2022 Fall EE5183 FinTech - Homework 3

Deep Learning Model: Bidirectional Recurrent Neural Network, Transformer

Due: Nov 15, 2022

INSTRUCTIONS

1. Please only use PyTorch and scikit-learn to build the model.
2. You should write your codes independently. Plagiarism is strictly prohibited.
3. Report can only be written in English
4. All the figures are just examples. You do not need to be the same as the figures.
5. You must turn in hw3_student_ID.pdf and hw3_student_ID.ipynb (only one .ipynb file) and zip the files as hw3_student_ID.zip. TAs will grade your python code by Google Colab. Please ensure your code can be run on the Google Colab GPU environment. The wrong format will not be graded.

PROBLEMS

With the rise of the internet and mobile devices, push notifications have become a powerful marketing tool for e-commerce and media websites. Push notifications are short messages that pop up on the user's mobile or desktop, nudging them to take some action. There are usually three important stages: push(an app owner sends a push notification to users) →impression(user receives the push notification on their device) →click (user clicks or not after receiving the impression.) And click-through rate (CTR) is one of the most important push notification performance metrics. In this homework, our goal is to predict the probability that a user will click after receiving each push notification.

**Click-through rate (CTR) represents the number of notifications that were clicked out of the total notifications that were successfully received.*

In this homework, the dataset is the impression data of the online news media, Newtalk, provided by AviviD innovative media company.
There are two kinds of datasets we will use.

News_list: The content of all push notifications(news) was collected in this dataset.

	news_index_encode	articleSection	keywords	datePublished	title	url	weekday	articleSection_encode
0	0	國際	中國,美國,裴洛西訪台,訪台,裴洛西	2022-07-30 06:24:37	裴洛西亞洲行將敗程 白宮：未見中國即將對台灣採取軍事行動的證據	https://newtalk.tw/news/view/2022-07-30/793561...	6.0	9
1	1	政治	民進黨,林智堅,論文抄襲,陳明通,邱顯智	2022-07-30 05:29:33	民進黨「集體瘋狂加入抄襲國家隊」邱顯智：毀黨沒意見 誤國罪擊深重	https://newtalk.tw/news/view/2022-07-30/793555...	6.0	2
2	2	政治	民進黨,林智堅,論文抄襲,陳明通,邱顯智	2022-07-30 05:29:33	民進黨「集體瘋狂加入抄襲國家隊」邱顯智：毀黨沒意見 誤國罪擊深重	https://newtalk.tw/news/view/2022-07-30/793555...	6.0	2
3	3	國際	中國,美國,裴洛西訪台,訪台,裴洛西	2022-07-30 06:24:37	裴洛西亞洲行將敗程 白宮：未見中國即將對台灣採取軍事行動的證據	https://newtalk.tw/news/view/2022-07-30/793561...	6.0	9
4	4	生活	低壓帶,雷雨,中央氣象局,高溫資訊	2022-07-30 07:22:39	高屏花東高溫36度 低壓帶接近午後留意雷陣雨	https://newtalk.tw/news/view/2022-07-30/793565...	6.0	17

- news_index_encode: ID of each news.
- articleSection: Category of news.
- keywords: Keywords of the news.

- datePublished: The date the news was published.
- title: Headline of the news
- url: URL of the news
- weekday: What day of the week is the news published?
  {0:'nan',1:'Mon',2:'Tue',3:'Wed',…,7:'Sun'}
- articleSection_encode:
  { 0:'nan',1:'科技',2:'政治',3:'中國',4:'財經',5:'旅遊',6:'網紅',7:'社會',8:'娛樂',9:'國際',10:'環保',11:'電競',12:'遊戲',13:'專欄',14:'新奇',15:'創夢',16:'體育',17:'生活',18:'選舉'}

Impression: This dataset contains impression records from 2022/08/01 to 2022/08/10.

	token	impression_time	news_index_encode	target_news_onehot	all_click_seq	click_seq_10	click_seq_10_onehot	y
131267	55714	2022-08-01 11:00:00	643	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, …	[]	[]	[[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…	0
131268	55714	2022-08-01 13:00:00	676	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, …	[]	[]	[[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…	1
131269	55714	2022-08-01 13:00:00	702	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, …	[676]	[676]	[[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,…	0
131270	55714	2022-08-01 14:00:00	714	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, …	[676]	[676]	[[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,…	0
131271	55714	2022-08-01 14:00:00	732	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, …	[676]	[676]	[[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,…	0
131272	55714	2022-08-01 14:00:00	764	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, …	[676]	[676]	[[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,…	1
131273	55714	2022-08-01 15:00:00	785	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, …	[676, 764]	[676, 764]	[[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,…	1
131274	55714	2022-08-01 18:00:00	805	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, …	[676, 764, 785]	[676, 764, 785]	[[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,…	0
131275	55714	2022-08-01 18:00:00	813	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, …	[676, 764, 785]	[676, 764, 785]	[[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,…	0
131276	55714	2022-08-01 18:00:00	836	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, …	[676, 764, 785]	[676, 764, 785]	[[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,…	0

- Token: Regard as each device id (or regard as each user id)
- impression_time: The time a user receives the push notification(news).
- news_index_encode: ID of push notification(news).
  (This is our target, we want to predict whether the token will click this news.)
- all_click_seq: Click record from start date(2022/08/01) to present.
- click_seq_10: The last 10 clicks record from now on.
- Y: Click or not(ground truth).

We can transform each piece of news as a one-hot vector of length 27(The first 19 elements (blue) are for the 'article section', and the last 8 elements (green) are for 'nan' & Mon-Sun).
For example, a piece of news published on Monday and belongs to {1:'科技'} category, can be expressed as [0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0].

- click_seq_10_onehot: Ten one-hot vectors for the last 10 clicked news
- target_news_onehot: One-hot vector of the target news.

1. Predict the probability of clicking the push notification

(i) (5%) Please split the impression dataset from 2022/08/01-2022/08/07 as the training part, the data from 2022/08/08-2022/08/09 as the validation part, and the data from 2022/08/10 as the test part.

(ii) (20%) Please construct a click prediction model with a bidirectional-LSTM cell for predicting the probability of click according to the binary cross entropy. Please explain how you design your model.

Hint: Please concatenate 'target_news_onehot' after 'click_seq_10_onehot' as input of bidirectional-LSTM cell. In brief, the input size of your model is (batch size,11, input dimension).

(iii) (10%) Please plot the training loss and validation loss like in Figure 1.

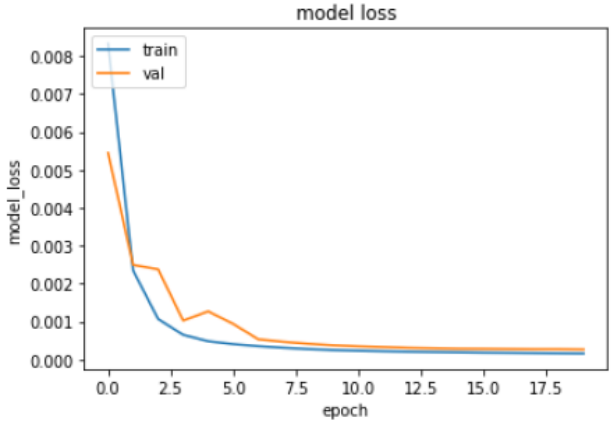
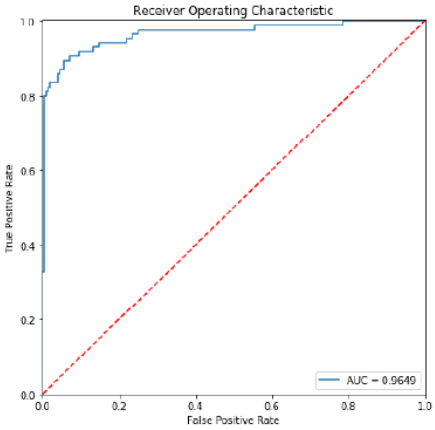


Figure 1: Example of loss curve. (The figure is just an example)

(iv) (15%) You have to plot the receiver operating characteristic curve (ROC), and precision-recall curve (PRC) with their area-under-curve (AUROC and AUROC), as shown in Figure 2 on the training and validation set. And show the AUROC score and AUROC score of your training, validation, and test set. Which metric do you think is more suitable for this task? Please explain why.



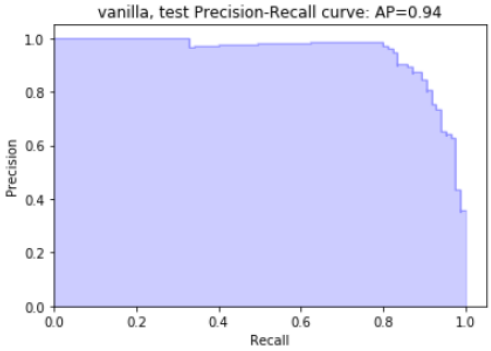


Figure 2: ROC curve and Precision-recall curve.

(v) (15%) Please replace the bidirectional-LSTM cell in your click prediction model with a bidirectional-GRU and repeat (iii), (iv).

(vi) (15%) Please replace the bidirectional-GRU cell in your click prediction model with a transformer encoder block and repeat (iii), (iv).
Reference document:
https://pytorch.org/docs/stable/generated/torch.nn.TransformerEncoderLayer.html

(vii) (5%) Please use the metric you choose in (iv) to compare these three models (Bidirectional-LSTM, Bidirectional-GRU, Transformer) and briefly describe what you found.

(viii) (15%) In the beginning, we mentioned in the impression dataset, 'click_seq_10' means the last 10 click records from now on, and 'click_seq_10_onehot' is their one-hot vector. Please try to replace this feature with the last 3 and 5 click records. (You need to generate the two datasets by yourself.) Then train your model again. (You can choose one from Bidirectional-LSTM, Bidirectional-GRU, or Transformer) and compare the performance of the model.

(ix) (Bonus 10%) In this homework, we only use 'articleSection' and 'weekday' as news-related features. However, some information like 'keywords' and 'title' are also important features we can use. Please use your creativity to add these features to the model. (you can explain your ideas or directly implement them, the latter gets more points.)