《人工智能算法与系统》第二次作业说明

2022 年 10 月 10 日

一、数据来源

反欺诈是金融行业永恒的主题，在互联网金融信贷业务中，数字金融反欺诈技术已经得到广泛应用并取得良好效果，这其中包括了近几年迅速发展并在各个领域得到越来越广泛应用的图神经网络。本项目基于的数据集 DGraph，是大规模动态图数据集的集合，由真实金融场景中随着时间演变事件和标签构成。数据集详细信息可见官网（https://dgraph.xinye.com/introduction）。为了更好进行节点异常检测任务，我们需要利用图结构的信息。同学可以应用授课中学习到的图神经网络以更好的捕捉网络结构信息来获得更好的表现。

二、评分说明

我们提供的数据集包含三个，一个是同学在项目中看到的训练集，另一个是平台上提交测试时用到的测试集，最后一个是用于评分的评分数据集（仅用于最终评分）。这三个数据集是随机拆解而来。

平台上显示的分数是同学们所提交模型在测试集上运行结果 **ROC 乘以 100 的数值**，并不是本次作业的分数。

本次作业的具体赋分规则如下（**按十分制记**）：

1. 提交模型且成功运行，且提交文档完整，即可获得 6 分基础分。
2. **ROC 大于 0.72（模型在评分数据集上运行结果），即可再获得 2 分，即总分 8 分。**
3. 在原模型的基础上，尝试不同的模型，尽可能地提高 ROC。**最终 ROC 大于 0.73，即可获得 9 分；最终 ROC 大于 0.74，即可获得 10 分。**

三、作业时间

作业截止时间为 11 月 14 日晚 12 点

四、常见问题

问题 1：模型训练时显存溢出/不够用怎么办？

答：可以把中间层维度降低，或者更换网络类型等方法。

问题 2：图神经网络的测试时间太长，没有结果怎么办？测试时内存溢出怎么办？

答：系统测试是在 CPU 上进行的，因此计算的效率会不如 GPU，时间上会比较久。以下方法可供参考减小内存占用并缩短测试时间：

如果大家用图神经网络的话，经常跑一次正向就要喂全部数据，跑出所有数据的预测标签，如果测试时每个数据都跑一遍正向，当然会慢。实际上只需要第一次跑一遍正向就可以了，然后保存下来数据（全局变量、文件啥的），之后直接根据 node_idx 查找结果就可以了，不会影响性能的。测试的时候不需要反向传播，参数不会改变的。

另外，测试集合的节点 node_idx 同学们其实是看得到的，只是 label 缺失，同学可以在 GPU 上把这些没有 label 的节点预测出来，然后保存下来当去做测试。

问题 3：Python 环境包的版本由于自己修改，导致无法运行怎么办？

答：可以点击项目右上方的重启按钮，重置环境，若仍然没有解决可以联系助教。

问题 4：报告一定要按照模板写吗？

答：不是的，模板只是用作参考。

问题 5：提交测试后如何停止测试？

答：测试弹窗不需要开在那里等待，结果可以关闭后在测试记录里查看，如果想要停止测试，可以关闭测试弹窗后再次点击“重新测试”按钮，出来的弹窗中会出现“取消测试”的按钮。