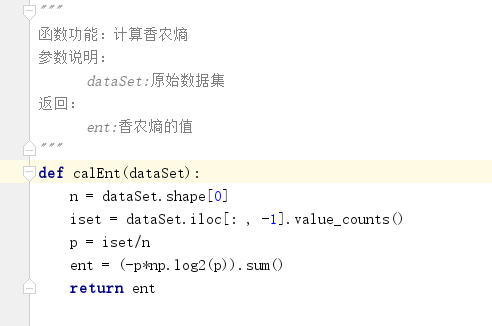
**决策树算法分析与实现**

**一．决策树概念**

1. 通俗来讲，决策树算法就是通过一些feature进行分类，每一个节点提一个问题，通过判断，将数据分为两类，再继续提问。这些问题是根据已有数据学习出来的，再投入新数据的时候，就可以根据这棵树上的问题，将数据划分到合适的叶子上。
2. 决策树是有监督学习中的一种算法，并且是一种基本的分类与回归的方法，也就是说，决策树有两种，一种是回归树，一种是分类树。
3. 首先，我们大致要理解三个概念：1.根节点：没有进边，只有出边；2.中间节点：既有进边也有出边，但进边有且仅有一条，出边也可以有很多条；3.叶节点：只有进边，没有出边，且进边有且只有一条。每个叶节点都是一个类别标签。 4.父节点和子节点：在两个相连的节点中，更靠近根节点的是父节点，另一个则是子节点。两者是相对的。
   1. **决策树的构建准备工作**
4. 特征选择

特征选择就是决定用哪个特征来划分特征空间，其目的在于选取对训练数据具有分类能力的特征。这样可以提高决策树学习的效率。如果利用一个特征进行分类的结果与随机分类的结果没有很大的差别，则称这个特征是没有分类能力的，经验上扔掉这些特征对决策树学习的精度影响不会很大

香农熵的python代码



我们以海洋生物数据为例，构建数据集，并计算其香农熵

|  |  |  |  |
| --- | --- | --- | --- |
| No. | No surfacing | flippers | fish |
| 1 | 1 | 1 | Yes |
| 2 | 1 | 1 | Yes |
| 3 | 1 | 0 | No |
| 4 | 0 | 1 | No |
| 5 | 0 | 1 | no |

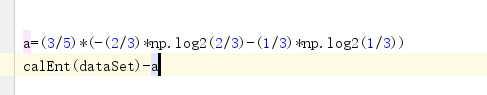


熵越高，信息的不纯度就越高。也就是混合的数据就越多。

1. 信息增益

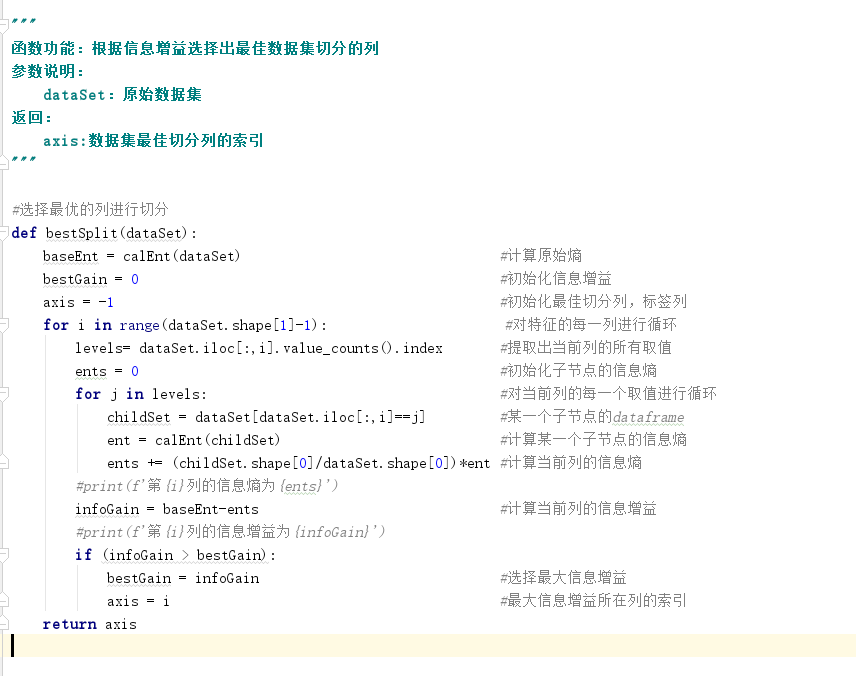
信息增益（Information Gain）的计算公式其实就是父节点的信息熵与其下所有子节点总信息熵之差。但这里要注意的是，此时计算子节点的总信息熵不能简单求和，而要求在求和汇总之前进行修正。

假设离散属性a有V个可能的取值{a1,a2,……,aV}{a1,a2,……,aV} \{a^1,a^2,……,a^V\}{a1,a2,……,aV},若使用a对样本数据集D进行划分，则会产生V个分支节点，其中第v个分支节点包含了D中所有在属性a上取值为avav a^vav的样本，记为DvDv ，D^vDv.我们可根据信息熵的计算公式计算出DvDv D^vDv的信息熵，再考虑到不同的分支节点所包含的样本数不同，给分支节点赋予权重∣Dv∣/∣D∣∣Dv∣/∣D∣ |D^v|/|D|∣Dv∣/∣D∣,这就是所谓的的修正。



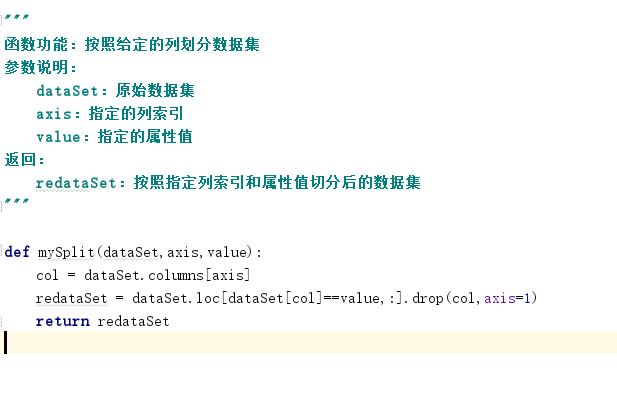
数据集最佳切分函数：

​ 划分数据集的最大准则是选择**最大信息增益**，也就是信息下降最快的方向。



1. **按照给定列切分数据集**

通过最佳切分函数返回最佳切分列的索引，可以根据这个索引，构建一个按照给定列切分数据集的函数



验证函数，以axis=0，value=1为例

mySplit(dataSet,0,1)

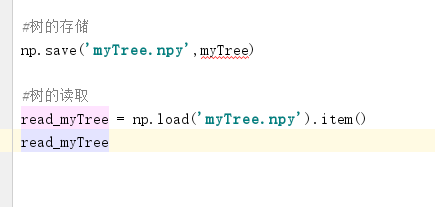
* 1. **递归构建决策树**



查看函数运行结果：myTree = createTree(dataSet)

myTree

* 1. **决策树的存储**

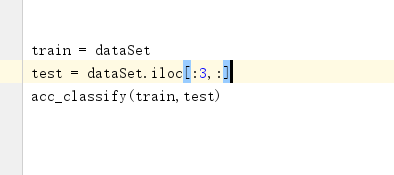


**五．使用决策树执行分类**





测试函数



**使用SKlearn中graphviz包实现决策树的绘制**

