Desicion Tree 作业

Problem 1

Consider a data set comprising 400 data points from class C1 and 400 data points from class C2. Suppose that a tree model A splits these into (300, 100) at the first leaf node and (100, 300) at the second leaf node, where (n, m) denotes that n points are assigned to C1 and m points are assigned to C2. Similarly, suppose that a second tree model B splits them into (200, 400) and (200, 0).
Evaluate the misclassification rates for the two trees and hence show that they are equal.
Similarly, evaluate the cross-entropy and Gini index for the two trees and show that they are both lower for tree B than for tree A.

Programming

实现决策树算法，并且在Sogou Corpus数据集上测试它的效果（数据集详情见readme）。
要求：
1. 不能调用已有的机器学习包
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，交叉验证集上选择超参数，用选出的最好模型在测试集上给出测试结果。因此，在报告中说明算法的超参数有哪些，在不同的超参数设置下，训练集和交叉验证集的分类正确率，最好模型的超参数设置，以及最后的测试正确率。
3. 请结构化代码，必须包含但不限于如下几个函数（请从代码中分离出来，有明确的这几个函数，函数参数可以有所变化）：

main()
 # 要求main函数在运行中，逐个测试不同的超参数，然后打印出每个超参数的设置，该设置下的训练、验证正确率（就是上面第二点提到的要出现在报告中的结果）。
GenerateTree(args)
生成树的总代码，args为各种超参数，包括但不限于下面的thres，或者其他会影响树性能的超参数，自由发挥。
SplitNode(samplesUnderThisNode, thresh, ...)
对当前节点进行分支，samplesUnderThisNode是当前节点下的样本，thresh是停止分支的阈值，停止分支的条件请在实验报告中说明。
SelectFeature(samplesUnderThisNode, ...)
对当前节点下的样本，选择待分特征。
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。
Decision(GeneratedTree, XToBePredicted)
使用生成的树GeneratedTree，对样本XToBePredicted进行预测。
Prune(GeneratedTree, CrossValidationDataset, ...)
对生长好的树GeneratedTree（已经经过stopped splitting）进行剪枝：考虑所有相邻的叶子节点，如果将他们消去可以增加验证集上的正确率，则减去两叶子节点，将他们的共同祖先作为新的叶子节点。或者实现其他的剪枝方法，如有，请在实验报告中说明。