**随机森林 实验报告**

姓名：曹蕊

学号：15331009

班级：15软工计应

一、实验要求

针对kaggle第二次实验，选择逻辑回归、随机森林、GBDT其中一个方法加以实现，内容着重在如何实现并行化，提高运行速度，突现cache友好。进行对比实验，并以直观的方式展示实验结果。

二、实验环境

我选择的题目是随机森林。

1、硬件：个人笔记本，CPU: Intel Core i7-4710MQ四核处理器；8GB内存。

2、操作系统：win10

3、运行环境：python3.6.5

三、实验设计

1、文件结构

文件结构如下图所示，在根目录random\_forest下，有三个文件：data文件夹用以存放数据（数据的路径也可以在myForest.py中进行修改）；myTree.py 是我定义的决策树class；myForest.py是我定义的随机森林class。

random\_forest

|- data

|- train.txt

|- test.txt

|- myTree.py

|- myForest.py

1、数据结构

设样本数为m，特征数为n。在该数据集中，m = 1719692，n = 201。摒弃了定义一个m\*n的二维list存储样本和标签，我定义了一个数据结构。

摒弃了定义一个

2、树的建立（复杂度建立）

3、森林的建立

2、内存友好

3、并行化

四、实验结果

1、建树和森林的效果：

2、内存友好前后对比

3、并行化的速度对比（共享内存）

五、有待改进和提升之处

六、实验总结与心得