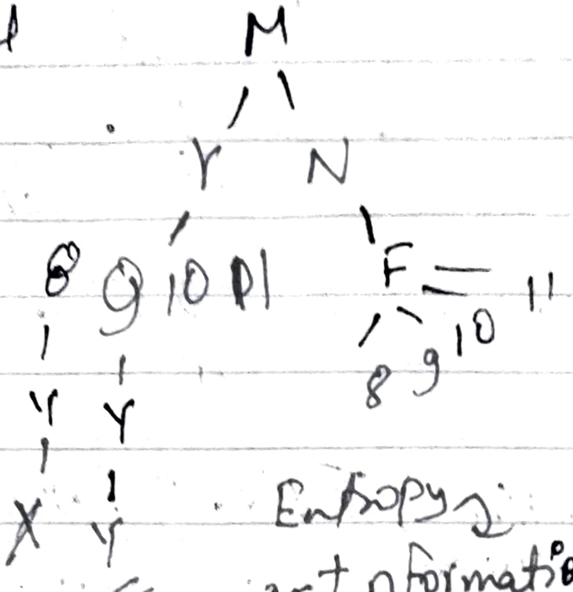
FPR

Decision Tree

classifier as well as Regression task.
Powerful than both Linear Regr. & Logic. Regr.

Class	Gender	Stay in Hostel
9	M	Y
10	F	N
8	F	Y
8	F	N
9	M	Y
10	M	N
11	F	Y
11	M	Y
8	F	Y
9	M	Y
11	M	N
10	F	N
10	M	Y
8	F	?



Entropy
Information Gain
Genny.

PS How to build a tree?

$$Ginni = 1 - \sum_{i=1}^{c} (P_i)^2$$

Class	S1H	Total Value	P(Y)	P(N)
8	Y=2 N=1	3	2/3	1/3
9	Y=2 N=1	3	2/3	1/3
10	Y=1 N=3	4	1/4	3/4
11	Y=3 N=1	4	3/4	1/4
		T= 14		

Ginni for 8th Class $= 1 - [P(Y)^2] - [P(N)^2]$

$$= 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 4/9$$

$$G(9) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 4/9$$

$$G(10) = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = \frac{6}{16} = \frac{3}{8}$$

$$G(11) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = \frac{6}{16} = 3/8$$

$$G(\text{entire colomn}) = \sum \frac{\text{No. of instance of Class}}{\text{Total No. of instance}} \times G(\text{respected class})$$

or G(C)

$$G(C) = \frac{n_8}{T} \cdot G(8) + \frac{n_9}{T} \cdot G(9) + \frac{n_{10}}{T} \cdot G(10) + \frac{n_{11}}{T} G(11)$$

$$= \frac{3}{14} \cdot \frac{4}{9} + \frac{3}{14} \cdot \frac{4}{9} + \frac{4}{14} \cdot \frac{6}{16} + \frac{4}{14} \cdot \frac{6}{16}$$

$G(C) = 0.404.$

Ginni = Ginni Impurity.

Gender	S/H	TV	P(Y)	P(N)
M	Y=5 N=3	8	5/8	3/8
F	Y=3 N=3	6	1/2	1/2
		14		

$G(M) = 1 - \left(\frac{5}{8}\right)^2 - \left(\frac{3}{8}\right)^2 = 15/32 = 0.468$

$G(F) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$

~~G(c) =~~

$G(\text{Gender}) = \frac{8}{14} \cdot 0.468 + \frac{6}{14} \cdot \frac{1}{2} = \underline{0.482}$

⇒ Select Ginni Impurity which is low.
(select class column).

Root node = Less Ginni value = class column.

Class

8 9 10 11

8	F	Y
8	F	N
8	F	Y

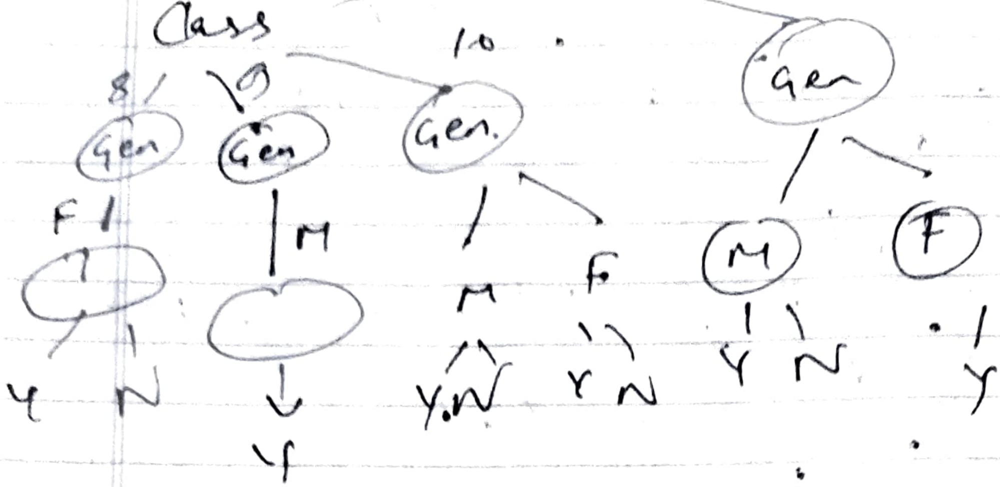
9	M	Y
9	M	Y
9	M	Y

10	F	N
10	M	N
10	F	N
10	M	Y

11	F	Y
11	M	Y
11	M	N
11	M	Y

Calculate Ginni for all

G



So.

8 F ? = 8 F

11 F ? ≡ 11 F Y

11 M ? ≡ 11 F Y

(According to Weightage)

Entropy & Information Gain

$$E = -\sum_{i=1}^{c} P \log_2(P) \qquad IG = \underbrace{E_{before}}_{\text{Label column}} - E_{After}$$

Gaining information after making a DT.

14 record $\quad n(Y) = 8 \quad n(N) = 6$

$$E(L) = -P(Y)\log_2(Y) - P(N)\log_2 P(N)$$

$$= -\frac{8}{14}\log\frac{8}{14} - \frac{6}{14}\log\frac{6}{14} = 0.98522$$

$E(8) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$

$E(9) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log 1/3 = 0.918$

$E(10) = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} = 0.811$

$E(11) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.811$

$I(\text{Class}) = \frac{3}{14} \times 0.918 + \frac{3}{14} 0.918 + \frac{4}{14} \times 0.811 + \frac{4}{14} \times 0.811 = \underline{0.8574}$

$IG = 0.98522 - 0.8574 = 0.1278.$
(class column)

$IG_{\text{Gender}} = 0.98522 - 0.974 = 0.01$

$E(M) = -\frac{5}{8}\log_2\frac{5}{8} - \frac{3}{8}\log_2\frac{3}{8} = 0.954$

$E(N) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$

IG(Gender)

(Before Gender) $= \frac{8}{14} 0.954 + \frac{6}{14} \times 1 = \underline{0.974}$ final.

ID3 C4.5 CART

$IG_{Gender} = 0.01$ $IG_{class} = 0.1278$.

* More IG = that is rootnode = Class Column

Which Class has high IG less randomness ?

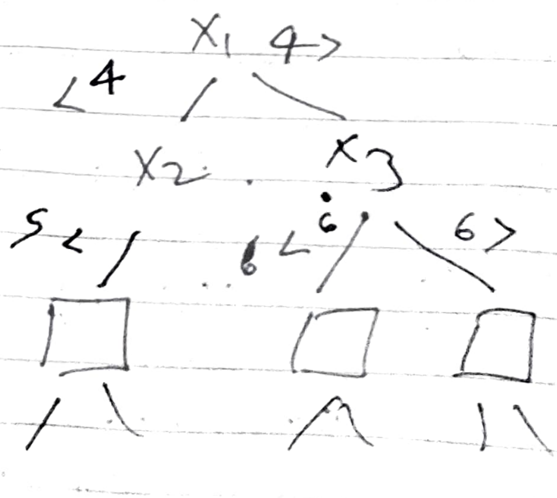
3 possibility

- Feature → categorical — Outcome → cat → classi
- Feature → conti — Outcome → cat → class
- Feature → conti. — Outcome → contin → Regression

ID3 ↔ Iterative dictomizer (Classifier)
C4.5 ↔
CART ↔ classi. & Regression Tree.

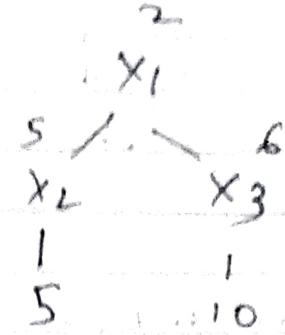
avg. <4< <5< <6<

X_1	X_2	X_3	Y
1.1	7.5	2.5	A
2.2	8.8	5.5	B
3	9.2	6	A
3.6	5.1	6.7	A
5	5.4	7	B
5.8	2	8.9	B
8	1	9.1	A



$\hat{y}$ is avg. of Record(y)

X_1	X_2	X_3	Y
1	2	1	3 3
2.4	5.8	6	10
1.9	6.1	7.2	12.8
5.8	2.8	9.6	13.9
7.2	7.8	8.9	19.8



$\Sigma(y-\hat{y})^2 + \Sigma(y-\hat{y})$

CART According to errors both side rearrangement of Tree occurs

In all these above algos. we have to determine threshold first (different mechanism for that).

To control under/over fitting we using pruning.

Q) Prepruning ~~prepruning~~ Post pruning ~~post pruning~~.

(backward)
Post pruning :- Build a comple dT then identify which branch is contributing to over/under fit issues. using Cross validation.

Cross validation :- Tune a parameter & removal of branches.

Backward pruning (Pre) :- Stop DT to create insignificant branches.

$\boxed{*}$

Ensemble - Techniques

* Combination of multiple algos. { Multiple. (Multidecision Maker) }

Q) Bagging   Boosting   Stacking

Q) Bootstrap Aggregation (Bagging).

Q) Pasting?

Q) Random Forest?

Bootstraping

Bagging
- SVM
- KNN
- NB
- DT
- RF