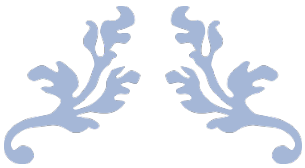

HOMEWORK 4

Jie Tang

NOVEMBER 5, 2020

1. Write a program to solve the node classification problem by Node2vec, and test it on Cora. Report your parameter settings of Node2vec, and the accuracy of your prediction. You can use scikit-learn package for classification.

The Node2Vec algorithm

The Node2Vec algorithm introduced in [1] is a 2-step representation learning algorithm. The two steps are:

1. Use second-order random walks to generate sentences from a graph. A sentence is a list of node ids. The set of all sentences makes a corpus.

2. The corpus is then used to learn an embedding vector for each node in the graph. Each node id is considered a unique word/token in a dictionary that has size equal to the number of nodes in the graph. The Word2Vec algorithm [2], is used for calculating the embedding vectors.

For the parameter settings, I tried multiple combinations, (p = 1, q = 1, num_walk=10, walk_length = 40), (p = 0.5, q = 2, num_walk=10, walk_length = 40), (p = 2, q = 0.5, num_walk=10, walk_length = 40). And then I used logistic regression to evaluate it. The final results are shown below:

p	q	num_walk	walk_length	accuracy	time
1	1	10	40	0.756	10.6s
0.5	2	10	40	0.716	10.9s
2	0.5	10	40	0.766	11.5s

From the above result, we can know that DFS perform better than BFS in our case, even though it takes more time.