

Solution

Step 1 : discover users' friend relationships

Step 2 : analyze the influence from dropout friends quantitatively

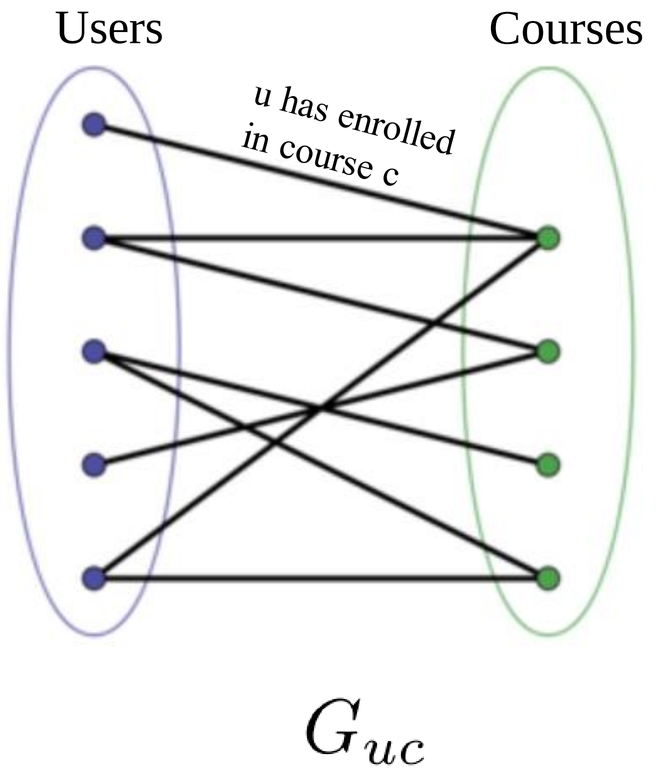
Users’ learning behavior

Step 1 : discover users’ friend relationships

1) DeepWalk , to learn a low dimensional vector for each user node.

2) Compute the cosine similarity between users who have enrolled a same course.

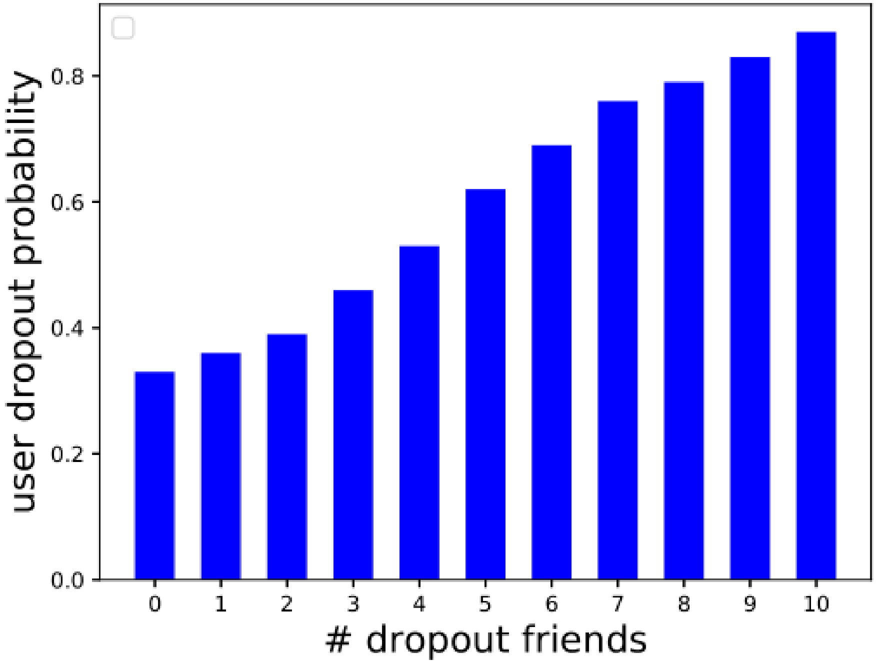
3) Those users with high similarity score, i.e., greater than 0.8, are considered as friends.



Users' learning behavior

Step 2 : analyze the influence from dropout friends quantitatively

P(users' dropout probabilities | the number of dropout friends)

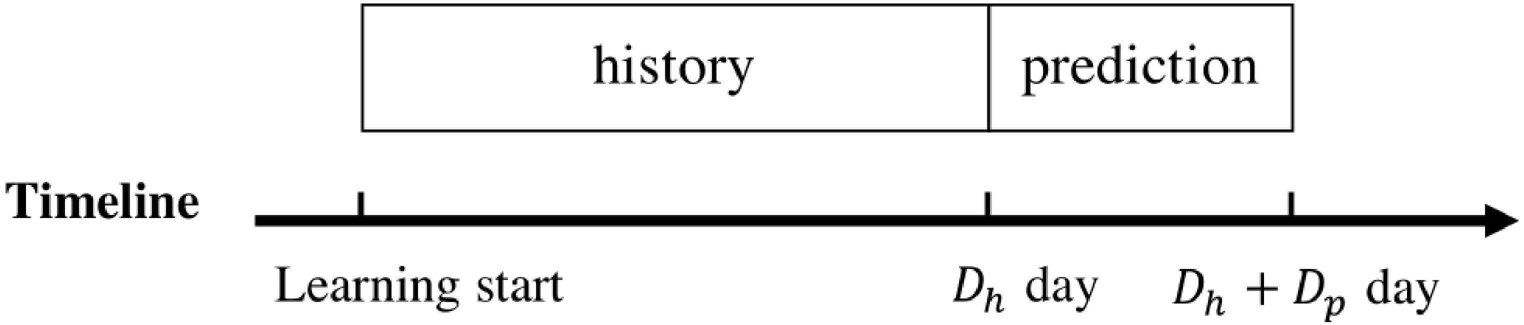


Users' dropout probability increases monotonically from 0.33 to 0.87 when the number of dropout friends ranges from 1 to 10.

This indicates that a user's dropout rate is greatly influenced by her/his friends' dropout behavior.

Methodology

For Q3 : Context-aware Feature Interaction Network (CFIN) to deal with the dropout prediction problem



$$f : (\mathbf{X}(u, c), \mathbf{Z}(u, c)) \rightarrow y_{(u,c)}$$

$$y_{(u,c)} = \begin{cases} 1, \text{u has not taken activities on c in the } \textit{prediction} \text{ period} \\ 0, \textit{otherwise} \end{cases}$$

Methodology

Definition

Definition 2 ***Enrollment Relation*** *:* E $, denote\ the\ set\ of\ all\ enrollments, i.e., \{(u,c)\}$

Definition 3 ***Learning Activity*** *:* X(u,c) = $\left[x_1(u,c), \cdots, x_{m_x}(u,c)\right]$

where $x_i(u,c)$ is a continuous feature value associated to u's learning activity in a course c.

Those features are extracted from user historical logs, mainly includes the statistics of users' activities.

Definition 4 ***Context Information*** *:* Z(u,c) = [user information, course information]

User information is represented by user demographics (i.e. gender,age, location, education level) and user cluster; Course information is the course category.

The categorical information (e.g. gender, location) is represented by a one-hot vector, continues information (i.e. age) is represented as the value itself.

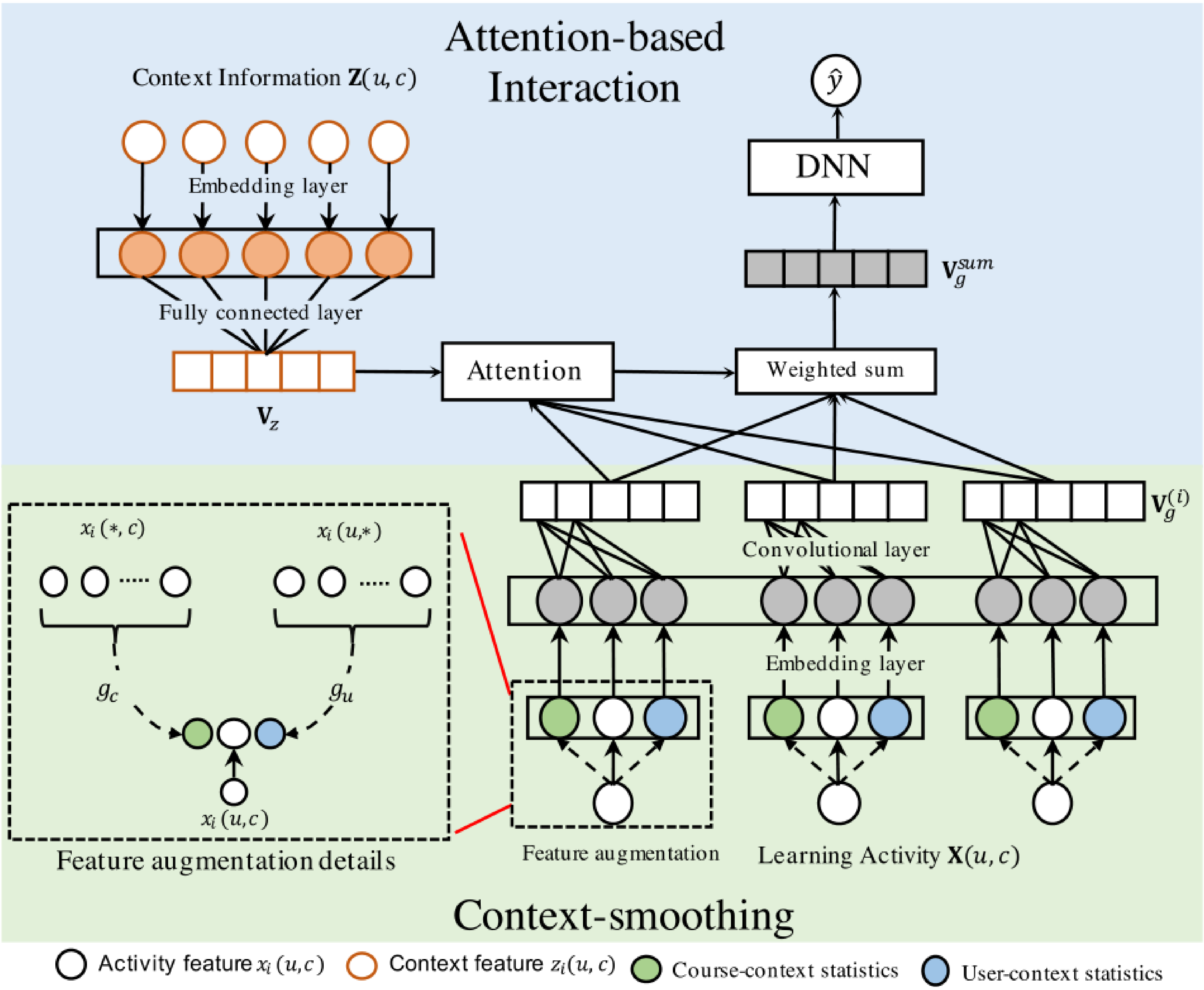
Methodology

The architecture of CFIN

From prior analyses, we find users' activity patterns in MOOCs have a strong correlation with their context.

So, the value of learning activity vector X(u,c) is highly sensitive to the context information Z(u,c).

How to tackle this issue?

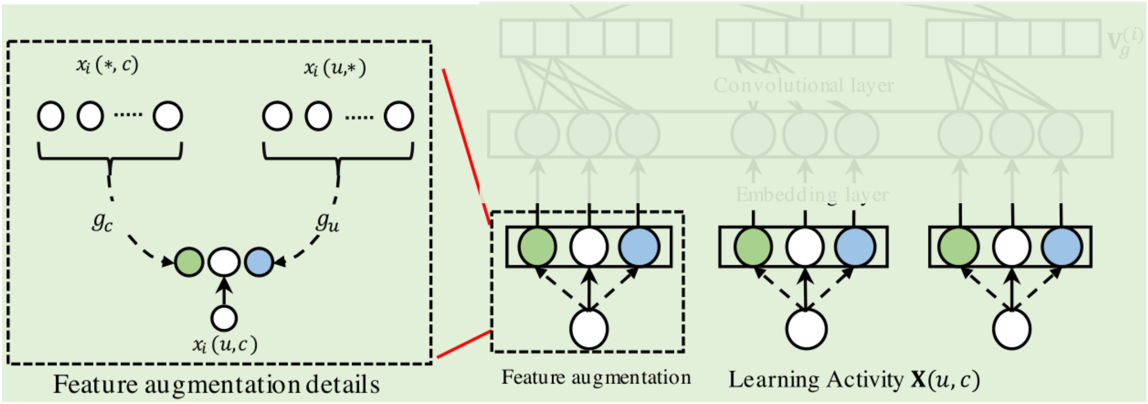


Attention-based Interaction
Context Information $\mathbf{Z}(u,c)$
Embedding layer
Fully connected layer
$\mathbf{V}_z$
Attention
Weighted sum
$\mathbf{V}_g^{sum}$
DNN
$\hat{y}$
$\mathbf{V}_g^{(i)}$
Convolutional layer
Embedding layer
$x_i(*,c)$
$x_i(u,*)$
g_c
g_u
$x_i(u,c)$
Feature augmentation details
Feature augmentation
Learning Activity $\mathbf{X}(u,c)$
Context-smoothing
Activity feature $x_i(u,c)$
Context feature $z_i(u,c)$
Course-context statistics
User-context statistics

Methodology

Context-Smoothing.

STEP 1 : From X(u,c) $to\ \widehat{\mathrm{X}}\ \ by\ \ feature\ \ augmentation$



$\widehat{\mathrm{X}} \in R^{m_g m_x}$

$\mathrm{X(u,c)} \in R^{m_x}$

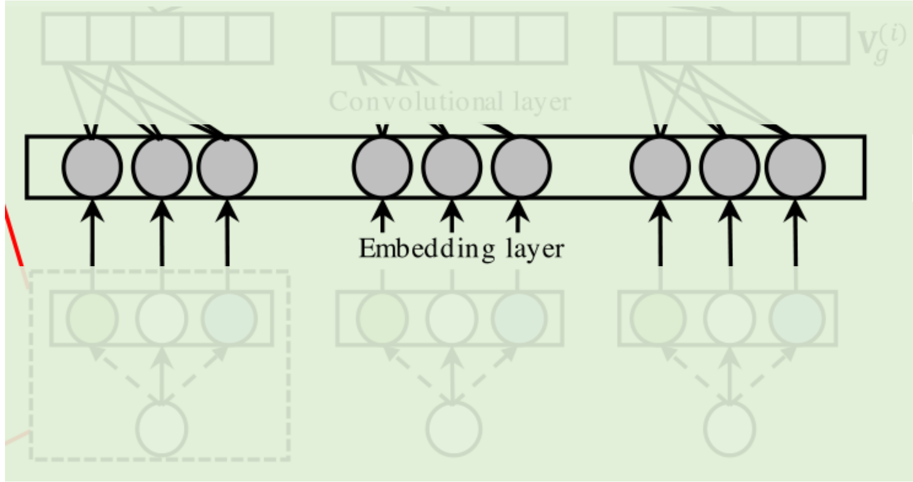
$$g_u \quad : \quad x_i(u,c) \rightarrow [\mathrm{avg}(\{x_i(u,*)\}), \max(\{x_i(u,*)\}), \ldots]$$

$$g_c \quad : \quad x_i(u,c) \rightarrow [\mathrm{avg}(\{x_i(*,c)\}), \max(\{x_i(*,c)\}), \ldots]$$

Methodology

Context-Smoothing.

STEP 2 : each $\widehat{x} \in \widehat{X}$ is converted to a dense vector through an embedding layer.



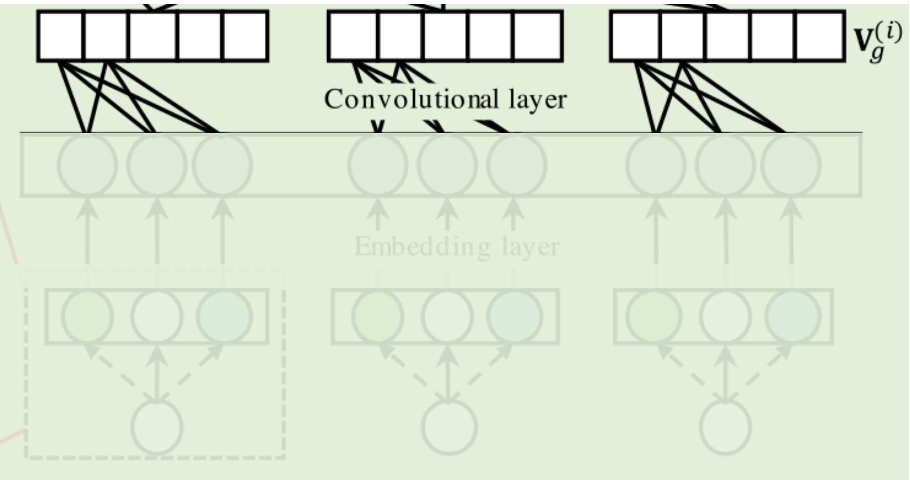
$\mathbf{E}_x \in \mathbb{R}^{m_g m_x \times d_e}$

a parameter vector $\mathbf{a} \in \mathbb{R}^{d_e}$: $\mathbf{e} = \hat{x} \cdot \mathbf{a}$.

Methodology

Context-Smoothing.

STEP 3 :feature fusion. compress each $\mathbf{E}_g^{(i)} (1 \leq i \leq m_x)$ to a vector.



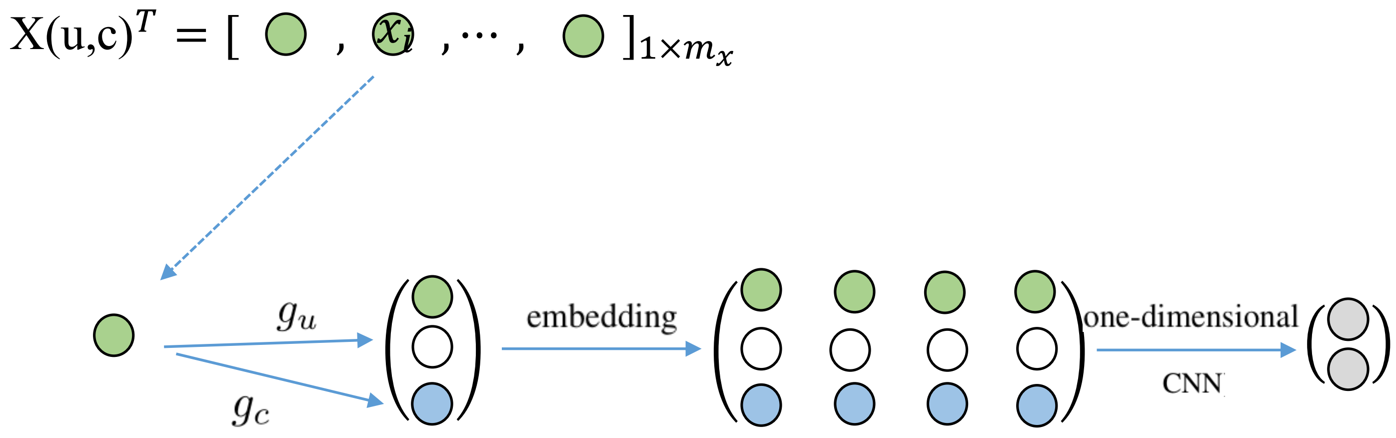
More formally, a vector $\mathbf{V}_g^{(i)} \in \mathbb{R}^{d_f}$ is generated from $\mathbf{E}_x^{(i)}$ by

$$\mathbf{V}_g^{(i)} = \sigma(\mathbf{W}_{conv} \delta(\mathbf{E}_g^{(i)}) + \mathbf{b}_{conv}), \quad (2)$$

where $\delta(\mathbf{E})$ denotes flatting matrix $\mathbf{E}$ to a vector, $\mathbf{W}_{conv} \in \mathbb{R}^{d_f \times m_g d_e}$ is convolution kernel, $\mathbf{b}_{conv} \in \mathbb{R}^{d_f}$ is bias term.

Methodology

Context-Smoothing.



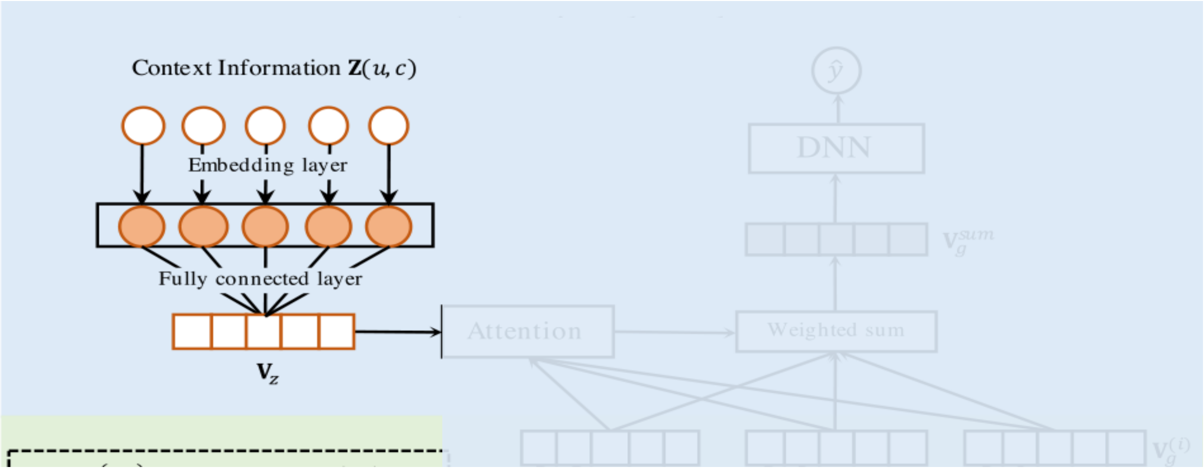
expanded with its <u>user and course-context statistics</u>

It can be seen as the context-aware representation of each x_i with <u>integrating its context statistics.</u>

Methodology

Attention-based Interaction

STEP 1 : the representation of Z

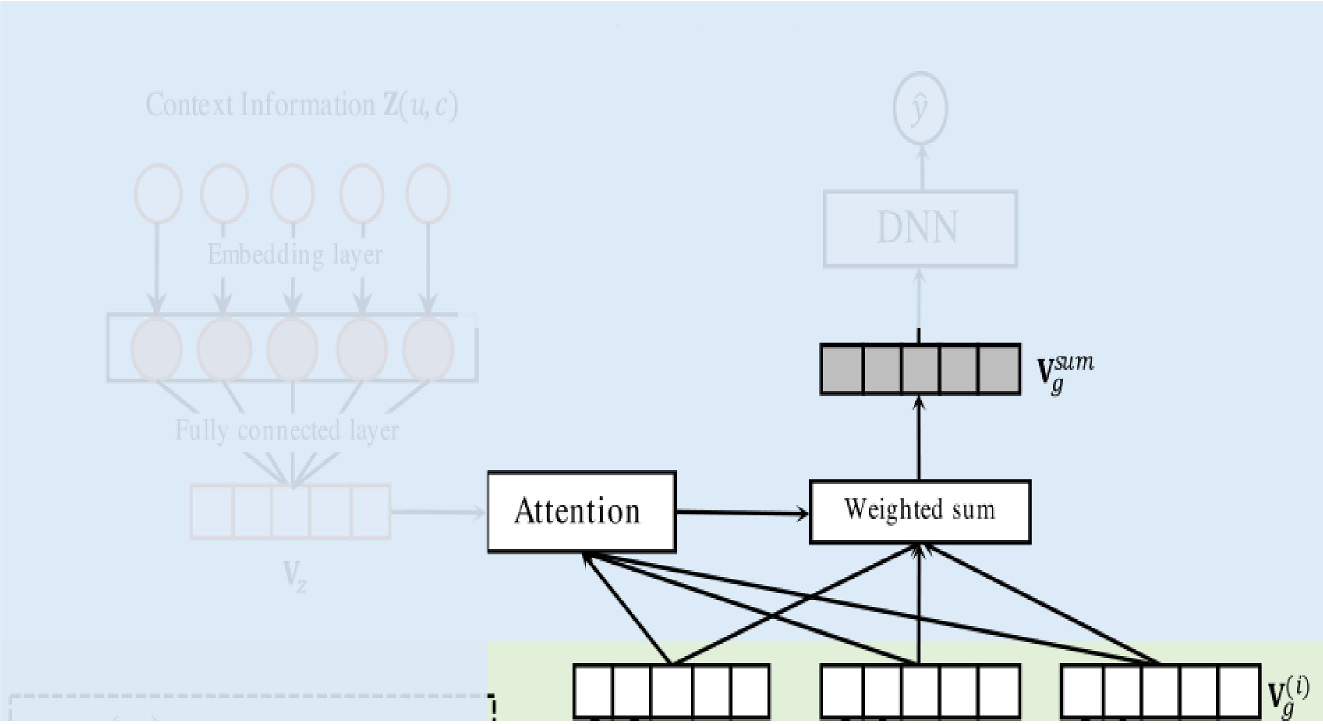


$$\mathbf{V}_z = \sigma(\mathbf{W}_{fc}\delta(\mathbf{E}_z) + \mathbf{b}_{fc}) \in \mathbb{R}^{d_f}$$

Methodology

Attention-based Interaction

STEP 2 : Attention



V_g^{sum}* can be seen as the context-aware representation of *X

$$\mathbf{V}_g^{sum} = \sum_{1 \le i \le m_x} \lambda_i \mathbf{V}_g^{(i)}.$$

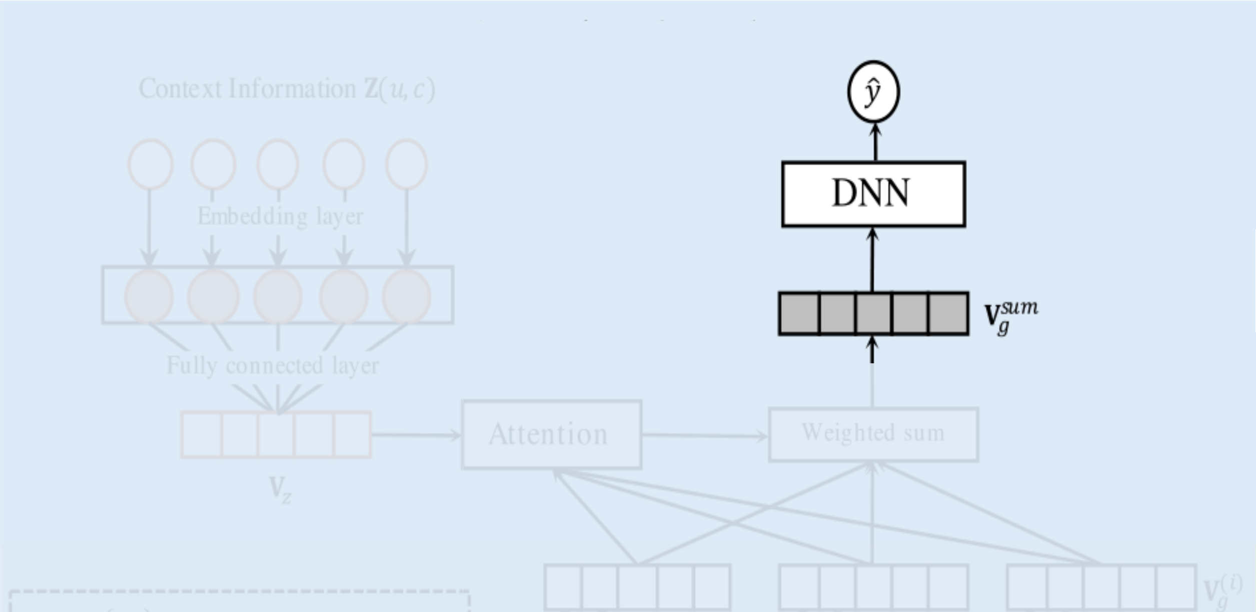
$$\hat{\lambda}_i = \mathbf{h}_{attn}^{\mathrm{T}} \sigma(\mathbf{W}_{attn}(\mathbf{V}_g^{(i)} \oplus \mathbf{V}_z) + \mathbf{b}_{attn})$$

$$\lambda_i = \frac{\exp(\hat{\lambda}_i)}{\sum_{1 \le i \le m_x} \exp(\hat{\lambda}_i)},$$

Methodology

Attention-based Interaction

STEP 3 : to learn the interactions of features *by DNN*



$$\hat{y}_{(u,c)} = \frac{1}{1 + \exp(-\mathbf{h}_{sigmoid}^{\mathrm{T}} \mathbf{V}_d^{(L-1)})}$$

$$\mathbf{V}_d^{(l+1)} = \sigma(\mathbf{W}_d^{(l)} \mathbf{V}_d^{(l)} + \mathbf{b}_d^{(l)})$$

Methodology

Model Ensemble

For further improving the prediction performance, we also design an ensemble strategy by combining CFIN with the XGBoost (Chen and Guestrin 2016), one of the most effective gradient boosting framework. Specifically, we obtain $\mathbf{V}_d^{(L-1)}$, the output of DNN's $(L-1)^{th}$ layer, from a successfully trained CFIN model, and use it to train an XGBoost classifier together with the original features, i.e., $\mathbf{X}$ and $\mathbf{Z}$. This strategy is similar to Stacking (Wolpert 1992).

Experiments

Implementation Details

L2 regularization

Features normalized

Rectified Linear Unit (Relu)

Adam

TensorFlow

cross-entropy cost function

Experiments

Comparison Methods & Prediction performance

Table 4: Overall Results on KDDCUP dataset and IPM courses of XuetangX dataset.

Methods	KDDCUP AUC (%)	KDDCUP F1 (%)	XuetangX AUC (%)	XuetangX F1 (%)
LRC	86.78	90.86	82.23	89.35
SVM	88.56	91.65	82.86	89.78
RF	88.82	91.73	83.11	89.96
DNN	88.94	91.81	85.64	90.40
GBDT	89.12	91.88	85.18	90.48
CFIN	90.07	92.27	86.40	90.92
CFIN-en	**90.93**	**92.87**	**86.71**	**90.95**

Experiments

Feature Contribution (feature ablation experiments)

Table 5: Contribution analysis for different engagements on KDDCUP dataset and IPM courses of XuetangX dataset.

	KDDCUP		XuetangX	
Features	AUC (%)	F1 (%)	AUC (%)	F1 (%)
All	90.07	92.27	86.50	90.95
- Video	87.40	91.61	84.40	90.32
- Forum	88.61	91.93	85.13	90.41
- Assignment	86.68	91.39	84.83	90.34

On KDDCUP, assignment plays the most important role.

On XuetangX, video seems more useful.

Experiments

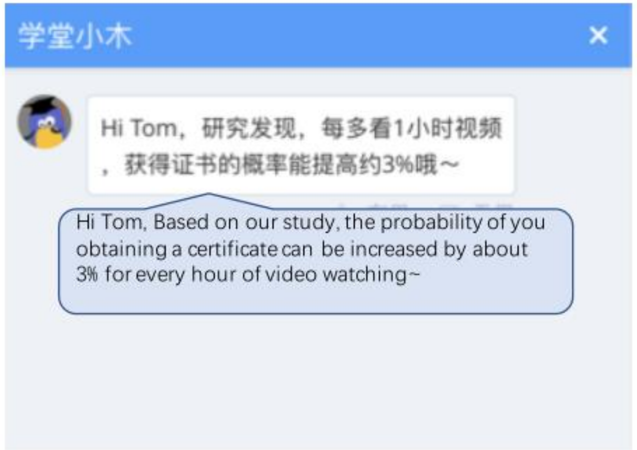
Feature Contribution
(fine-grained analysis for different features on different groups of users)

Table 6: Average attention weights of different clusters. C1-C5 — Cluster 1 to 5; CAR — average correct answer ratio.

Category	Type	C1	C2	C3	C4	C5
video	#watch	0.078	0.060	0.079	0.074	0.072
	#stop	0.090	0.055	0.092	0.092	0.053
	#jump	0.114	0.133	0.099	0.120	0.125
forum	#question	0.136	0.127	0.138	0.139	0.129
	#answer	0.142	0.173	0.142	0.146	0.131
assignment	CAR	0.036	0.071	0.049	0.049	0.122
	#reset	0.159	0.157	0.159	0.125	0.136
session	seconds	0.146	0.147	0.138	0.159	0.151
	count	0.098	0.075	0.103	0.097	0.081

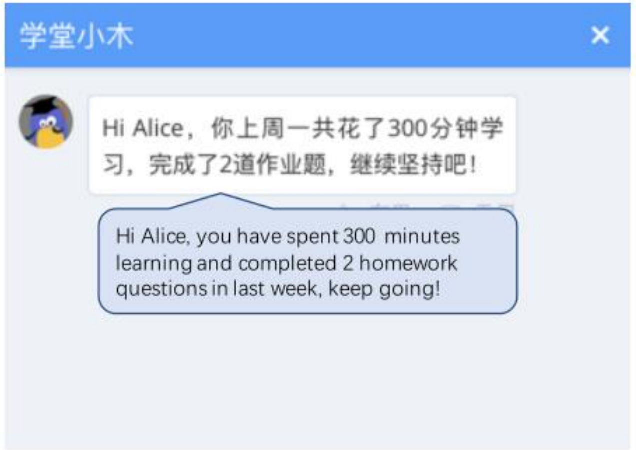
Experiments

From Prediction to Online Intervention (A/B test)



(a) Strategy 1: Certificate driven

(b) Strategy 2: Certificate driven in video



(c) Strategy 3: Effort driven

4 courses : Financial Analysis and Decision Making; Introduction to Psychology; C++ Programming; Java Programming

Experiments

From Prediction to Online Intervention (A/B test)

Table 7: Results of intervention by A/B test. WVT — average time (s) of video watching; ASN — average number of completed assignments; CAR — average ratio of correct answers.

Activity	No intervention	Strategy 1	Strategy 2	Strategy 3
WVT	4736.04	4774.59	5969.47	3402.96
ASN	4.59	9.34*	2.95	11.19**
CAR	0.29	0.34	0.22	0.40

*: $p-$value ≤ 0.1, **: $p-$value ≤ 0.05 by $t-$test.

Strategy 1 and Strategy 3 can significantly improve users' engagement on assignment.

Strategy 2 is more effective in encouraging users to watch videos.

Some Thoughts

Context Information can involve more information:

①Someone's dropout history:
one dropouts frequently has bigger probability to dropout,
and influence of course correlation.

②Friends' dropout situation.

Thank You!

QUESTIONS?