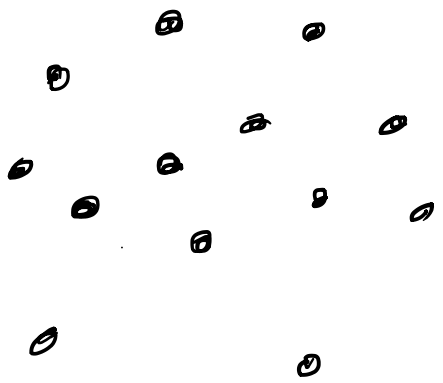
Decision Tree

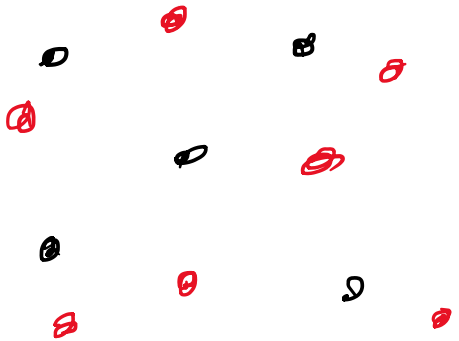
Classification & Regression

- For a given dataset, there can be multiple split conditions
- But then how does decision tree decide which one to choose
- To human eye it visible but how to mathematically do that
- We use concept of Entropy and Information Gain

Entropy: The measure of randomness



Low entropy

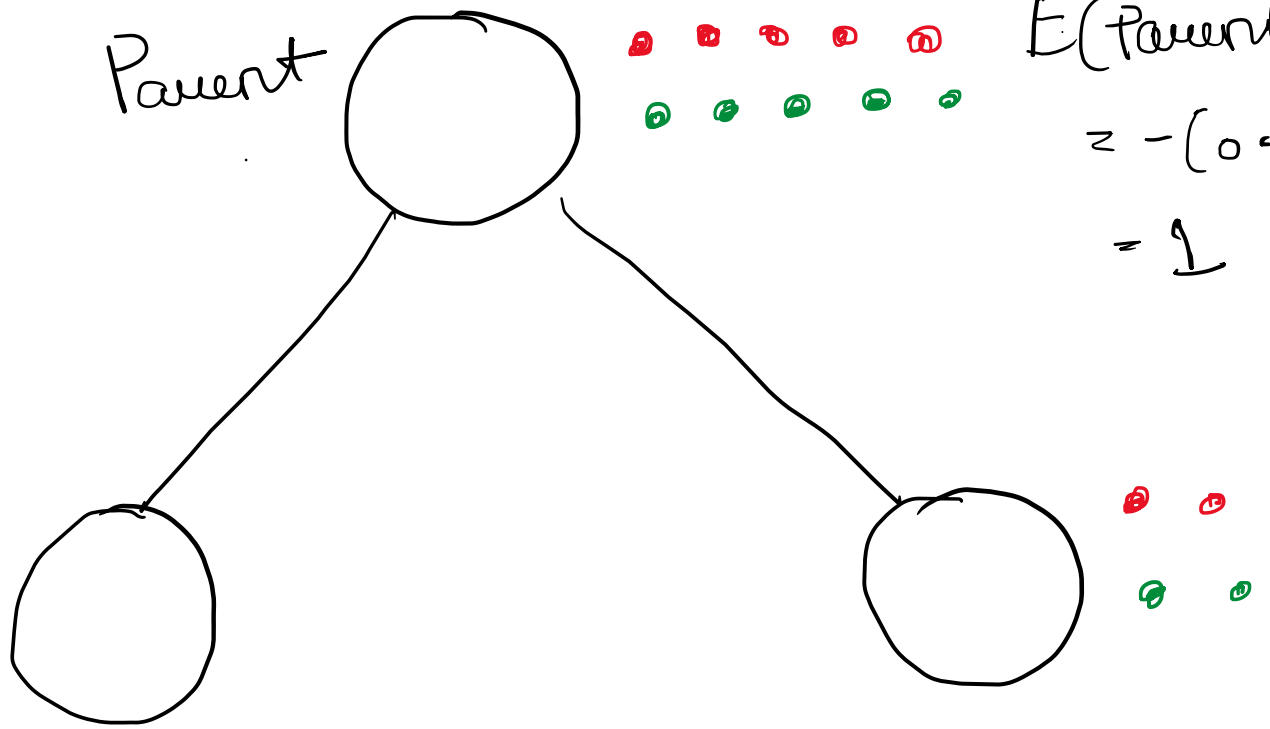


High entropy

$$\text{Entropy} = \sum_{i=1}^{C} -p_i \log(p_i)$$

Example

1st possibility



$E(\text{Parent}) = -(0.5)\log(0.5) - (0.5)\log(0.5)$

$= 1$

$E(C_1) = -\frac{3}{6}\log\left(\frac{3}{6}\right) - \frac{3}{6}\log\left(\frac{3}{6}\right)$

$= 1$

$E(C_2) = -\frac{2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log\left(\frac{2}{4}\right)$

$= 1$

$\therefore$

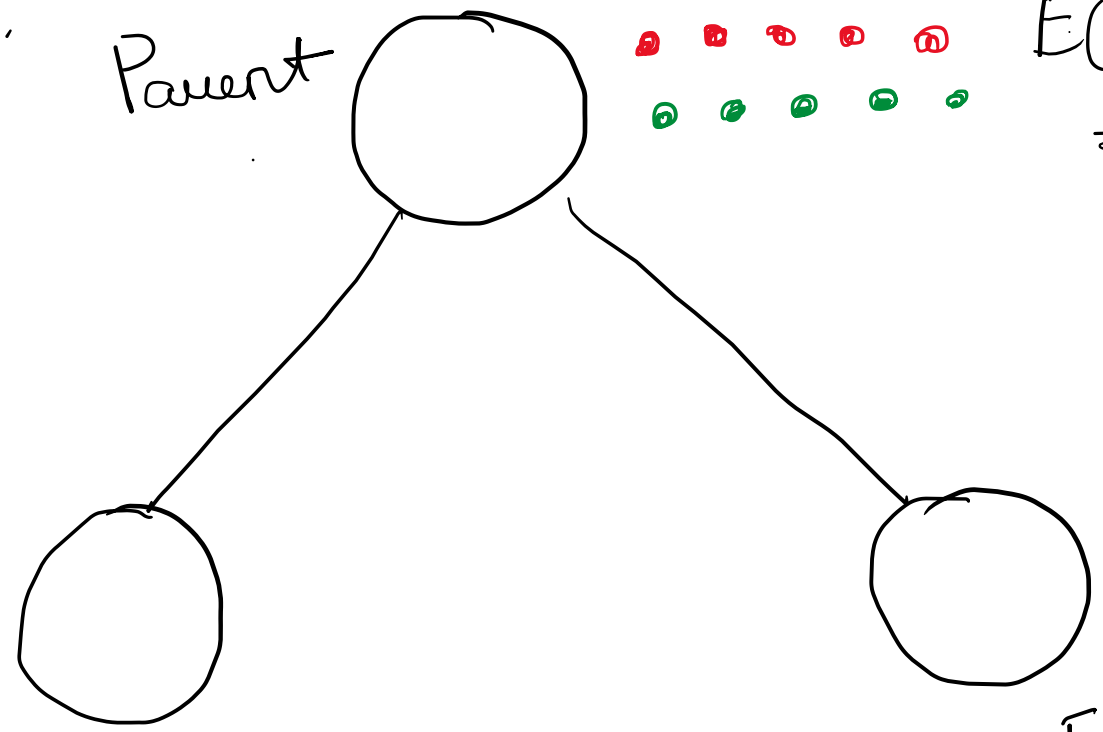
$$\text{Information Gain} = E(\text{parent}) - \sum w_i E(\text{child}_i)$$

$$= 1 - \left[\frac{6}{10} \times 1 + \frac{4}{10} \times 1 \right]$$

$$= 1 - \left[\frac{6}{10} + \frac{4}{10} \right] = 1 - \frac{10}{10} = 1 - 1 = 0$$

Example

2nd possibility



$E(\text{Parent}) = -(0.5)\log(0.5) - (0.5)\log(0.5)$

$= 1$

$E(C_1) = -1\log(1)$

$= 0$

$E(C_2) = -\frac{1}{6}\log\left(\frac{1}{6}\right) - \frac{5}{6}\log\left(\frac{5}{6}\right)$

$= -(0.17)(-0.78) -$

$(0.83)(-0.08)$

$= 0.132 + 0.06$

$= 0.192$

Information Gain $= E(\text{parent}) - \sum w_i E(\text{child}_i)$

$$= 1 - \left[0 \times \frac{4}{10} + 0.192 \times \frac{6}{10} \right]$$

$$= 1 - 0.1152 = 0.8848$$

$\therefore$ $IG(1^{st}) < IG(2^{nd})$ $\therefore$ 2nd split it is

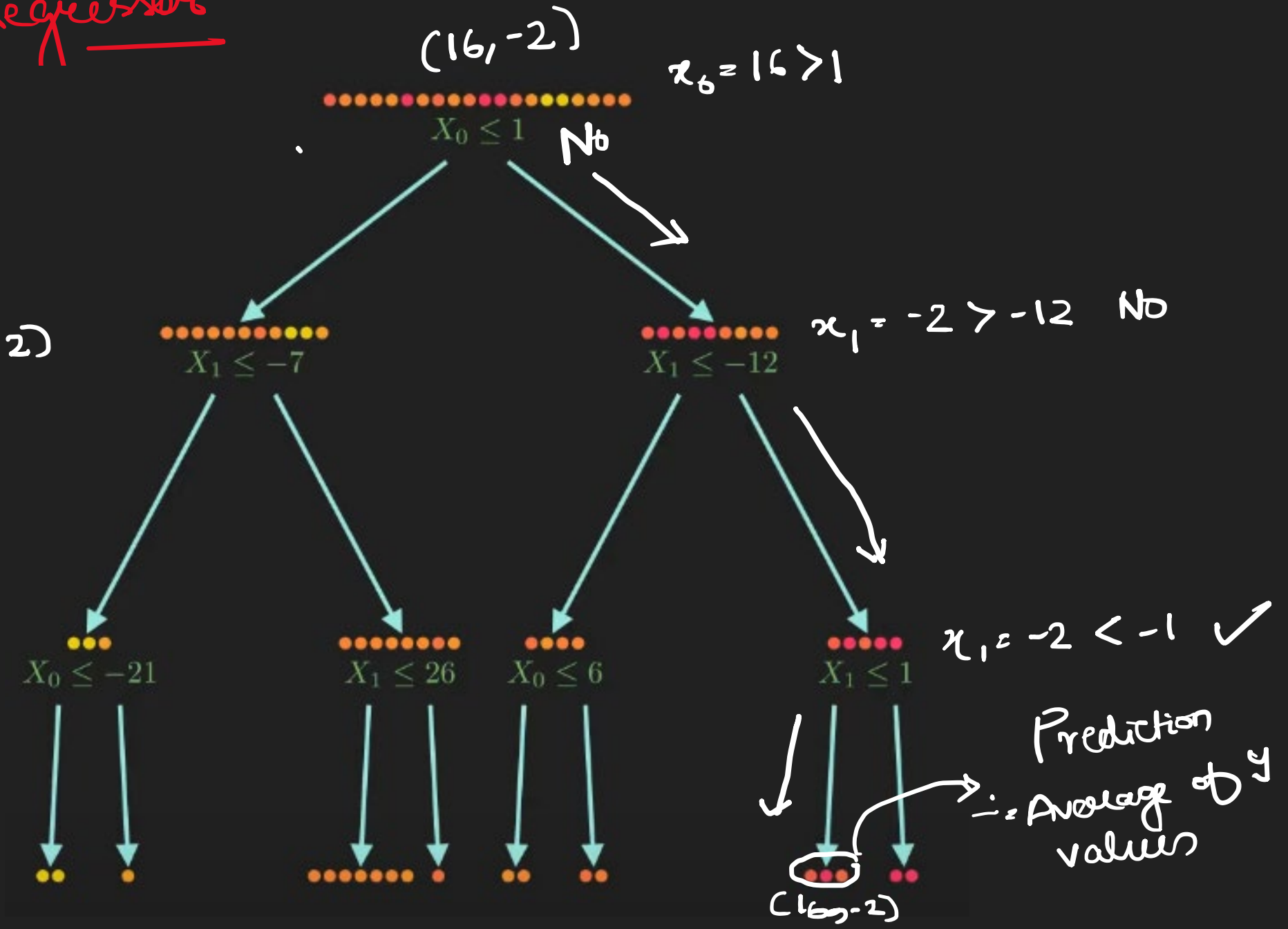
This way the algorithm calculates all the way for all possible splitting possibilities.

∴ Hence we get our best split and decision tree

Decision tree Regressor

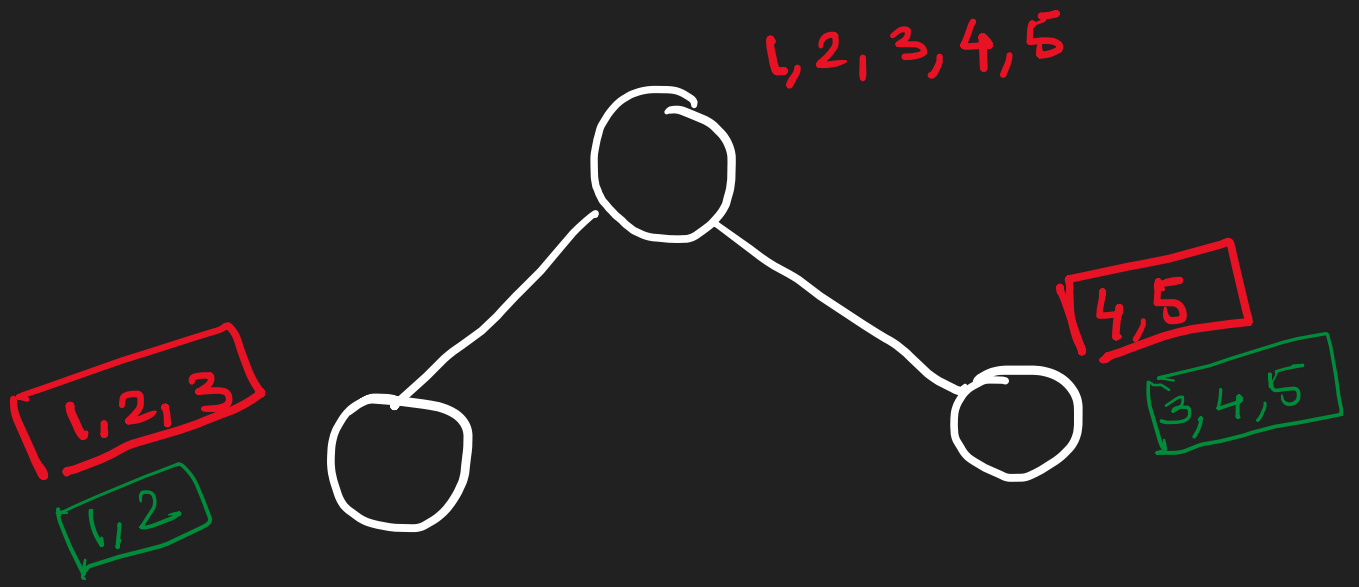
Suppose we have a new point

$(x_0, x_1) = (16, -2)$

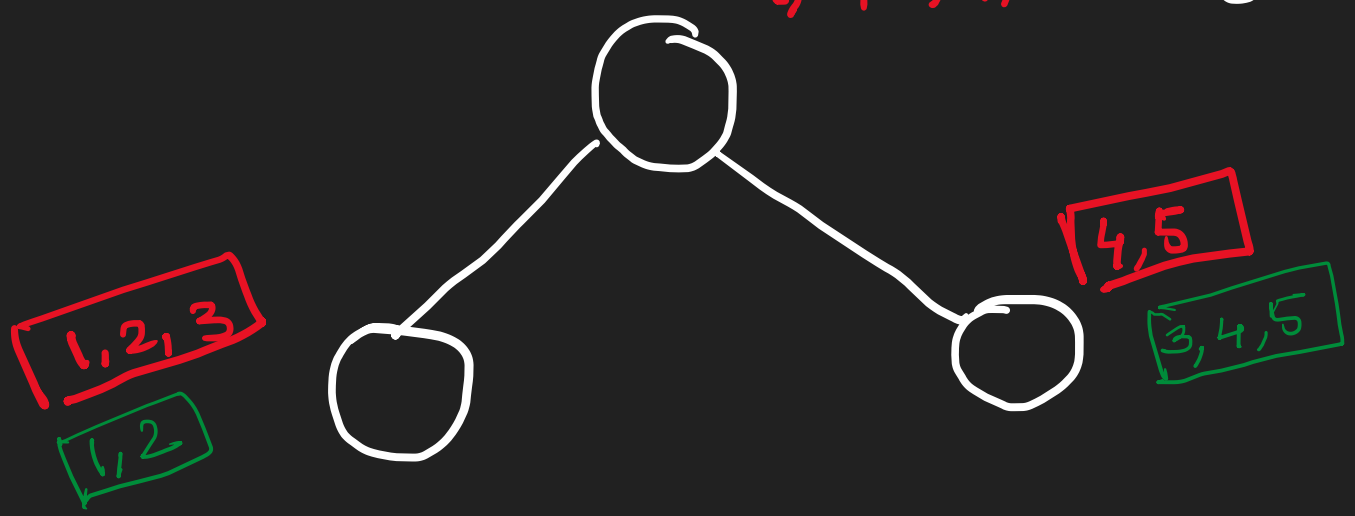


But how to split

We use a parameter called variance reduction



$$\bar{y} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



$$\text{Var (Parent)} = \frac{1}{n}\sum (y_i - \bar{y})^2 = \frac{1}{5}\left[(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2\right]$$

$$= \frac{1}{5}\left[4+1+0+1+4\right] = \frac{10}{5} = 2$$

1,2,3 → V_1

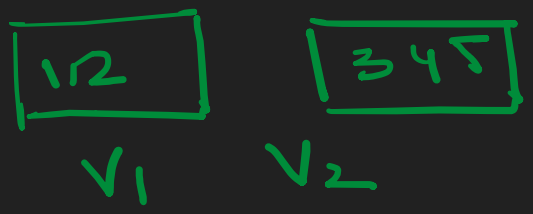
4,5 → V_2

V_1 mean $= \frac{1+2+3}{3} = 2$

V_2 mean $= \frac{4+5}{2} = 4.5$

$$V_1 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{3}\left[(1-2)^2+(2-2)^2+(3-2)^2\right] = \frac{1}{3}\left[1+0+1\right] = \frac{2}{3} = 0.67$$

$$V_2 = \frac{1}{2}\left[(4-4.5)^2+(5-4.5)^2\right] = \frac{1}{2}\left[0.25+0.25\right] = \frac{1}{2}\times 0.5 = 0.25$$



V_1 mean $= \frac{1+2+3}{2} = 3/2 = 1.5$

V_2 mean $= \frac{3+4+5}{3} = \frac{12}{3} = 4$

$$V_1 = \frac{1}{2}\left[(1-1.5)^2 + (2-1.5)^2\right] = 0.25$$

$$V_2 = \frac{1}{3}\left[(3-4)^2 + (4-4)^2 + (5-4)^2\right] = \frac{1}{3} * 2 = \frac{2}{3} = 0.67$$

For split 1

$$\text{Var Red} = \text{Var}(\text{Parent}) - \sum w_i V(\text{child}_i)$$

$$= 2 - \left[\frac{3}{5} \times 0.67 + \frac{2}{5} \times 0.25 \right]$$

$$= 2 - [0.402 + 0.1]$$

$$= 2 - 0.502 = 1.498$$

For split 1

$$\text{Var Red} = \text{Var}(\text{Parent}) - \sum w_i V(\text{child}_i)$$

$$= 2 - \left[\frac{2}{5} \times 0.25 + \frac{3}{5} \times 0.67 \right]$$

$$= 2 - [0.402 + 0.1]$$

$$= 2 - 0.502 = 1.498$$

So Variance Reduction for both splits are same. So we can select any one of them for our split.