《复杂网络分析》课程作业

课程编号：U10M11181

课程类别：专业选修课程

学时/学分：32 学时/2 学分

一、作业目的

通过本课程作业加深对复杂网络基础理论的认识和了解，锻炼定量、定性分析复杂拓扑结构特征，并解决问题的能力。通过对复杂网络进行建模、分析、验证，基本掌握复杂网络数据挖掘的基本过程，为后续学习大数据挖掘、人工智能等奠定基础。

二、作业题目

- 作业题目：癌症基因识别

  本题目要求实现一个癌症基因识别算法。人类基因组有数万个基因，部分基因和癌症发生发展密切相关，请实现一个二分类算法，挖掘癌症相关基因，并在标注数据中进行交叉验证，使用具体的分类指标说明方法性能。

三、作业说明及要求

- 数据集：数据来源于 EMOGI(https://github.com/schulter/EMOGI)，包含 13627 个基因的组学特征向量（64 维），以及 CPDB 的 PPI 网络。
- 样本分布：数据集中包含 796 个正样本（癌症基因，对应 label 值为 1）、2187 个负样本（非癌症基因，对应 label 值为 0）以及 10644 个未知样本（未知基因，对应 label 值为-1）
- demo 说明：

  project/demo: 是个二分类 demo 算法，随机预测样本的类别（划分了训练集、

测试集）。

project/features.csv: 13627 个基因的组学特征。

project/labels.csv: 13627 个基因的标签，1 表示正样本（癌症基因），0 表示负样本（非癌症基因），-1 表示未知样本（未知基因）。

project/adj.csv: 13627 个基因的对应的 PPI 网络，使用邻接矩阵存放。

运行方式：安装好必要的包后，在 project 目录中运行`python src/main.py`

其他文件：“.joblib”后缀的文件可以使用 joblib 库（版本号>=1.3.1）的 load 方法直接导入数据，导入数据集时，选择 csv 文件和 joblib 文件均可。

- 基本要求：基于网络数据，实现一个分类算法，能够将基因分为癌症基因和非癌症基因两个类别，可以参考 python 实现的 demo 框架，可以使用其他编程语言实现，不限制采用的方法。请在标签数据（标签为 1 和 0）上进行训练并测试模型性能（注意将数据集进行划分，切忌出现标签泄漏），请使用 AUPRC、ROCAUC 以及其他常见的分类指标在测试集上测试，量化方法性能。要求使用 5 次 5 折交叉验证给出分类指标的平均结果（demo 中有交叉验证的示例模块）。如果实验过程中尝试了多种方法，请在实验报告中说明，比较不同方法的性能差异。

  实验相关的代码需要提交，请在代码的根目录中加入说明文件，说明如何运行代码。实验完成后，按照模版完成实验报告，并将代码和实验报告打包后提交。
- 提示：建议充分结合多组学数据，只利用单一网络拓扑信息可能效果不佳。
- 提交方式及截止时间

  提交方式：将代码和实验报告打包为 zip、tar.gz、rar 等常见格式的压缩包，发送到邮箱`cnanwpu@163.com`，文件命名方式为`**学号+姓名+复杂网络分析大作业**`。

  截止时间为 2024 年 7 月 4 日 23:59