# 生成决策树

ID3, C4.5生成算法：

ID3使用信息增益，c4.5使用信息增益比

C4.5生成算法描述：

1. 先从特征集A中选取信息增益比的特征Ag，
2. Ag的信息增益比小于阈值，将D作为一个单节点，返回
3. 按照Ag可能的取值ai将训练数据D进行分割Di。
4. 将分割后的Di的实例数中最大的类，作为A的label。
5. 按照每个可能的ai作为一个子节点构造树。子节点的D为Di，特征集为A-{Ag}
6. 如果不满足停止条件，对每个子节点，进行（1-5）的操作

停止条件：

* Di的数据为同一类
* 特征集为空
* 算法中的第2步也算是停止条件之一。

关于ID3和C4.5的历史：

**ID3**算法（Iterative Dichotomiser 3 迭代二叉树3代）是一个由[Ross Quinlan](http://zh.wikipedia.org/w/index.php?title=Ross_Quinlan&action=edit&redlink=1)发明的用于[决策树](http://zh.wikipedia.org/wiki/%E5%86%B3%E7%AD%96%E6%A0%91)的[算法](http://zh.wikipedia.org/wiki/%E7%AE%97%E6%B3%95)。

[C4.5](http://zh.wikipedia.org/w/index.php?title=C4.5&action=edit&redlink=1)算法，它同时也是ID3的升级版

ID3 采用自顶向下不回溯的策略搜索全部的属性空间，它建立决策树的算法简单，深度 小，分类速度快。

传统的 ID3 算法选择某个属性 A 作为测试属性的原则是使得信息增益 Gain(A)最大。

研究表明这种方法存在一个弊端：**算法往往偏向于选择取值较多的属性**，因 为加权和的方法使得实例集的分类趋向于抛弃小数据量的数据元组， 然而取值较多的属性却 不总是最优的属性，即按照使熵值最小和信息增益最大的原则被 ID3 算法列为应该首先选 取的属性在现实情况中却并不那么重要，也就是说对这些属性进行测试不会提供太多的信 息，例如：在股票市场，个股分析需要对某些少量的元素组有足够大量重视，而用 ID3 则 会忽略个股的重要属性[7-8]。

C4.5 也是对 ID3 的改进算法 ID3 算法不能处理数值属性、残缺值等问题，C4.5 算法从这几个方面很好的弥补了 ID3 算法的不足，并且还加入了剪枝等处理方法，使得精度有了很大的提高。

# cart决策树剪枝

损失函数



T叶节点个数，H(T)经验熵，反应对训练数据一个节点的不确定度



C(T)表示了模型对训练数据的预测误差



正则化的经验损失理解：

C(T)最小化误分类和

a|T|降低模型复杂度

剪枝，就是当a确定时，选择损失函数最小的模型

# CART算法

## 回归树：（最小二乘回归树）



使用平方误差来表示对于训练数据的预测误差



Cm的最优值是Rm上的所有输入实例xi对应的输出yi的平均值（公式5.17）。

最小二乘回归数，就是对不同的xi进行切分

切分公式（5.19）



J为切割分量，s为切割点。先固定j求s，然后对于所有的结果，求最小值。

**生成算法：**

1. 变量所有j，求出固定j对应的最小值的s，多有所有的j，s组合，求最小值
2. 按照j，s划分训练集，并求c1，c2，
3. 重复1,2，直到满足条件（例如。平方误差小于某个值，小于某个值）
4. 得到M个R区间和M个C值

## 分类树

基尼指数：

特征A条件下集合D的基尼指数定义为：（公式5.25）

基尼指数也表示的集合的不确定性。

Cart生成算法：

（1）对一个特征A可能的取值a，把D分割为D1和D2，并求分割后的基尼系数

（2）对训练数据集合D，每一个特征，以及特征的每一个可能取值，求基尼系数，并去最小的，然后将D划分为D1,D2

（3）对D1和D2使用（1），（2），直到满足停止条件

（4）这样就得到了一颗树。(叶节点的label，对应其训练集合D中，最多的label)

## Cart剪枝

计算每一种可能子树，取最优的结果

子树损失函数：



C(T)为训练数据的预测误差（如基尼指数）

对于固定的a，Ca(T)最小子树Ta是唯一的。

a大时，Ta子树偏小，a小时，Ta子树偏大。

裁剪一个以t为根的子树，对应的a的计算方式：



Tt表示以t为根的子树。

G(t)表示剪枝后的整体损失函数减少的程度

进行剪枝时，以多数表决的方式决定t节点label，

剪枝算法：

1. 对决策树的每一个内部节点，计算一个a值g(t)，这样就得到一个关于a的序列
2. 将a的序列按照从小到大进行排列
3. 递归从a1到an剪枝对应的以其内部节点的子树，得到一个子树序列
4. 使用交叉验证的方式验证子树序列，取最优的。（k折交叉验证，附件ppt）