【Practice of Social Media Analytics】

HW1 – Link Prediction

<Kaggle> https://www.kaggle.com/t/a57615e692e449b49f8400926ba2d71b

說明:

這次作業是要預測**node pair (node1, node2)**有沒有隱藏關聯**(hidden edge)**。會提供大家約24000條的edge，讓大家建立出social network以及訓練資料集。資料集是一個「有向網路」，因此每組node pair代表一條具有方向性的邊，例如(9112，38149, 0)代表節點9112指向節點38149沒有邊，(38751, 38824, 1) 代表節點38751指向節點38824有邊。

此social network中有大約3000條hidden edge，這是你們要預測的關係。可以使用任何的方法預測，直接使用呼叫的function也行，作業要上傳到Kaggle評分，寫完後要繳交程式檔和說明文件到Moodle。說明文件中請簡述演算法流程，並說明如何執行你的程式。

作業成績評分方式:60%是performance(Kaggle分數)，40%是report

new_train_data.csv 是拿來training的資料

node1	node2	label
9112	38149	0
38751	38824	1
23013	7184	0
38000	38145	1
37109	8452	0
18794	22228	0
38041	38100	1

new_test_data.csv 是要predict是否正確的node pair (node1, node2)

node_pair_	node1	node2
0	40963	40966
1	4544	49357
2	38726	38760
3	4636	38678
4	26789	5842
5	38192	38180
6	38628	38631
7	38736	38670

sample_submit.csv 是上傳到Kaggle的格式

node_pair_id	ans
0	
1	
2	
3	
4	
5	
6	
7	
8	

node_pair_id是單純的index，從0開始依序排到5999，ans就是預測結果0或1。(0表示你預測此node pair沒有hidden edge，1表示你預測此node pair有hidden edge)

Description:

This assignment is to **predict whether a node pair (node1, node2) has hidden relation (i.e., hidden edge)**. There are about 24,000 edges provided for you to reconstruct the social network and the training dataset. (This is a directed network, so each node pair represents a directed edge. E.g., (9112, 38149, 0) represents node 9112 to node 38149 have no edge, while (38751, 38824, 1) represents an edge from node 38751 to node 38824 .

The social network has about 3,000 hidden edges. These are the relationships you are asked to predict. You can use any prediction method, and you can use any functions/libraries/packages directly. The result should be uploaded to the Kaggle platform for evaluation. You also need to upload the program files and the report document to Moodle. In your report, please briefly describe the algorithm you use, and provide instructions about how to execute your program.

Homework scoring method: 60% is performance (Kaggle score), 40% is report

new_train_data.csv is the information used for training

node1	node2	label
9112	38149	0
38751	38824	1
23013	7184	0
38000	38145	1
37109	8452	0
18794	22228	0
38041	38100	1

new_test_data.csv is the node pair (node1, node2) to be predicted

node_pair_	node1	node2
0	40963	40966
1	4544	49357
2	38726	38760
3	4636	38678
4	26789	5842
5	38192	38180
6	38628	38631
7	38736	38670

sample_submit.csv is the format uploaded to Kaggle

node_pair_id	ans
0	
1	
2	
3	
4	
5	
6	
7	
8	

node_pair_id is index, starting from 0 to 5999, ans is the prediction result 0 or 1. (0 means there is no hidden edge for this node pair; otherwise the ans is 1.)