读书报告

71117203-袁佳怡

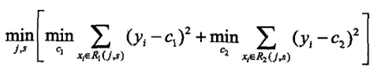
1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：5.5中，CART算法在决策树生成中，基于训练数据集生成决策树生成的决策树要尽量大是为什么？

讨论后的理解：CART树的划分本质是一棵二叉树，每个特征用来划分的特征值比较少，树的每个节点划分出来的分支只有两个，这样的划分很容易导致划分力度不够导致分类不明确。因此呢，选择CART树尽量大一些，可以让分类更加准确，决策更加有效。

1. 提出的问题2：式子5.31表示了剪枝后整体损失函数减少的程度，但为什么说是在T0中剪去g(t)最小的T，不应该剪去损失函数减少的多的吗，怎么理解？

讨论后的理解：α的值是一个变大的过程，对于某一结点a，当总体损失函数中的α=g(t)时，剪和不剪总体损失函数是一样的，这是如果α稍微增大，那么不剪的整体损失函数就大于剪去的，即α大于g(t)时应该剪枝，剪了能够降低整体的损失函数；反之就不要剪枝。

当α从0逐渐缓慢增大的过程中，总有某棵子树该被剪枝，其他子树不该剪，即阿尔法超过了某个节点的g(t)但没有超过其他节点的g(t)。所以每次剪掉的都是当前α下g(t)最小的那一棵子树。

1. （必填）别人提出的问题的理解（选择几个问题罗列，并给出理解）：
2. 问题3：CART的回归决策树生成算法中：

这里的求最小值中的c1和c2也是平均值吗？

自己的理解：我觉得是，分类的均值对应的应该是让上式达到极值，如果让c值是平均值的，带入上面的式子，可以发现得到的是两个方差的和，这样找到了使得方差最小的j和s。

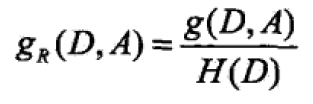
1. 问题4: 证明CART剪枝算法中，当α确定时，存在唯一的最小子树Tα使损失函数Cα(T)最小。

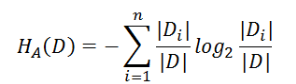
自己的理解：利用反证法，如果存在不唯一的最小子树，即有最优子树不唯一，那么也就是说对于原来的树中有多个树要被删除的，而且删去的部分一定是不同的；对于一棵最优子树，不一定会保留另一棵最优子树中被删除的部分，而这一部分仍然可以删除，这样的话，删除后得到的子树才是更小的，就与题设的最优子树构成了矛盾。所以只能有唯一的最小子树。

1. 问题5：在5.4决策树的剪枝中，对于式子(5.14)而言，为什么较大的α促使选择较简单的模型，较小的α促使选择较复杂的模型？

自己的理解：在损失函数的计算公式里，节点数目|T|越大，也就是复杂度越高，所以增加α的值的话，会让损失函数更大，会倾向于选择复杂度较低的树。这里的损失函数其实是一个正则化的计算方式，α越大对复杂度高的模型惩罚越大，这里是为了防止过拟合。

1. 问题6：ID3算法和C4.5算法的区别，我认为信息熵增益最大不是应该就意味着相对信息熵增益最大吗，毕竟信息熵是相同的

自己的理解：书上的公式是但事实上这里的H(D)指的是在以某个特征A为前提下的训练数据的熵值，是，对于样本集合D，将当前特征A作为随机变量（取值是特征A的各个特征值）求得的经验熵；这里的分母并不是之前把集合类别作为随机变量，而是把某个特征作为随机变量，按照此特征的特征取值对集合D进行划分

计算的公式是：

信息增益比的本质应该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。

1. （必填）读书计划

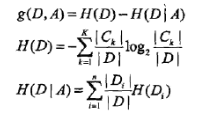
1、本周完成的内容章节：《统计学习方法》第五章

2、下周计划：《统计学习方法》第六章

1. 读书摘要
   1. 决策树是一种基本的分类与回归方法，它既能够实现分类也能够实现回归。所以可以分别它们称为分类树和回归树。分类决策树模型是基于特征对实例进行分类的过程，可以被认为是if-then规则的集合，也可以认为是定义在特征空间与类空间上的条件概率分布。
   2. 将决策树转化为if-then 规则的过程：由决策树的根节点到叶节点的每一条路径构建一条规则，路径上内部节点的特征对应着规则的条件，而叶节点的类对应着规则的结论。这些都基于一个重要的性质：互斥并且完备。
   3. 决策树算法学习包括三部分：特征选择，树的生成、树的剪枝。常用的算法由ID3、C4.5、CART。
   4. 特征选择在于选取对训练数据具有分类能力的特征，这样可以提高决策树学习的效率。下面介绍几个名词的定义：

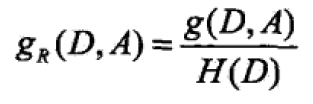
A、熵：表示数据的不确定性，当熵值越大，表示数据越混乱。

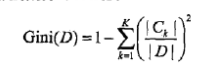
B、信息增益：在选取一个分类特征前后，数据不确定性的差值，这个值越大表示数据变得更加分类正确，不确定度下降最快。样本集合D对特征A的信息增益表达为：

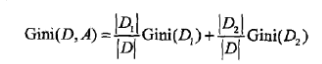


式中，H(D)是数据集D的熵，H(D|A)是数据集D对于特征A的条件熵。

在ID3的算法中就是利用了信息增益去衡量选取哪一个分类的特征值来使得分类之后不确定度下降最多，让数据能够变得更纯，让这棵树的分类效果变得更好。

C、信息增益比：，因为信息增益值的大小是相对于训练数据集而言的，没有绝对意义，在分类问题困难时，信息增益会偏大。对造成一定的误差所以在C4.5算法中就用了信息增益比来选择特征，降低过拟合的风险。

D、基尼指数：样本集合的基尼指数：

特征A条件下集合D的基尼指数：

基尼指数表示集合D的不确定性，Gini（D,A）表示经过A=a分割后集合D的不确定性，基尼指数值越大，样本集合的不确定性也越大。

* 1. 理解CART算法：

CART，又名分类回归树，是在ID3基础上进行优化的决策树。当CART是分类树时，采用GINI值作为节点分裂的依据；当CART是回归树时，采用样本的最小方差作为节点分裂的依据；CART是一棵二叉树。分类树的作用是通过一个对象的特征来预测该对象所属的类别，而回归树的目的是根据一个对象的信息预测该对象的属性，并以数值表示。

分裂的目的是为了能够让数据变纯，使决策树输出的结果更接近真实值。在CART中，如果是分类树，CART采用GINI值衡量节点纯度；如果是回归树，采用样本方差衡量节点纯度。节点越不纯，节点分类或者预测的效果就越差。