Practice of Social Media Analytics

HW1 – Link Prediction

執行程式方法:

1. 使用 jupyter(此檔案為 ipynb)
2. 安裝套件 numpy、pandas
3. 執行 Preprocessing 下的每個 code block
4. 可分別執行 SVM、RandomForest、xgboost 下的 code block
   - 分數最高為: RandomForest
5. 產生 csv 檔再放到 kaggle 即可

演算法流程:

1. 將訓練資料及測試資料讀進來
2. 將訓練集中要預測的 label 刪掉(drop)並將 label 存至變數 y
3. 將訓練資料分為訓練集以及驗證集
4. 測試資料集只留下特徵: node1、node2
5. 分別將訓練資料丟入 SVM、RandomForest 以及 Xgboost 去訓練
6. 使用測試資料去預測最後結果
7. 將預測結果與 sample_submit.csv 合併產生完整可交至 Kaggle 的檔案

預測結果之 public leaderboard 分數(Random Forest)

RandomForest

- 一種集成式學習(Ensemble Learning)的方法
- 由多棵決策樹(decision tree)組成
- 方法
  - 從資料集中挑 n 筆資料
  - N 筆資料中隨機挑選 k 個做為樣本
  - 重複 m 次，產生 m 個決策樹
  - 分類，以多數決制進行預測