Task3

写在前面：这次的任务量略高，算法本身不难理解，相信我这绝对不是“大跃进”。

1 学习数据结构树的相关知识

你需要掌握：

- 什么是二叉树、森林？数据结构如何定义？
- 树的前序遍历、中序遍历、后序遍历、层序遍历。
- 根据有序序列建立一棵合理的儿茶排序树(BST)并实现中序遍历。

2 学习机器学习的评价指标

- ROC、AUC；
- Recall、Recall@K、Precision、Precision@K。

前两个任务时间4天。

！！！务必确保在完成以上任务后继续。

3 阅读论文

论文: **Isolation Forest**

这是由南京大学周志华教授团队LAMDA提出的一种异常检测算法，发表在2008年的数据挖掘国际顶级会议ICDM'08上。该算法以simple but work著称。在机器学习风起云涌的今天，虽然有非常多的异常检测相关的论文在国际顶会/顶刊上发表，但是Isolation Forest算法仍然以强大的鲁棒性和低时间复杂度而在当今的工业界广泛使用。

大家阅读论文的时候不要感到畏惧，要知道这个算法真的是very simple但是又very work。在现在的这个阶段，对论文的公式要不求甚解，主要体会论文的motivation、背后的思想。刚开始的时候没有必要花大量的时间在看论文的实验上。**在阅读完论文后，应该要能回答以下问题：**

- 算法的输入输出是什么？是向量？是标量？是分数？是相似度？表示什么？等等。
- 该算法的流程是怎么样的?
- 算法的鲁棒性、收敛性如何保证?
- 什么样的样本是异常样本?
- 异常检测难在哪里？异常检测本身也是一个二分类问题，和普通的二分类问题有什么区别?

在实在看不懂论文的情况下，借助知乎等平台，看看人家的论文阅读笔记。

第3项任务5天。

4 实验

根据算法，对论文进行复现，可选用数据集：论文中Table2的数据集任选进行复现，注意论文中关于树棵数、树深度等参数的设置。调整这些参数，看看对效果和运行时间的影响，用Recall@K、Precision@K、AUC值衡量，记录成表格，并绘制ROC曲线。

实验技巧：自己先生成一组小的数据集测试一下有没有写错然后再拿去跑实验。Github上有一些代码可以适度参考。

第4项任务5天。

希望大家完成后会对这个算法产生惊叹！并增加今后阅读顶会论文、复现顶会论文的信心，顶会并不是高不可攀。