**第一次小作业：Random Forest**

在本次作业中，需要手动实现一个简单的Random Forest分类器，并使用给定数据集评估分类器的性能。有关Random Forest的基础知识可参考课上PPT文件。

作业要求：

1. 提交截止日期：2020年12月20日23:00，命名格式：“学号\_姓名\_第一次小作业.zip”，提交至课程中心。
2. 建议语言：Python 3.5或以上版本。
3. 软件包要求：模型训练和优化部分不允许调用现有的软件包，但允许使用numpy/scipy等类似工具包进行数学运算。对于数据读入、划分、预处理等则不做要求。
4. 模型评估：课程提供训练数据集与评价指标。数据集包含特征和标签，可自行使用交叉验证等方法对模型预测性能进行评估与调优，注意防止过拟合。
5. 模型保存：由于作业包含课上测试部分，实验课上将公布测试集进行现场测试，建议实现模型保存和读取部分的代码，并将训练完成的模型参数提前保存，以免实验课上现场训练出现耗时过长的情况。
6. 提交内容：包含两部分，即相关代码文件和说明文档。说明文档需包含两个部分i) 阐明如何通过运行提交的代码文件，完成从数据输入到预测结果输出的过程，并尽量保证结果的可复现性；ii) 对代码中各部分的作用进行简要介绍。

数据集和评价指标

本次作业采用的数据集为某分类数据集，原始数据包含24个变量，变量的组成如下：

* id：数据的唯一识别码，int型
* label：待预测变量（1：正类，-1：负类）
* 22个与调查对象相关的统计特征，命名为x1-x22，对应变量类型为
  + 类别属性：
    - 标称类型Nominal：x1-x22

模型预测性能的评价指标为F1分数。记label为1的样本为正样本，label为-1的样本为负样本，则F1的计算方法如下：

其中，分别表示精确率和召回率；、、、分别对应真正例（预测为正样本且实际为正样本）、假正例（预测为正样本但实际为负样本）、真负例（预测为负样本且实际为负样本）和假负例（预测为负样本但实际为正样本）的预测样例数目。

以下提供了一些相关文献及博客内容，以供参考

1.《决策树（ID3、C4.5、CART、随机森林）》，这是一篇中文博客，比较详细地分析了决策树和随机森林算法的过程。<https://blog.csdn.net/gumpeng/article/details/51397737>

2. <https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf>