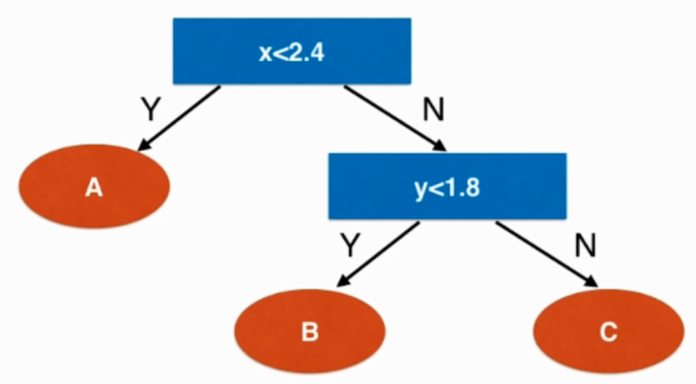
# 决策树分类及回归:

将样本依次在不同维度的不同值上进行区分,将各样本划分成不同的类型.从而达到分类的效果.如果需要变成回归,只需要将划分为同一类的样本进行各维度的平均值计算即可完成回归预测的值.

决策树有两个东西需要确定:

1.按哪个维度进行区分.

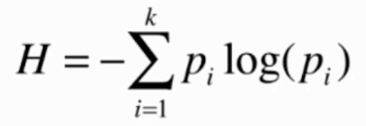
2.按这个维度的哪个值进行区分.



# 信息熵:

熵在信息论中代表随机变量不确定度的度量,信息熵越大,数据的不确定性越高, 信息熵越小,数据的不确定性越小.可以连线物理热力学中的熵(热能量)和分子运动不确定性的关系.在数学中,各值概率都相差不大或者都相等,熵越大(不确定度越高不知道会落哪个),各值概率相差很大,比如有一个值的概率为99% 则熵越小,即样本不确定度越低.

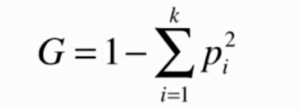
熵的公式,其中p为各样本个数所占总体的数量比例.



有了信息熵的概念,我们可以以划分后两个类别的信息熵和(y标)最小.对于划分后有分类信息熵如果不为0 则表明可以继续划分下去.

# 基尼系数:

公式:



基尼系数和熵的几何形态和信息熵类似,都是抛物线一样(信息熵为类抛物线),在通常情况下,基尼系数和信息熵的预测效果相同,但计算复杂度稍小.

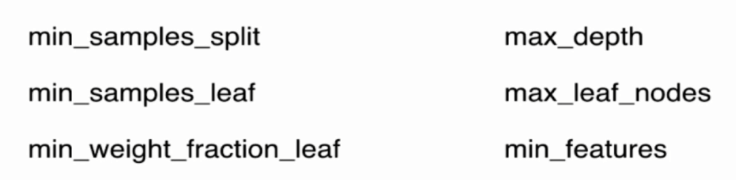
# cart(classification and regression tree)

在生成决策树的过程中,可以有cart,C4.5等方法不同的方法意味着有不同的超参数来进行调节,决策树是一种非常容易过拟合的算法,一般都要求设置参数来防止过拟合.

min\_samples\_split:最小样本分类(到达最小样本后不再继续分)

max\_depth: 树的最大深度

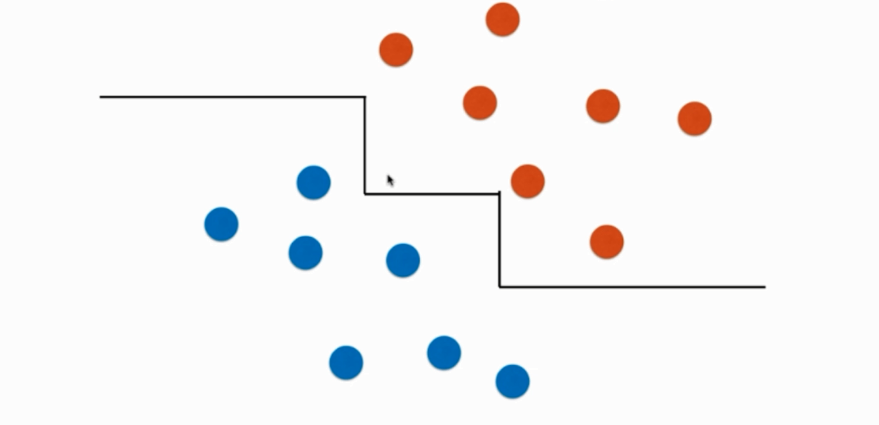
等等:sklearn中实现的决策树有以下参数可以调节.



# 决策树缺点:

决策树的分类和回归效果一般来说都不会很好,有很多的制约性主要是由以下缺点:

1.由于决策树分类是根据一个维度的一个值进行的分类,所以其决策边界只能是平行各维度的直线,如果样本的分类是条横线,决策树可能只能用楼梯形状来进行分类.



2.容易受极值点影响(所有非参数学习的通病)

决策树单纯的机器学习算法效果一般,但其思想是非常重要的,在集成学习中得到广泛的应用并得到了很好的效果.