AI3602 实验二 链路预测算法

实验介绍

大家好，我们的第二次实验作业是Link Prediction（链路预测算法）。本次实验的数据集共包含16863个节点和46116条边，你可以自行划分训练集和验证集来学习一个模型，使得该模型能够预测出一对给定节点间存在连边的概率。训练后的模型会被用由从原网络中获取的点对组成的测试集来进行测试，你需要给出测试集中每个给定的点对在原网络中相连的概率p（$\mathrm{p} \in (0,1)$，小数点后保留四位）。

格式要求

1．数据集共包含两个 .csv文件：1) lab2_edge.csv 是网络的边信息；2) lab2_test.csv 是测试集信息，共包含10246个点对。你可以用node2vec或者deepwalk的方法进行实现。
2．你需要把你的预测结果写入prediction.csv文件中，共分两列，第一列是id，第二列是预测的概率（第一行不需要写列名，直接写第一个结果）。测试集中的每个点对都有一个相应的id，你的预测结果也要按照相应的id顺序进行呈现。比如点对(0, 100, 400)，如果你的模型预测节点100和节点400之间存在连边的概率为0.8000，那么你上传的结果文件中对应行应该是0，0.8000。
3．我们会用AUC指标来衡量你的预测结果：

$$\mathrm{AUC(f)} = \frac{\sum_{t_0 \in D^0} \sum_{t_1 \in D^1} 1[f(t_0) < f(t_1)]}{|D^0| \cdot |D^1|},$$

其中$\mathrm{f(t)}$ 是点对t存在连边的概率；$1[\cdot]$是指示函数；D^0是标签为0的集合；D^1是标签为1的集合。

环境要求

请使用Python 3.8来撰写代码。

上传要求

1．你上传的文件需要被命名为“学号_lab2.zip”，里面包含一个同名的文件夹。
2．同名文件夹中需要包含

- “src”文件夹（放代码）
- “data”文件夹（放数据）
- Prediction.csv（你的预测结果）
- Readme.txt（介绍如何使用你的代码进行复现）
- Requirements.txt（可选，如果有一些需要的特殊包及版本号需要在此说明，如果没有就不需要）

3 “/src/main.py”

．复现需要的主代码请放到　　中。
4．请务必在作业截至日期(**2021.12.8** 晚上**12**：**00**)前将zip文件上传到oc.sjtu.edu.cn上面，延期提交会适当扣分。
5. 如需使用GPU资源，请参考 https://notes.sjtu.edu.cn/s/KEbYCpbce 。

注意事项

为了保证公平性，我们做了以下几点要求，任何违反要求的行为都会被扣分。
1．不能使用任何封装好的API（比如Networkx中的Link Prediction的函数）；
2．不能抄袭别人的代码（我们会对所有作业进行查重）；
3．不能使用其他的数据集来训练你的模型；
4．不能使用预训练模型；
5．请按照环境要求进行实验（由于环境问题导致的复现失败后果自负）。