DecisionTree from Scratch

1. What are decision tree?

Statement!

DO YOU WANT TO LEARN DECISION TREES?

If yes, start reading notes

If no, still read you don't have option :)

Decisions!

* 2 types of decision trees

1) Classification trees

Same ques.

YES → STUDY!!

NO → CHILL!!

Here we've classified into categories

2) Regression trees

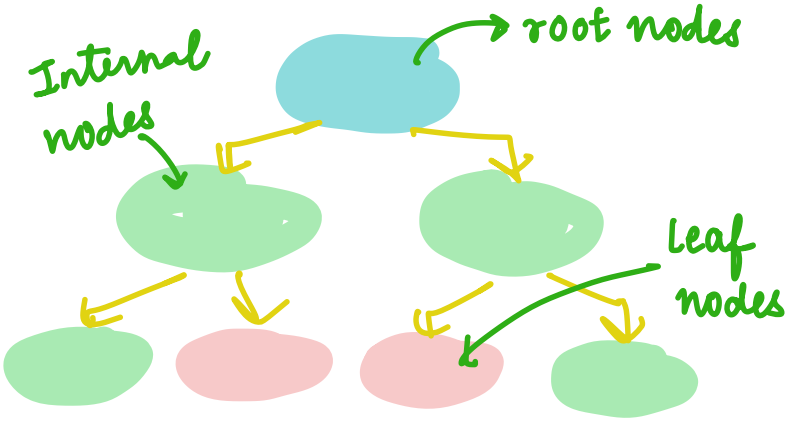
Same ques.

YES → AGE (15-40)

NO → OTHER THAN THAT

Predicting numerical value

* Terminologies *

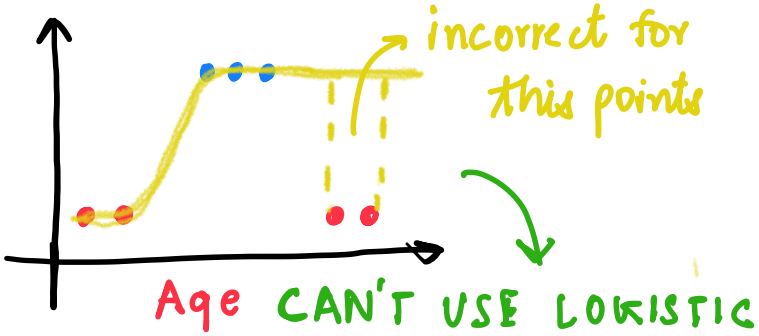


Example!

20 questions game.

What we're going to cover?

1. Build Classification from scratch (No code)
2. Build Regression ... "
3. Python hands on decision tree.

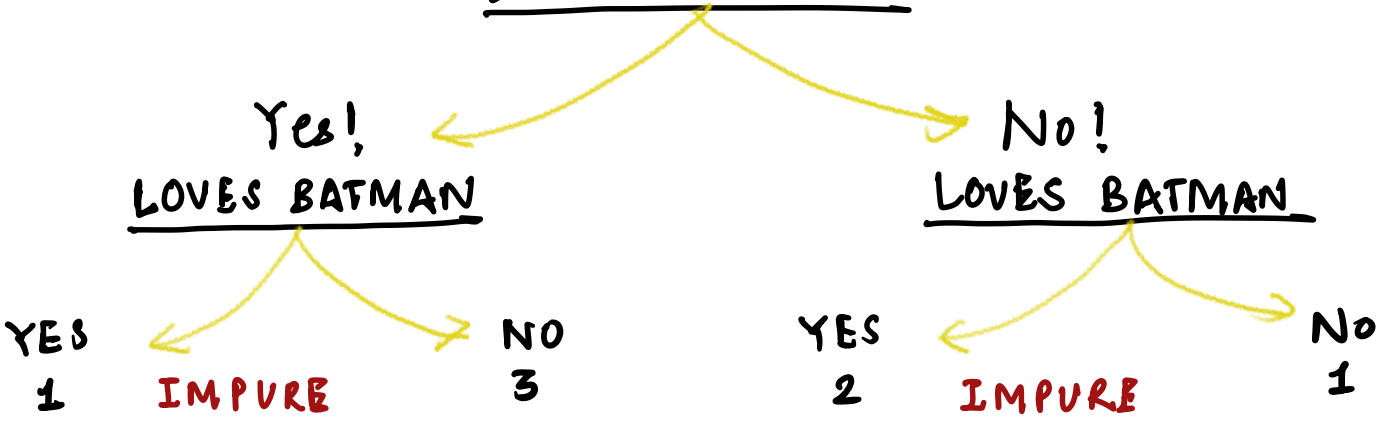


2. Classification trees

Aim : Given the training dataset, build a classification that uses loves movies, loves cartoons and age to predict whether someone likes batman or not.

Step 1 : Which questions to ask root node ??

Step 2 : Let's start with loves movies :



* DATASET *

Loves Movies	Loves Cartoons	Age	Loves batman?
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

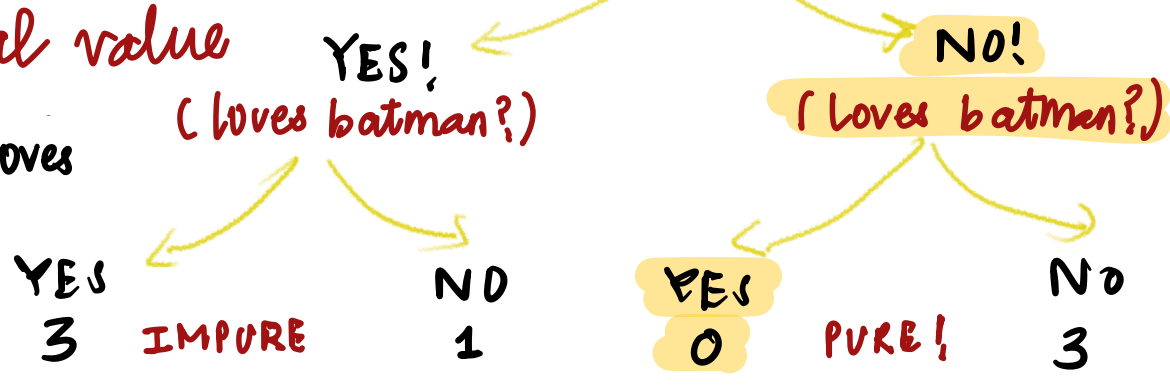
mixture of discrete and continuous data

used for prediction if someone loves batman

LOVES CARTOONS

optimal value

* We'll take loves cartoons as a root nodes.



- Leaves which consists mixture of classifation is impure leaves

BUT HOW WE'LL QUANTIFY THIS ??

GINI IMPURITY

Formula for gini impurity : $1 - (\text{Probability of YES})^2 - (\text{Prob. of NO})^2$

① for leaf 1 :

Prob. of yes = $\frac{1}{4}$

Prob. of NO = $\frac{3}{4}$

→ Gini Impurity :-

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

$= 0.375$

② for leaf 2 :-

Prob. of yes = $\frac{2}{3}$

Prob. of NO = $\frac{1}{3}$

→ Gini impurity

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

$= 0.444$

Loves movies

GI : 0.375 GI : 0.444

TOTAL GI = Weighted of GI leaves avg

$= \left(\frac{4}{7}\right)^* 0.375 + \left(\frac{3}{7}\right)^* 0.444$

$= 0.405$

Similary, for loves cartoon → TOTAL = 0.214

LOVES CARTOON < LOVES MOVIES

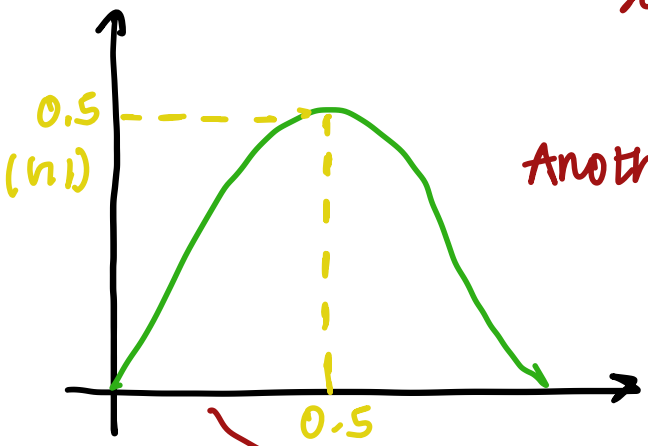
loves cartoon need not intuition, better classifying people.

* VISUALIZATION *

Let's consider eq^n as :-

$$GI = 1 - \underbrace{(\text{Prob. of YES})^2}_{x} - \underbrace{(\text{Prob. of NO})^2}_{1-x^2}$$

$$= 1 - x^2 - (1-x)^2$$



Highest impurity for $x = 0.5$

Another measure of impurity is 'ENTROPY'

formula = ...

$$\text{ENTROPY} = -[\text{Prob. of YES}] * \log[\text{Prob. of YES}] + [\text{Prob. of NO}] * \log[\text{Prob. of NO}]$$

Leaf 1

$$= -\left[\frac{1}{4} * \log\left(\frac{1}{4}\right) + \frac{3}{4} * \log\left(\frac{3}{4}\right)\right]$$

$$= 0.8113$$

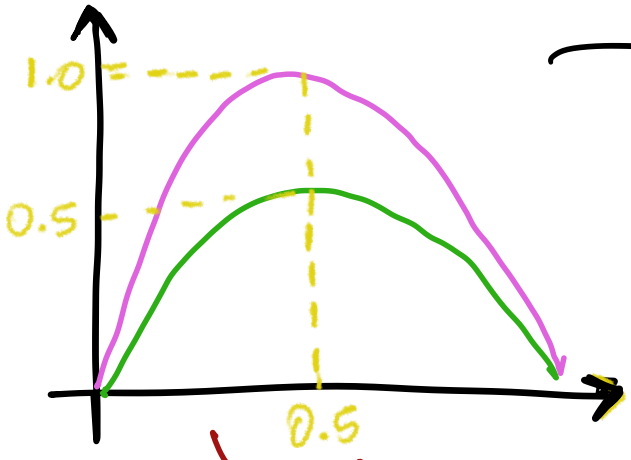
Leaf 2

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

$$= 0.9183$$

TOTAL ENTROPY (I)

$$= \frac{4}{7} * (0.8113) + \frac{3}{7} * (0.9183) = 0.857$$



$P_{True} = 0.5$ (Highest)

→ * **ENTROPY v/s IMPURITY**

● → Entropy

● → Impurity

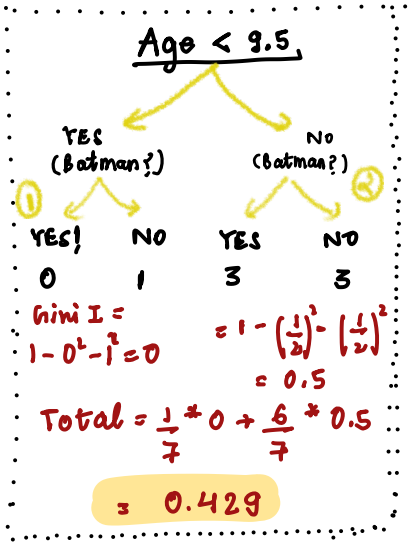
Gini Impurity for numeric data (ages)

Step 1: Sort age from lowest to highest

Step 2: Calculate avg. for adjacent rows.

Step 3: Calculate Gini impurity

Ages	Avg
7	9.5
12	15
18	26.5
35	36.5
38	44
50	66.5
83	



Similarly for the rest we have

Ages	Avg	Gini. Imp.
7	9.5	0.429
12	15	0.343
18	26.5	0.476
35	36.5	0.476
38	44	0.343
50	66.5	0.429
83		

lowest Gini impurity = 0.343

So, now we can say that :

Loves movies — GI = 0.405

Loves cartoon — 0.214 — lowest of them all

Age < 15 — 0.343

When we take Age < 12.5 (GI = 0)

* So, because GI of loves cartoon is the lowest, we will choose that as root node!

NOW WE HAVE TO FIND LEAF NODES

LOVES CARTOON — root node!

* So, the next pure is age < 12.5 ↓ That will be next node

YES! — Next node?? (GINI IMPURITY)

NO! (Pure, no need to split)

Age < 12.5

Loves batman — YES (0) PURE! — NO (1)

Loves batman — YES (3) PURE! — NO (0)

We do not need to go further as we got our decision tree

Final decision tree →

LOVES CARTOON ← from scratch!!

YES! → Age < 12.5

YES! → doesn't love batman

NO! → loves batman

NO! → does not love batman → 100% Accuracy

But there are some problems :-

① leaf 1. there is only one person (small data)

TWO WAYS TO SOLVE

Pruning

(3 or more people) Put limits on how trees grow

② If we ignore this part, then our decision tree will be ←

LOVES CARTOON → with 75% ACCURACY (less)

YES! → loves batman

NO! → does not love batman

What is Pruning?

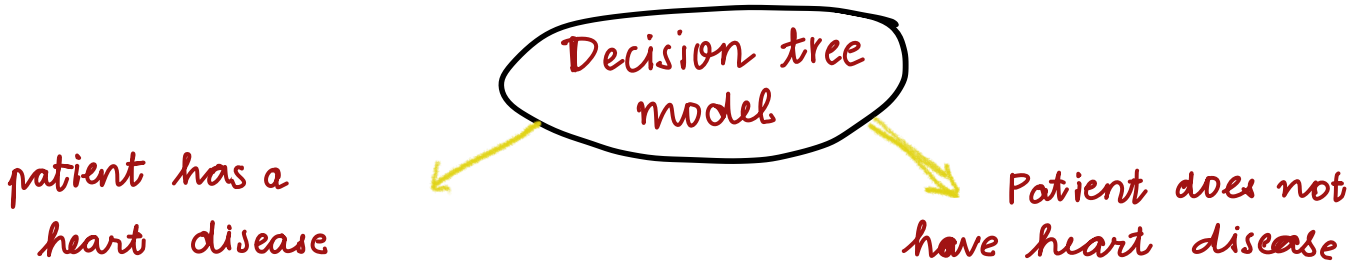
remember when we had only 1 person here, which is overfitting.
& reducing no. of leaves is called pruning.

TREE SCORE = total Impurity + α * no. of leaves

"Pruning is similar as regularization" → pruning parameter

* We're going to call this as CCP (cost complexity pruning)
* cross validation to find α → Why??

building decision tree on dataset



Step 1 : understand dataset

13 columns

1. **age:** The patient's age in years.
2. **sex:** The patient's gender (1 = male, 0 = female).
3. **cp:** The type of chest pain the patient experiences (0-3 scale, higher indicates more typical angina).
4. **restbp:** The patient's resting blood pressure measured in millimeters of mercury (mm Hg).
5. **chol:** The patient's serum cholesterol level measured in milligrams per deciliter (mg/dl).
6. **fbs:** Indicates if the patient's fasting blood sugar is above 120 mg/dl (1 = true, 0 = false).
7. **restecg:** Results of the patient's resting electrocardiogram (0-2 scale).
8. **thalach:** The highest heart rate the patient achieves during exercise.
9. **exang:** Indicates if the patient experiences angina (chest pain) induced by exercise (1 = yes, 0 = no).
10. **oldpeak:** The amount of ST segment depression induced by exercise relative to rest.
11. **slope:** The slope of the ST segment during peak exercise (0-2 scale).
12. **ca:** The number of major blood vessels visible under fluoroscopy (range: 0-3).
13. **thal:** Results from the thallium heart scan (3 = normal, 6 = fixed defect, 7 = reversible defect).
14. **hd:** The diagnosis of heart disease (0 = no heart disease, 1 = heart disease).

2.