Decision tree (for classification algorithms or regression (contineous)

↓

uses in

classification (mostly used) (which output to which class its belong)

regression.

* it has two version

C5.0

CART

* in decision tree there two nodes

Decision node (have branches)

leaf nodes (have not branches.

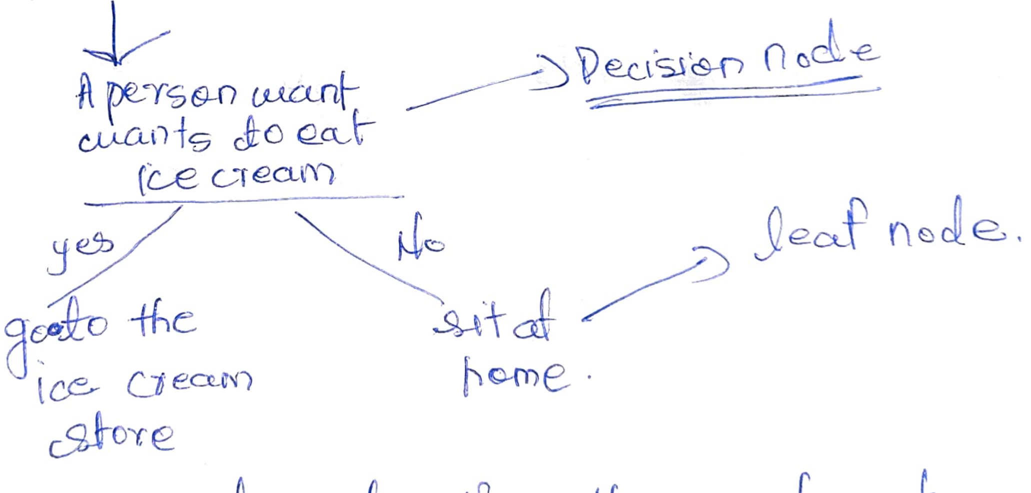
* decision will happen with the help of

yes

no

simple example of decision tree

↓



* when decision tree classifiers things into ~~[illegible]~~

categories

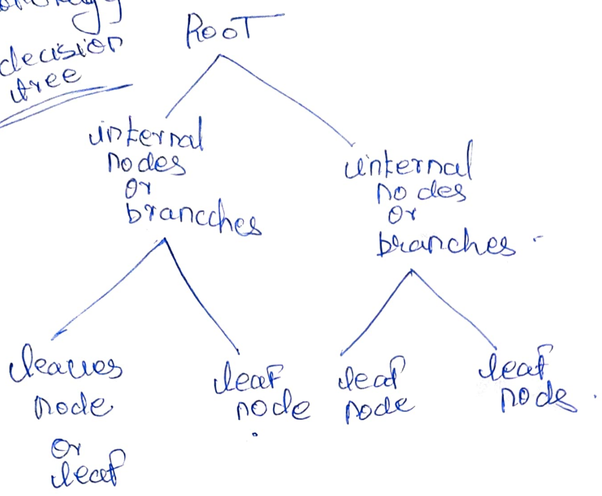
numeric values

classification tree

Regression tree.

* we will focus on decision tree (classification)

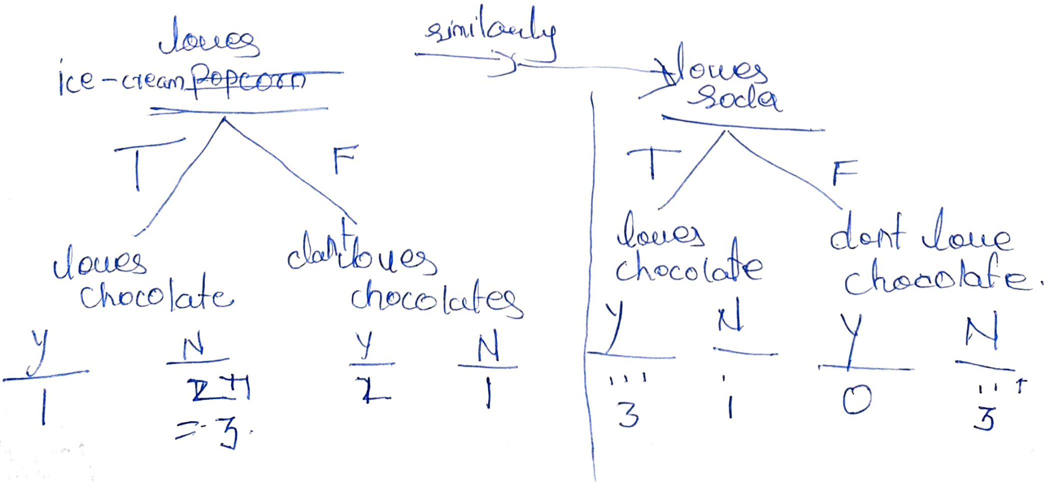
termonology on decision tree



* lets take an example for decision tree.

loves ice cream	loves soda	Age	loves chocolate
Y	Y	7	N
Y	N	12	N
N	Y	18	Y
No	Y	35	Y
Ye	Y	38	Y
Y	N	50	N
N	N	83	N

Y — yes
N - No



★ how to indentify Root nodes

→ by looking independent variable.

★ in previous example you can see the split hasn't happened into equal parts (yes or No) these splits are called impurity.

→ to solve this using methods

Gini impurity (mostly used)

Entrophy

information gain.

Gini impurity

how to calculate gini impurity

★ Gini impurity for a leaf $= 1 - (\text{the probability of yes})^2 - (\text{the probability of No})^2$

take previous eaxample of ice-cream (leftside)

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

$$= 0.375$$

and calculate similarly for right side.

$$= 0.444.$$

Now calculating the gini impurity

total Gini impurity = weighted average of Gini impurities for the the leaves

$= \left(\frac{4}{4+3}\right) 0.375 + \left(\frac{3}{4+3}\right) 0.444$

$\left(\frac{\text{total impurity in leave on left}}{\text{total impurity in both leave}}\right)$ Gini impurity + ("similarly" for right) Gini impu

conclusion is Gini impurity for loves ice cream = 0.405.

this is we have done Just for categoriEs what about Numeric data?

from the pervisous example we will take age column

step ① calculate average of adjecent date
② then calculate Gini impurity for each ~~for~~ value.

Age

7
12
18
35
38
50
83

$\frac{7+12}{2} = 9.5 \rightarrow 0.429$ — for example → Age<9.5

15 → 0.343

~~25~~ 15.5 → 0.476

36.5 → 0.476

44.0 → 0.343

66.5 → 0429

Age<9.5

T: soda
Y 0
N. 1

F: soda
3 3

So total gini impurity = 0.429

So the conclusion is 0.343 is lowest so ~~we~~ there are two 0.343 we will pick any one.

Similarly we took out gini Impurity for age So we we take out ice-cream and Soda and we have fount out Soda has the lowest Ginni impurity values. is 0·214

so likewise ~~we~~ we will put Soda on the top and according to gini impurity we will distribut categories and whoever got the highest impurity ~~Soda~~ will be the leaf node and thoes who got the highest majority values will win example

Soda

Y 1 N 0

love soda

Soda

Y 0 N 1

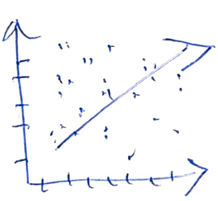
dont love Soda

problem statement (regression tree)

what if data is Not linear and we have to predict the values.

Decision — Classification

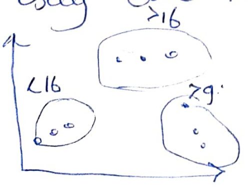
Decision — Regression (numeric)



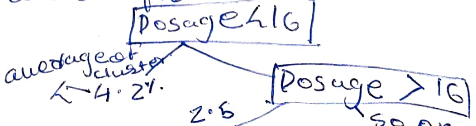
? how solution is regression tree

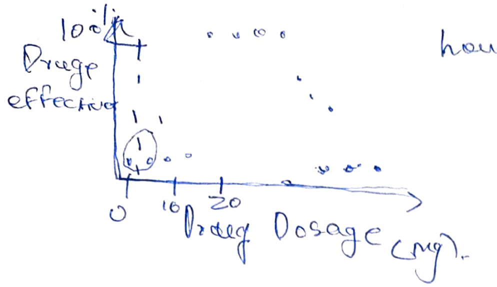
So How regression work?

lets say we have one example.



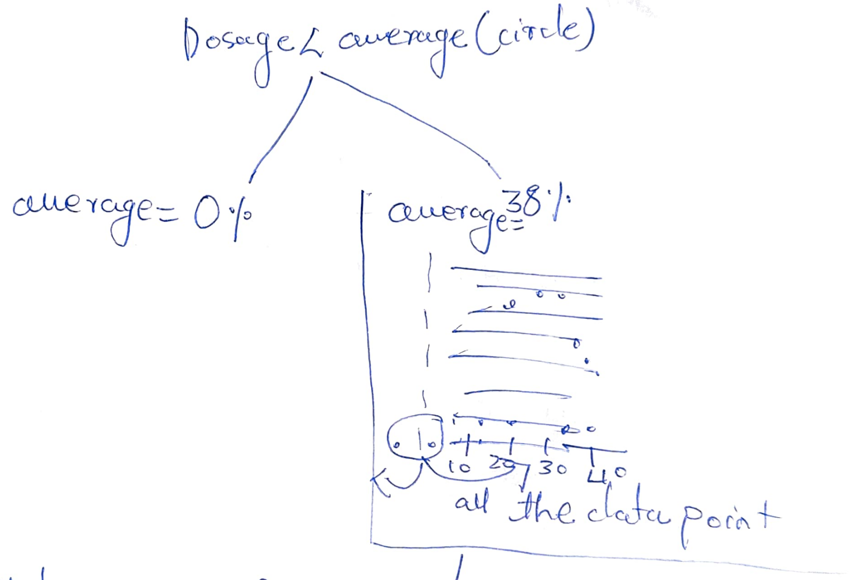
so you can see there is three part in graph. each has threshold value so we will plot tree on the basis of this





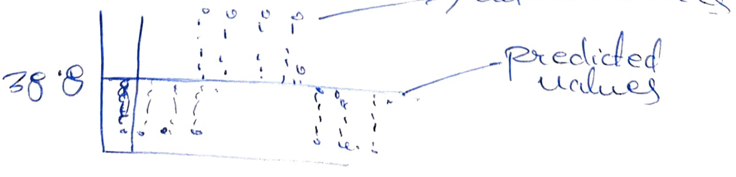
how to build tree for regression.

in above graph we ~~can~~ have took only two datapoint (circle(s)) the line is average values on that basis we can build tree



and if the values come (dosage 20) ~~are~~ so according graph it showing 100% but according to Decision tree it showing 38% so it really a poor prediction

★ So we can we visualize the predicted values and actual values.



what the Solution then ?

=> we can ~~Not~~ not be accurate but we can improved the decision tree.

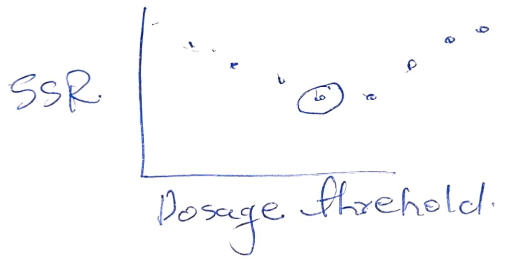
to improve.

① take any threshold datapoint and ~~sub~~ minus it to all datapoint by square like this.

$(datapoin - datapoint1)^2 + (datapoin - datapoint2)^2 + (datapoint - datapoint n)^2$

= Ans.

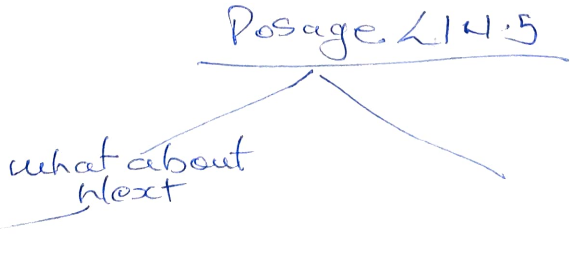
whatever you gothe ans plot this in graph.



So whoever gothe the lowest SSR that threshold we win. because it has lowest SSR.

so that threhold will be the root.

for example



do the same process to calculate threshold values again.

So do split until the Drug effectiveness is same or you dont have data point to split.

but this process can lead to model variance so divide the data according to groups or cluster not by data point. what we there are not 1 there many independend variable.

* 1) step calculate threshold for every column.
2) and pick that threshold (clowmn) whoever got the lowest values.
3) then make branch (asending to decenting on the basis of threeshold)

Decision tree

(work on

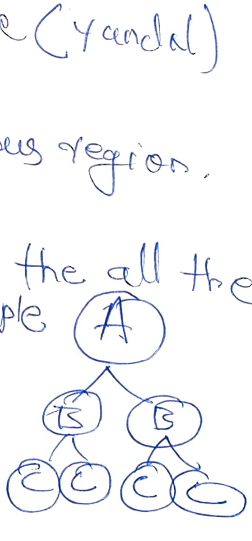
Divide and conquer technique (yandal)

↓

every Node is purly homogeneous region.

↓

at the root node it represent the all the data set and so on for example



you can check Accuracy of Decision tree by Holdout method.

Decision tree (C5.0).

in C5.0 to select feature at branch it uses something called entrophy.

data with high entrophy ↓ diverse, little information

↓ low entroph ↓ homogeneous region

most important feature will determine by entrophy score.

$$entrophy(S) = \sum_{i=1}^{c} -p_i \log_2(p_i).$$

p = proportion.

i = features.

to calculate entrophy we have to create random split for example (ABCD)

(subset) ① (ABC)

② (~~C~~DB)

③ (CDB)

② then take 1st random and so on calculate entrophy just say

~~co~~ play

| y
| y y n |

$$\frac{3}{4} - \log_2(3/4)$$

$$= 6.4$$

③ calculate entropy for another random also.

for Purely humogeneous region = Entropahy will always be 0·

Now we will calculate Entrophy before by calculate-ing complete dataset.

before contineous we should understand why we pecking up important feature by entrophy method or any methode

① if dont select the decision tree will have nodes to much;

② if dont seled the decisiontree will dont predict values properly. (impurity will increase

entrophy :- measures the purity of split.

	N	Y	
wrost entrophy	50%	50%	=1 bits
best entrophy	100%	0%	= 0 bits.

entrophy ranges to 0-1 and entrophy unit is bits

* the split will go on till the leaf node is peerly homogeneous subset.

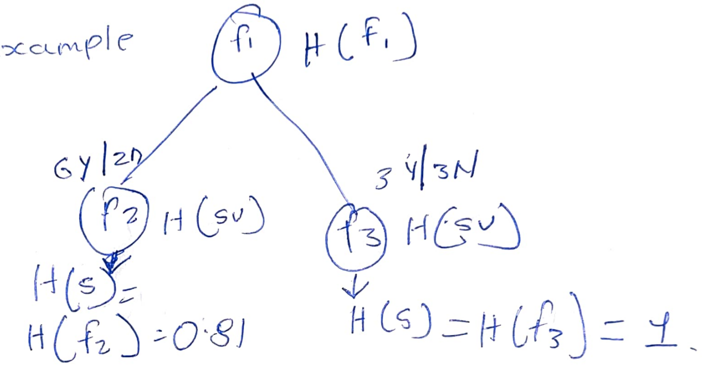
information gain.

thes will give the which split is best.

$$= Gain(S, A) = H(S) - \sum_{V \in VAL} \frac{|S_V|}{|S|} H(S_V)$$

$H(S)$ = entrophy for particular subset (after split what is the split.)

example



$$Gain(S, f_1) = H(S) - 8/14\ H(f_2) - \frac{6}{14} H(f_3)$$

$$= 0.91 - 8/14 \times 0.81 - \frac{6}{14} \times 1$$

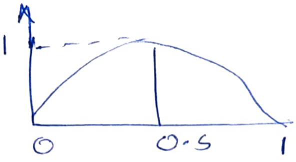
$$= 0.049$$

↓ what this values indicate.

if this values values is high it means this structure is best to use the machine algorithm will use differen Decision tree structure to identify which one is the best

~~to calculalate impud~~
to calculate purity in set we use
- entropy
- Gini impurity

① entrophy =



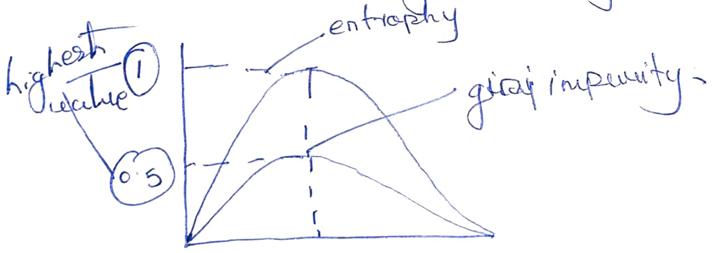
② Gini impurity,

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

$$= 1 - [(P+)^2 + (P-)^2].$$

but we are using both to calculate impurity then which one to use and what is the different.

graph for
entroph and gini impurity



Gini impurity takes less time to do operation.

* entrop before (~~after~~ before creating random set) original dataset

*

* information gain (lies in 1 to 0).

= (P)* ~~entro~~ - entropy before - entrophy after

* information gain (How useful is split).

0 - worst

1 - best.

* to avoid overfiting issue we will use **pruning**.

Deceision tree: C5.0

(alpha Here zero)
↓
research paper on decision tree.

* How long Decision tree.

= it will go indefinitly untill algorithm have'nt got homogeneous region.

* CART (classification and Regression trees