Hw1-1 report

陳偉 R06944043

Private Score | Public Score： 0.67554 | 0.67556

I. Preprocessing & negative sampling:
- validation 資料來自 train 10%，validation 部分的 vertex 不出現在 training 中
- negative labeled data 取 random walk step = 2 不構成邊的點

Final Features：
- 採用hidden-size = 128， walk-length = 40， window-size = 15 的 deep walk 得到 embedding ， grid search 得知 walk-length 和 window-size 開的越大，學到的 embedding performance 最好，究其原因，是因为增加了 training data 的数量。因開越大花費時間越多，因此選擇了 default parameter 跑 deep walk。
- 兩個 vertex neighbor 的 Intersection *2, union*2 (因 test edge 只釋出一半), jaccard

Classifier：

cosine similarity
- source 的 embedding 和 target 的 embedding 算 cosine similarity，得到的結果取中位數作為 threshold。
- 相較 discriminant classifier 如 SVM 會獲得更好的結果，因為 testing 和 training distribution 不一致，test vertex 的 的 edge 數只釋出一半，classifier 在 training 上得到對 feature 的 coefficient 在 test 上會有bias。事實上，SVM 得到的 testing prediction 對 1/0 的分佈也遠離 50%。

Experiment：

尝试了 MF（0.50225）, GCN（0.67402）, DeepWalk（0.67554）train embedding。
- MF，因一開始錯誤把非 1 的地方全設為 0，導致 MF train 不起來，得到的結果為 random，究其原因，還是對於 MF 缺少深入認知。
- GCN，input 有： adjacent list、feature matrix 為對角線為 1 的 identity matrix、degree 相等的 vertexes 擁有相同label 作為 label matrix (test vertexes 的 degree *2)。
- DeepWalk，hidden-size = 128， walk-length = 40， window-size = 15。