问题

决策树是怎么一步步生成的

决策树原理

决策树简单来说就是带有判决规则（if-then）的一种树，可以依据树中的判决规则来预测未知样本的类别和值。

决策树的学习本质上是从训练集中归纳出一组分类规则，得到与数据集矛盾较小的决策树，同时具有很好的泛化能力。决策树学习的**损失函数通常是正则化的极大似然函数**，通常采用启发式方法，近似求解这一最优化问题。

决策树生成流程

顾名思义，决策树是一种树结构，一般的，包含一个根节点、若干个内部结点和若干个叶节点。根节点包含样本全集；叶节点对应着分类结果，即在同一个叶节点中的样本被分类归属为一个类别；而内部结点中的每一个结点对应于属性测试，即当前内部结点是其父结点根据最优属性划分后的一个分支，在当前结点依然可分的情况下，继续寻找最优划分属性并划分为该属性所有取值的个数的分支，至于怎么选择最优划分属性，请参考机器学习文件夹中的问题“20_随机森林思想”。

决策树的生成是一个递归的过程。在决策树的基本算法中，有三种情况会导致递归返回：
（1）当前节点包含的样本全属于同一类别，无需划分；
（2）当前属性集为空，或是所有样本在所有属性上取值相同，无法划分，这时把当前结点标记为叶节点，并将其类别设定为该节点所含样本最多的类别；
（3）当前节点包含的样本集为空，不能划分。

拓展

决策树不仅可以用于分类问题，同样可以应用于回归问题。

回归树总体流程与分类树类似，不过在每个节点（不一定是叶子节点）都会得一个预测值，以年龄为例，该预测值等于属于这个节点的所有人年龄的平均值。分枝时穷举每一个feature的每个阈值找最好的分割点，但衡量最好的标准不再是最大熵，而是最小化均方差--即（每个人的年龄-预测年龄）^2 的总和 / N，或者说是每个人的预测误差平方和 除以 N。这很好理解，被预测出错的人数越多，错的越离谱，均方差就越大，通过最小化均方差能够找到最靠谱的分枝依据。分枝直到每个叶子节点上人的年龄都唯一（这太难了）或者达到预设的终止条件（如叶子个数上限），若最终叶子节点上人的年龄不唯一，则以该节点上所有人的平均年龄做为该叶子节点的预测年龄。

参考资料

《机器学习》周志华
决策树原理详解
决策树模型 ID3/C4.5/CART算法比较