

机器学习实验报告

题 目： 决策树

学生姓名： 张芮熙

学 号： 22354188

指导教师： 马倩

专业班级： 22智科3班

2021 年 5 月

本科生院制

目 录

* 实验目的与思路
* 核心代码及实现
* 心得体会

**第一部分——编写函数**

* 1. 实验目的、思路**：**



# 

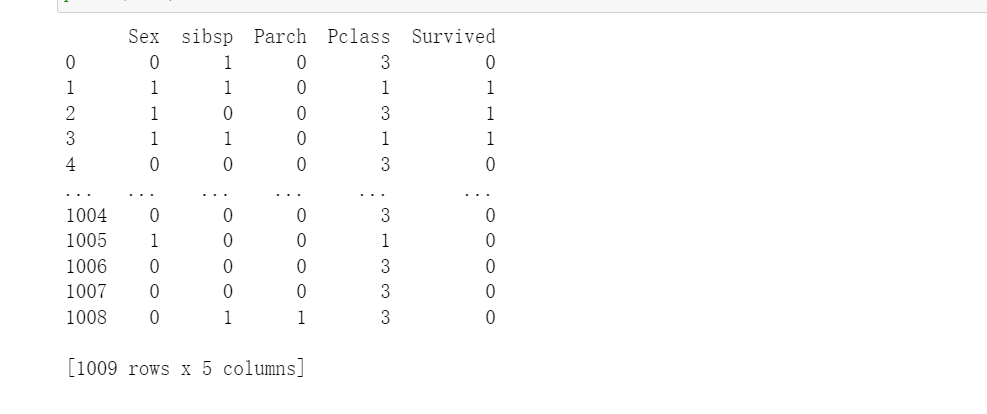
# 2.核心代码与实现

# （1）数据导入

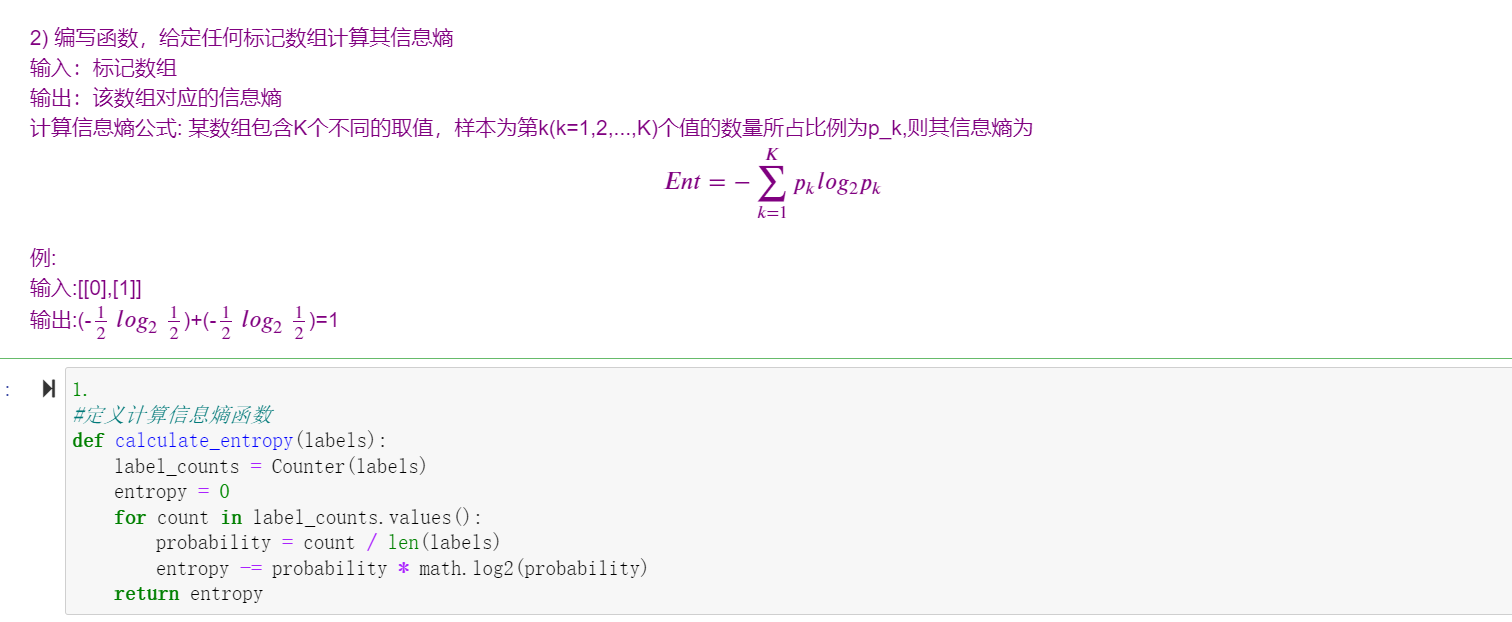
# 

# 运行结果：

输出了数据集Titanic，共有1009行，5列。







(3)定义划分不同数据集函数





（4）ID3算法

根据信息增益计算公式，遍历feature，把best\_gain（最大信息增益）和info\_gain（当前信息增益）进行比较，得到最大的信息增益和划分的dimension.



（5）C4.5算法

在ID3的基础上，增加计算信息增益率功能，信息增益率=（最大信息增益）/(信息熵)，遍历数据集，将gain\_ratio和best\_gain\_ratio进行比较，从而获得最佳的信息增益率和其对应的dimension。



（6）利用GINI指数进行判断

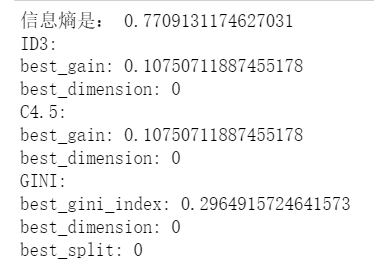
基尼指数=（1-遍历样本中重复的概率的平方求和），基尼指数越大，说明信息熵越大，样本的类型越多，基尼指数越小，说明信息熵越小，样本的类型越少，大多数集中在同一个类型，越接近最佳，在这个算法中，遍历样本，将best\_gini\_index进行迭代，求解最小的基尼指数和对应的dimension，对应的分类值。



（7）将ID3,C4.5,GINI三种判断算法综合利用，求解最优信息增益和其对应的dimension



运行结果：



可见，未进行处理的原数据的信息熵较大，说明样本的取值较为分散，用三种算法进行处理，都得到了第一个维度的信息增益最大，也就是对应样本中性别的信息熵最大，说明男女比例可能差距较大（经过打开数据集的Excel表格发现，在1009个样本中存在两类样本0,1代表性别，平均值只有0.36，说明性别比例较大）

3.心得体会

信息增益最大的特征值所对应的维度具有以下特点：样本之间差异大，对分类结果影响显著，具有代表性。但是需要注意的是，信息增益最大的特征值所对应的维度并不一定与label值最相关。

**第二部分——构建决策树**

1.实验目的、思路**：**

利用第一部分中的计算熵entropy函数，以及划分数据集的split函数，还要定义信息增益最大的最优划分函数，由于实验要求是根据ID3算法进行决策树构建并求出预测准确率，在这里，就只使用ID3算法而不是利用GINI指数和C4.5算法进行求解。

2.核心代码及实现：

1. 定义计算熵的公式，定义划分不同数据集的函数，定义信息增益最大的最优划分（和第一部分一致）





1. **建立树的结点（参考PPT上的伪代码）**



（3）构建决策树

在DTree类中定义一系列函数：生成结点函数，选择信息增益最大函数，将信息增益最大的作为划分决策树的结点。

1. 决策树的构建（参考PPT的伪代码）

在DTree中定义一个生成结点的函数Treegenerate，把数据集的标签值导出到label中，利用unique函数，如果label\_unique的长度是1，那么就说明样本属于同一个类型，可以生成结点，叶子结点标记为true。

反之，如果label\_unique的长度不是1，说明样本不属于同一个类型，那么就遍寻样本，找到样本最多的类型，并构建结点。



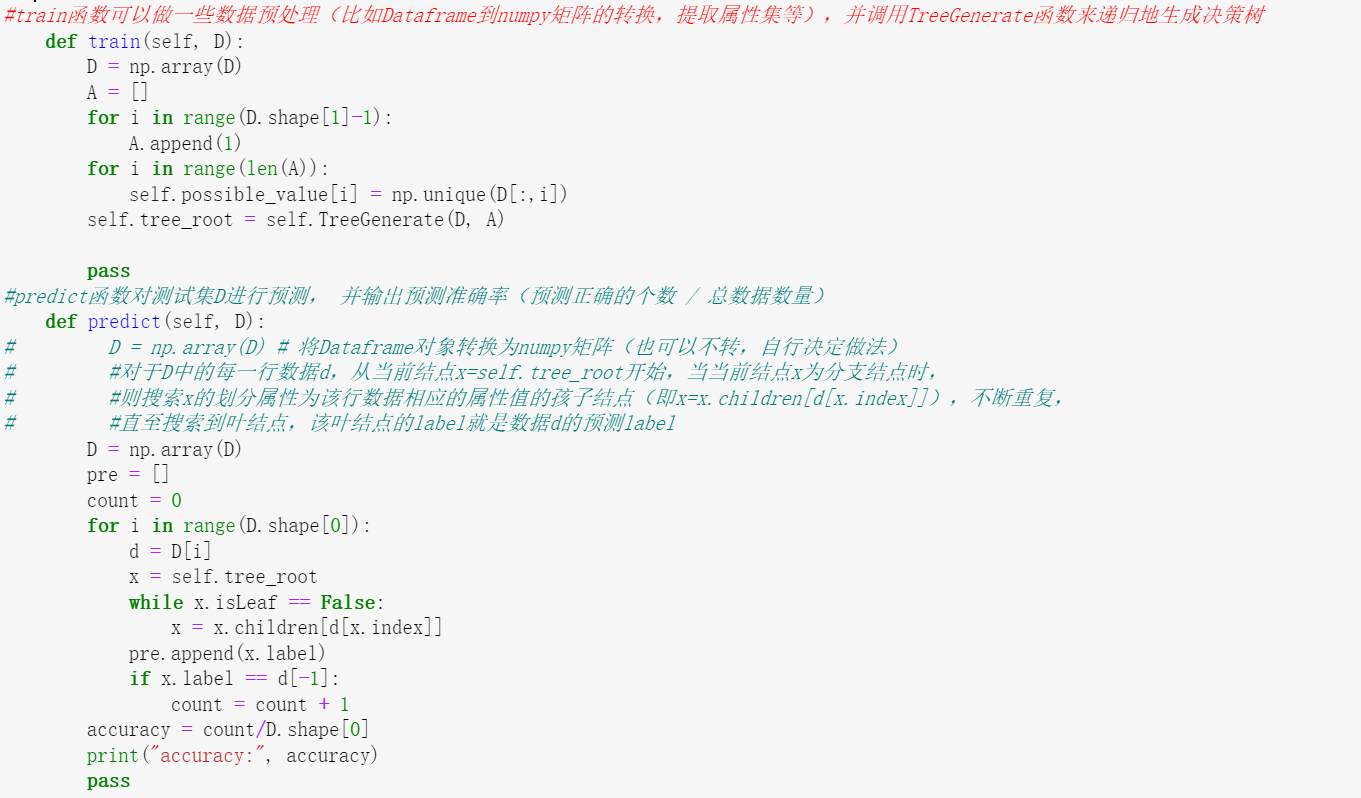
1. 通过最大信息增益划分决策树



3.train函数和predict函数

决策树的构建已经完成，接下来构造train函数，先生成根节点，再通过递归调用生成决策树。

Predict函数则将预测的样本类型和真实值进行对比，如果预测正确，count+=1,最后用count/len(D)得到准确率。



验证：

利用train中的样本构建一颗决策树，再用test进行预测，得到预测准确率。



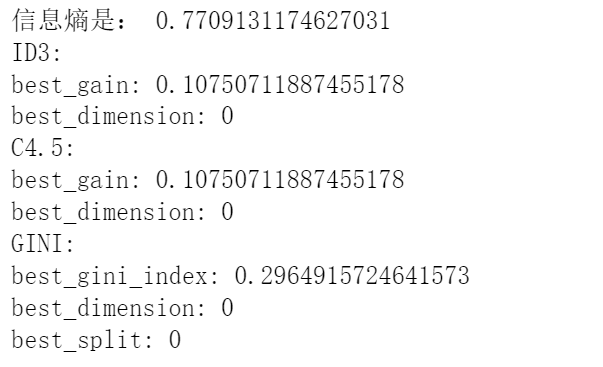
心得体会：

在构建决策树的过程中，遇到了不少困难。

1. 第一次构建的时候，输出的准确率是0.78，以为已经成功地构建了决策树，后来检查的时候发现决策树根本没有叶子结点（label全是-1），也就说明了根本就没有成功地构建起一棵决策树，不得不和同学讨论，推倒重建，修改后的代码经过了验证，可以构建一棵决策树了。
2. 由于之前写过了C4.5和cart算法，我在网上查到决策树的预测准确率大概在70%-90%左右，出于好奇心于是我又利用C4.5和cart算法又实现了一次，结果如下：（两种算法结果基本一致）



经过计算发现C4.5算法，cart算法和ID3算法在预测率上并没有明显的区别（只相差了0.01）经过一系列的探讨和上网查找，我从实验一函数的实现处得到了灵感，在最后输出最大信息增益时，ID3和C4.5算法并没有明显的差异，说明二者在处理该样本的时候可能并没有明显的区别。

（两者的best\_gain和best\_dimension一样）

而Cart算法也没有较大区别的原因可能是样本数据比较便于划分的原因，没有极端情况，所以三种方法没有明显的区别。经过上网查找，我得到了这三种算法的差异和优缺点：**（总而言之，ID3容易造成过拟合，Cart计算量相对较小，ID3和C4.5适合小样本的计算，如果样本过多可能造成过拟合，而Cart适合大样本量，对于小样本量，泛化误差会比较大，各有利弊。）**

1.划分标准的差异：ID3 使用信息增益偏向特征值多的特征，C4.5 使用信息增益率克服信息增益的缺点，偏向于特征值小的特征，CART 使用基尼指数克服 C4.5 需要求 log 的巨大计算量，偏向于特征值较多的特征。

2.使用场景的差异：ID3 和 C4.5 都只能用于分类问题，CART 可以用于分类和回归问题；ID3 和 C4.5 是多叉树，速度较慢，CART 是二叉树，计算速度很快。

3.样本数据的差异：ID3 只能处理离散数据且缺失值敏感，C4.5 和 CART 可以处理连续性数据且有多种方式处理缺失值；从样本量考虑的话，小样本建议 C4.5、大样本建议 CART。C4.5 处理过程中需对数据集进行多次扫描排序，处理成本耗时较高，而 CART 本身是一种大样本的统计方法，小样本处理下泛化误差较大。

4.样本特征的差异：ID3 和 C4.5 层级之间只使用一次特征，CART 可多次重复使用特征。