一、问题讨论

我的问题：

1.为什么要采用信息增益比？是比直接的信息增益算法更有效吗？

讨论后的理解：信息增益算法的缺点是，当特征的取值较多时，根据此特征划分更容易得到纯度更高的子集，因此划分之后的熵更低，由于划分前的熵是一定的，因此信息增益更大，因此信息增益比较，偏向取值较多的特征。为了解决这一问题，引入了信息增益比。

信息增益比是在信息增益的基础之上乘上一个惩罚参数。特征个数较多时，惩罚参数较小；特征个数较少时，惩罚参数较大。惩罚参数：数据集D以特征A作为随机变量的熵的倒数，即：将特征A取值相同的样本划分到同一个子集中（之前所说数据集的熵是依据类别进行划分的）

2.ID3和C4.5两种算法生成的决策树还需要后剪枝吗？

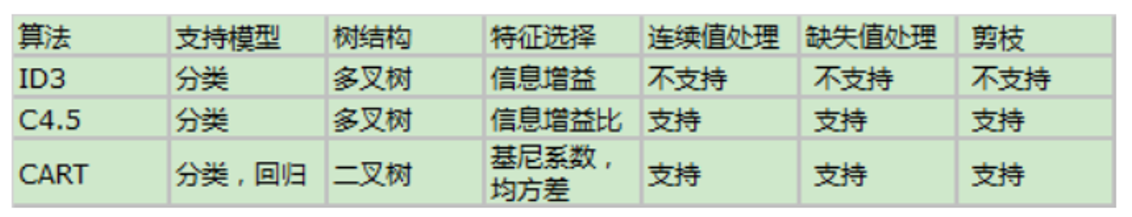
讨论后的理解：ID3和C4.5是预剪枝的算法，在算法运行时已经做出了限制，因此可以不需要进行后剪枝。

解决他人问题：

3.CART剪枝中，“以t为单节点树”得到的损失函数是什么含义？

讨论后的理解：具体地，从整体树T0开始剪枝，我们每次剪枝剪的都是某个内部节点的子节点，也就是将某个内部节点的所有子节点回退到这个内部节点里，并将这个内部节点作为叶子节点。因此在计算整体的损失函数时，这个内部节点以外的值都没变，只有这个内部节点的局部损失函数改变了，因此我们本需要计算全局的损失函数，但现在只需要计算内部节点剪枝前和剪枝后的损失函数。

4.ID3、C4.5和CART三种决策树的区别？

讨论后的理解：ID3和C4.5的主要区别是前者采用了信息增益算法，后者采用了信息增益比。相比CART而言，这两个决策树都是预剪枝，而CART是后剪枝。具体见下表：

二、读书计划

本周完成第5章

下周计划第6章

三、读书摘要或者代码实现

读书摘要：本章学习决策树

**什么是决策树？**

分类决策树模型是描述对实例进行分类的树形结构

决策树包括：

内部结点：表示特征

边：连接结点

叶结点：表示一个类

**决策树与if-then规则：**

由决策树的根摘要到叶结点的每一条路径制定一条规则

路径上的内部结点对应着规则的条件，而叶结点对应着规则的尺度

每一个实例都被一条规则覆盖

**信息增益（信息\空间增益）：**

概念：表示认识特征X的信息而适当类Y的信息不确定性的减少

定义：特征A对训练数据集D的信息增益g=（D，A），定义为集合D的经验熵H（D）与特征 A 给定条件下 D 的经验条件熵 H（D | A）之差，即： g（D，A）= H（D）-H（D | A）

**特征选择：**

概念：特征选择在于选取对训练数据具有分类能力的特征，通常特征选择的规则是信息增益或则信息增益比

**ID3算法：**

ID3算法是一种基于信息增益属性选择的决策树学习方法。核心思想是：通过计算属性的信息增益来选择决策树各级节点上的分裂属性，使得在每一个非叶子节点进行测试时，获得关于被测试样本最大的类别信息。基本方法是：计算所有的属性，选择信息增益最大的属性分裂产生决策树节点，基于该属性的不同属性值建立各分支，再对各分支的子集递归调用该方法建立子节点的分支，直到所有子集仅包括同一类别或没有可分裂的属性为止。由此得到一棵决策树，可用来对新样本数据进行分类。

**ID3算法优缺点分析：**

优点：构建决策树的速度比较快，算法实现简单，生成的规则容易理解。  
缺点：在属性选择时，倾向于选择那些拥有多个属性值的属性作为分裂属性，而这些属性不一定是最佳分裂属性；不能处理属性值连续的属性；无修剪过程，无法对决策树进行优化，生成的决策树可能存在过度拟合的情况。