機器學習與深度學習期中報告

TMU ML Course Team Battle 1

組別 ： 鎮瀾宮的骰子(Feat. 標太郎水餃)

組員 ： 黃彥鈞(m947108006)

石家宜(m946108003)

趙上涵(m946108004)

a）**資料集**

Data	Label
影評內文	Pos/ Neg

b）**競賽結果**

Public : 0.87298 (Rank 3)、Private : 0.86924 (Rank 6)

c）**基礎分類模型**

前處理 ： Data Representation : TFIDF

Models	Cross_validation	Private_scores	Public_scores
Multinomial NB	0.865	0.740	0.761
kNN	0.775	0.663	0.672
Decision Tree	0.704	-	-
Linear Regression	0.890	0.838	0.846
Random Forest	0.739	0.706	0.691
SVM (kern el = linear)	0.896	0.846	0.859
XGBoost	0.808	0.874	0.864

d) **Bert**

使用建構於 PyTorch 框架的 Bert 模型[1]，並將在原始文本資料經過轉換成 Bert 接受的.tsv 檔案，並進行特徵萃取。最後再讀入 Pre-trained 的 Bert 模型，進行 Tokenization 和 Fine Tuning 後，交付模型訓練。

[1] . Rajapakse, T. (2019, June 10). A Simple Guide On Using BERT for Binary Text Classification.

Model	Private scores	Public scores
Bert binary classification (3 epochs)	0.869	0.873
Bert binary classification	0.860	0.867

- 參數:

MAX_SEQ_LENGTH = 256

TRAIN_BATCH_SIZE = 24

EVAL_BATCH_SIZE = 32

LEARNING_RATE = 2e-5

NUM_TRAIN_EPOCHS = [1, 3]

RANDOM_SEED = 42

GRADIENT_ACCUMULATION_STEPS = 1

WARMUP_PROPORTION = 0.1

e) **Conclusion**

經歷這次報告深刻感受到運算硬體很重要，特別是在比較複雜的模型如 SVM、DNN (賽期中未成功架設完成)、BERT 等模型運算過程很吃記憶體，SVM 約莫 1 小時訓練時間，BERT 訓練一次大概花了整個晚上。

目前學習到許多不同的方法，例如 CNN, RNN, 特徵選擇方法等等，未來的操作方向可以結合這些方法與 BERT、XGB 等等目前結果不錯的模型來更加優化測試結果。

f) **Reference**

1. Rajapakse, T. (2019, June 10). A Simple Guide On Using BERT for Binary Text Classification [Web blog message]. Retrieved from https://medium.com/swlh/a-simple-guide-on-using-bert-for-text-classification-bbf041ac8d04