INTRODUCTION TO MACHINE LEARNING

FINAL REPORT

Dcard 關鍵字分析

0316048 蘇炳立
0316027 郎宇傑
0316327 吳俊賢

I. INTRODUCTION

Dcard（狄卡），是台灣一個提供社群網路服務的網站, 有感於大部分同學，都只在生活圈交友，沒有接觸到更多的人脈，因此設立的論壇，包括各種看板，類似 PTT，但多了交友的功能。在這次的 Final Project 當中，我們選擇了"**心情版**"和"**西斯版**"做研究，分析這個版上的關鍵字，訓練各種不同的 Model 以及演算法，最後預測此篇文章**是否可以突破一千個讚數**。

在分析訓練之前，首先要做的是資料的前處理，使用 Python 語言寫**爬蟲**程式，對接 Dcard 的 API，拿到關鍵字的資料，整理之後，利用 Python 和 MATLAB 機器學習的模型，我們選擇了 **KNN**、**RegressionTree**、**Decision Tree**、**B ayse Model** 等，演算法的部分使用 **AdaBoost**、**LogiBoost**、**TotalBoost**。最後使用"正確率"、"命中率"、"陽性預測值"、還有關鍵字與 Target 的相關係數作分析。

II. 資料的前處理

A. 爬蟲程式對接 Dcard 的 API

抓取選定看板的文章（Ex: mood, sex)

B. 篩選出關鍵的 TAG(Dcard 提供)

每篇文章依據 tag existence 增加 feature 抓取發文者性別 圖片數量 是否為回覆某篇文章

III. 資料集

- 心情版 30000 篇文章 meta data，其中有 827 篇破千讚的文章
- (暗黑板) 14250 篇文章 meta data，其中有 604 篇破千讚的文章

IV. 關鍵字

gender,hasSchool,reply,numImg,withNickname,過得,怎麼,認識,覺得,任何,客人,朋友,小姐,家屬,或許,哥哥,照片,吃飯,文長,看見,真的,情侶,幾個,出來,不要,這樣,fb,自己,阿姨,旁邊,對話,女友,女生,一種,他們,同學,想說,那種,一樣,幾天,這次,因為,發現,圓,家人,叔叔,幫忙,女兒,這些,沒什麼,身體,爸爸,一切,之後,事情,女孩,我們家,擔心,遇到,阿嬤,姊姊,台灣,大家,店員,家長,時間,為什麼,病人,飲料,一些,男生,怎樣,畢業,醫院,家裡,事件,差點,一間,人生,不會,新聞,告訴,雖然,閨蜜,狼師,還是,高鐵,那個,高中,心情,男友,老師,妹妹,強暴,減肥,結果,別人,影片,記得,粉絲,然後,房間,總是,室友,聽到,兩個,電話,學長,已經,弟弟,媒體,來說,還有,黑特,運動,廁所,離開,學妹,位子,當下,拜託,的人,她說,姐姐,男朋友,大學,謝謝,成績,瑞雪,隔壁,國小,孩子,這麼,的話,這個,問題,一個,班上,爸媽,個人,醫生,受傷,更,感覺,大哥,應該,現在,自殺,關於,趕快,學生,曾經,有沒有,小孩,其實,每個,當時,手機,難過,媽媽,時候,你們,父母,活動,對不起,回家,00,發生,後來,知道,有時候,社會,沒有,這件,限時,兒子,討厭,的說,對方,群組,youtuber,句話,ig,大學生,下車,件事,奶奶,學校,那麼,三個,各種,看著,心裡,這種,餐廳,報警,全部,些人,東西,身邊,喜歡,國中,我們,開始,什麼,甚麼,開心,她們,外婆,likeCount

v. 選用的 MODEL

A. *k-nearest neighbor*

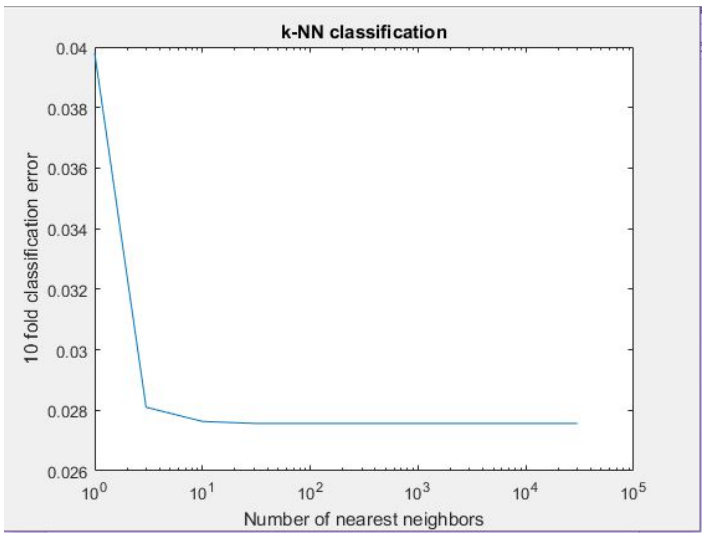
- 訓練的方式

Python: KNeighborsClassifier
MATLAB: fitcknn

兩者的參數均為 neighbor 的數量、演算法是否用 KDTree、以及算距離的方法(曼哈頓、歐基里德...)

- 選擇一個適合的 N

N 選太小，會造成 Overfiting，N 選太大則會造成整個 Model 對於所有的文章都只預測其不會破千讚。所以選擇 3、5 做 KNN 的訓練。



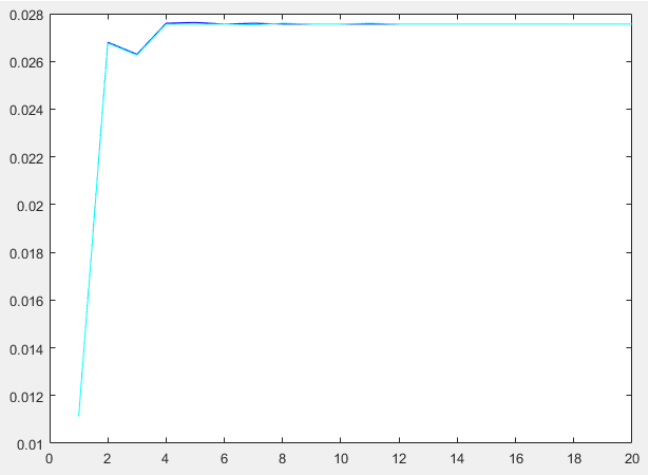
- KNN 可能的隱憂:

1. 取越多的鄰居，而鄰居都是沒有破千讚的(20000 多篇文章是沒有破千讚的，造成 **underfiting**，若 K 值只取一則會造成 **overfitting**，因此，在後面的結論當中，雖然 KNN 的**正確率** Accuracy 很高，但是**命中率 TPR(實際破千讚，成功的被預測出來)以及陽性命中率 PPV(預測千讚中，正確的比率)普遍都偏低。**

2. 由於資料的維度太高，因此在建 kd-tree 的時候很耗時

- KNN 的比較

在此次訓練之中，無論選擇曼哈頓距離，抑或是歐基里德距離，彼此的 error rate 相差不到 1%，因此固定選擇歐基里德距離配合 kd-tree 演算法



B. *Bayes Model (Smoothing)*

- 訓練的方式

Python: sklearn.naive_bayes
MultinomialNB() (Smoothing)
bay.fit(X,Y)
MATLAB: Mdl = fitcnb(X,Y)

- 貝氏模型對於本次的訓練之中，依據理論來說，二元分類(binary classification)會比其他 model 準確，但是由於 noise 偏多，導致準確率不如預期。

C. Random Forest

- 從 dataset 中取 subset 去做給定特定數量的決策樹形成 forest，然後從各個樹中投票決定分類(每棵樹沒做 pruning)
- 權重調配 : 千讚與非破千讚比例的倒數

D. Decision Tree

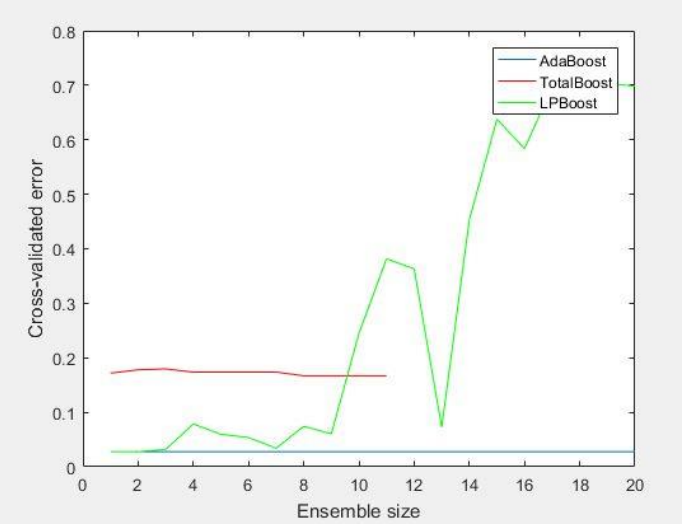
- 訓練的方式

Python: tree.DecisionTreeClassifier
Tree.fit(X_train,y_train)
MATLAB: fitctree(X,Y)
prune(tree)
crossval(tree)

- 由於樹的構造過於複雜，造成 overfiting，但是 Decision Tree 對於 noise 有很好的抗性，因此陽性命中率較其他的 model 為高。

- Boosting

fitensemble(x3,Y,'AdaBoostM1',T,'Tree')
在 Decision tree 的 model 當中，分別使用 AdaBoost、LPBoost、TotalBoost 作比較，由圖可以發現，AdaBoost 對於 noise 太多的問題仍無法解決。



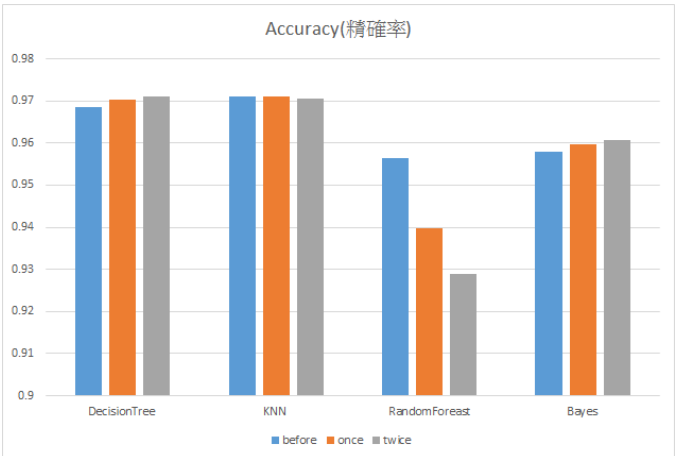
VI. 訓練的改善(特徵的選取)

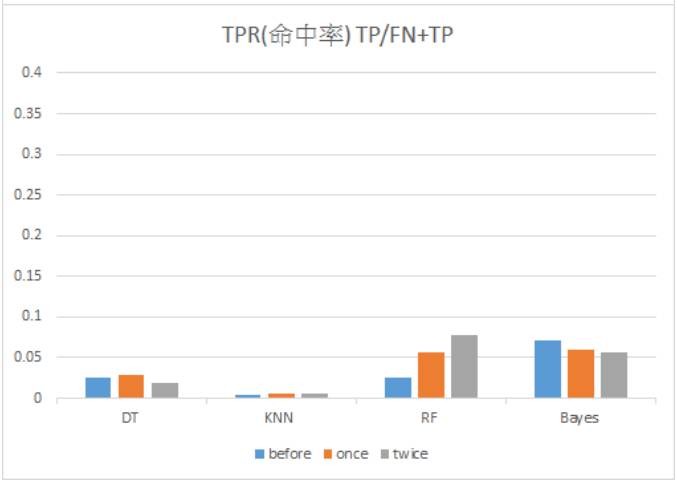
```
from sklearn.ensemble import
ExtraTreesClassifier
from sklearn.feature_selection import
SelectFromModel
```

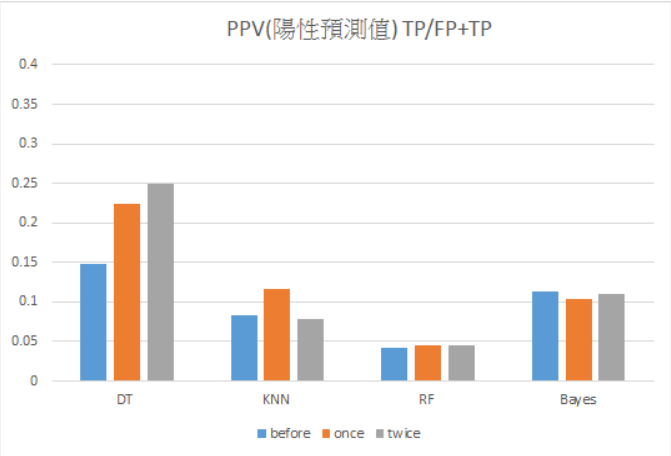
原理:ExtraTreesClassifier 是將資料分為數個子集合，分別建造 decision tree，根據 gini importance(資料分類的純度)，選取深度較淺的 node 所代表的 feature

VII. Model 的比較與結果

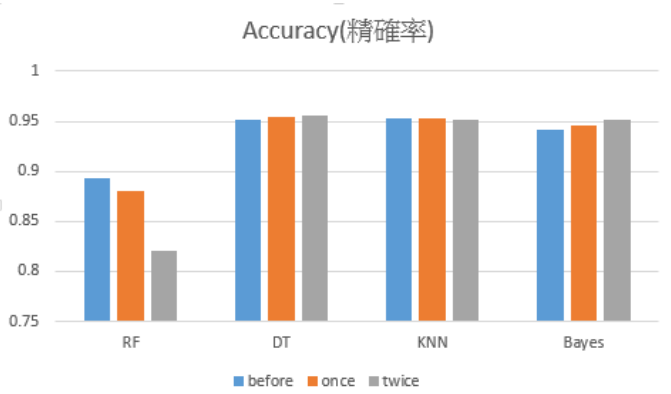
A. 心情版的預測

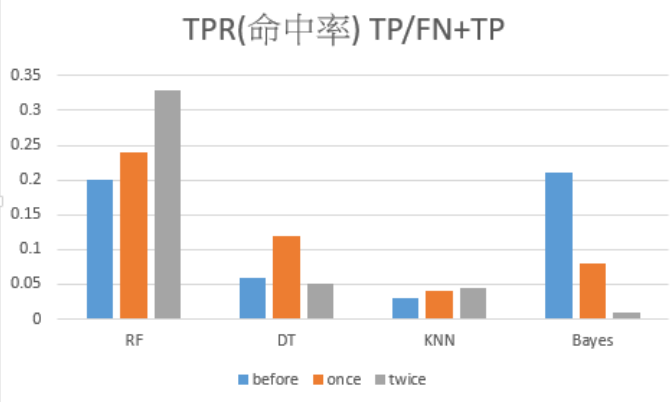


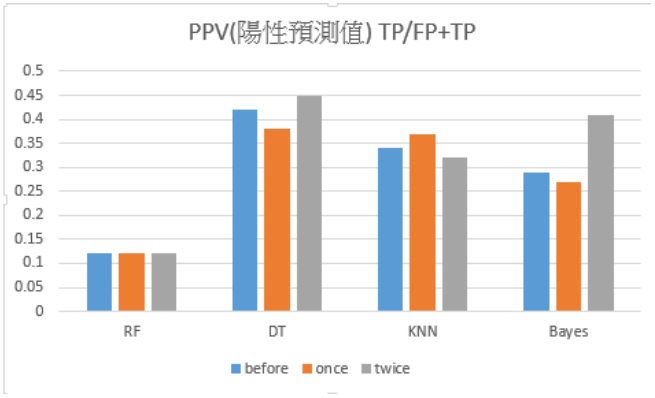




B. 黑暗版的預測







C. 討論

1.為何陽性預測值比命中率還要高?
我們預測破千讚的的文章準確率比較高

代表 model 從破千讚文章中有學習到的特性是準確率高的 ，但其中 model 還是有很多沒學習到破千讚的特性所以命中率較低

2. 為什麼預測破千讚的結果不佳

由於大多數的文章沒有破千讚，導致許多 model 沒有學到關鍵的 feature(可能一開始即沒有選到)，很多其實是破千讚的文章，我們卻預測它並沒有破千讚

VIII. 關鍵字與 Target 的相關係數

拿來訓練的 features，似乎不是決定會不會破千讚的關鍵
各個 descriptive feature 與 target 的相關係數都不高 (最高為圖的數量約 0.0627)

媽媽的相關係數

1.0000	0.0525
0.0525	1.0000

學長的相關係數

1.0000	0.0187
0.0187	1.0000

朋友的相關係數

1.0000	-0.0108
-0.0108	1.0000

真的的相關係數

1.0000	-0.0049
-0.0049	1.0000

客人的相關係數

1.0000	0.0175
0.0175	1.0000

圖的相關係數

1.0000	0.0627
0.0627	1.0000

IX. 訓練成果

1 最常出現的關鍵字:

媽媽、學長、朋友、真的、客人

標題: (圖多)媽媽的客人真的是學長的朋友

標題: #圖 朋友的客人真的是學長的媽媽

2、 發文附圖

X. 內部分工

0316048 蘇炳立：資料前處理、提供主要方向、程式重構、程式結果整理

0316027 郎宇傑： 選用 model 研究、圖表製作

0316327 吳俊賢：報告及 PPT 製作、matlab 模型套用、圖表製作