作业6：决策树和继承学习

助教邮箱:wanghc17@mails.tsinghua.edu.cn

1 Bagging

在课堂上我们已经了解到，Bagging方法可以减小模型的误差。我们现在从理论上来证明这一点。首先我们在数据集合D中有放回的抽样生成了m个数据集$\{D_i\}_{i=1}^m$，在每个数据集合D_i上训练分类器$h_i(x)$，假设我们现在有n个待预测的新样本$\{x_j\}_{j=1}^n$。对于其中的任意一个样本x，Bagging方法的预测值可以定义为多个分类器的预测平均值：

$$h_B = \frac{1}{m}\sum_{i=1}^{m} h_i(x)$$

若对于样本x真实的预测值为$y(x)$,则可以知道每个分类器$h_i(x)$的误差ϵ为:

$$\epsilon_i(x) = h_i(x) - y(x)$$

对于m个单独的分类器，它们的平均均方误差可以定义为：

$$E_h = \frac{1}{m}\sum_{i=1}^{m}\{\frac{1}{n}\sum_{j=1}^{n}[\epsilon_i(x_j)]^2\}$$

对于Bagging分类器的均方误差可以定义为：

$$E_{h_B} = \frac{1}{n}\sum_{j=1}^{n}[\epsilon_B(x_j)]^2 = \frac{1}{n}\sum_{j=1}^{n}[h_B(x_j) - y(x_j)]^2$$

(1) 假设所有分类器的误差均值为零，而且互不相关，即：

$$\frac{1}{n}\sum_{j=1}^{n}\epsilon_i(x_j) = 0.(i \in \{1, 2, ..., m\})$$

$$\frac{1}{n}\sum_{j=1}^{n}\epsilon_i(x_j)\epsilon_k(x_j) = 0.(i, k \in \{1, 2, ..., m\})$$

请证明：

$$E_{h_B} = \frac{1}{m}E_h$$

(2)但在实际情况中，往往它们的误差是高度相关的，请在(1)条件不满足的情况下证明：

$$E_{h_B} \leq E_h$$

2 决策树

实现决策树算法，并且在Sogou Corpus数据集上测试它的效果。

（注：附件中，*Sogou−webpage.mat* 存储有wordMat和doclabel两个变量。前者为特征矩阵，大小为14400 * 1200，即包含14400个数据，每行数据包含1200维特征；后者为14400个数据的标签。可以使用predeal.py完成数据载入)

要求：

1. 请自己编写一种决策树算法。

2. 将数据**随机**分为3：1：1的三份，分别为训练集、交叉验证集、测试集。请在训练集上训练，交叉验证机上选择超参数，并在测试集上给出测试效果。因此，需在报告中给出超参数的选择，以及不同超参数下，训练集、交叉验证集的分类正确率，给出最好的超参数设置，并在测试集上给出测试效果。

3. 请在编写程序时，必须包含但不限于以下的几个函数：

(1) **GenerateTree(args)**:

#生成树的总代码，args为各种超参数，请自由选择各类影响树性能的超参数。

(2) **SplitNode(samlesUnderThisNode,thre,...)**:

#对当前节点进行分支，**samlesUnderThisNode**是当前节点下的样本，**thre**是停止分支的阈值，停止分支的条件应在实验报告中说明。

(3) **SelectFeature(samlesUnderThisNode,...)**:

#对当前节点下的样本，选择待分特征。

(4) **Impurity(samples)**:

#给出样本**samples**的不纯度，请在实验报告中说明采用的不纯度度量。

(5) **Decision(GeneratedTree, SamplesToBePredicted)**:

#使用生成的树**GeneratedTree**，对样本**SamplesToBePredicted**进行预测。

4. 请同学们尝试使用sklearn中的DecisionTreeClassifier与RandomForestClassifier函数，应用于该数据集中，将自己编写的决策树与这两种方法的测试集正确率进行对比，并做简要分析。

5. 有兴趣的同学，可以对树进行剪枝操作，实现一种剪枝方法，提升树的分类能力。（此项不做要求）