

Answer Correctness Prediction:

Track knowledge states of 1M+ students using Machine Learning

利用機器學習建立知識追蹤模型預測學生學習成效

組別：B74

組員：林姮汝

指導教授：孫民

Introduction

COVID-19迫使大多數國家暫時關閉學校，進一步推遲了學習機會和智力發展，使教育變得非常艱難，這也迫使我們重新思考當前的教育系統。Rid Labs是一家人工智慧解決方案供應商，它使全球教育參與者重新思考利用AI的學習方式，希望能根據不同學生提供個人化的學習體驗。在本次專題中，我將創建「知識追蹤」的演算法，搭配該公司提供的EdNet（人工智慧教育開放資料庫，包含超過1億個學生互動）對學生隨著時間的知識表現進行建模。

我共訓練了四個著名的模型：XGBoost[1], LightGBM[2], LSTM[3], Transformer[4][5]，用以預測學生未來學習表現的判別器(classifier)，並比較不同模型的分類效果。訓練完成的模型將有助於應對教育面臨的全球挑戰，在後COVID-19世界中建立更好、更公平的教育模式。

Method

首先對資料集進行處理，移除與答題情況無關的特徵欄，接著根據學生ID與問題ID新增統計特徵，例如：各用戶的平均回答正確率、各用戶的回答正確率之標準差、各問題回答正確率偏差、各問題平均回答正確率等。最後的新資料集包含了共12個特徵加上1個target（answered_correctly），如圖一。

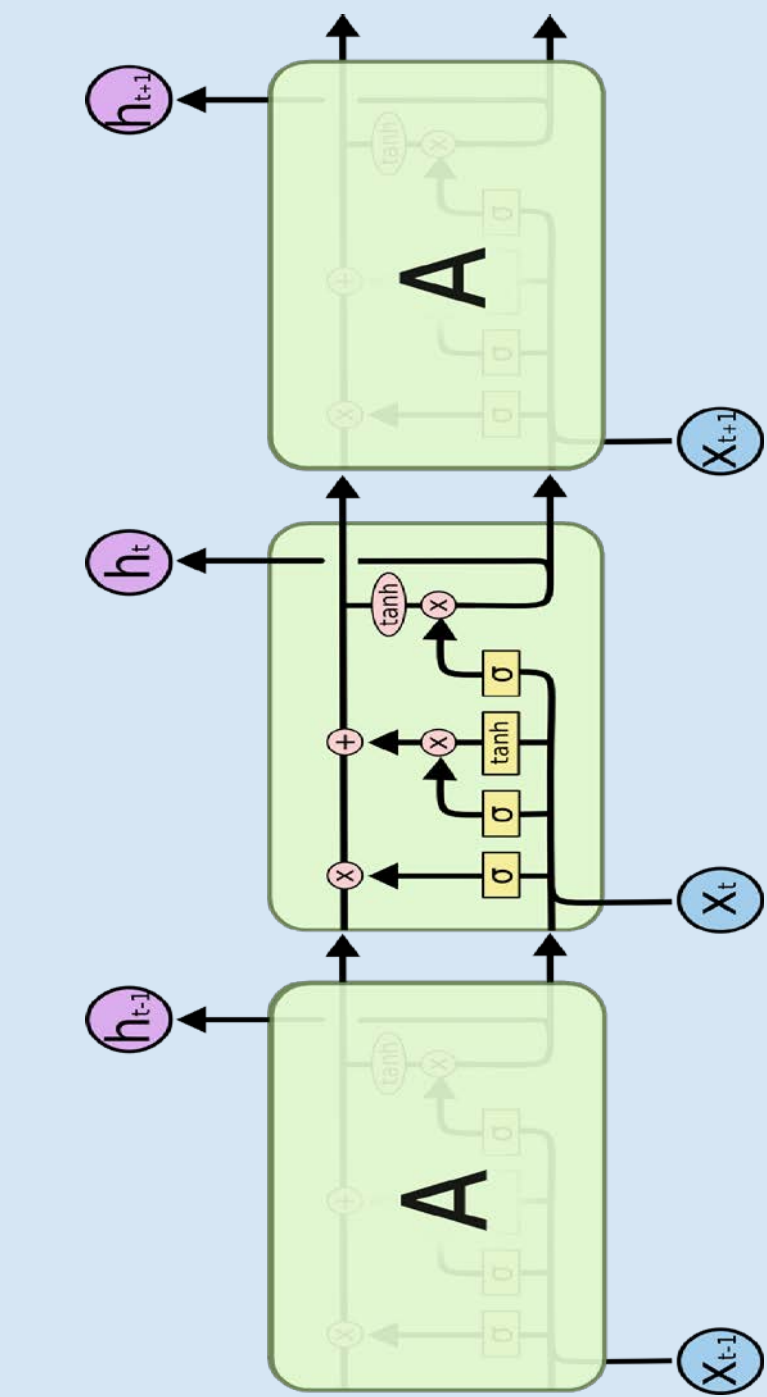
user_id	content_id	prior_question_elapsed_time	prior_question_had_explanation	mean_user_accuracy	questions_answered	
0	115	5692	0.5	0	0.695652	46
1	115	5716	37000.0	0	0.695652	46
2	115	128	55000.0	0	0.695652	46
3	115	7860	19000.0	0	0.695652	46
4	115	7922	11000.0	0	0.695652	46
std_user_accuracy	skew_user_accuracy	mean_accuracy	question_asked	std_accuracy	skew_accuracy	answered_correctly
0.465215	-0.879359	0.727740	1752.0	0.445250	-1.024143	1
0.465215	-0.879359	0.743266	1188.0	0.437015	-1.115184	1
0.465215	-0.879359	0.973442	979.0	0.160869	-5.898109	1
0.465215	-0.879359	0.959854	1096.0	0.196391	-4.691604	1
0.465215	-0.879359	0.953878	954.0	0.209858	-4.334655	1

▲圖一、新資料集示意圖
（12個特徵+1個target）

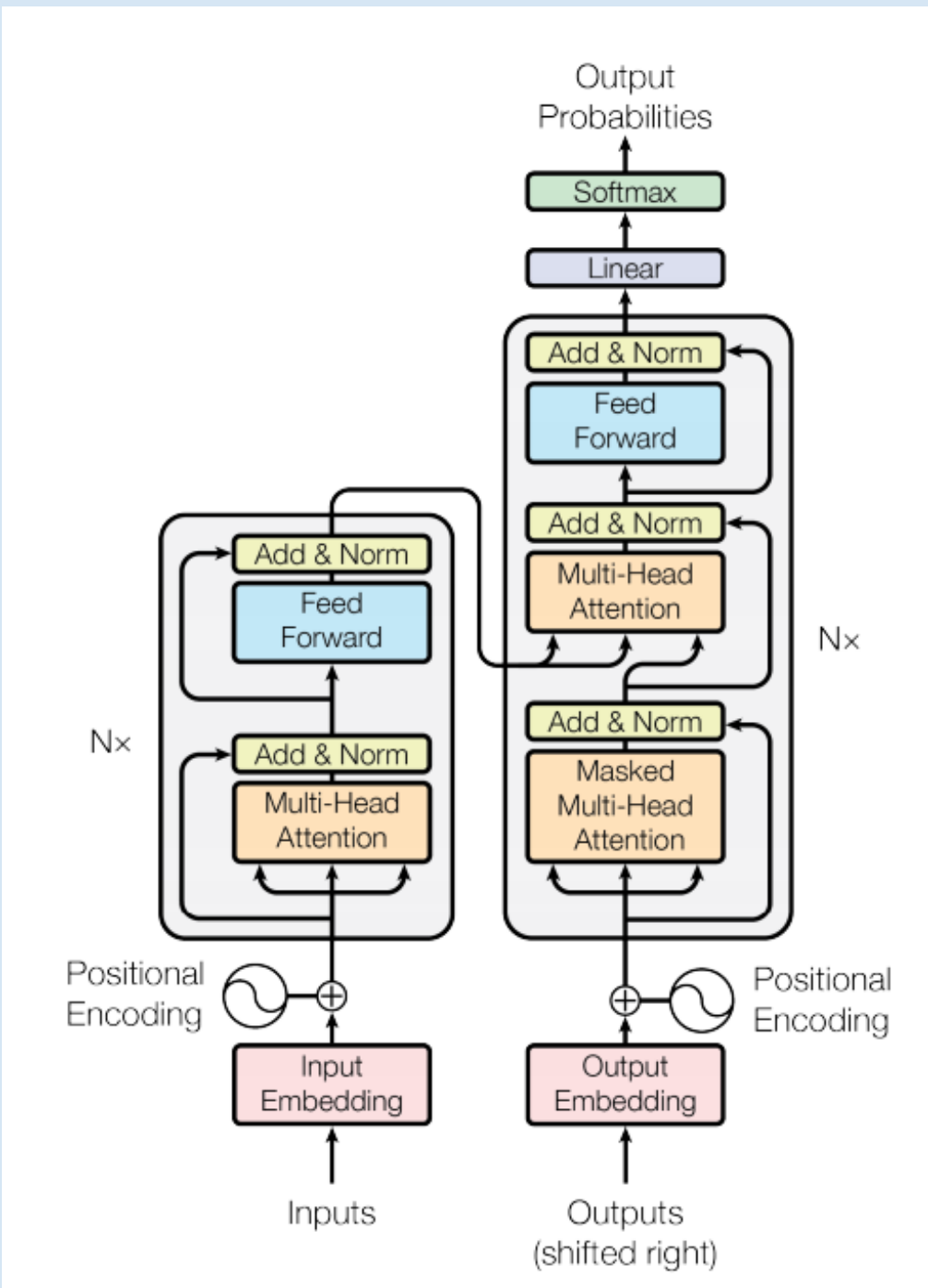
我先採用了著名的分類模型：XGBoost[1] 和 LightGBM[2]，接著嘗試深度學習模型 LSTM[3] 及 Transformer[4][5]，分別訓練後進行各個網絡的效能比較。

先將training data以8:2的比例分成training set和 validation set，接著做feature scaling，將特徵值進行標準化，使資料不過度發散，提高模型訓練結果。

XGBoost[1]和LightGBM[2]皆是基於決策樹演算法，將多個弱分類器集結成為一個強分類器。接著對此兩個模型，我進行了feature selection，比較提取較重要特徵前後模型的表現。至於LSTM[3]（圖二）和Transformer[4][5]（圖三）皆適用於處理與時間相關的事件，但是比起前者用hidden state來記錄過去的歷史，Transformer[4][5]使用了multihead attention，可以把任意一個時間的資訊傳到下一層的任何一個node，改善了LSTM[3]進行recurrent state運算時會忘記以前資訊的問題。因此針對此專題中涵蓋長時間事件的資料集，Transformer[4][5]的表現理應比LSTM[3]佳。



▲圖二、LSTM架構示意圖



▲圖三、Transformer架構示意圖

Result

整個專題研究中，我對4個不同的網絡建立模型，並比較其結果。這裡採用的模型評估方法是AUC-ROC曲線，因為training set有class imbalance的現象，不能用單純的accuracy評估模型的表現。表一為從feature importance提取較重要特徵前後，XGBoost[1] 和 LightGBM[2] 的表現；表二則為4個模型在All features（12個特徵）下的表現。

Method	All Features	After Feature Selection
XGBoost	0.7586	0.7568
LightGBM	0.7645	0.7641

▲表一、特徵提取前後，驗證集的AUC-ROC

Method	AUC-ROC
XGBoost	0.7622
LightGBM	0.7645
LSTM	0.7610
Transformer	0.5332

▲表二、AUC-ROC of Validation Set

Conclusion

由實驗結果，LightGBM[2]的AUC-ROC達到76.45%，在實作上的效果在4個網絡中是最優異的，其次為XGBoost[1]。然而根據Reference[5]，Transformer[4][5]的結果是目前研究界的SOTA，但在本次專題實作上，效果卻是最差的。除此之外，其餘三個模型的表現都還有進步空間，因此我有以下結論：

對原資料集中的特徵加上資料處理後的平均值、標準差、偏差等統計特徵組合成的新資料集，四個模型的表現都還能提升。未來可以針對不同組合的特徵與統計方式形成多個資料集，分別丟進模型中進行訓練與驗證。如此一來便可找出隨時間變化最相關的特徵組合，提高模型效能，實際運用於教育層面便能針對不同學生的學習狀況提供最適合的模式。

Reference

- [1] Tianqi Chen, Carlos Guestrin: XGBoost: A Scalable Tree Boosting System. arXiv:1603.02754, 2016
- [2] Guolin Ke, Q. Meng, Thomas Finley: LightGBM: A Highly Efficient Gradient Boosting Decision Tree. NIPS 2017.
- [3] Ralf C. Staudemeyer, Eric Rothstein Morris: Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks. arXiv:1909.09586, 2019.
- [4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: Attention Is All You Need. arXiv:1706.03762, 2017.
- [5] Youngduck Choi, Youngnam Lee, Junghyun Cho, Jineon Baek, Byungsoo Kim, Yeongmin Cha, Dongmin Shin, Chan Bae, Jaewe Heo: Towards an Appropriate Query, Key, and Value Computation for Knowledge Tracing. arXiv:2002.07033, 2020.