Classification →

Ex Mail → Spam (+ve) / Not Spam (-ve)

Online Transaction → Fraudulent (+ve) / Non Fradulent (-ve)

Tumor → Malignent (+ve) / Benign (-ve)

where $y \in \{0, 1\}$ — 0 ⇒ -ve class; 1 ⇒ +ve class.

Here y has Discrete Value.

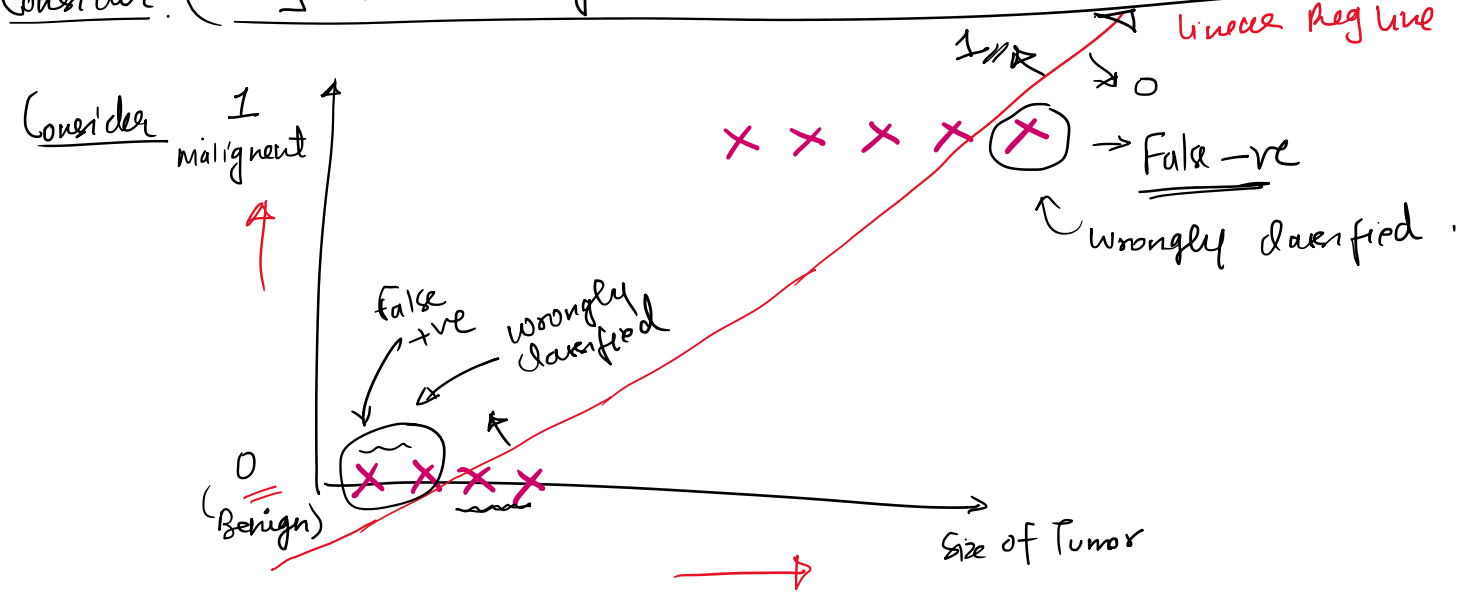
In above case o/p has only 2 class (0, 1) } Binary classification.

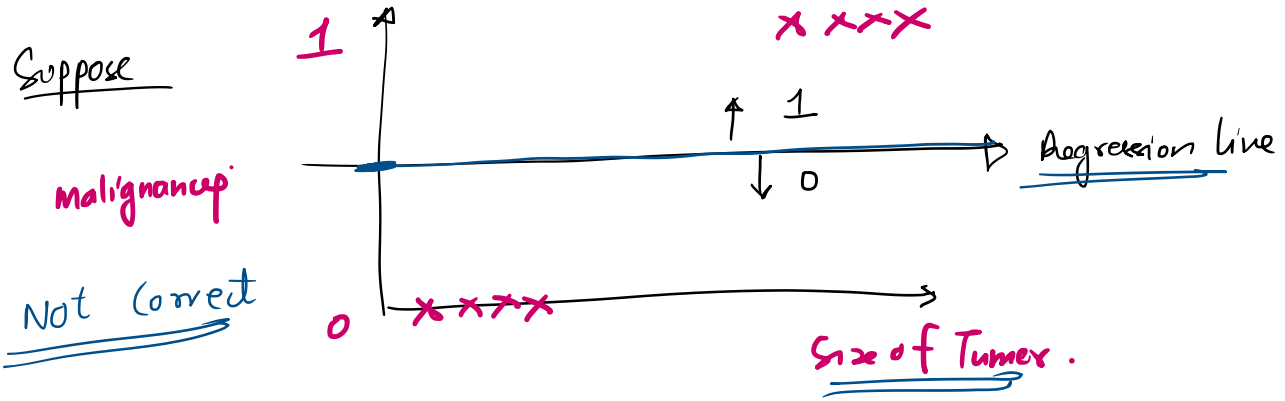
To check Weather.

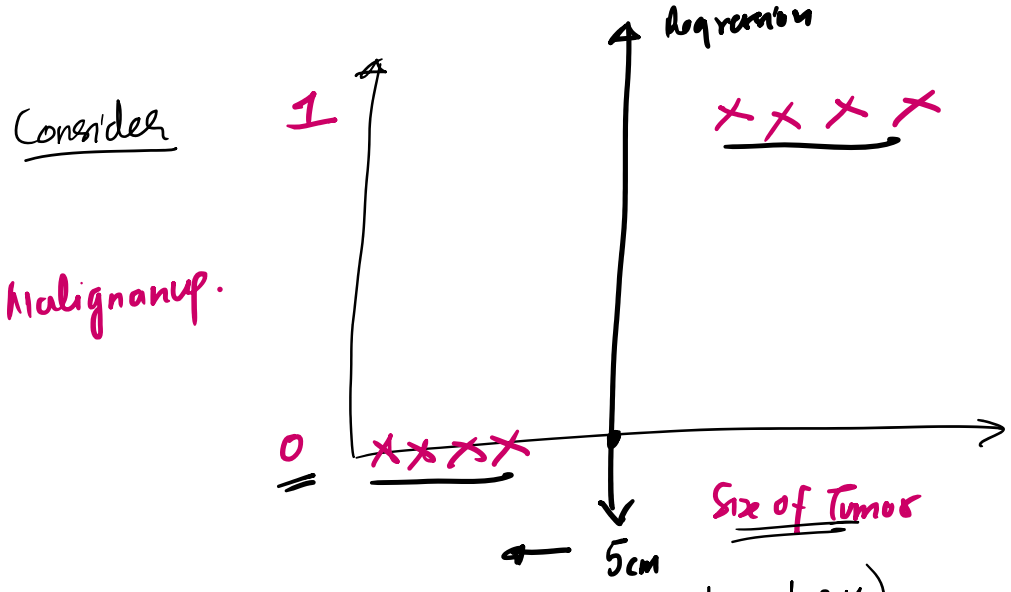
Possible o/p — Windy, Sunny, Cloudy, Rainy } o/p has more than one class [multi class classification]

* Classification bothers about label and not the Exact value.

Consider (Why linear Regression is not used for Classification)







Here we can observe for above Regression line

If tumor size $\leq$ 5cm $\rightarrow$ Yes (1) Malignant

If tumor size $\leq 5cm$ → Yes (1) Malignant

tumor size $> 5m$ → No (0) Benign

We can say let 'P' denote Probability that $y=1$ when $X=x$.

$$y = P_r(y=1 \mid X=x) = [\beta_0 + \beta_1 x]$$

In linear Reg

P = probability lies bet^n 0 to 1

But linear function are unbounded.

and Expected o/p here is 0 or 1

So we cannot use Regression to build Classifier

∴ linear Regression is not suitable for Classification.

For Classification we will use Logistic Regression.

* In Logistic Regression we get probability score.
* It predicts the probability of occurance of event

Odd = (Represents chances that the event will occur) = $\frac{\text{No of time the Event happens}}{\text{No of time the Event will not happen}}$.

Ex ⇒ If odd of India winning against W.I is 4:1 = $\frac{\text{No of India win}}{\text{No of India not win}}$

$= \frac{4}{1}$

Best Case ⇒ The odd of India winning against WI = ∞

If odd of W.I winning against India is 1:4

$= \frac{\text{No of W.I win}}{\text{No of W.I not win}} = \frac{1}{4}$

Worst Case ⇒ W.I is winning 0 match = Odd = 0

Range of value that odd can take = 0 to ∞

Relationship between Odd & Probability = Odd = $\frac{P}{1-P}$

$\therefore \quad odd = \frac{p}{1-p}$.

Here we know that odd has range 0 to ∞

→ There is no upper bound for odd.

→ But odd has lower bound.

+ To remove lower bound, to have symmetrical analysis

Ex odd = 1:6 1/6 = 0.167
    odd = 6:1 6/1 = 6
} not Symmetrical

But $\ln(1/6) = -1.79$
     $\ln(6/1) = 1.79$
} Symmetrical

By taking log we have overcommed the lower bound also

log(odd) → will not have upper bound
         → will not have lower bound.

$$\log(odd) = y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n.$$

Let $z = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$.

$$\therefore \log_e\left(\frac{p}{1-p}\right) = z$$

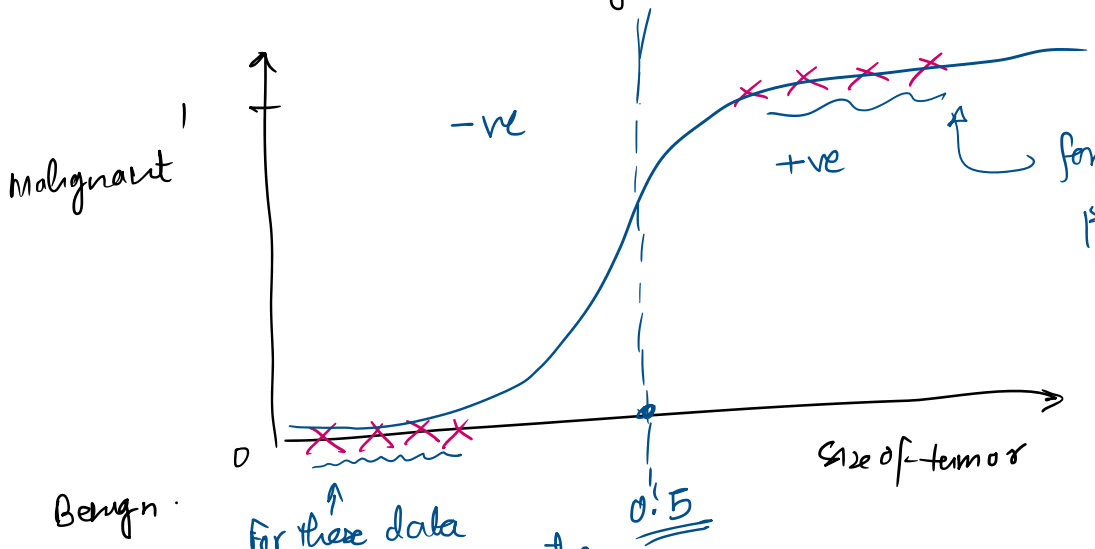
$$\frac{p}{1-p} = e^z$$

$$p = -e^z p + e^z$$

$$p(1+e^z) = e^z$$

$$p = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}}$$

Sigmoid fn.



Logistic Regression Model :

In linear reg $h_\theta(x) = \theta^T x$ → range $-\infty$ to $+\infty$

x ← feature vector.

θ ← parameter vector

$h_\theta(x)$ ← hypothesis (prediction).

In logistic Regression $0 \le h_\theta(x) \le 1$

$h_\theta(x)$ ← predicted value

$$h_\theta(x) = g(\theta^T x)$$

g ← Sigmoid

Sigmoid

let $z = \theta^T x$

$$h_\theta(x) = g(z) = \frac{1}{1+e^{-z}}$$

Estimated probability that $y=1$ on given input x.

prediction.

if $h_\theta(x) = 0.7$

this means there is 70% chance that tumor is malignant.

Since $h_\theta(x) = 0.7 > 0.5$

So $\boxed{y=1}$

if $h_\theta(x) = 0.3$

this means there is 30% chance of tumor being malignant

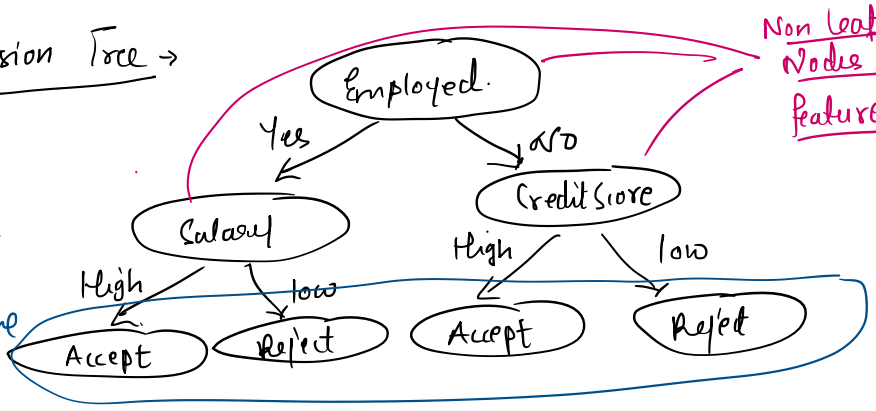
Since $h_\theta(x) = 0.3 < 0.5$

So $\boxed{y=0}$

Decision Tree :→

Consider → Bank → Details of Loan Applicant (Input) → Accept (1) / Reject (0)

A Sample Decision Tree →



Non leaf Nodes are features.

Here Independent Features → Employed → Salary → Credit Score

Outcomes are leaf node

Outcome → Accept → Reject

Q) Create Decision Tree for following using Gini Index (Classification & Regression Tree) → CART ✓

Weekend	Weather	Parent	Money	(y) Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

p

Solution → Independent features: Weather, Parent, Money

Decision/Outcome → Cinema, Tennis, Stay In, Shopping.

shopping.

Step 1 >

We will calculate Gini Index for Overall collection of Outcomes of Training Examples.

There are 4 possible outcomes for decision
- Cinema — 6 instances
- Tennis — 2 instances
- Stay In — 1 instance
- Shopping — 1 instance.

$$Gini_{(decision)} = 1 - \left(\left(\frac{6}{10}\right)^2 + \left(\frac{2}{10}\right)^2 + \left(\frac{1}{10}\right)^2 + \left(\frac{1}{10}\right)^2\right)$$
$$= 1 - \left(\frac{42}{100}\right) = 0.58$$

Note -> In Machine Learning, Gini index/coefficient is utilized as an impurity measures in decision tree for classification.

$$Gini = 1 - \sum_{i=1}^{n} (P_i)^2$$ where P_i probability of outcome of specific class

Step 2: To find Gini Index for money

$Gini_{(money)}$ → a possible value → Rich, poor

(1) $Gini_{(money = Rich)}$ → 7 instance → possible decision
- Cinema → 3 time
- Tennis → 2 time
- Stay In → 1 time
- Shopping → 1 time.

$$Gini_{(money=Rich)} = 1 - \left(\left(\frac{3}{7}\right)^2 + \left(\frac{2}{7}\right)^2 + \left(\frac{1}{7}\right)^2 + \left(\frac{1}{7}\right)^2\right)$$
$$= 0.694$$

(2) Gini → ... → Cinema.

(2) Gini (money=poor) → 3 instance → Decision → Cinema.

$= 1 - \left((3/3)^2\right) = 0$ //

Weighted Average $Gini_{(money)} = \left(Gini_{(money=Rich)} * \text{Proportion of rich}\right) + \left(Gini_{(money=poor)} * \text{Proportion of poor}\right)$

$= (0.694 * 7/10) + (0 * 3/10)$

$Gini_{(money)} = 0.485$

Step 3 Gini Index on Parent

For Parent feature → Possible values → Yes / NO

$Gini_{(parent=Yes)}$ = 5 instances → Possible Decision → Cinema

$= 1 - \left((5/5)^2\right) = 0$ //

$Gini_{(parent=NO)}$ = 5 instance → possible decision → Tennis → 2 times; Stay In → 1 times; Shopping → 1 time; Cinema → 1 time.

$= 1 - \left((2/5)^2 + (1/5)^2 + (1/5)^2 + (1/5)^2\right)$

$= 1 - (7/25) = 0.72$

Weighted Average of $Gini_{(parent)} = (0 * 5/10) + (0.72 * 5/10) = 0.36$

$Gini_{(parent)} = 0.36$

$Gini_{(parent)} = 0.36$

Step 4) Gini Index for Weather.

Weather → Possible values
- Sunny = 3 instances
- Windy = 4 instances
- Rainy ⇒ 3 instances.

Gini (weather = sunny) ⇒ 3 instances → possible outcomes
- Cinema → 1 time
- Tennis ⇒ 2 time

$= 1 - \left((1/3)^2 + (2/3)^2 \right) =$ 0.444

Gini (weather = windy) → 4 instances → possible outcome
- 3 time Cinema
- 1 time Shopping

$= 1 - \left((3/4)^2 + (1/4)^2 \right) =$ 0.375

Gini (weather = Rainy) → 3 instance → possible outcomes
- 2 cinema
- 1 Stay In

$= 1 - \left((2/3)^2 + (1/3)^2 \right) =$ 0.444

Weighted Average $Gini_{(Weather)} = (0.44 * 3/10) + (0.375 * 4/10) + (0.44 * 3/10)$

$= 0.414$

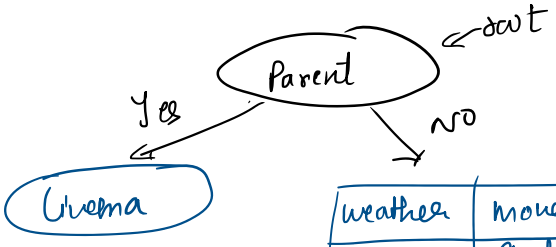
$= 0.486$

Gini (money) = 0.486
Gini (parent) = 0.36
Gini (weather) = 0.416

Minimum Gini ⟶ Minimum Impurity in Decision.

Here Minimum Gini Value = Gini (parent) = 0.36

So the root of Decision is Parent



weather	money	Decision
Sunny	Rich	Tennis
Rainy	Rich	Stayin
Windy	Poor	Cinema
Windy	Rich	Shopping
Sunny	Rich	Tennis

We need to find Gini (parent = No and weather)

Also Gini (parent = No and money)

Steps) Gini (parent = No and Weather) ⟶ possible value ⟶ Sunny – 2 time, Windy – 2 time, Rainy – 1 time

5 instances

Gini (parent = No and weather = Sunny) ⟶ 2 instance ⟶ Tennis = $1 - \left(\left(\frac{2}{2}\right)^2\right) = 0$ // possible Outcome

Gini (parent = No and weather = Windy) ⟶ 2 instances ⟶ Cinema, Shopping ⟶ $= 1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right) = 0.5$ possible Outcome

Gini (parent = No and Weather = Rainy) → 1 instance → Stay in = $1 - ((1/1)^2) = 0$ — possible outcome

Weighted Average Gini (parent = No and Weather) = $0.5 \times 2/5 = 0.2$.

Step 6) Gini (parent = NO and Money) — 5 instance

- Rich = 4 time
- poor = 1 time

Possible values.

Gini (parent = NO and Money = Rich) → possible outcome (4)

- Tennis – 2
- Stay In – 1
- Shopping – 1

$= 1 - \left((2/4)^2 + (1/4)^2 + (1/4)^2 \right) = 0.625$

Gini (parent = NO and Money = Poor) → possible outcome

Cinema $= 1 - ((1/1)^2)$

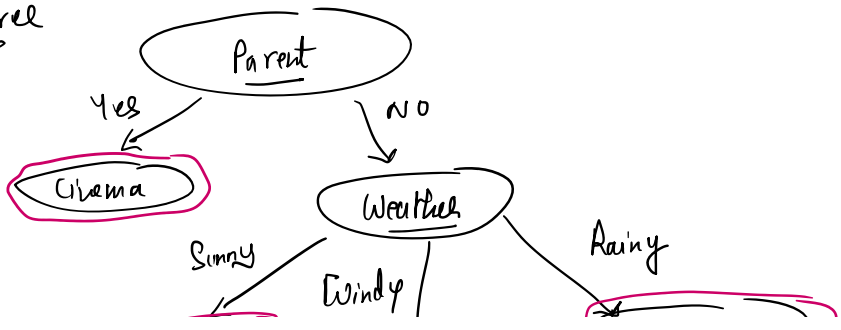
$= 0$

Weighted Average Gini (parent = No and Money) = $0.625 * 4/5 = 0.5$

Here Gini (parent = NO and Weather) has smallest value

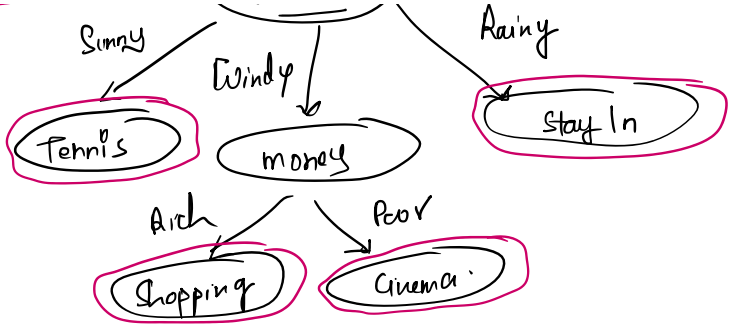
So now Next Node = Weather

* Updated Decision Tree



Ans.

Ans



HW

Construct an optimal Decision Tree for following.

Outlook	Temperature	Humidity	Windy	Play (Decision)
Sunny	Hot	High	False	NO
Sunny	Hot	High	True	NO
Overlast	Hot	High	F	Y
Rainy	mild	High	F	Y
Rainy	Cool	Normal	F	Y
Rainy	Cool	Normal	T	N
Overlast	Cool	Normal	T	Y
Sunny	mild	High	F	N
Sunny	Cool	Normal	F	Y
Rainy	mild	Normal	F	Y
Sunny	mild	Normal	T	Y
Overlast	mild	High	T	Y
Overlast	Hot	Normal	F	Y
Rainy	mild	High	T	N