# 机器学习实验报告

**实验一：决策树**

在学习决策树的过程中，我发现决策树根据划分属性的不同可以分为以下三种：ID3，C4.5，CART。他们之间的区别如下：

(1) ID3算法以信息增益为准则来进行选择划分属性，选择信息增益最大的；

(2) C4.5算法先从候选划分属性中找出信息增益高于平均水平的属性，再从中选择增益率最高的；

(3) CART算法使用“基尼指数”来选择划分属性，选择基尼值最小的属性作为划分属性.

为了探究他们的区别，分别使用三种算法对同一个数据集进行测试。

**数据集**：通过四个特征属性：年龄段，工作，有无房子，信贷情况，来判断是否应该给予贷款。为了方便属性的描述，我们对数据进行标注。

年龄：0为青年，1为中年，2为老年

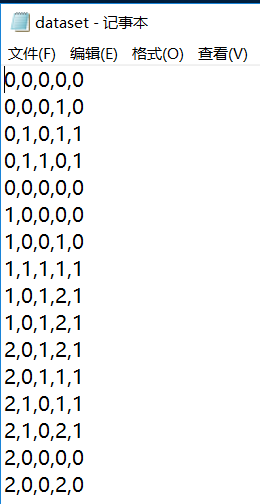
有无工作：0代表无工作，1代表有工作

有无房产：0代表无，1代表有

信贷情况：0代表一般，1代表好，2代表很好

是否给予贷款：0代表不给，1代表给

将他们存入记事本中，如下所示：

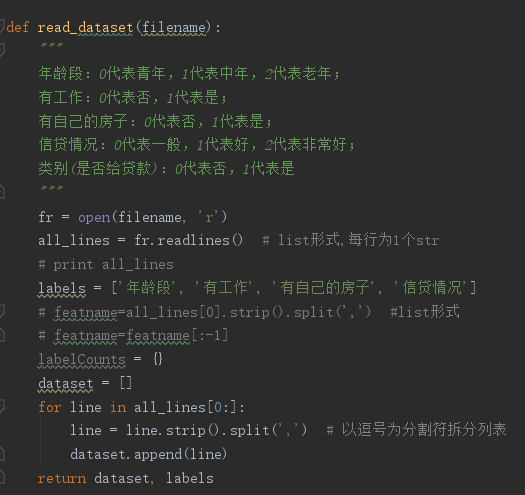


随后分别用上述三种算法对数据集进行决策树的分类。

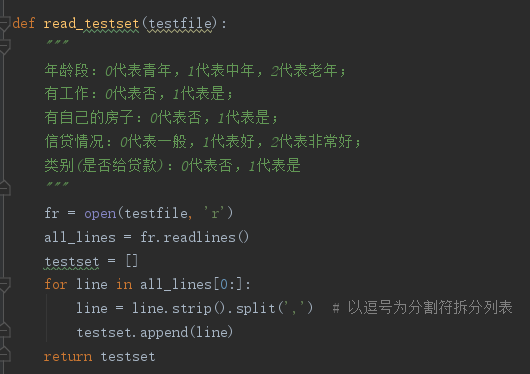
**实验代码**

**Tree.py**

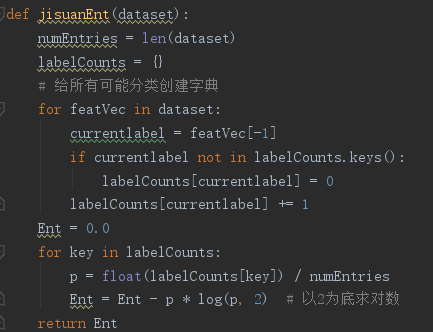
**读取数据集函数**



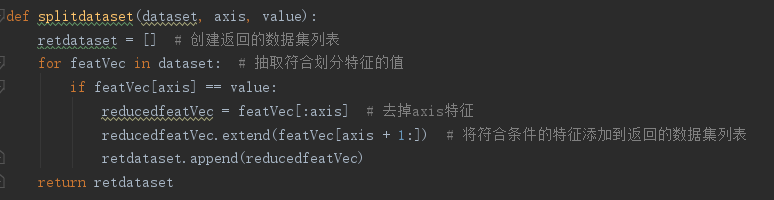
**读取测试集函数**



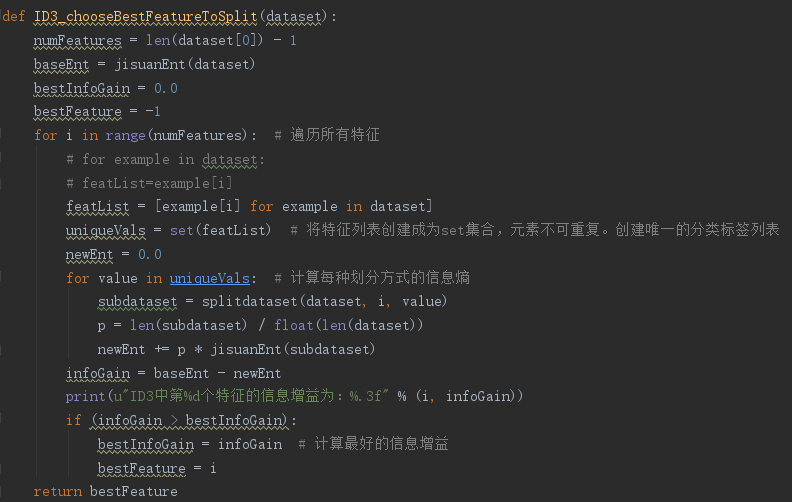
**计算信息熵函数**



**划分数据集函数**



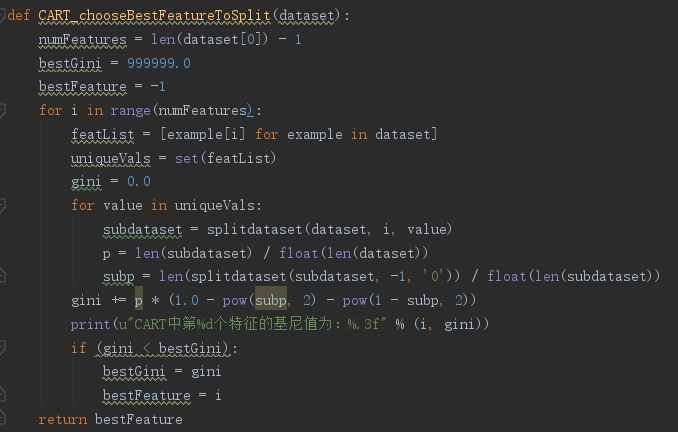
**ID3算法**



**C4.5算法**

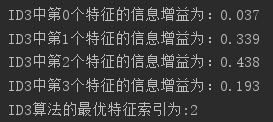


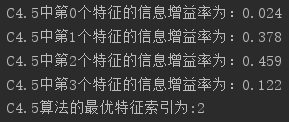
**Cart算法**

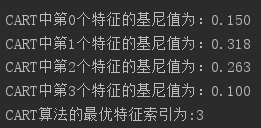


**实验过程：**

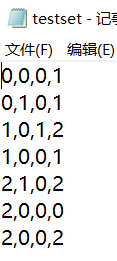
（1）首先读入数据，让三种算法分别计算出第一个最优特征。计算结果如下所示，可以看到ID3和C4.5首个最优特征索引为2，即有无房产；而CART首个最优特征索引为3，即信贷情况。可见这三种算法的区别在第一步就已经够体现出来了。



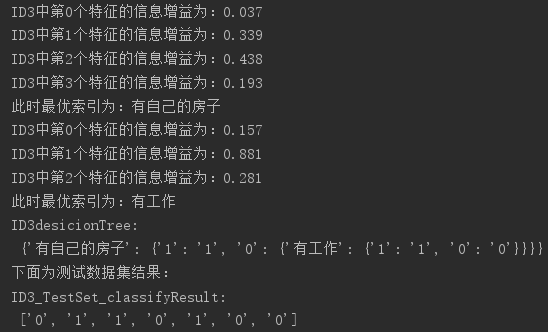




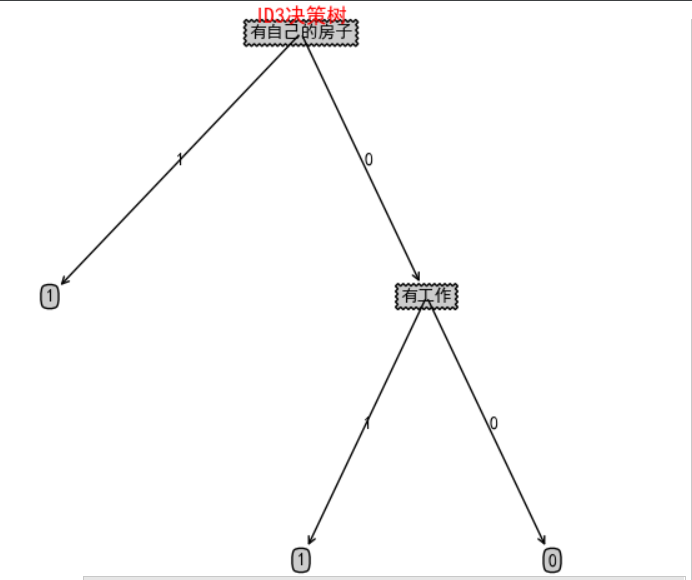
（2）分别创建三种算法的决策树，并用测试集进行测试。测试集数据如下：



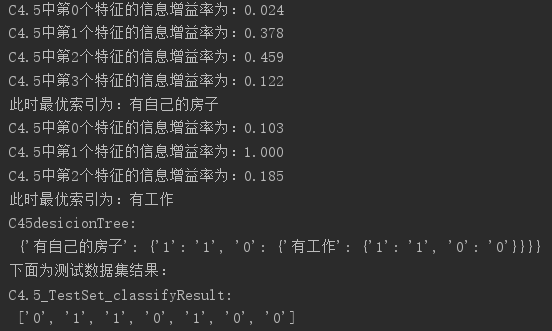
**ID3算法结果：**

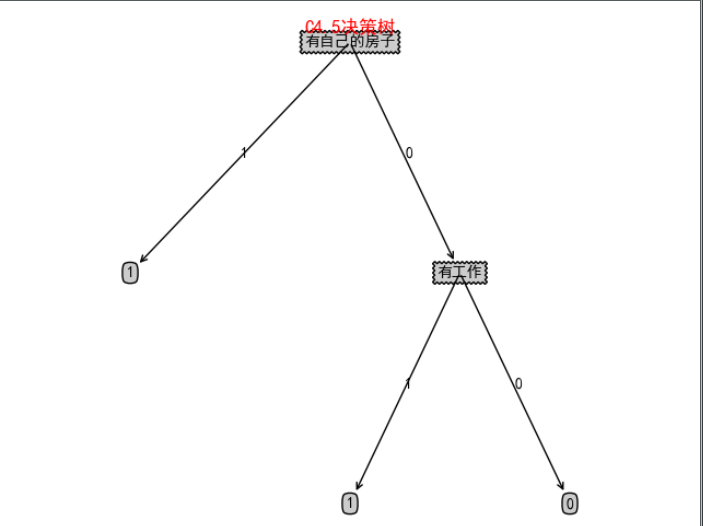


ID3决策树如下图所示



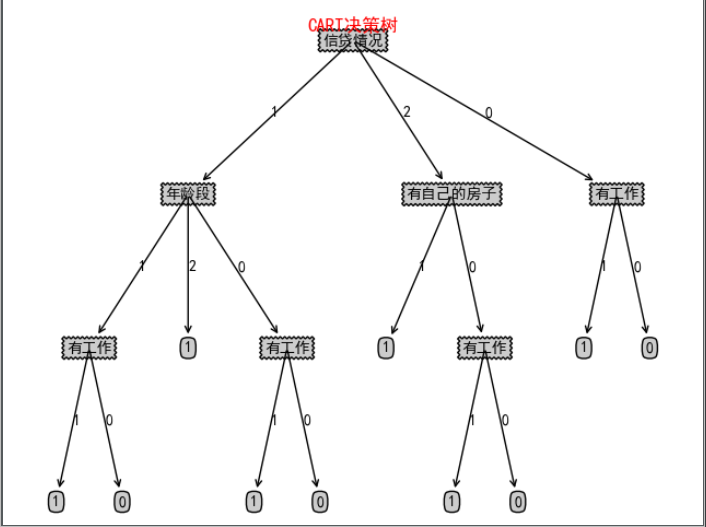
**C4.5结果：**





**Cart结果：**





**分析：**

（1）ID3和C4.5的最优索引以及决策树形图是相同的，而CART的最优索引以及决策树形图与前面两者不同，这与它们的选择标准以及训练集有关；

（2）但同时我们也发现，三种算法对测试集的测试结果是相同的，经过后期手动匹配，结果完全正确，这说明我们的决策树实验结果是正确的。