12/11/22.

Decision Tree Regressor

1. Build DecisionTree with numeric values
2. DT regressor
3. Visualize DT
4. Pre Pruning and Post Pruning
5. DTR Practical Implementation.

Black Box model :- The model in which we can't visualize the mathematical analysis.

White Box model :- We can clearly visualize the mathematical analysis like max_depth, min_sample_split, criterion, mse, etc see DT.

Steps.

1. Sort the values
2. Avg. of the adjacent value
3. W.R.T. to every avg. value, need to find out "gini impurity" / entropy → I.G.

Ex

numeric

Weight	heart disease
220	Y
180	Y
225	Y
190	N
155	N

After sorting values

weight	heart disease
155	N
180	Y
190	N
220	Y
225	Y

0.10 avg. = 167.5

0.10 = 185

0.10 = 205

0.37 222.5 avg.

IG1 is more gives more info.

Gender	heart disease
M	N
F	Y
M	N
F	Y
F	Y

3Y/2N

Gender

M → [0Y/2N]

F → [3Y/0N]

- Correct threshold with low impurity

Root.

W [3Y/2N]

≤ 167.5 → [0Y/1N] left

> 167.5 → [3Y/1N] right

↓ leaf nodes

Gini impurity :- $1 - \sum_{i=1}^{n} p_i^2$

Root node :- weight : $1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right]$

$= 1 - \frac{9+4}{25} = \frac{25-13}{25} = \frac{12}{25} = 0.48.$

Left node.

$0y/_{1N} \Rightarrow 1 - \left[\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right]$

$= 1 - 1 = 0.$

Right node :-

$3y/_{1N} = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right]$

$= 1 - \left[\frac{9+1}{16}\right] = 1 - \frac{10}{16} = \frac{6}{16} = 0.375.$

Information Gain $= H(S) - \sum \frac{|S_v|}{|S|} H(S_v)$

Information Gain = GI (ROOT) $- \sum_{v=value} \frac{|S_v|}{|S|}$ GI (child) ← weighted impurity

$= 0.48 - \left[\frac{1}{5} \times 0 + \frac{4}{5} * 0.375\right]$

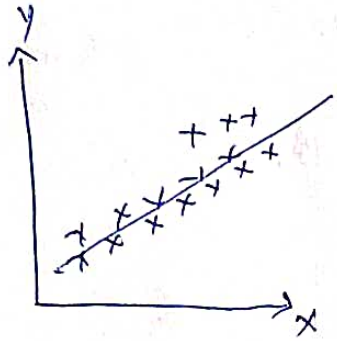
IG. (167.5) $= 0.48 - \left[\frac{4 \times 0.375}{5}\right] = 0.18.$

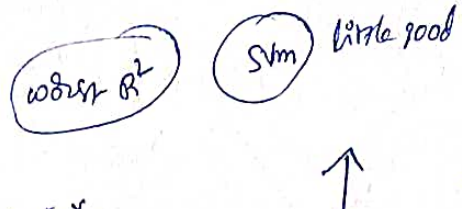
G.I	weight	healthy
	155	N
0.18	167.5	Y
	180	N
0.20	185	Y
	190	Y
0.30	205	Y
	220	
	225	

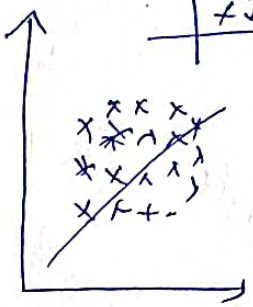
0.30 → low impurity

Conclusion :- gain should be High ; impurity is low(less)

Decision tree - Regression :-





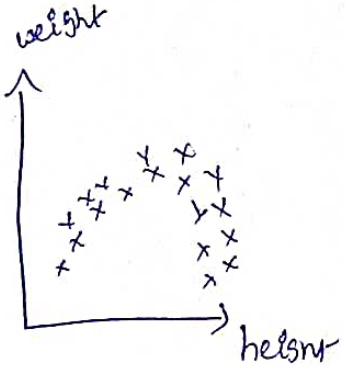


if R^2 <<<

worst

R^2 will near to 1

good ..



height	weight
165	65
160	50
180	90
170	80
175	70

1. Sorting values

	height	weight
Avg.	160	50
162.5	165	65
167.5	170	80
172.5	175	70
177.5	180	90

Classification Tree :

→ Entropy

→ Gini Impurity

→ Information gain

Regression :- Target variable in numeric value

1. Mean
2. MSE/MAE/RMSE
3. Reduction Variance

1st threshod — height [50, 65, 85, 70, 90] → mean is 72

MSE = Variance. → ... best threshold

< 162.5 → Var (left) [50]

> 162.5 → [65, 85, 70, 90] → mean = 77.5 → Var (Right).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y-\hat{y})^2 \rightarrow \text{variance}$$

(MSE → error)

$$\text{height (Variance) root (node)} = \frac{(72-50)^2 + (72-65)^2 + (72-85)^2 + (72-70)^2 + (72-90)^2}{5}$$

$(18)^2 + 7$

$= 206.$

Variance (left) = 50.

$$\text{MSE Varian Right} = \frac{(77.5-65)^2 + (77.5-85)^2 + (77.5-70)^2 + (77.5-90)^2}{4}$$

$= 106.25.$

$$\text{Reduction in variance} = \text{Var (root)} - \sum_{i=1}^{n} w_i \cdot \text{Var (child)}$$

$= 206 - \{\frac{1}{5} 50 + 106.25$

used '0' insted of 50

$= 206 - \{\frac{1}{5} \cdot 0 + \frac{4}{5}(106.25)\}$

RV = 121

Conclusion :- Here we need 2 choose MSE ↓ R² ↑

low in errs (MSE) ⇒

Best threshold → Min MSE

Pruning → cutting

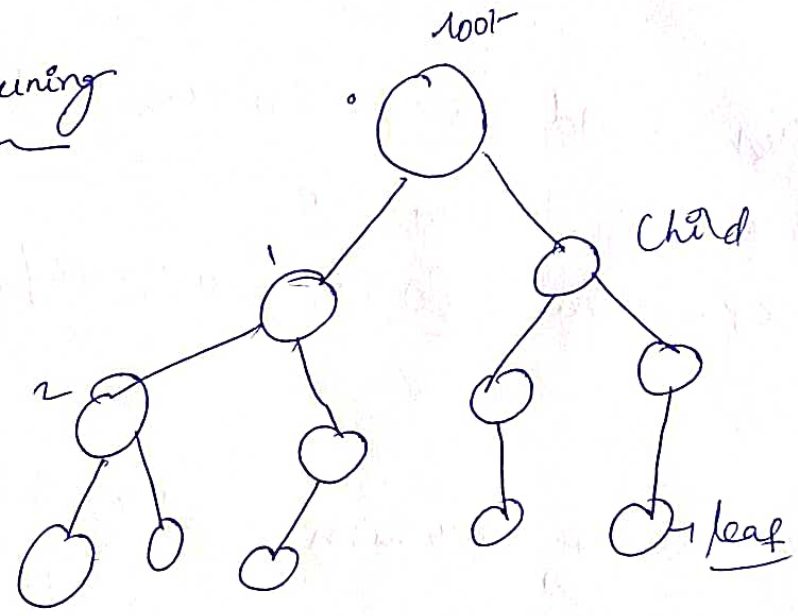
Prepruning and Post Pruning

↓

Overfitting

↓

trimming the data



Pre-Pruning :- before building a decision tree.

while I am going to

Create a decision tree.

max_depth = 2

min_sample_leaf =

min_Sample_Split , max_feature

} Hyper Palameters

responsible for pre pruning =.

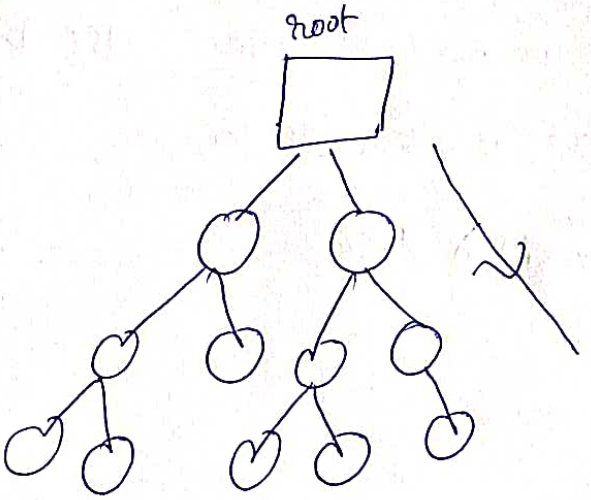
(14)

12Y/2N

12Y/0N 0Y/2N → overfitting → Perform well with training data

low bias

Post Pruning :-



1. first build the decision tree till last
2. Cut the decision tree
3. ccp - alpha = [0.5, 0.6, 0.1, 0.2] ⇒ threshold with gini / threshold entropy.

if ccp-alpha ≤ 0.4 ⇒ very high = DT - height → high (depth)

my - DT height (low)
0.2

Assignment :- Graduation Admission

Individual household Prediction

Census income dataset

Decision tree regressor :-