Social Media Analytics #HW1

**Link Prediction**

M11007601 周沂潔

1. 說明如何執行程式(並附上程式碼檔案)

程式撰寫在**Google Colab**中，

匯入 (1) data\_train\_edge.csv (2) predict.csv (3) ans500\_ground\_truth.csv 後，

可以直接執行，結果會顯示出 predict accuracy，confusion matrix以及匯出一個result file。

1. 簡介你所使用的程式架構及演算法流程(如果有進行前處理也請解釋原因)
2. Strategy to Solve the Link Prediction Problem

Diagram

Description automatically generated with medium confidence

Fig.1 Original Graph

* 原本的Graph的連線狀態，在不影響Graph的connected component數量及原本node數量情況下，將幾條link拔掉，作為training的data。

A picture containing text, clock, vector graphics

Description automatically generated

Fig.2 Graph after remove some links

* 將移除的link與原本就沒有連線的link結合，並加上label。
* 0 unconnected
* 1 connected

Table

Description automatically generated

Fig.3 Features and Labels

會發現label的數量差異很大，而在實際情況也是unconnected node pairs會佔多數。

1. 程式架構及演算法流程

Import libraries

Text

Description automatically generated

讀取資料集

Text

Description automatically generated

將沒有連線到的點記錄下來

Text

Description automatically generated

建立adjacency list

Text

Description automatically generated

建立Graph以利後續計算

Text

Description automatically generated

依adjacency list來找出所有unconnected node pairs

Text

Description automatically generated

將所有unconnected node pair存入data中，並加上label 0。

Text

Description automatically generated

取出所有不影響Graph中number of connected components 和initial node count的link。

Text

Description automatically generated

將剛剛移除的link加入剛剛的data。

Text

Description automatically generated

並建立移除後的新的graph\_new。

Text

Description automatically generated

使用node2vec建立出graph\_new的feature。

* num\_walks : 為Graph中每一個node來走多少random walk的次數
* walk\_length : 為每一個random walk走的長度

Graphical user interface, text

Description automatically generated

依feature建立出training和testing的資料集。

Graphical user interface, text

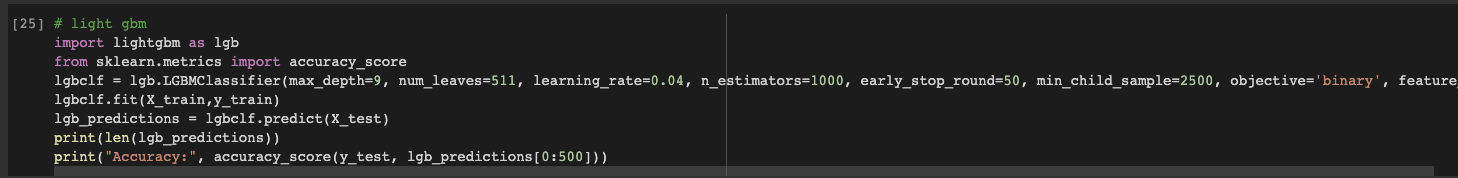
Description automatically generated

做特徵縮放，將標準差調為1平均調為0

Text

Description automatically generated

使用light gbm



* learning\_rate : 學習速率，選擇小一點的學習速率會獲得較穩定的模型
* n\_estimators : boosting的迭代次數，選擇較大的值配合early\_stopping\_round來讓模型自動選擇最好的迭代次數
* mean\_child\_sample : 根據資料量來決定，資料量大時提昇讓葉子結點分不穩定
* max\_depth : 模型的最大深度
* num\_leaves : 一顆樹上的葉子節點，一般設置為 < 2 ^ max\_depth - 1

顯示出Confusion Matrix，查看前五百筆資料預測結果。

Text

Description automatically generated

匯出預測結果，會在資料夾中產生新的csv檔。

Text

Description automatically generated