机器学习总结：

决策树：

Hands-On Machine learning 第171页

+ 统计的学习方法 CART算法

CART training algorithm:

Classification and regression tree:

cost function the algorithm tries to minimize:

$$J(k, t_k) = \frac{m_{left}}{m} G_{left} + \frac{m_{right}}{m} G_{right}$$

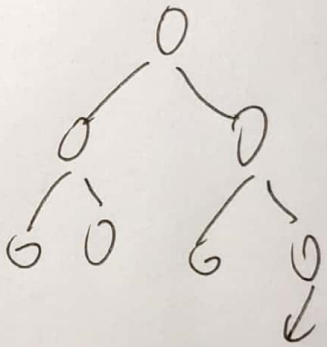
↳ left/right: measures the impurity of the left/right subset.

左边第data个数 / 整个data个数

决策树的基石：

↳ Gini impurity

$$G_i = 1 - \sum_{k=1}^{n} P_{i,k}^2$$



里面有5个元素

属于类1的：0.

类2：49

3：5.

the ration

$$G = 1 - \left[\left(\frac{0}{54}\right)^2 + \left(\frac{49}{54}\right)^2 + \left(\frac{5}{54}\right)^2\right]$$

$$= 0.168$$

↓

不纯的度量 Gini impurity 为 0.168.

CART算法：先把 training set 分成两个 subset.

怎么分呢，优化 $J(k, t_k)$ 见上. 使其 minimize.

→ 再分别将左右两颗树分成2个 subsets.

直到达到最大深度和或不能减少 impurity 3.

另一种衡量 impurity 的方法：entropy

$$H_i = -\sum_{k=1,\ P_{i,k}\neq 0}^{n} P_{ik} \log(p_{i,k})$$

比较 entropy 和 Gini impurity：

区别不大。

Gini 倾向于将最频繁出现的 class 分出到一个单独的分支中。

entropy 倾向于产生均衡一点的 tree。

两边 number 差不多。