Homework 3 for "Machine Learning"

Instructor: Prof. Jie Tang
Course ID. #80245013

The most important feature/standard of a high-qualified research paper is about its "repeatability". In this task, you should re-implement several network embedding algorithms. Network representation learning (NRL) is learning vector representation for nodes in networks. It is also called network embedding and graph embedding.

You should implement at least four network embedding algorithms. You are required to implement DeepWalk[1] and LINE[2]. They are two representative NRL algorithms and advance the research of NRL successfully. You are free to choose the other two algorithms. You can refer to some useful resources such as [3], but your choices are not limited to them.

Dataset: Cora (citation dataset) and Tencent Weibo (following network)
- The two datasets and data description can be found here.

- Cora is used for semi-supervised learning on graphs. Specifically, you will do multi-class classification for research papers. We also provide a script (data_utils_cora.py) to specify how to split dataset to train, validate and test. You should use the same data partitioning method to evaluate your codes. For DeepWalk and LINE, you should use logistic regression as classifier.

- Dataset for Tencent Weibo is used for unsupervised learning on graphs. Specifically, you will do link prediction task for nodes in the network. That is to say, you should infer whether there exists an edge between two given nodes. We have already divided this dataset into several parts for you to train, validate and test. You should use the same divided datasets to evaluate your codes. You are recommended to use cosine similarity or inner product to measure node similarity. Then you can predict links based on node similarity (You don't need to apply other link prediction algorithms).

Evaluation:
- Cora: you should evaluate classification accuracy on test data.
- Tencent Weibo: you should obtain AUC score on test edges (including positive and negative edges). Since we evaluate AUC score, validation set is not used. You can train embeddings on training set and adjust threshold for evaluation on test set.

NOTE1: Word2vec, Tensorflow, PyTorch and other deep learning frameworks can be used, but you can't call off-the-shelf packages of these algorithms directly.

NOTE2: Each dataset should be used in at least one of your algorithms. You are not required to test each of the four algorithms on both datasets.

Submission:
- You should submit all your codes here. Compress your codes in a file in zip format and the file is named <your_student_id>_hw3.zip, such as 2017xxxxxx_hw3.zip.
- You should specify how to run codes of each algorithm.

- Please fill in Google sheet with your results. The links are Cora, Tencent Weibo.
- You DO NOT need to submit this task on learn.tsinghua.edu.cn.

References:

[1] DeepWalk: Online Learning of Social Representations. Bryan Perozzi, Rami Al-Rfou, Steven Skiena. KDD 2014.
[2] LINE: Large-scale Information Network Embedding. Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, Qiaozhu Me. WWW 2015.
[3] Must-read papers on network representation learning: https://github.com/thunlp/NRLPapers