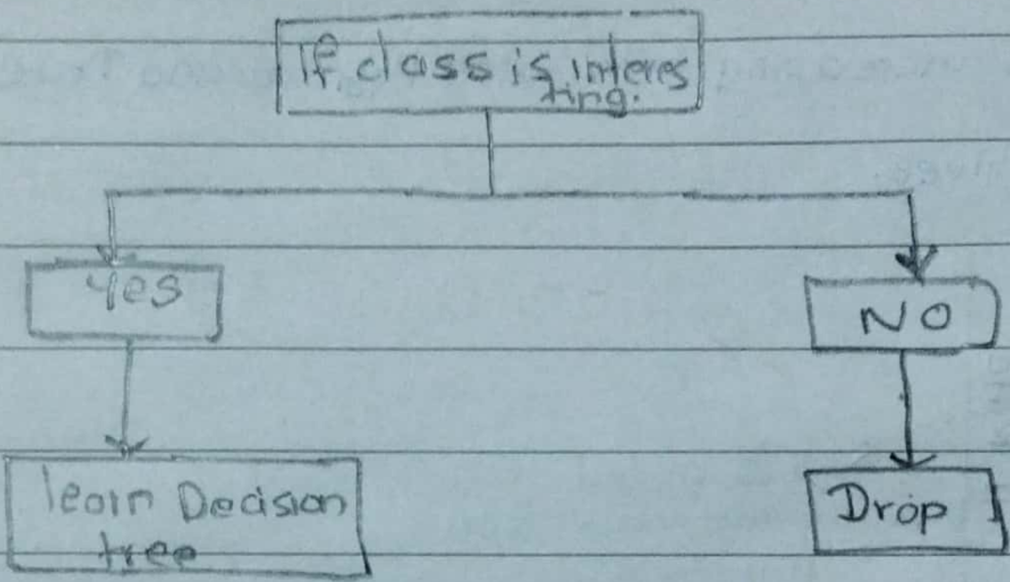
Decision Tree



→ Covers both Regression and classification

In Decision tree the major challenge is to identification of the attribute for the root node in each level. This process is known as attribute selection.

We have two popular attribute selection measures:-

① Information gain → use entropy to make decisions

② Gini Index

Mo Tu We Th Fr Sa Su

Memo No. ______

Date / /

Entropy :

Entropy is the measure of uncertainty in the data. The effort is to reduce the entropy and maximize the information gain.

* Feature having the most information is considered important by the algorithm and is used for training the model.

$$\boxed{Entropy = -\sum_{i=1}^{n} P_i \times \log(P_i)}$$

for binary

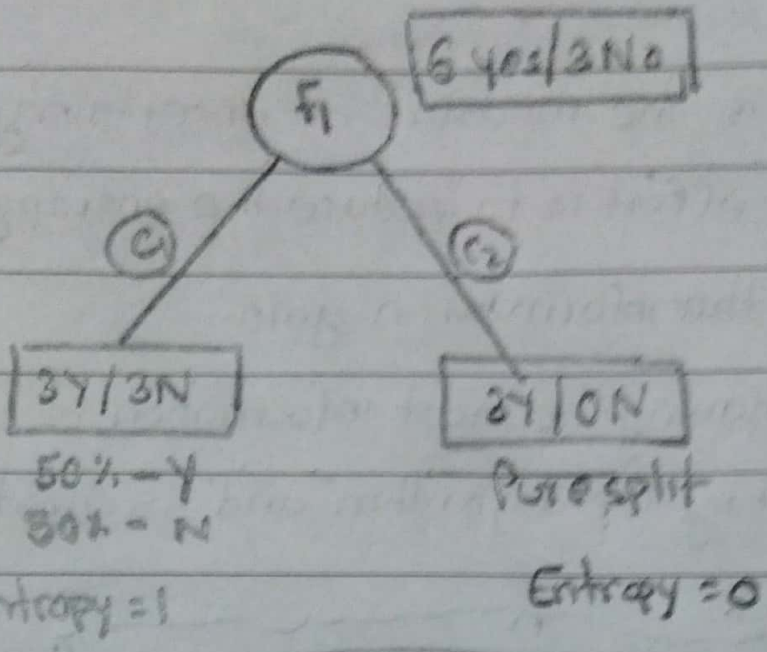
$$Entropy = -P_y \log(P_y) - P_N \log(P_N)$$

for multiclass

$$Entropy = -P_{c_1} \log(P_{c_1}) - P_{c_2} \log(P_{c_2}) - P_{c_3} \log(P_{c_3})$$

Memo No.
Date / /

Example



Entropy for C_1

class here yes & No

$$H(S) = -\sum_{i=1}^{n} P_i \times \log_2(P_i)$$

$$= -P_y \times \log_2(P_y) - P_N \times \log_2(P_N)$$

$$= -\frac{3}{6} \times \log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log\left(\frac{3}{6}\right)$$

$$= 1$$

Entropy for C_2

$$H(S) = -\sum_{i=1}^{n} P_i \times \log_2(P_i)$$

$$= -P_y \times \log_2(P_y) - P_N \times \log_2(P_N)$$

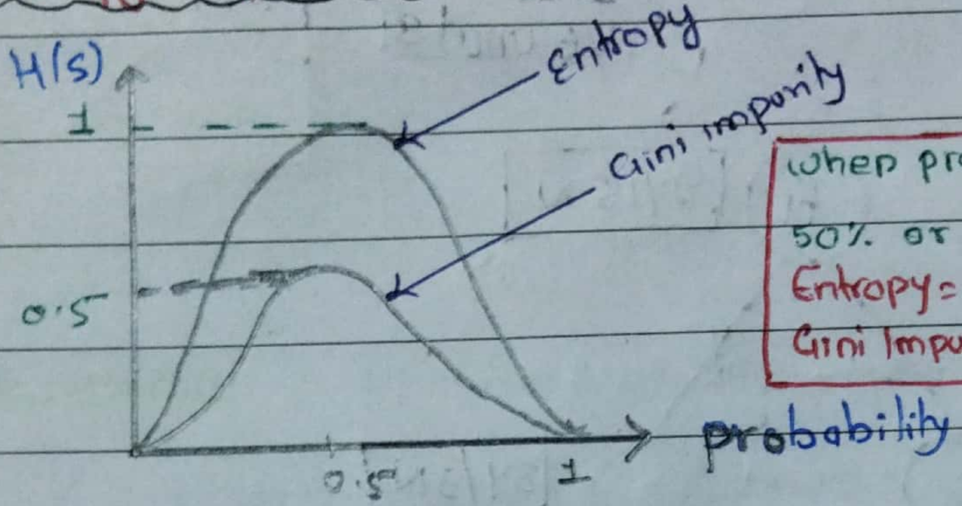
$$= -\frac{3}{3} \times \log_2\left(3/3\right) - 0$$

$$= 0$$

Memo No. ______

Date / /

Entropy Graph with Respect to Probability :



when probability is 50% or 0.5
Entropy = 1 (max)
Gini Impurity = 0.5 (max)

Entropy:

$H(S) = 1 \Rightarrow$ Very impure split

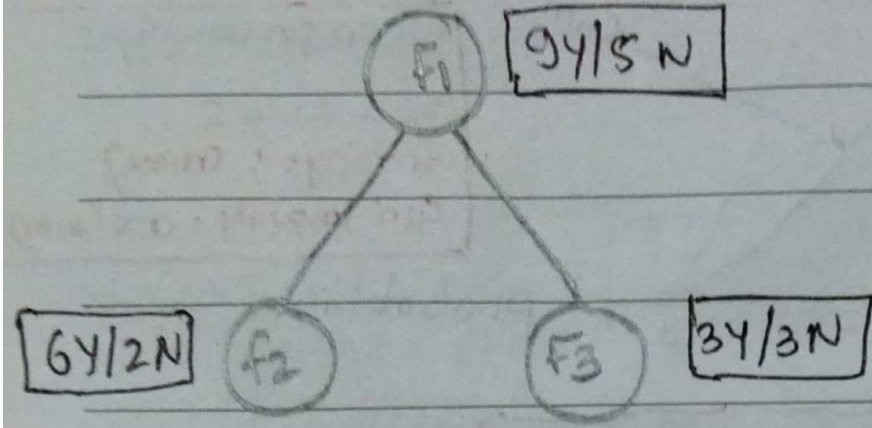
$H(S) = 0 \Rightarrow$ pure split

Information Gain

→ Information gain is used in decision trees & random forest to decide the best split. Thus, more the information gain the better the split and this means lower the entropy. The entropy of a dataset before and after a split is used to calculate information gain.

Entropy of root — split size — Entropy of split

$$Gain(S,A) = H(S) - \sum_{v \in val} \frac{S_v}{|S|} H(S_v)$$



$$H(S) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{9}{14}\right)$$

$$= 0.94$$

$$H(f_2) = -\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)$$

$$= 0.81$$

$H(f_3) = 1$ [since equal split]

Highly impure.

Memo No. ______

Date __/__/__

Now,

$$Gain(S,A) = 0.94 - \frac{8}{14} \times 0.81 - \frac{6}{14} \times 1$$

$$= 0.0486$$

Explanation: In decision tree, many sets of splits are possible. So, for this, information gain of every split are calculated and one with maximum information is considered.

Gini Index or Impurity

→ It is a measurement used to build decision tree to determine how the features of a dataset should split nodes to form the tree.

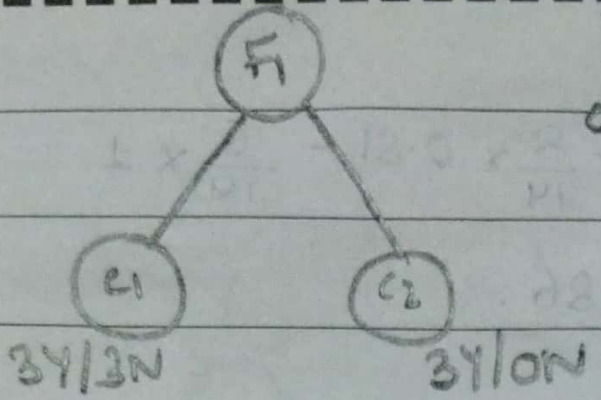
* The lower the Gini impurity, the better the split is

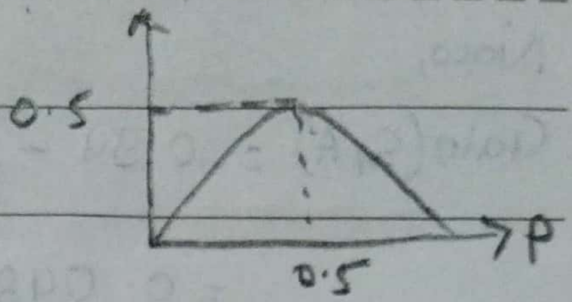
$$\boxed{Gini\ Index(G.I) = 1 - \sum_{i=1}^{n} (P_i)^2}$$

Mo Tu We Th Fr Sa Su

Memo No. ______

Date / /





$$GI = 1 - \sum_{i=1}^{n} (P)^2$$

(for c1)

$$= 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right]$$

$$= 0.5$$

Why Gini impurity over Entropy?

→ The range of entropy lies in between 0 to 1 and range of ~~ent~~ G.I is 0 to 0.5. Hence, G.I is computationaly faster.

Post-Prunning and Pre-Prunning in Decision Tree:

→ Prunning is a process of removal of selected part. In decision tree, prunning overcomes the overfitting condition of ~~technique in~~ decision tree.

Memo No.
Date / /

a) Post Prunning: (backword prunning).

→ This technique is used after construction of decision tree.

→ This technique is used when decision tree will have large depth and will show overfitting of model.

Control the branches of decision tree that is max-depth and min-sample-split.

b) Pre-Prunning (

→ This technique is used before construction of decision tree.

→ Pre-Prunning can be done using Hyper-parameter tunning.

→ Overcome the overfitting issue.

→ we can use GridsearchCV.

When we use Post Prunning?

→ If the dataset is small, only then we can use it.

Decision Tree Regression

- o/p will be continuous value

So, in classification, we used to calculate entropy, gini impurity and information gain to split and construct decision tree.

dataset:

Exp	Gap	salary
2	Yes	40k
2.5	Yes	42k
3	No	50k
4	No	60k
4.5	Yes	56k

We have to perform binary split.

steps:

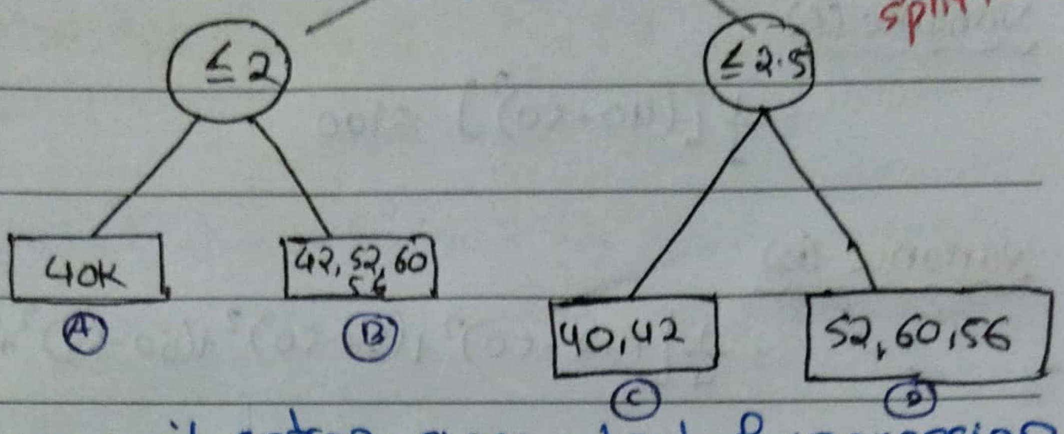
1) sorting in ascending order. [our is already sorted]

$\bar{y} = 50$

[40, 42, 52, 60, 56]

We can split in any way.

~~But how to~~ But how to know which split to use??



Since, we can't entropy since output of regression is continuous and o/p of entropy is binary.

So, we use variance reduction

$$\text{Variance} = \frac{1}{n}\sum_{i=1}^{n}(y-\bar{y})^2$$

Explanation:
We calculate variance of each node of the split. and which has higher variance reduction is selected.

for root (2)

$$\text{Variance(Root)} = \frac{1}{5}\left[(40-50)^2+(42-50)^2+(52-50)^2 + (60-50)^2+(56-50)^2\right]$$

$$= \frac{1}{5}\left[100+64+4+100+36\right]$$

$$= 80.8$$

Memo No.
Date / /

Variance (A)

$$= \frac{1}{1}\left[(40-50)^2\right] = 100$$

Variance (B)

$$= \frac{1}{4}\left[(42-50)^2+(52-50)^2+(60-50)^2+(56-50)^2\right]$$

$$= \frac{1}{4}\left[64+4+100+36\right]$$

$$= 51$$

Variance Reduction :

$$= \mathrm{Var(Root)} - \sum w_i \, \mathrm{Var(child)}$$

$$w_i = \frac{\text{total element in child}}{\text{total no of element in parent node}}$$

$$= 60.8 - \left[\frac{1}{5}(100) + \frac{4}{5}(51)\right]$$

$$= 60.8 - [20 + 40.8]$$

$$= 0$$

Similarly
Variance for root(a.5) = 60.8

Variance (C) = 82

Variance (D) = 46.66

Memo No.

Date / /

variance reduction $= 60.8 - \frac{2}{5} \times 82 - \frac{3}{5} \times 46.66$

$= 0.304.$

* We have to choose the split on which variance reduction is higher.

Important:

When we have to prune in decision tree, if our leaf node consist 2 or more element, then our output will be the average of those elements.

min_sample_leaf: minimum number of sample required to be at a leaf node.

min_sample_split: minimum number of samples required to split an internal node.