Chiehmin Wei
2020/04/16

Usage

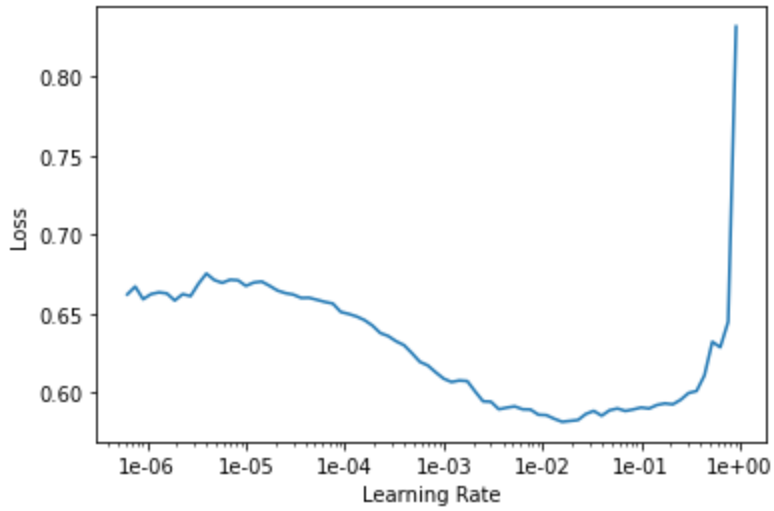
train.py
usage: python train.py {database_host} {model_filepath}
e.g. python train.py 35.187.144.113 models/trained_model.pkl

predict.py
usage: python predict.py {database_host} {model_filepath} {output_filepath}
e.g. python predict.py 35.187.144.113 models/trained_model.pkl predictions.csv

Methodology

從資料庫裡面抓出來 經過整理後
訓練集共有 103259222 個unique post
在如此龐大的數據量下
計算量太過龐大 原本想先做的baseline例如SVM以及random forest都做不了
決定直接使用深度學習的方式建模
其中訓練時由於使用的Google Colab instance會有OOM的問題
因此在訓練集裡隨機取樣50%作為實際模型訓練時的訓練集
並取2000個post作為validation set

由於input為tabular data
決定使用簡單的2層全聯接MLP模型
Using a 2-level MLP. Sizes are [200, 100]
使用f1 score為metric
batch size=64
用標準的adam optimizer並參照論文 Fixing Weight Decay Regularization in Adam
Learning rate的設置和warmup上參照Leslie Smith's paper 提到的one-cycle policy來選擇



Chiehmin Wei
2020/04/16

值得一提的是對特徵的處理
除了各種count之外 我認為po文的時間也會對一篇文是否上熱門有很大的影響
例如深夜發文可能沒什麼人看 或是是平日還是週末發文 可能都有影響
經分析後決定
將原始table裡的created_at_hour這個column由timestamp轉為兩個categorical特徵:
created_at_hour (0~23)以及created_at_weekday(0~6)
分秒太細決定不包括 年太粗略決定不包括
月份在檢視訓練集資料時發現沒有包含所有1年12個月的資料 決定不包括
這些categorical variable各經過embedding layer輸入進模型裡
embedding size分別為{'created_at_Dayofweek': 10, 'created_at_Hour': 15}

至於剩下的count variables 就用很基本的standardization來處理
讓他們的值都在0~1之間 幫助訓練並解決scale不同的問題 (例 愛心人數通常比分享人數高很多)

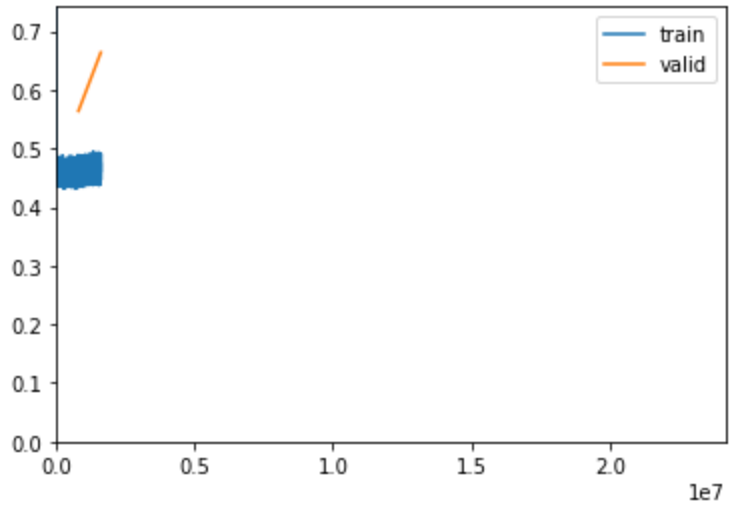
另外following the spec
我將like_count_36_hour根據是否大於1000來作為是否is_trending的標準
將regression問題簡化為binary classification問題 相信比較好訓練
最後input為以下table

```
#   Column                Dtype
---  ------                -----
 0   created_at_hour       int64
 1   share_count           int64
 2   comment_count         int64
 3   like_count            int64
 4   collect_count         int64
 5   is_trending           int64
 6   created_at_dayofweek  int64
```

很可惜這週在期中考 開始做的比較晚 今天下午才開始弄
沒想到數據量這麼大 模型訓練的很慢 一些原本想跑的實驗(e.g.換參數)沒時間完成
原本打算訓練30個epoch 使用early stopping (threshold=0.01, patience=5 epochs)
實際上由於時間不足 1個epoch需要訓練2小時 只訓練了2個epoch

epoch	train_loss	valid_loss	f_beta	time
0	0.467352	0.563821	0.633263	2:02:04
1	0.472983	0.663502	0.429027	2:02:29

Chiehmin Wei
2020/04/16



並且loss diverge了 也許需要跟低的learning rate比較好
或是調整一下模型的參數 可惜這次太晚開始做

在Testing Data上的表現

在testing data上22929481個unique post的f1-score為0.3333533360924349