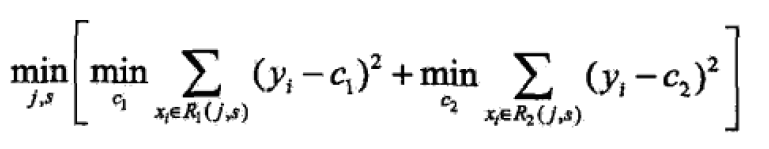
方骏-2020年4月5 日-读书报告

1. 自己提出的问题的理解：
2. 提出的问题1：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？

讨论后的理解：是的。因为要求的是平方误差，所以分类的均值对应的分类值理应是让上式到达极值的情况。那么如果c值是平均值得话，代入上式，可以发现两边的式子都是方差，因此这样就是找到方差最小的j和s，是合情合理的。

1. 提出的问题2：证明CART剪枝算法中，当α确定时，存在唯一的最小子树使损失函数最小。

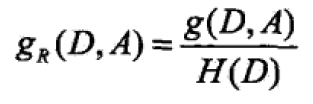
讨论后的理解：如果假设两个相同，有两个最优子树，那么也就是说对于原来的树中有两个子树是要被删掉的，而且删去的部分一定是不同的，对于一颗最优子树，是不是一定会保留另一颗最优子树中被删掉的部分，而这一部分由于测试误差仍然是可以减小的（因为该部分是最优子树），所以仍然可以删去这一部分，那么显然，删去后得到的子树是更小的，所以这两个子树就不是最优的，矛盾。因此存在唯一的最小子树。

1. 别人提出的问题的理解：
2. 问题3：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：节点数|T|越大的话，也就是复杂度越高，所以如果增加α的值的话，就会让损失函数更大，更倾向于选择复杂度低的树，α越大对复杂度高的模型的惩罚越大，也可以看作是权重，α越大，那么与α相乘的|T|就应该相应的变小，所以模型复杂度就变小了。本质上，这个α就是起到一个惩罚机制。

1. 问题4： 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：书上的公式没有表示清楚，这里的：



中的H(D)并不是全体训练数据的熵，而是在以A为特征的前提下，在多个A的取值下训练数据的熵，即：

这样的分母才有意义，可以限制住当A特征取值太多，导致A特征的熵太大的情况，这样用作为分母就可以平衡这种情况，让信息熵评价更加准确。

1. 问题5： 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？

自己的理解：应该是因为CART树毕竟是一个二叉树，用来作为划分的特征依据其实太少了，肯定不如ID3这样的多叉树划分的情况多，这样的话二叉树的分支就太少，可能会导致划分不够具体从而分类不准确。因此CART树生成的决策树应该尽量大一些，过拟合出现的情况比正常的决策树要好一些。

1. 读书计划

1、本周完成的内容章节：《统计机器学习》第五章

2、下周计划：《统计机器学习》第六章

四、读书摘要及理解

1、决策树是一种基本的分类与回归方法。决策树模型呈树形结构，在分类问题中，表示基于特征对实例进行分类的过程。它可以认为是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。其主要优点是模型具有可读性，分类速度快。决策树学习通常包括3个步骤：特征选择、决策树的生成和决策树的修剪。

2、分类决策树模型是一种描述对实例进行分类的树形结构。决策树由结点和有向边组成。结点有两种类型：内部结点和叶结点。内部结点表示一个特征或属性，叶结点表示一个类。用决策树分类，从根结点开始，对实例的某一特征进行测试，根据测试结构，将实例分配到其子结点；这时，每一个子结点对应着该特征的一个取值。如此递归下去，直至到达叶结点。最后将实例分到叶结点的类中。

值得注意的是，每一个实例都被一条路径或一条规则所覆盖，而且只被一条路径或一条规则所覆盖。这里所谓覆盖是指实例的特征与路径上的特征一致或实例满足规则的条件。决策树还表示给定特征条件下类的条件概率分布。这一条件概率分布定义在特征空间的一个划分上。将特征空间划分为互不相交的单元或区域，并在每个单元定义一个类的概率分布就构成了一个条件概率分布。

决策树学习的目标是根据给定的训练数据集构建一个决策树模型。决策树学习本质上是从训练数据集中归纳出一组分类规则。我们需要的是一个与训练数据矛盾较小的决策树，同时具有很好的泛化能力，即对未知数据有很好的预测。决策树学习用损失函数表示这一目标，通常是正则化的极大似然函数。

决策树学习的算法通常是一个递归地选择最优特征，并根据该特征对训练数据进行分割。开始构建一个根结点，然后选择一个最优特征，按照这一特征将训练数据集分割成子集，使得各个子集有一个在当前条件最好的分类。如此递归，直至所有训练数据子集被基本正确分类或没有合适的特征为止。对于有可能发生的过拟合现象，需要对已生成的树自下而上进行剪枝，将树变得更简单。决策树的生成只考虑局部最优，相对地，决策树的剪枝则考虑全局最优。

3、特征选择在于选取对训练数据具有分类能力的特征。通常特征选择的准则是信息增益或信息增益比。熵是表示随机变量不确定性的度量。随机变量X的熵定义为：



条件熵表示在已知随机变量X的条件下随机变量Y的不确定性。条件熵的定义为：



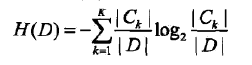
信息增益表示得知特征X的信息而使得类Y的信息的不确定性减少的程度。

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的经验条件熵之差，即：



信息增益依赖于特征，信息增益大的特则具有更强大的分类能力。根据信息增益准则的特征选择方法是：对训练数据集D，计算其每个特则的信息增益，并比较它们的大小，选择信息增益最大的特征。

公式如下：



信息增益比就是信息增益g(D,A)于训练数据集D的经验熵之比：



4、决策树有两种主要的生成方法：ID3和C4.5方法，两者大同小异。

ID3算法的核心是在决策树各个结点上应用信息增益准则选择特征，递归地构建决策树。具体方法是：从根结点开始，对结点计算所有可能的特征的信息增益，选择信息增益最大的特征作为结点的特征，由该特征的不同值建立子结点；再对子结点递归地调用以上方法，构建决策树；直到所有特征的信息增益均很小或没有特征可以选择为止。最后得到一个决策树。

C4.5算法与ID3算法相似，C4.5算法对ID3算法进行了改进。C4.5在生成的过程中，用信息增益比来选择特征。

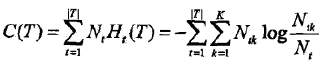
5、在决策树中将已生成的树进行简化的过程称为剪枝。决策树的剪枝往往通过极小化决策树整体的损失函数或代价函数来实现。决策树学习的损失函数可以定义为：



其中经验熵为：



将上式右端第一项记作：



这时有：



参数α控制预测误差和复杂度两者之间的影响。

当α确定时，选择损失函数最小的模型，即损失函数最小的子树。决策树生成学习局部的模型，而决策树剪枝学习整体的模型。先计算每个结点的经验熵，递归地从树的叶结点向上回缩，设一组叶结点回缩到其父结点之前与之后的整体数分别为与，如果：



那么剪枝，父结点变为新的叶结点。递归如上。

6、CART算法也是由三个必要步骤组成，CART假设决策树是二叉树，内部结点特征的取值为“是”和“否”，这样的决策树等价于递归地二分每个特征。

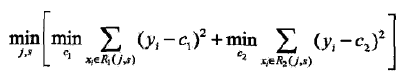
6.1、回归树用平方误差最小化准则，回归树模型可表示为：



当输入空间划分确定时，可以用平方误差表示回归树对于训练数据地预测误差。每一个单元上地最优分类值是均值：



选择第j个变量和它取的值s，作为切分变量和切分点，分成两个区域，然后寻找最优切分变量j和最优切分点s。求解如下：



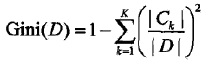
对固定输入变量j可以找到最优切分点s：



6.2、分类树的生成依靠基尼指数，概率分布的基尼指数定义为：



对于给定的样本集合D，其基尼指数为：



将集合D根据特征A是否取一可能值ａ被分割成两部分，那么在特征A的条件下，集合D的基尼指数定义为：



同样，基尼指数越大，样本集合的不确定性也就越大。遍历每一个特征和每一个特征的所有取值，找到基尼指数最小的特征和对应的切分点即可。

6.3、剪枝过程中，子树的损失函数仍然是：



对固定的α，一定存在使损失函数最小的子树。可以用递归的方法对树进行剪枝，将α从小逐渐增大，产生一系列区间，剪枝得到的子树序列对应着α的区间。

以t为根结点的子树的损失函数是：



当α=0及α充分小时，有不等式：



当α增大时，在某一α有：



此时的α就是我们要找的那一个划分α值。计算：



表示减少的程度，将得到的子树作为，同时将最小的设为，为区间的最优子树，就这样得到一系列子树，利用独立的验证数据集，测试这些子树的平方误差或基尼指数，最小的就是最优的决策树。