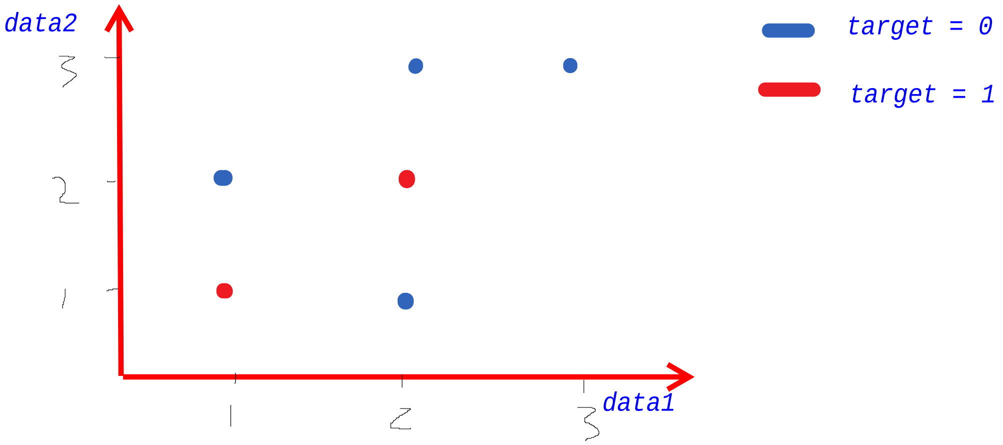
decision tree and random forest



无限分隔，直至每个区域只有一个点，然后可以准确判断他们的target

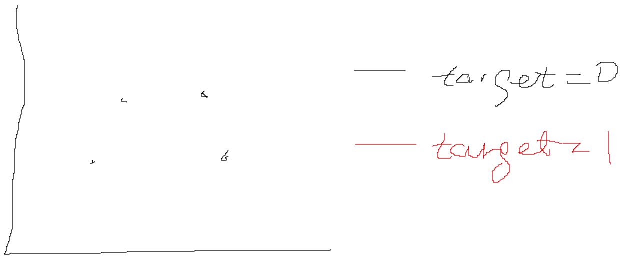
$$I_G(n) = 1 - \sum_{i=1}^{J} (p_i)^2$$

Gini Impurity

$$I_{root} = 1 - \left[\underbrace{\left(\frac{2}{6}\right)^2}_{\text{if target}=1} + \underbrace{\left(\frac{4}{6}\right)^2}_{\text{if target}=0} \right] = 1 - \frac{20}{36} = \frac{4}{9}$$

without classification line

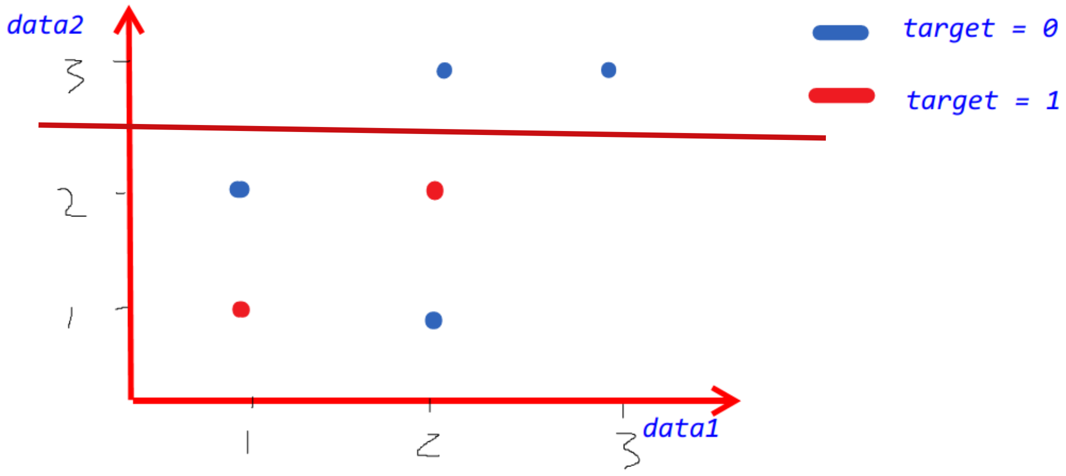
if there's only one type of data in the group, then I(root) = 0. for example: target=0



$$I_{root} = 1 - \left[\left(\frac{4}{4}\right)^2 + \overbrace{\left(\frac{0}{4}\right)^2}^{\text{target}=1} \right]$$

$$= 1 - 1$$

$$= 0$$



if we have a classification line(red line)

gini impurity for the upper part   lower part

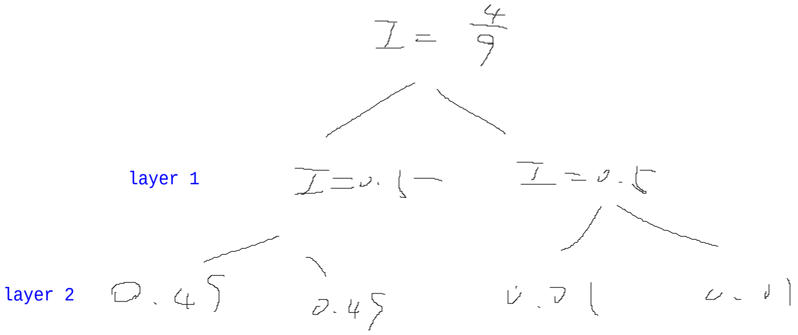
$$I = \left(1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right]\right)\times\frac{2}{6} + \left(1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right]\right)\times\frac{4}{6}$$

$$= \frac{1}{3} < \frac{4}{9}$$

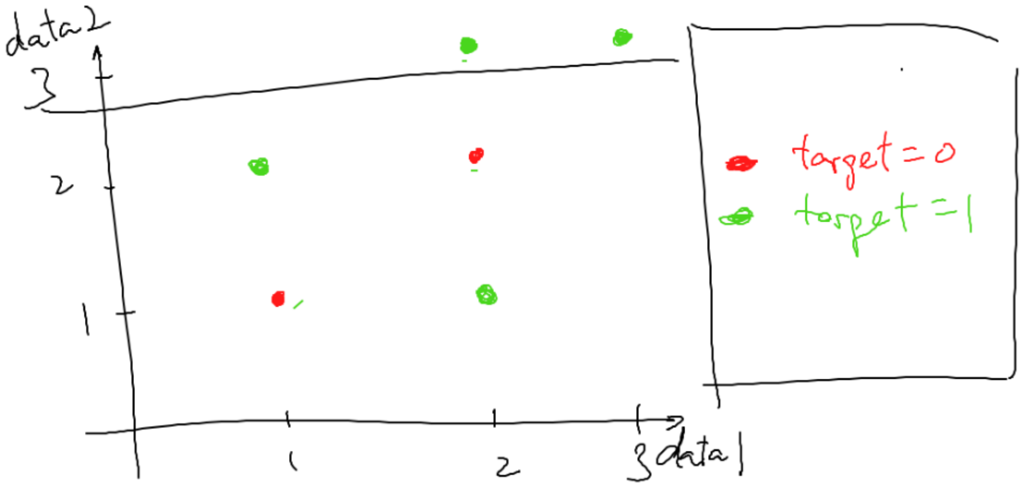
so, every time you draw a classification line, it'll be better

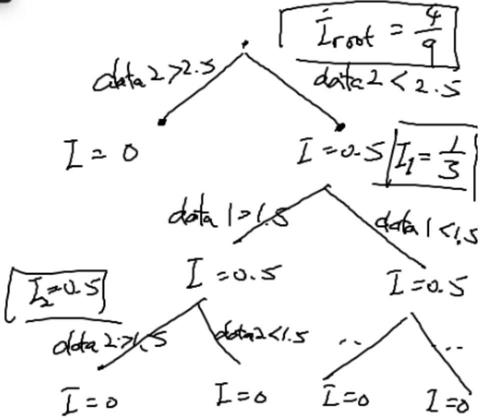
Notice, sometimes, we don't need to let I = 0. (overfitting)
if two points are different fundamentally, then it is necessary to add the line,
however, if they are different because other noise, then there's no need to add the line.

In practice, if you find the next level of Gini impurity improves a little, then there's no need to add line any more.



normally, we can limit the number of layer in each tree to solve over fitting problem. ie, if we allow 1 layer, then, layer 2 will not appear.





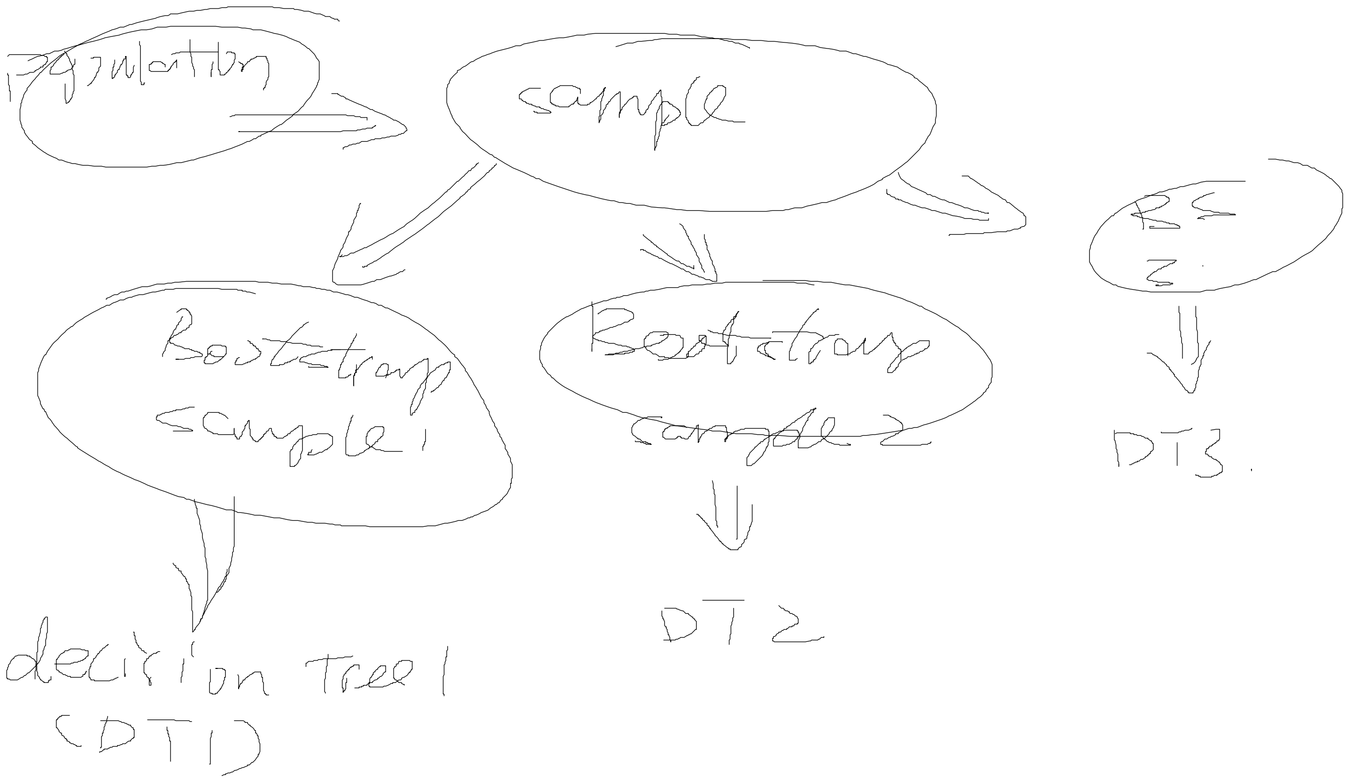
Gini Impurity

$$\boxed{I_G(n) = 1 - \sum_{i=1}^{J} (p_i)^2}$$

$$I_{root} = 1 - \left(\left(\frac{2}{6}\right)^2 + \left(\frac{4}{6}\right)^2\right) = 1 - \frac{20}{36} = 1 - \frac{5}{9} = \frac{4}{9}$$

$$I_1 = \left(1 - \left(\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right)\right) \cdot \frac{2}{6} + \left(1 - \left(\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right)\right) \cdot \frac{4}{6}$$

$$= \frac{1}{3}$$



as bootstrapping is random, each of them have a decision tree. these decision tree are random. this is random forest