西南科技大学

计算机实验报告

课程名称：数据挖掘应用技术与实践

实验名称： ID3算法实现

学 号：5120180269

学生姓名：李若昊

班 级： 计科1803

指导教师：吴珏

评 分：

实验日期： 2020年9月19日

实验目的：

通过编程实现决策树算法，信息增益的计算、数据子集的划分、决策树的构建过程。加深对相关算法的理解过程。

实验内容：

1. 分析决策树算法的实现流程
2. 分析信息增益的计算、数据子集的划分、决策树的构建过程
3. 根据算法描述编程实现算法，调试运行；

实验要求：

找出天气数据集中数据的内在关系，得到在什么天气情况下可以运动的“规则”。然后，根据这个“规则”，对给定的新的天气情况给出是否可以运动的判断。

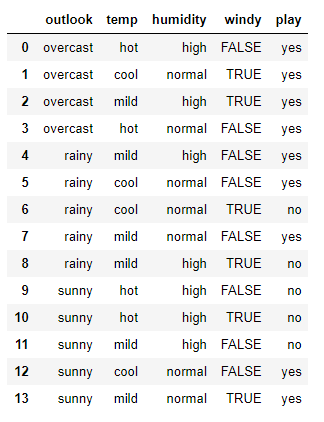
实验步骤：

首先介绍本次实验我选取的实验数据集：天气数据集。其例举了在何种天气条件下可以进行体育运动，数据集中的样本由五个属性值来表示，通过测量不同天气的四个指标得到样本。

天气问题的四个指标是：天气趋势（outlook）、温度（temperature）、湿度（humidity）和刮风（windy）。最后一个属性表示样本的类别，即在知道四个天气指标的前提下得到是否可运动（play）的结论。

其中，天气趋势的属性值分别为sunny、overcast和rainy；温度的属性值分别为hot、mild、cool；湿度的属性值分别为high、normal；刮风的属性值分别为true、false；是否可以运动的属性值分别为yes和no。

考虑到第一个实验只是体验了解机器学习的方法，所以四个指标和一个目标属性都只采用标称符号表示，并不采用具体数值，以简化形式。因为如果包含了数值型属性，学习方案可能需要对此类属性建立不等式，从而使得到的规则会比较复杂。



所利用的库：numpy和pandas.

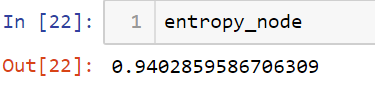
其中，利用pandas创建dataframe（类似于关系型数据库中的一张表）

创建本次实验决策树的过程：

1. 计算整个数据集的熵（Entropy）

记S为天气数据集，有14个样本。类别属性play有2个值，14个样本中有9个样本的play属性为yes，5个为no。说明yes在S中出现的概率为9/14，no出现的概率为5/14。因此S的熵为：

Entropy(S) = Entropy(9/14,5/14) = -9/14\*log2(9/14) - -5/14\*log2(5/14) = 0.94



1. 再计算按各个指标划分S后的样本集的熵和信息增益值（Information Gain）：

a) 分割出样本子集，并计算样本子集的熵：

以指标windy为例，它有两个可能的取值{true, false}，仅考虑windy的取值，将windy取值相同的划在一组，从而形成2个子集S1，S2，分别有8个和6个样本。

从而对样本子集S1，play = yes 有6个样本，play = no有2个样本，则

Entropy(S1) = -6/8\*log2(6/8) - -2/8\*log2(2/8) = 0.811

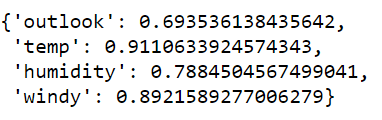
对样本子集S2，play = yes 有3个样本，play = no有3个样本，则

Entropy(S2) = -3/6\*log2(3/6) - -3/6\*log2(3/6) = 1

则S在用windy划分的情况下的熵为：

Entropywindy(S) = (|S1|/|S|)\* Entropy(S1) + (|S2|/|S|)\* Entropy(S2)

= (8/14)\* Entropy(S1) + (6/14)\* Entropy(S2) = 0.891



b) 计算样本集的信息增益值

S在用windy划分的情况下的信息增益值（Gain）为：

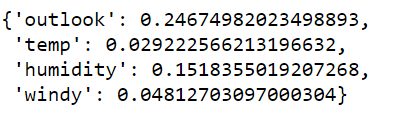
Gain(S, windy) = Entropy(S) - Entropywindy(S) = 0.049

1. 重复步骤2，同理得到S在被其他指标划分下的信息增益值：

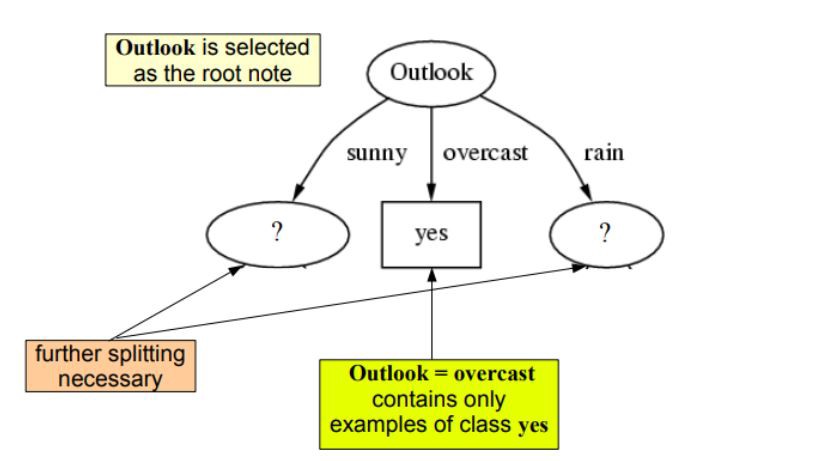
Gain(S, outlook) = 0.246

Gain(S, temperature) = 0.029

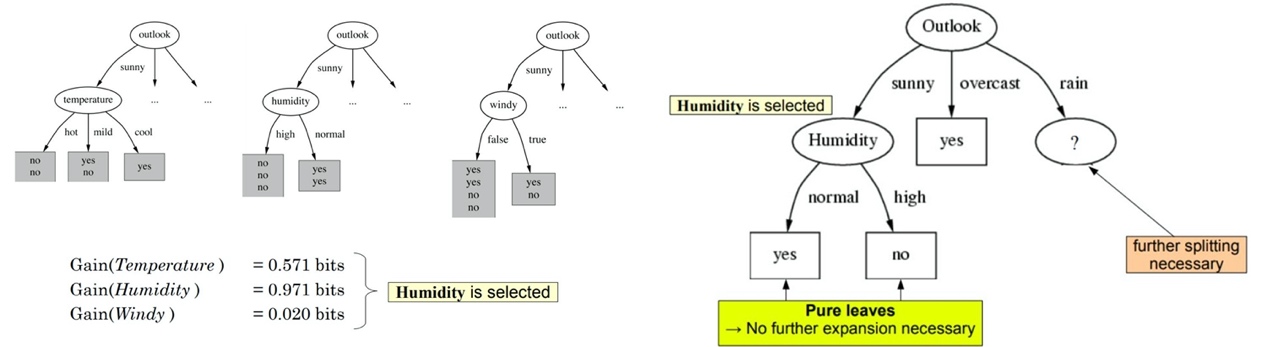
Gain(S, humidity) = 0.152



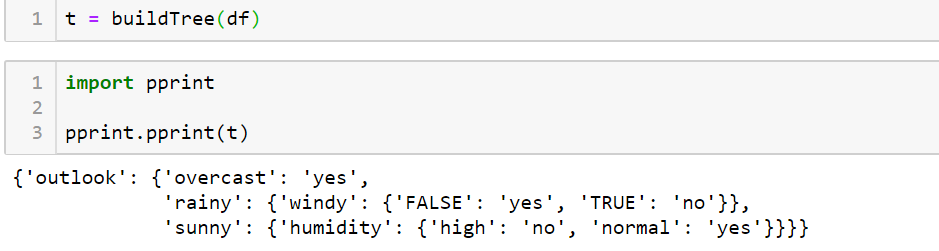
1. 将信息增益值最大的那个指标作为分裂节点，例如一开始选择outlook指标作为树的根节点，随即产生3个分支

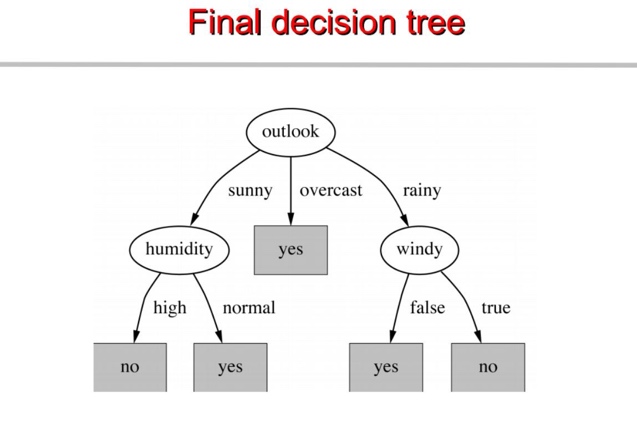


1. 对根节点的三个分支递归调用以上方法，各求其子树



1. 直到尾部全为叶节点（可以直接给出答案的节点）





实验思考：

1. 信息增益的计算

见上实验步骤2b)

1. 选择相关字段后根据相关字段的取值对数据集进行划分

见上实验步骤2a)

1. 决策树构建的终止条件

见上实验步骤6

实验体会：

在上学期的机器学习课程中已经利用一些工具直接使用过该算法（的改进版），现在我知道了算法的原理，可以帮助我在应用时更好地理解结果和参数的意义。

感谢吴老师的悉心指导。