**[分类与回归树（CART）](http://www.cnblogs.com/kukri/p/8458236.html)**

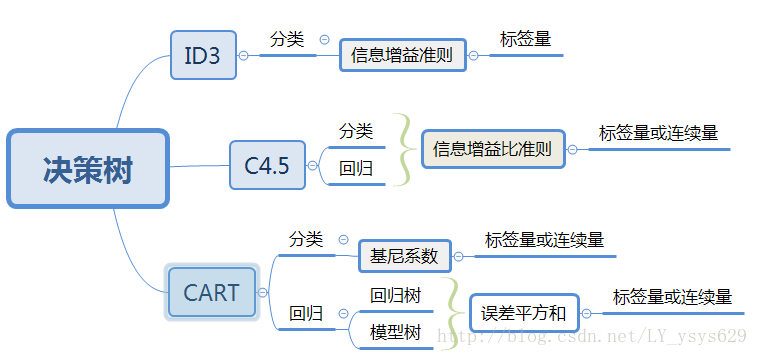
<https://www.cnblogs.com/kukri/p/8458236.html>

该文章说明了对特征取值数>=3时的二分处理方法（组合）及特征连续取值时的处理方法（根据已存在的取值切分，提高效率）。

**一、CART简介**

　　分类与回归树（calssification and regression tree，CART）是决策树算法中的一种，与其他决策树算法相同，同样由特征选择，树的生成与剪枝组成。CART被广泛应用，且被用于树的集成模型，例如，GBDT、RF等集成算法的基学习器都是CART树。决策树是典型的非线性模型，GBDT和RF因此也是非线性模型。

　　决策树的经典算法包括ID3、C4.5、CART算法，其应用领域及所使用的准则，如下图所示。



　　CART算法的重要基础包含以下三个方面：

（1）二分(Binary Split)：在每次判断过程中，都是对观察变量进行二分。

CART算法采用一种二分递归分割的技术，算法总是将当前样本集分割为两个子样本集，使得生成的决策树的每个非叶结点都只有两个分枝，CART二分每个特征（包括标签特征以及连续特征），如果标签特征有3个属性，可以将其中的两个属性归为一类，另一个属性归为一类。因此CART算法生成的决策树是结构简洁的二叉树。因此CART算法适用于样本特征的取值为是或非的场景，对于连续特征的处理则与C4.5算法相似。

（2）单变量分割(Split Based on One Variable)：每次最优划分都是针对单个变量。

（3）剪枝策略：CART算法的关键点，也是整个Tree-Based算法的关键步骤。

剪枝过程特别重要，所以在最优决策树生成过程中占有重要地位。有研究表明，剪枝过程的重要性要比树生成过程更为重要，对于不同的划分标准生成的最大树(Maximum Tree)，在剪枝之后都能够保留最重要的属性划分，差别不大。反而是剪枝方法对于最优树的生成更为关键。

**二、CART生成算法**

　　CART的全称是classification and regression tree，意为分类回归树。也就是说这类决策树既可以解决分类问题，也能解决回归问题。因此针对分类树和回归树，CART有两种生成方式。针对任何一种机器学习模型，其训练过程都大同小异，目的都是为了使训练集数据尽可能被较好地分类和拟合。CART也不例外，无论是分类树还是回归树，无外乎都需要在生成过程（训练过程）使得某些误差最小，最恰当的安排训练数据。下面分别说明。

**1. 分类树**

分类树用基尼指数（gini index）最小化准则进行特征选择生成二叉树 ，基尼指数G(D,A)表示经A=a分割后集合D的不确定性，基尼指数越大，样本集合的不确定性就越大。

**基尼指数:**

对于一份lable为分类变量的训练集，存在一个指标叫做基尼指数，只需要知道基尼指数是概率分布的其中一种表示

https://images2015.cnblogs.com/blog/965227/201609/965227-20160909142320879-1702885481.png

pk表示某一类别的样本数占总样本数的比例。再说一次，只要是分类性质的样本集，本身就存在基尼指数这个属性。

然后就是到底按照哪个特征的哪个值进行二分训练集，注意，CART树的每一步都是二分，没有多分情况。

https://images2015.cnblogs.com/blog/965227/201609/965227-20160909142641785-1041913678.png

https://images2015.cnblogs.com/blog/965227/201609/965227-20160909142706894-915288620.png

没有简单的办法，只有穷举，用每个特征（这里是A）的每个可能的值（这里是a）进行二分，每分一次就会产生两份样本，然后按照上式求每一次的综合基尼指数，最后选择基尼指数最小的那个分法就好了。为什么选择最小的呢？我们看看基尼指数的含义，实质上表示的是一份样本集的纯净程度，拿二分类类型的样本集来说，https://images2015.cnblogs.com/blog/965227/201609/965227-20160909142320879-1702885481.png这个算式在什么情况下最大？最小？当所有样本都是一类的时候，最小，为0；当一半一半时，最大，为2\*0.5^2。所以，越小的基尼指数，意味着这次二分最大程度保证了分开的两份样本集各自的纯净度，这也正是我们想要的。

**对离散分布，且特征值>=3的特征的处理：**

正是因为CART树是二叉树，所以对于样本的有N>=3个取值的离散特征的处理时也只能有两个分支，这就要通过组合人为的创建二取值序列并取GiniGain最小者作为树分叉决策点。如某特征值具有['young','middle','old']三个取值,那么二分序列会有如下3种可能性(空集和满集在CART分类中没有意义):

[(('young',), ('middle', 'old')), (('middle',), ('young', 'old')), (('old',), ('young', 'middle'))]

采用CART算法，就需要分别计算按照上述List中的二分序列做分叉时的Gini指数，然后选取产生最小的GINIGain的二分序列做该特征的分叉二值序列参与树构建的递归。如果某特征取值有4个，那么二分序列组合就有7种，5个取值就有15种组合

**对连续特征的处理：**

先把连续属性转换为离散属性再进行处理。虽然本质上属性的取值是连续的，但对于有限的采样数据它是离散的，如果有N条样本，那么我们有N-1种离散化的方法：<=vj的分到左子树，>vj的分到右子树。计算这N-1种情况下最大的信息增益率。另外，对于连续属性先进行排序（升序），只有在决策属性（即分类发生了变化）发生改变的地方才需要切开，这可以显著减少运算量。

**分类树算法步骤：**

1）计算现有样本D的基尼指数，之后利用样本中的每一个特征A以及A的每一个特征取值a，根据特征的属性（离散or连续）分为两部分，并计算Gini(D,A)的值。

2）找出对应基尼指数最小的Gini(D,A)的最优切分特征及取值，并判断是否切分停止条件，否，则输出最优切分点

3）递归调用1）2）

4）生成CART决策树

切分停止条件：结点中的样本个数小于预定阈值，或样本集的基尼指数小于预定阈值（样本基本属于同一类），或者没有更多特征。

**2. 回归树**

对应与分类树的最小化基尼指数准则，回归树该怎么确定哪个分法才最好呢？最小二乘！经典的最小二乘算法！

之所以称为最小二乘回归树，是因为，回归树以误差平方和为准则选择最优二分切点，该生成算法在训练数据集上所在的输入空间中，递归的将每个区域划分为两个子区域并决定每个子区域的输出值，在这里分为两种情况，一是输出值为子区域输出值的均值该种情况下为回归树，二是输出值为子区域输入与输出的线性回归，输出值为回归系数，该种情况下为模型树。

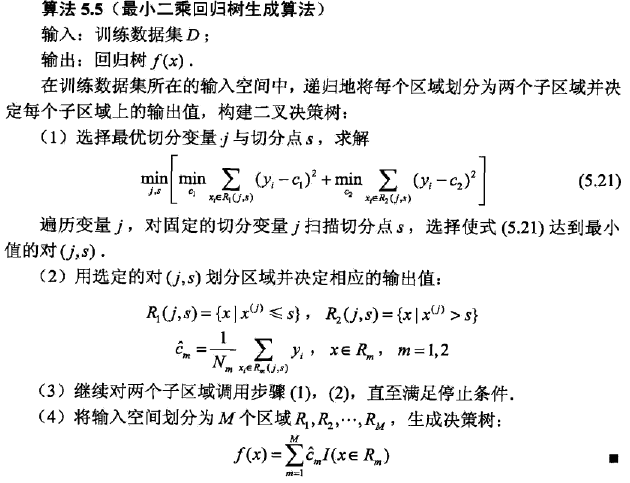
**算法实现步骤：**

1）选择最优切分特征J与切分点s，按照如下原则：

minj,s[minc1∑(yi−c1)+minc2∑(yi−c2)]

c1,c2分别为左右子区域输出的均值（模型树时是输出变量的回归值），可通过遍历每个变量的每个可能取值来切分数据集找出最优切分点。   
2）用切分点划分区域并决定相应的输出值 （如果是回归树的话输出值为子区域的输出值的均值，如果是模型树的话为输入与输出的线性回归，输出值为回归系数）  
3）递归调用1）2）直到满足停止条件   
4）利用字典，递归调用创建二叉树，生成决策树

停止条件：可以认为设定当某个树杈包含的样本个数或者方差小于某个阈值时停止。



**三、CART剪枝**

由于CART生成算法是让生成的决策树自然生长，尽量大，所以容易造成过拟合，因此采取剪枝策略避免过拟合。

CART采取Cost-Complexity Pruning(CCP、代价复杂度)剪枝法：

CCP方法包含两个步骤：  
1：从原始决策树T0开始生成一个子树序列{T0、T1、T2、...、Tn},其中Ti+1是从Ti总产生，Tn为根节点  
2：通过交叉验证法在独立的验证数据集上对子树序列{T0、T1、T2、...、Tn}进行测试，从中选择最优子树。

**参考链接：**

https://www.cnblogs.com/lvlvlvlvlv/p/5856986.html

http://blog.csdn.net/LY\_ysys629/article/details/72809129

http://blog.csdn.net/suipingsp/article/details/42264413

http://blog.csdn.net/u011067360/article/details/24871801

李航《统计学习方法》