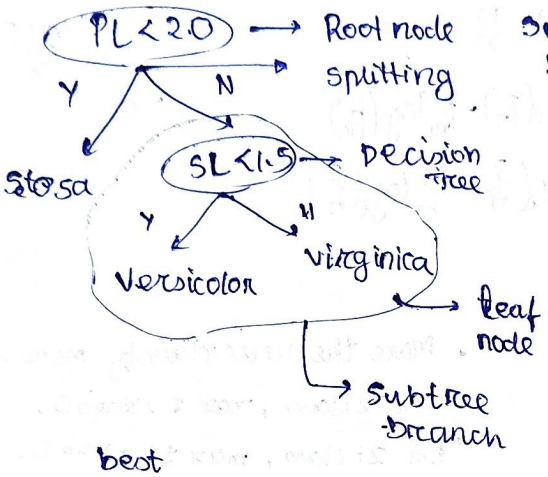
Decision Tree

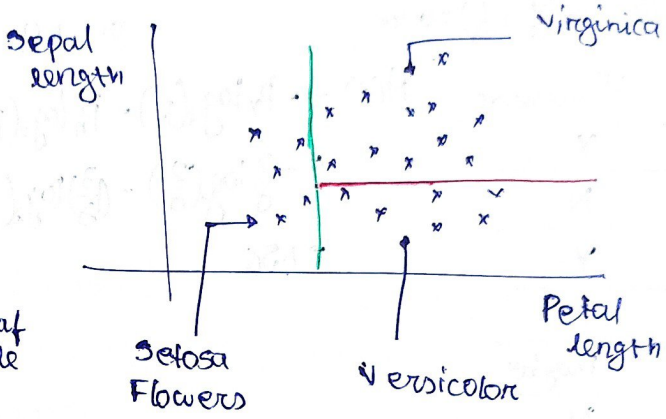
P-1

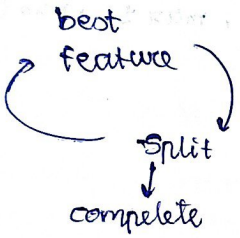
8:49 pm
Thu may 1

Gender	Occupation	Suggestion
F	student	Game
F	programer	Github
M	"	whatsapp
F	"	github
M	student	Game
M	Game	"

```
if occupation == Game
    print("Game")
Else
    if gender == Female
        print (Github)
    Else
        print(whatsapp)
```







Q
- which is best features for splitting.
- splitting criteria ?

CART - Classification & Regression Tree

Entropy

- measure of purity/impurity

$$E(S) = \sum_{i=1} - p_i \log_2 p_i$$

Ex- if data have two classes

$$E(D) = -p_{yes} \log_2(p_{yes}) - p_{no} \log_2(p_{no})$$

where

p_i is simply the frequentist probability of an element/class 'i' in our data

Bggg	Purchase
21	Y
51	N
69	Y
86	N
76	N

Purchase
Y
N
N
N
N

Pur
N
N
N
N
N

$H(d) = -P_Y \log_2(P_Y) - P_n \log_2(P_n)$

$= -2/5 \log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5})$

$= 0.97$

$G_1 = 1 - (\frac{4}{25} + \frac{9}{25}) = 0.48$

$H(d) = -P_Y \log_2(P_Y) - P_n \log_2(P_n)$

$= -\frac{1}{5}\log(\frac{1}{5}) - \frac{4}{5}\log_2(\frac{4}{5})$

$= 0.72$

$G_1 = 1 - (\frac{1}{25} - \frac{16}{25}) = 0.72$

$H(d) = -P_Y \log_2 P_Y - P_Y \log_2 P_Y$

$= -\frac{0}{5}\log(\frac{0}{5}) - \frac{5}{5}\log_2(\frac{5}{5})$

$= 0$

Purchase
Y
N
Y
N
maybe
N
maybe
maybe

$H(d) = -P_Y \log_2(P_Y) - P_n \log_2(P_n) - P_m \log_2(P_m)$

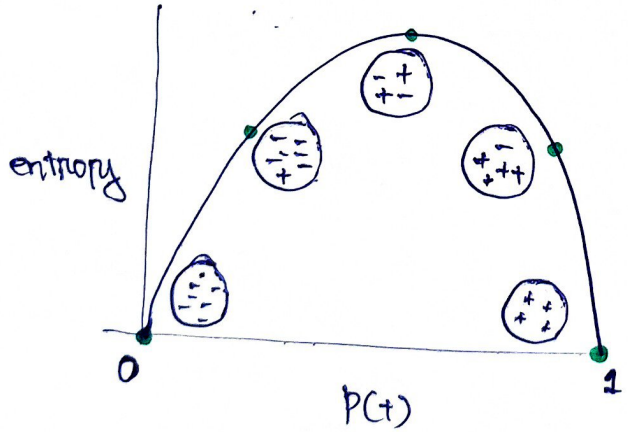
$= -\frac{2}{8}\log(\frac{2}{8}) - (\frac{3}{8})\log_2(\frac{3}{8}) - \frac{3}{8}\log_2(\frac{3}{8})$

$= 1.56$

- More the uncertainty more is entropy
- For 2 class, max 1, min 0.
- For 2+ class, max 1+, min 0.

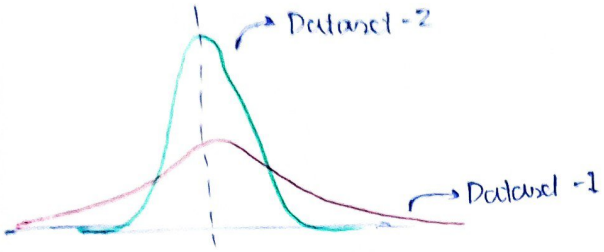
Y → 2
N → 3
m → 3

Entropy V/s Probability



Entropy for Continuous variables

1. The more you know about the data then less entropy.



which datasets have higher entropy?
- whichever is less peaked.

Information Gain

- is a metric used to train Decision Trees.

$$\text{information gain} = E(\text{Parent}) - \{\text{weighted Avgerage}\} - E\{\text{children}\}$$

Step 1 Entropy of Parent means full dataset. $E(P)=0.97$

S2 Calculate for child

Outlook

Sunny	Overcast	Rain
$E(S)=0.97$	$E(O)=0$	$E(S)=0.7$

S3. weighted Entropy of children

$$W.E(\text{children}) = \frac{5}{14}0.97 + \frac{4}{14}\times 0 + \frac{5}{14}\times 0.97$$

$$= 0.69$$

S4 Info. Gain $= 0.97 - 0.69 = 0.28$

So information Gain, when we split this data on the basic of outlook column is 0.28

S7. highest Entropy split on that column

Recursively

when Entropy $=0$ then stop

Gini Impurity

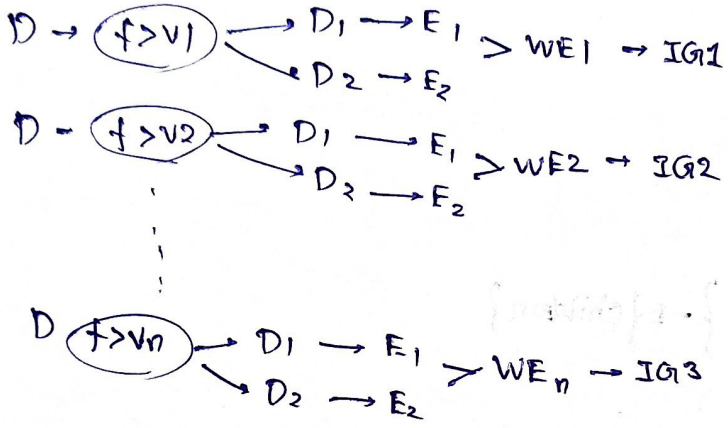
$$G_I = 1 - (P_Y^2 - P_N^2)$$

Computionaly fast

Numerical Data

S1. Sort data on the basis of one column

S2. split entire data on the basis of every value of that column



S3. Max {IG1, IG2 .... IGn}

S4. Do recursively

Expensive

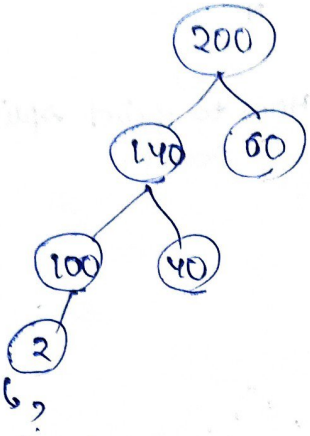
- in Train time
- not in test time

Hyperparameters

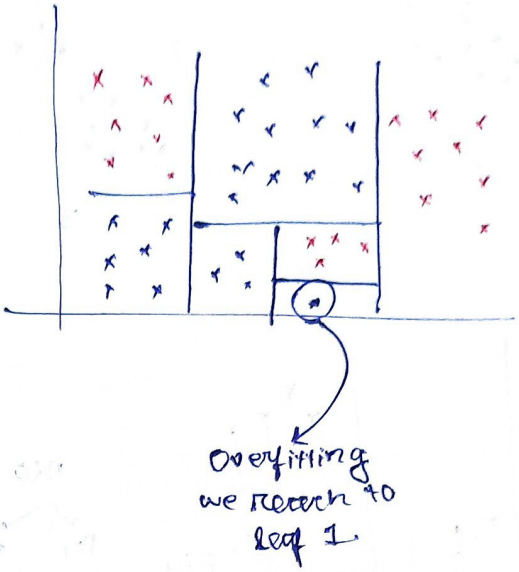
Overfitting

Performs well : Train
badly : Test

max_depth = None



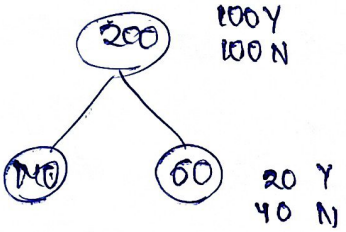
Noisy / Outlier
Erroneous



Overfitting we reach to leaf 1.

Underfitting

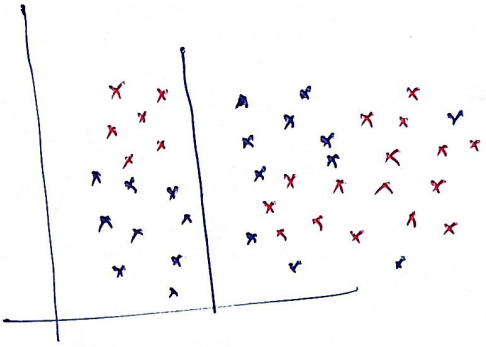
max_depth = 1



100 Y
100 N

80 Y
60 N

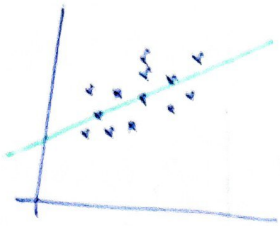
20 Y
40 N



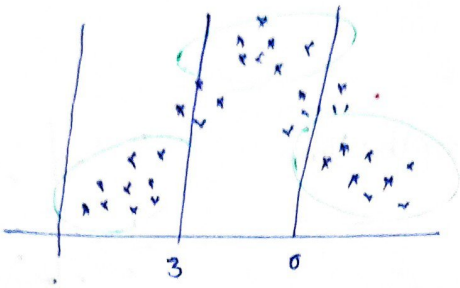
P3

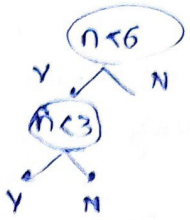
858 AM
Fri May 2

Regression Trees

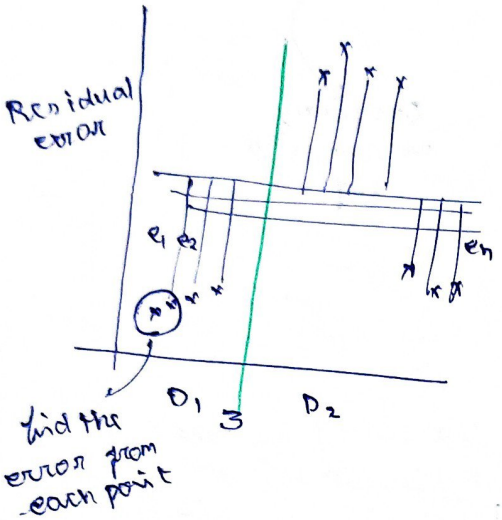


non-linear





How to find splitting point;

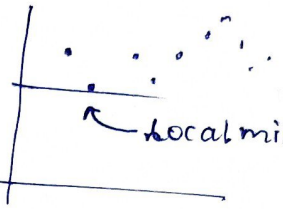


n<0.5, 1.5, 3

D1 D2

mean

$SSE_1 = e_1^2 + e_2^2 + \dots e_n^2$



↳ Here we do splitting