

机器学习实验报告

ID3/C4.5实现决策树

姓 名： 倪长好

班 级： 硕士班

学 号： SA19219093

指导老师： 何劲松

1. 实验目的

理解并编程实现ID3算法实现决策树分类，可以通过使用训练好的决策树对输入数据进行分类。

1. 实验原理与设计

1、实验原理

　ID3决策树是一种非常重要的用来可以处理分类问题的结构，通过把实例从根节点排列到某个叶子节点来分类实例，叶子节点即为实例所属的分类。树上的每一个节点说明了对实例的某个属性的测试，并且该节点的每一个后继分支对应于该属性的一个可能。

构造好的决策树的关键在于如何选择好的逻辑判断和属性。对于同一组例子，可以有很多组决策树符合这组例子。人们研究出，一般情况下或具有较大概率地说，树越小则预测能力越强。要构造尽可能小的决策树，关键在于选择恰当的属性或逻辑判断。由于构造最小的树是NP问题，因此只能采用启发式策略选择好的逻辑判断或属性。用信息增益度量期望熵最低，来选择属性。公式为：

2、实验设计

1. 对当前例子集合，计算各属性的信息增益;
2. 选择信息增益最大的属性;
3. 把在处取值相同的例子归于同一子集，取几个值就得几个子集;
4. 对既含正例又含反例的子集，递归调用建树算法；
5. 若子集仅含正例或反例，对应分支标上P或N，返回调用处。
6. ID3实验主要代码

实验环境：Windows10操作系统，Python3

1. 数据集：训练集来自dna.data，测试集来自dna.test。由于数据集中代表核苷酸的符号变量（仅A，G，T，C），被3个二进制指标变量代替。 因此原来将60个符号属性更改为180个二进制属性。另外，数据集中的原数据名称已被删除。所以需要对数据进行一系列预处理。

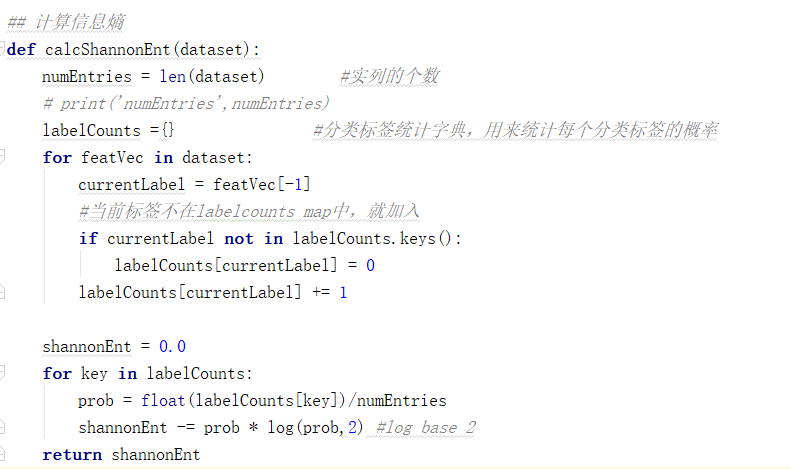


测试集中的数据都是离散的，同时以3个二进制转成一个类别，添加自己设定的标签。

1. 构建决策树

2.1计算香农熵

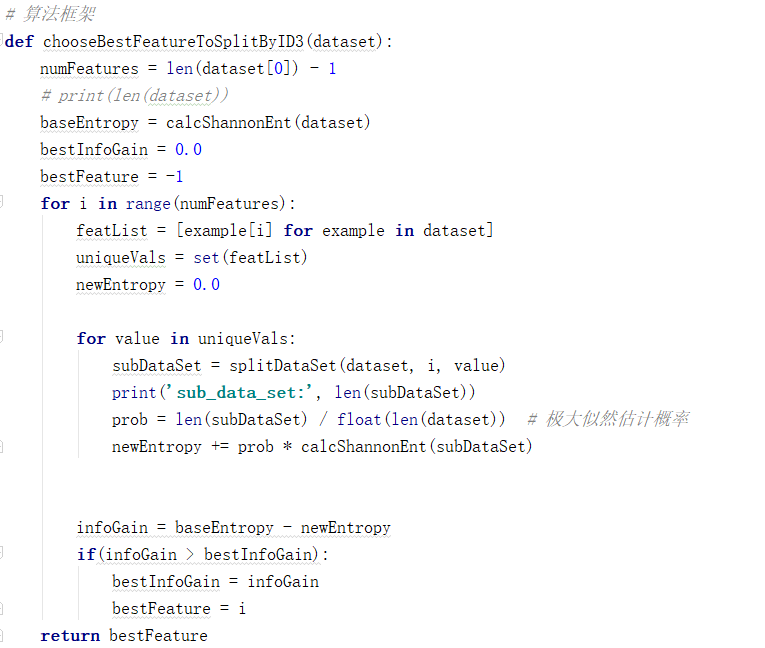
获取数据集样本实例的个数,得到不同标签样本的个数，进行香农熵的计算。

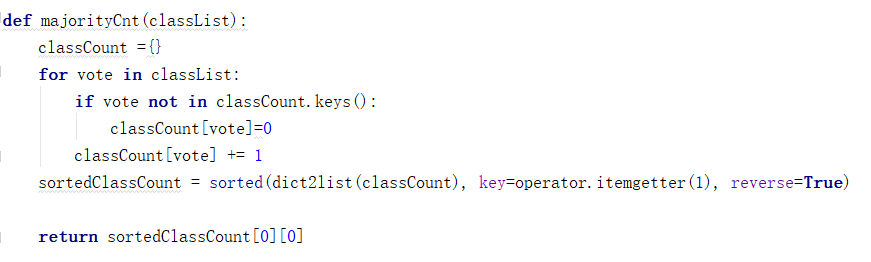


* 1. 对数据进行划分并计算各个属性的信息增益，选出最大的信息增益

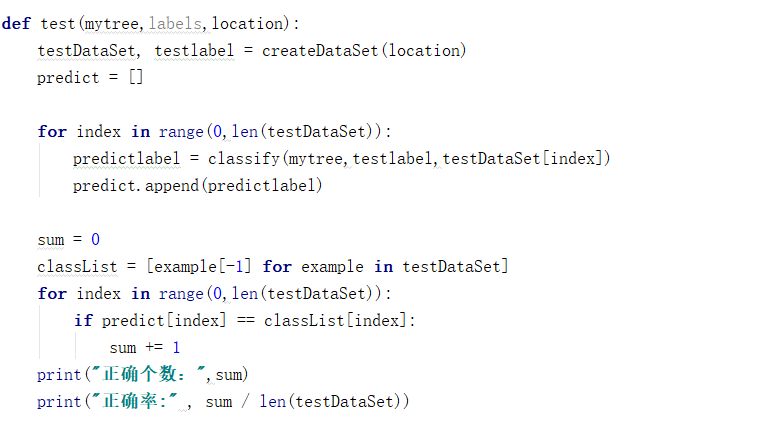
将数据集划分得到某一属性下的样本的值，计算他们的信息熵和信息增益，选出信息增益最大的属性作为下一个节点。





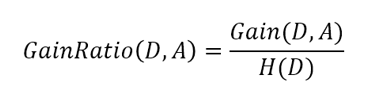


1. 对数据集进行测试



四、C4.5主要代码

C4.5算法与ID3算法很相似，C4.5算法是对ID3算法做了改进，改进主要分为4个方面，第一就是用信息增益率来选择属性，克服了用信息增益选择属性时偏向选择取值多的属性的不足。第二就是在树构造过程中进行剪枝。第三就是能处理非离散的数据。第四就是能处理不完整的数据。定义：特征A对训练数据集D的信息增益比GainRatio(D,A)定义为其信息增益Gain(D,A)与训练数据集D的经验熵H(D)之比：



C4.5算法过程跟ID3算法一样，只是选择特征的方法由信息增益改成信息增益比。

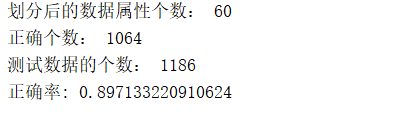


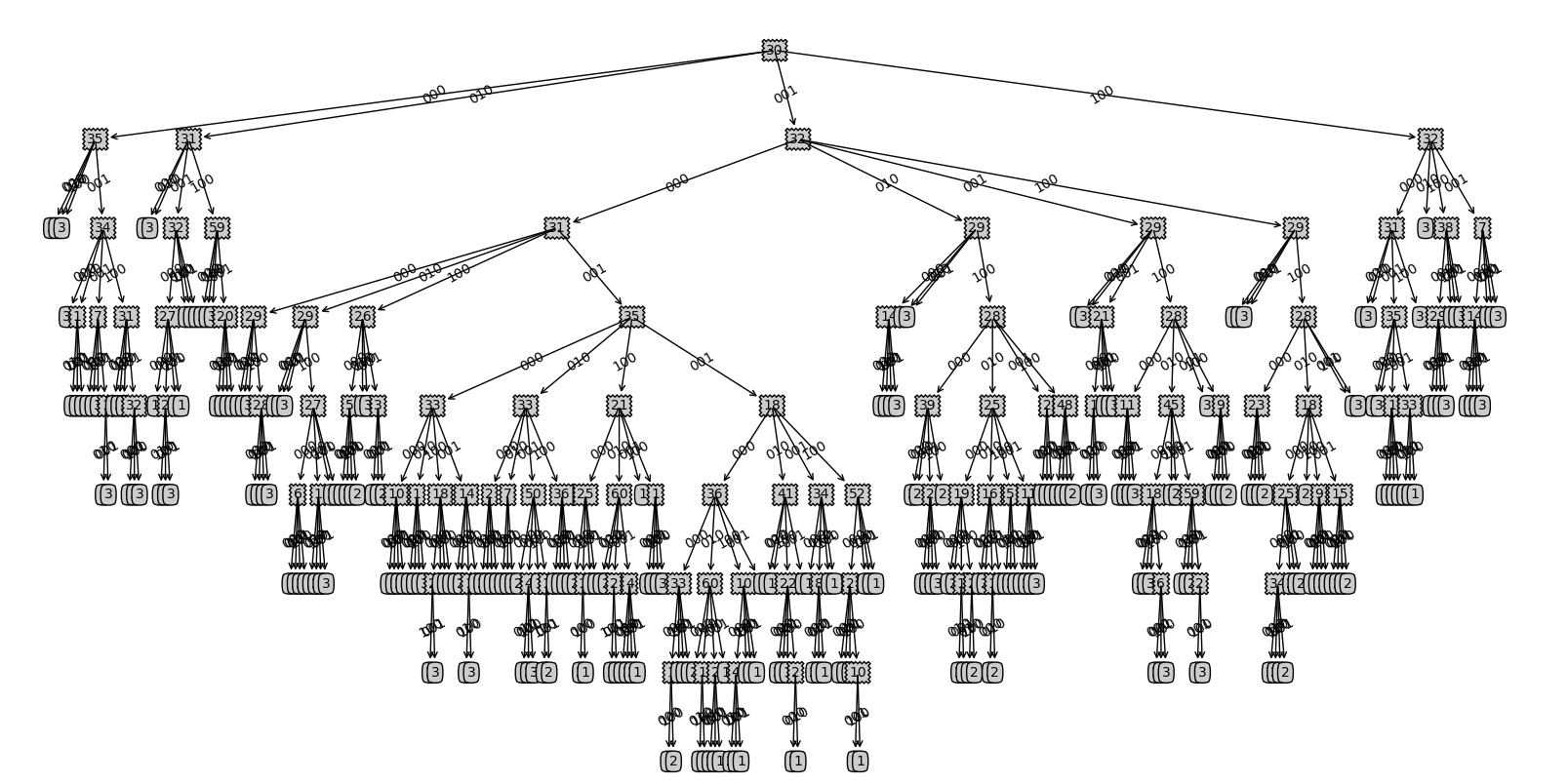
五、实验结果

5.1、ID3实验结果

数据集样本个数共2000个，经过数据的预处理划分为60个属性，构建了一个自己的ID3决策树，测试集数据一共1186个，其中正确分类得有1064个，同样经过数据预处理也划分为60个属性，正确率为0.897.

实验结果如下图所示：





5.1、C4.5实验结果

C4.5采用和ID3同样的训练集和测试集，测试集数据一共1186个，其中正确分类得有1073个，同样经过数据预处理也划分为60个属性，正确率为0.904。

