Data Science and Artificial Intelligence

Machine Learning

Decision Tree

Lecture No. 2

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

Topic Entropy

Topic GI Index

Topic Info. Gain

Topic

Topic

Topics to be Covered

Topic — Entropy Vs GI

Topic — Variance

Topic — Pruning Stopping Criteria in DT.

Topic

Topic


Hold the vision,
Trust the process.
-author unknown
Your Beautiful Life

REVISION

Gini Impurity Index and Entropy

Probab of misclassification

$$\Rightarrow 1 - \sum_{i=1}^{c} P_i^2$$

$$\left\{ \sum_{i=1}^{c} P_i \log_2 \frac{1}{P_i} \right\}$$

REVISION

Information gain

$$\Rightarrow \left\{ \text{Impurity}^{\text{Parent}} - \underbrace{\text{Impurity}^{\text{Child}}}_{\Downarrow \text{ with help of weighted avg.}} \right\}$$

Entropy Vs Gini Index

- 2 Class Case

  Class1 : P

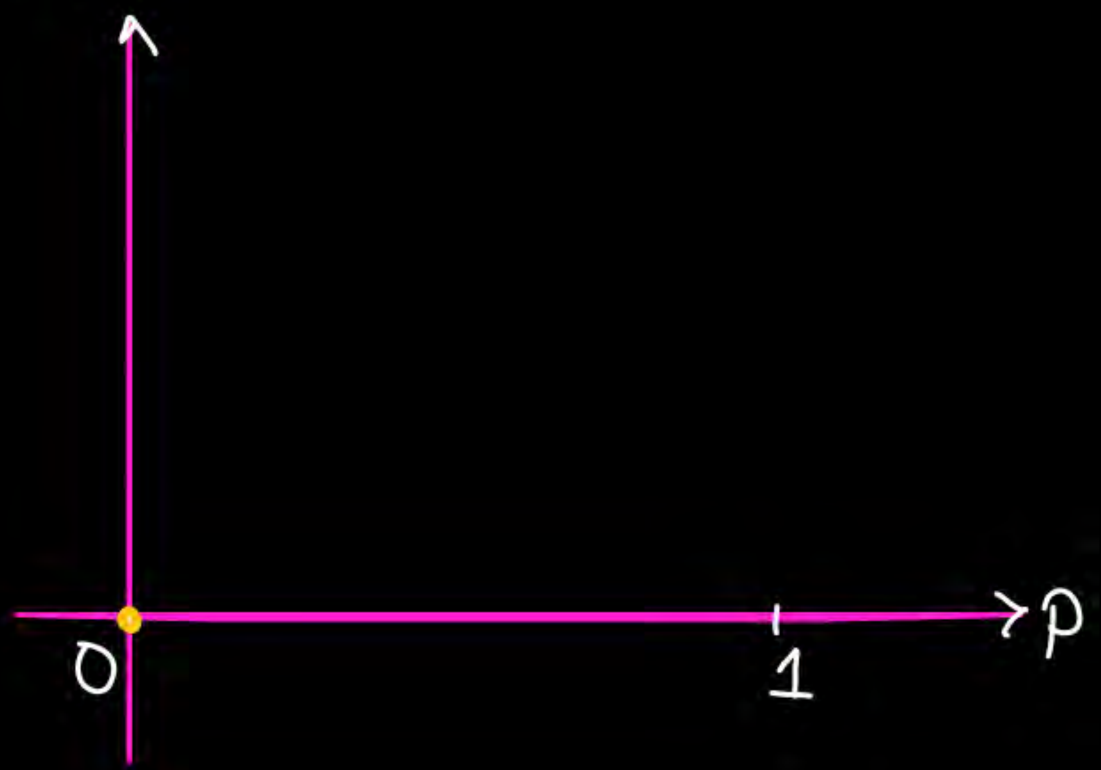
  Class2 : $1-P$

$$\text{Entropy} \Rightarrow \left\{ P\log_2\frac{1}{P} + (1-P)\log_2\frac{1}{1-P} \right\}$$

$$\text{GII} \Rightarrow 1-(P)^2-(1-P)^2$$

- @ $P=0$ Entropy $=0$
- @ $P=0$ GII $=0$
- @ $P=1$ Entropy $=0$

  GII $=0$



@ $P = 0.5$

$$\text{Entropy} \Rightarrow .5\log_2 \frac{1}{.5} + .5\log_2 \frac{1}{.5}$$

$$\Rightarrow .5\log_2 2 + .5\log_2 2$$

$$\Rightarrow .5 + .5 = 1.0$$

$$\text{GII} \Rightarrow 1 - (.5)^2 - (.5)^2$$

$$\Rightarrow 1 - \frac{1}{4} - \frac{1}{4}$$

$$\Rightarrow 1 - 1/2 = 1/2$$

Completely homogeneous

N points
Class1 → N
Class2 → 0

N points
Class 1 = 0
Class 2 = N

N points
Class1 → N/2
Class2 → N/2

← Completely nonhomo.

entropy.
1.0
•5
GI
P
O
•5
Non Homo.
1
Homogeneous

3class data

Npoints

Most non homogeneous

$P_1 = 1/3$

$P_2 = 1/3$

$P_3 = 1/3$

$$\text{Entropy} = \frac{1}{3}\log_2 3 + \frac{1}{3}\log_2 3 + \frac{1}{3}\log_2 3$$

$$= \log_2 3 \rightsquigarrow 1.58$$

$$GI = 1 - \sum_{i=1}^{3} (1/3)^2$$

$$= 1 - \frac{1}{9} \times 3$$

$$= 1 - 1/3$$

$$GI \Rightarrow 2/3 = 0.66$$

- Entropy → 0 to $\log_2$ No of Class
- GI (0 to 1)

Node

$P_1 = .9$

$P_2 = .05$ ✓

$P_3 = .05$ ✓

$$GI = 1 - (.9)^2 - (.05)^2 - (.05)^2 = 1 - .81 - .0025 - .0025$$

- So the classes of v. low Probability actually donot effect the GI

$$\text{Entropy} \Rightarrow .9\log_2\frac{1}{.9} + .05\log_2\frac{1}{.05} + .05\log_2\frac{1}{.05}$$

$$\Rightarrow .136 + .216 + .216$$

- GI take very less Computation → So Generally GI is used
- Entropy is Computationally extensive

But
Entropy give importance to low probabilities also
So it is better to use Entropy in Case of Imbalanced data

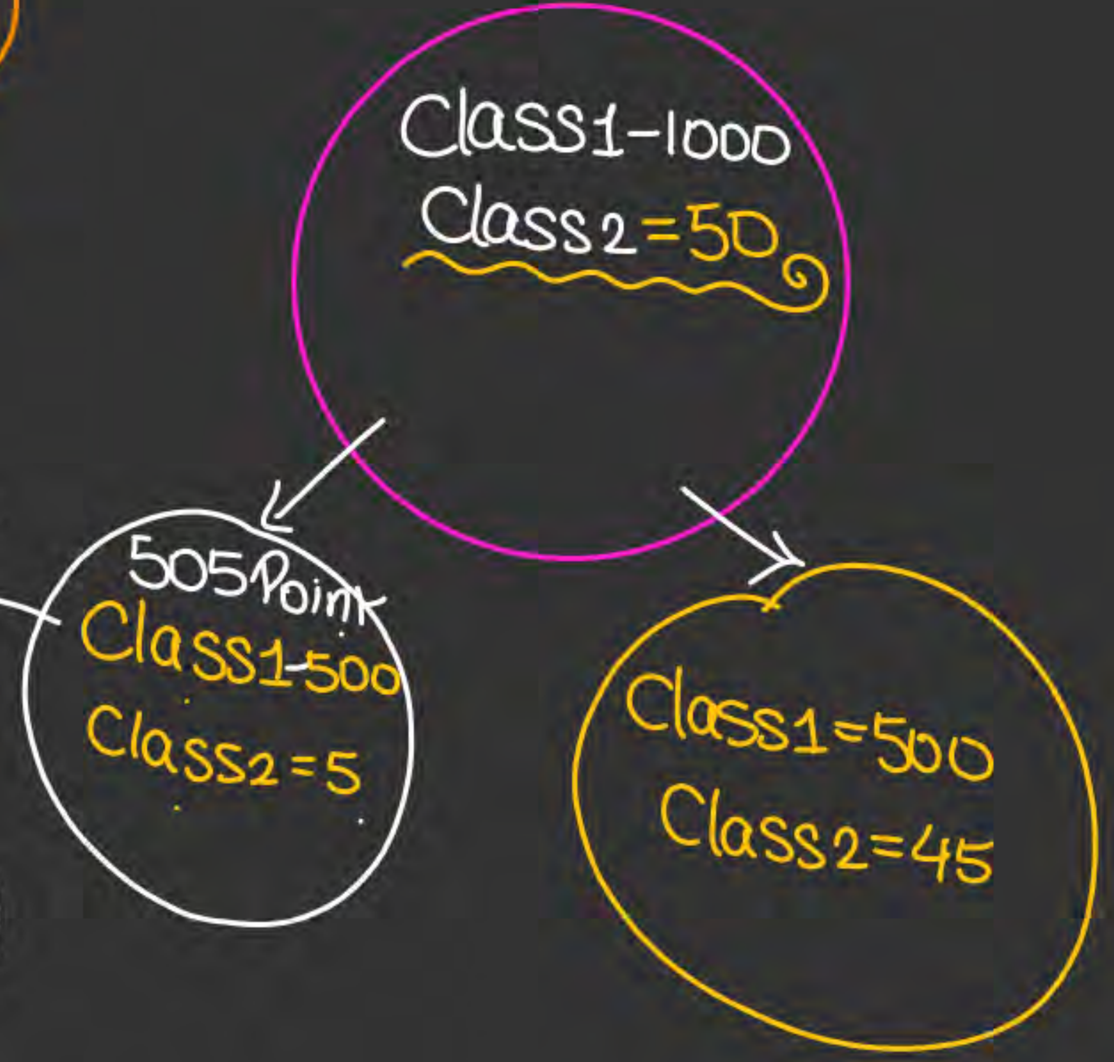
505 Points

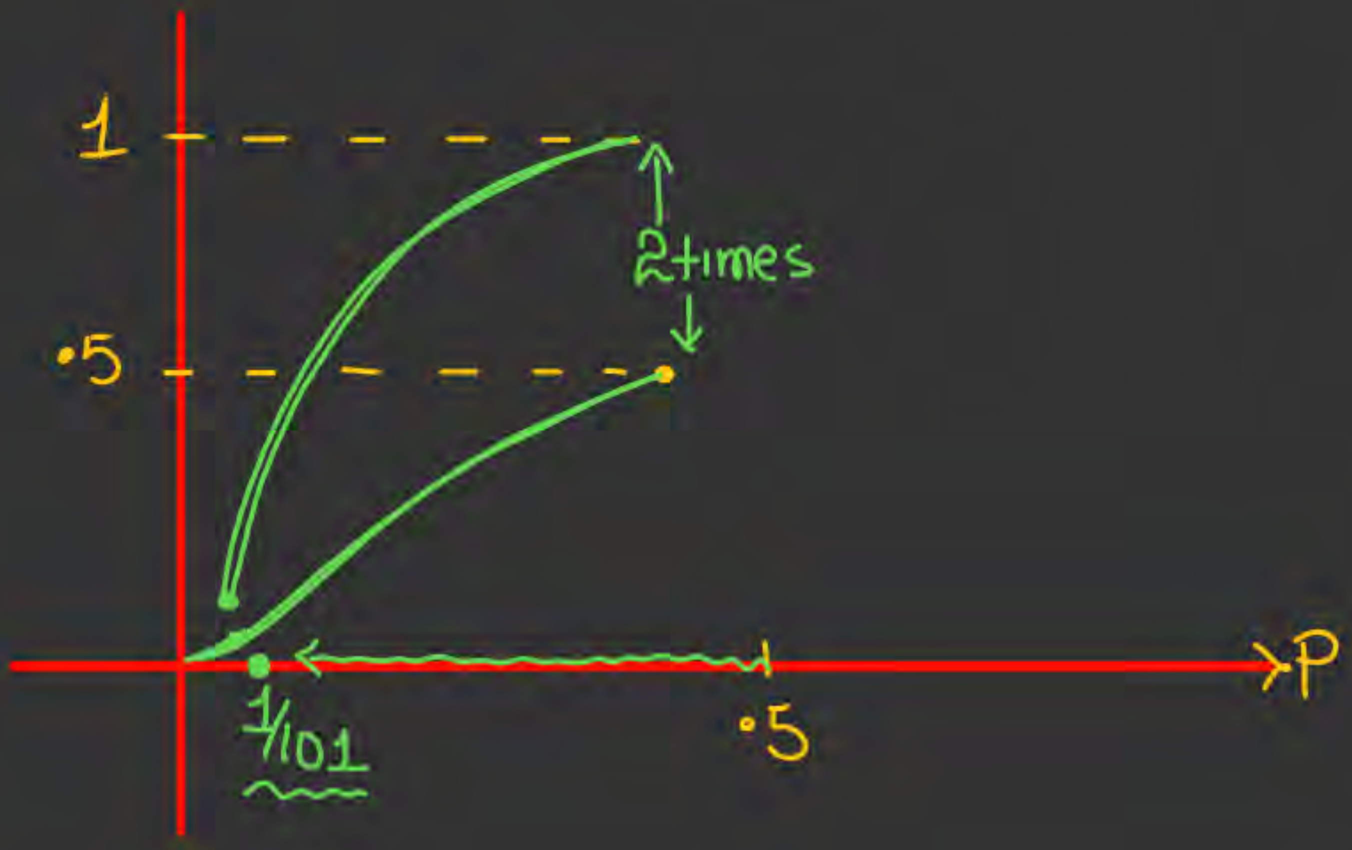
$P_1 = \frac{500}{505}$

$P_2 = 5/505$

GI = .0196

E = .0801





- Why Entropy is Sensitive to No of classes

Becoz max value of entropy = $\log_2 C$

Entropy Vs Gini Index

GI	Entropy
It is the probability of misclassifying a randomly chosen element in a set.	While entropy measures the amount of uncertainty or randomness in a set.
The range of the Gini index is [0, 1], where 0 indicates perfect purity and 1 indicates maximum impurity.	The range of entropy is [0, log(c)], where c is the number of classes.
Gini index is a linear measure.	Entropy is a logarithmic measure.
It can be interpreted as the expected error rate in a classifier.	It can be interpreted as the average amount of information needed to specify the class of an instance.
It is sensitive to the distribution of classes in a set.	It is sensitive to the number of classes.

How to select the attribute for splitting ?

Entropy Vs Gini Index

It is less robust than entropy.	It is more robust than Gini index.
It is sensitive.	It is comparatively less sensitive.
Formula for the Gini index is Gini(P) = $1 - \sum(Px)^2$, where Pi is the proportion of the instances of class x in a set.	Formula for entropy is Entropy(P) = $-\sum(Px)\log(Px)$, where pi is the proportion of the instances of class x in a set.

$\boxed{\text{GI} \Rightarrow \text{falls quickly as compared to Entropy on change of P}}$

(more sensitive to P)

(less sensitive to P)

Decision tree with numerical variables

Predictor

x	y
15	Y
20	N
33	Y
22	N
	N
25	N
18	N
19	Y
30	Y
40	Y Y

dimensions ⇒ Categorical ✓

So if dimension is numerical then all values have to be checked as threshold

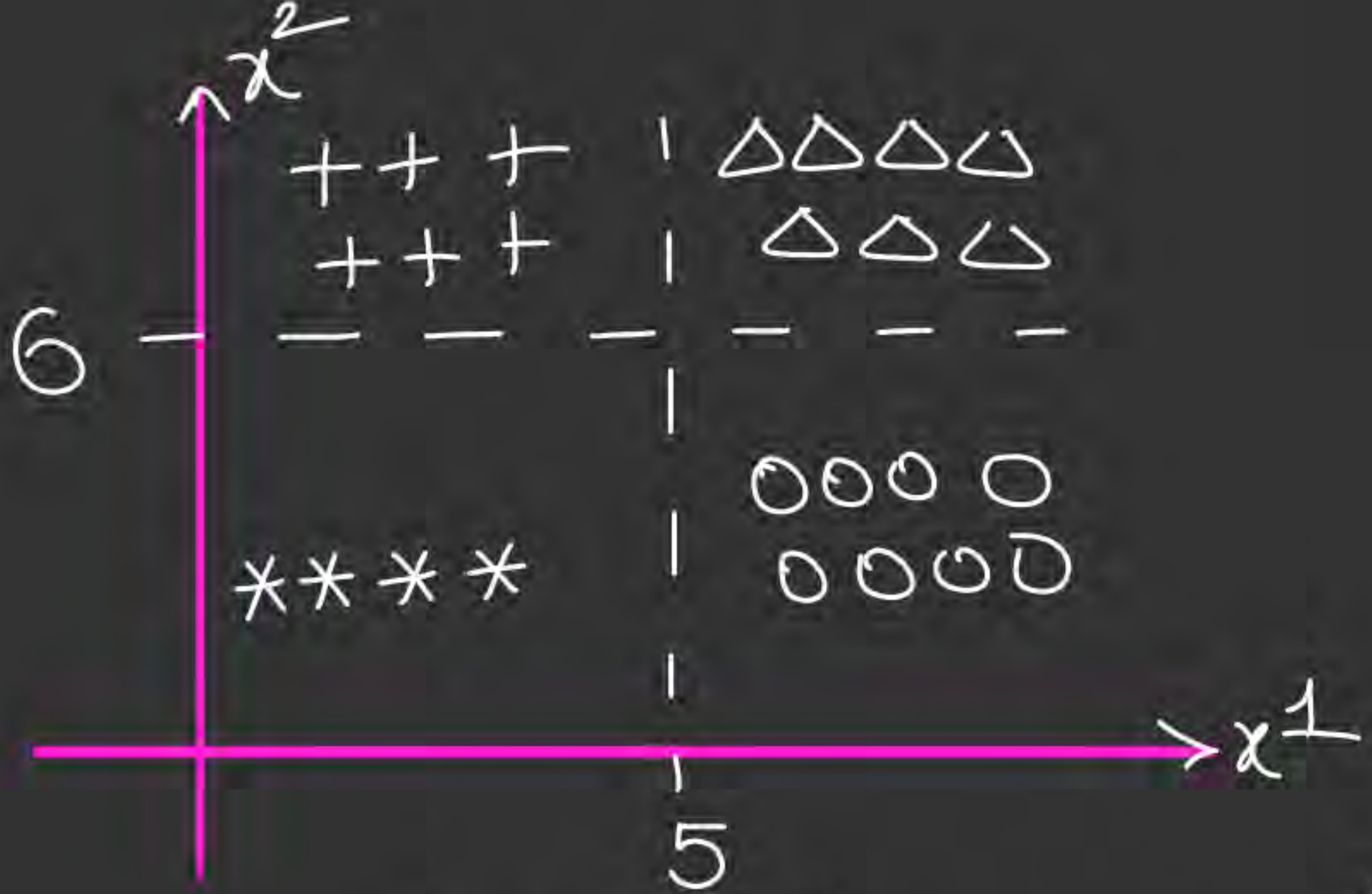
* we have to check for all

→ Very Complex.

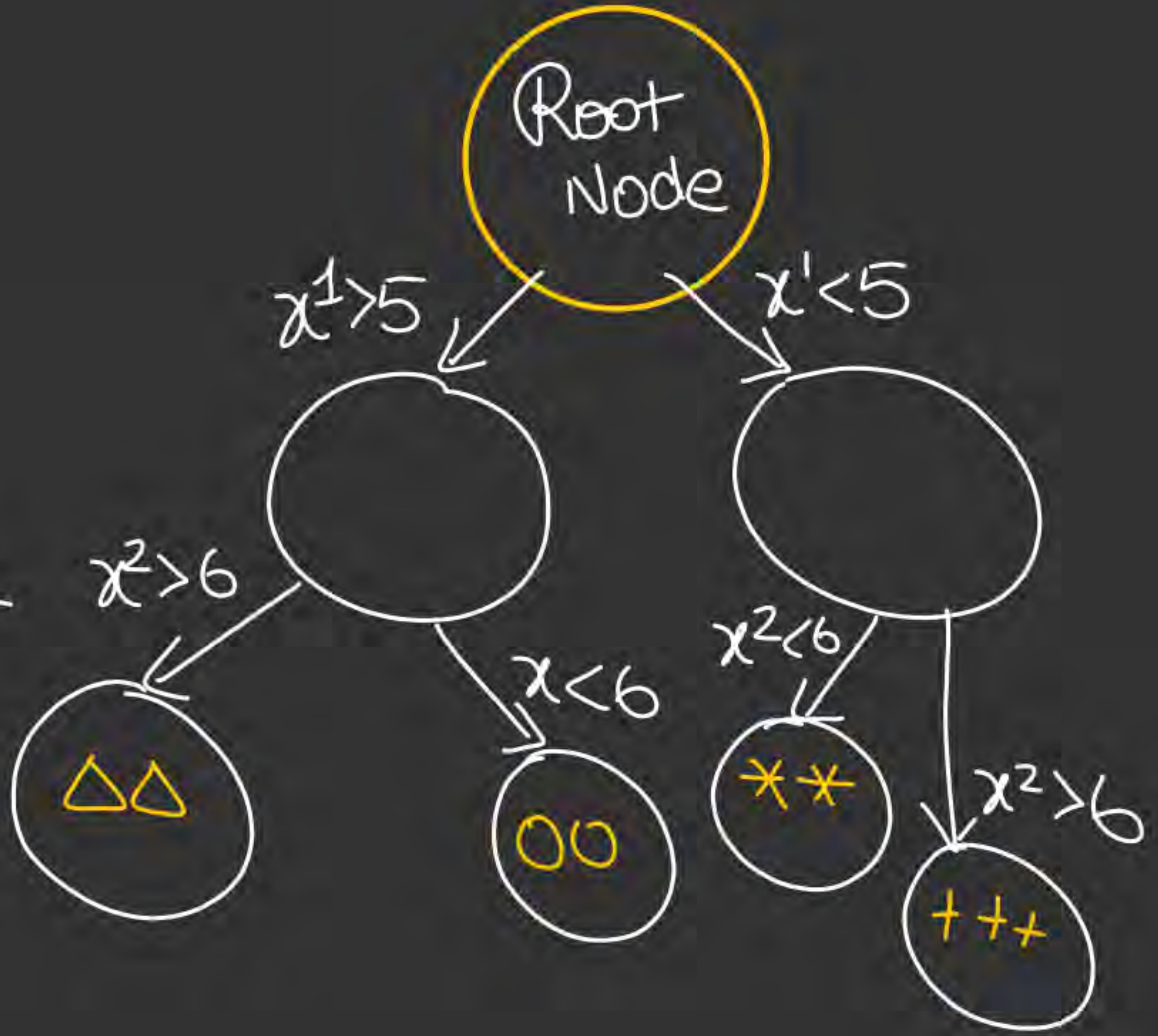
Root node

$x \geq 15$

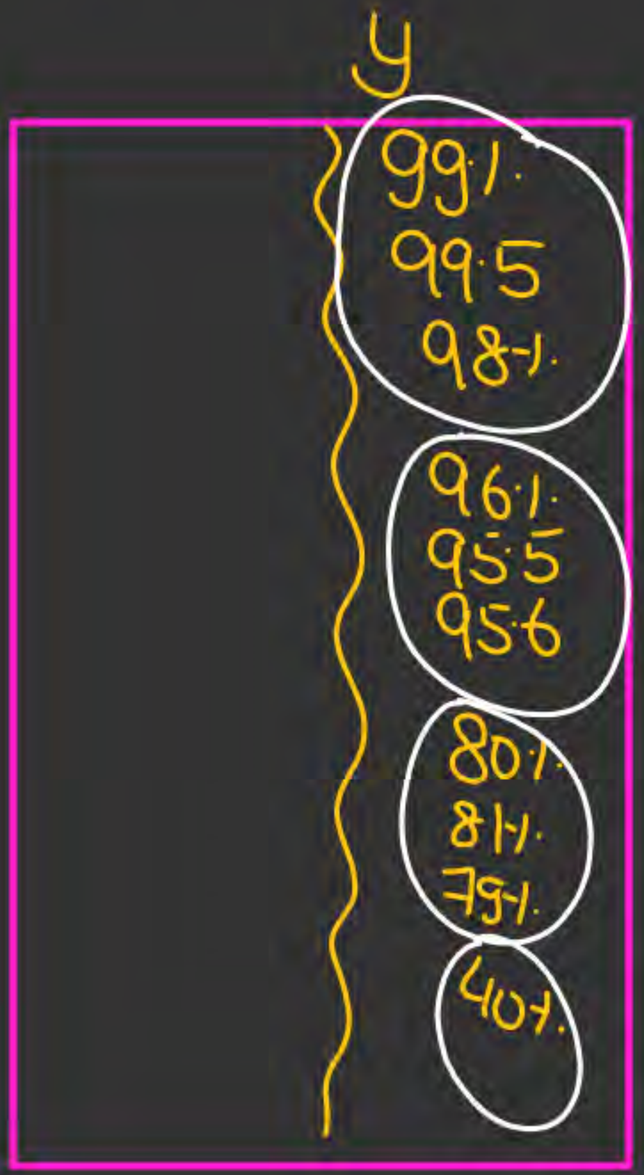
$x \leq 15$



Cont. Predictors



Regression



in decision tree Splitting/grouping

target ⇒ Homogeneous nodes

- In Regression case the homogeneous points are those which have Y values close to each other.
- Variance in Y is v. small.

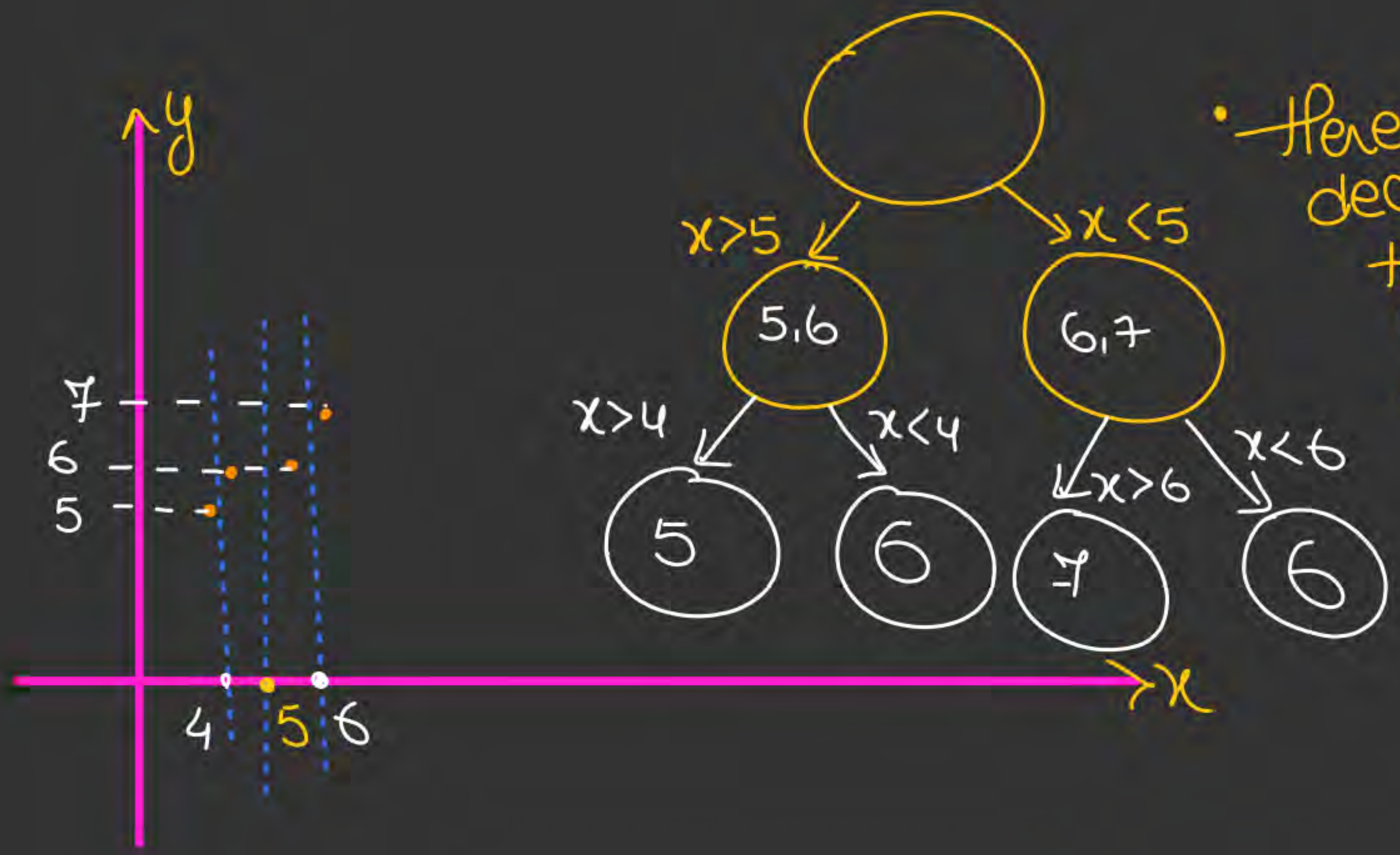
Regression

we measure Impurity of a node ⇒ Variance

Node has some Points ⇒ each Point has Y value ⇒

$$\text{Variance} = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{\text{No of points}}$$

$$IG = \left(\text{Variance}^{P} - \text{Variance}^{C}\right)$$



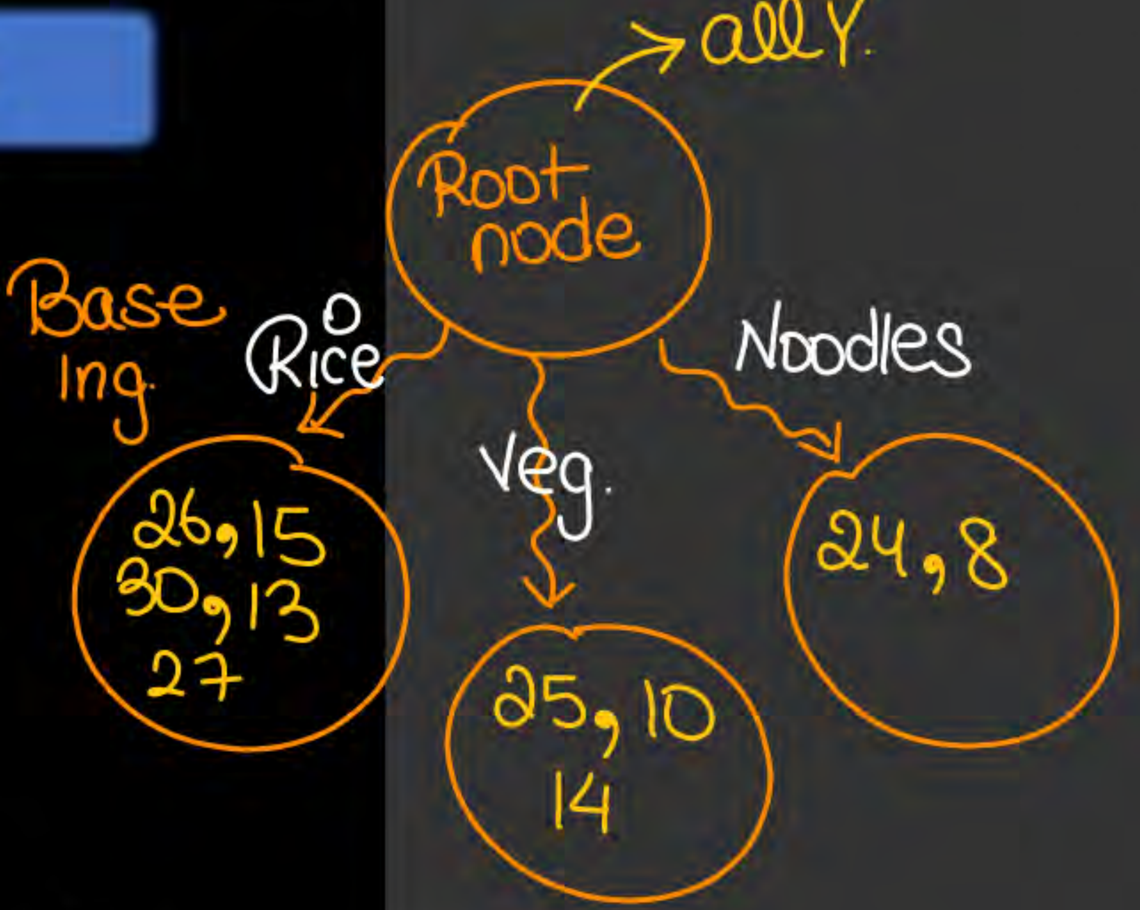
- Here decision tree is overfitting

Decision Tree

Variance as measure of impurity (Regression case)

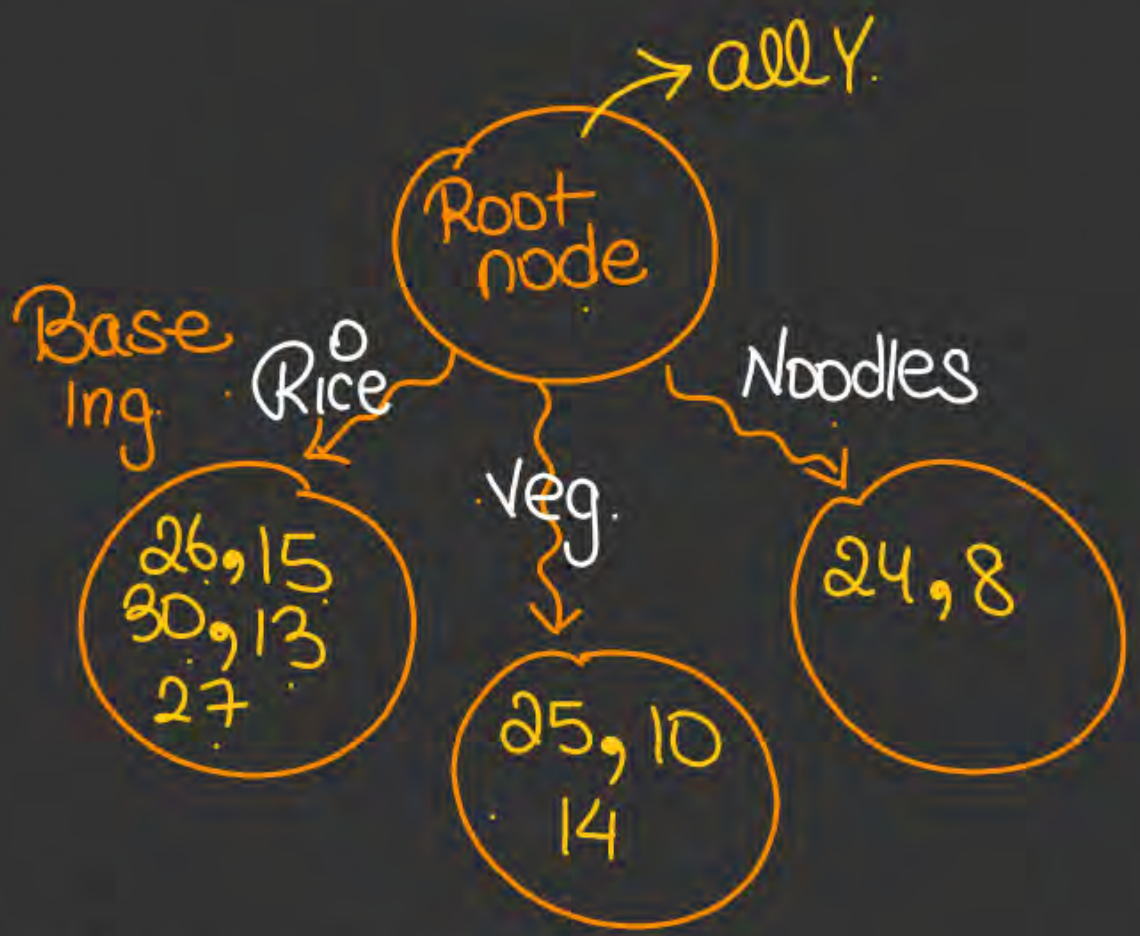
label

Type of Cuisine	Chilies	Cooked for Kids	Base Ingredient	Quantity of Dish	Quantity of Chili Powder
Indian	0	1	Rice	1300	26
Indian	1	1	Rice	800	15
Chinese	1	0	Vegetables	300	25
Thai	1	0	Rice	1500	30
Thai	1	0	Vegetables	980	10
Chinese	1	1	Noodles	1350	24
Indian	0	1	Rice	500	13
Indian	1	0	Noodles	200	8
Indian	1	0	Vegetables	450	14
Thai	1	0	Rice	1250	27



$Var_{Root} \to \bar{y} = 19.2$

$$\frac{\sum_{i=1}^{10} (y_i - 19.2)^2}{10} = 57.36$$



$$Var_{Rice} = \sum_{i=1}^{5} \frac{(y_i - 22.2)^2}{5} = 46.96$$

$$Var_{veg} = \sum_{i=1}^{3} \frac{(y_i - 16.33)^2}{3} = 40.22$$

$$Var_{Nood} = \sum_{i=1}^{2} \frac{(y_i - 16)^2}{2} = 64$$

$$Var^{C} = \frac{5 \times 46.96 + 3 \times 40.22 + 2 \times 64}{10} = 48.34$$

$$Var_{Root} \Rightarrow \bar{y} = 19.2$$

$$\sum_{i=1}^{10} \frac{(y_i - 19.2)^2}{10} = 57.36$$

Variance as measure of impurity (Regression case)

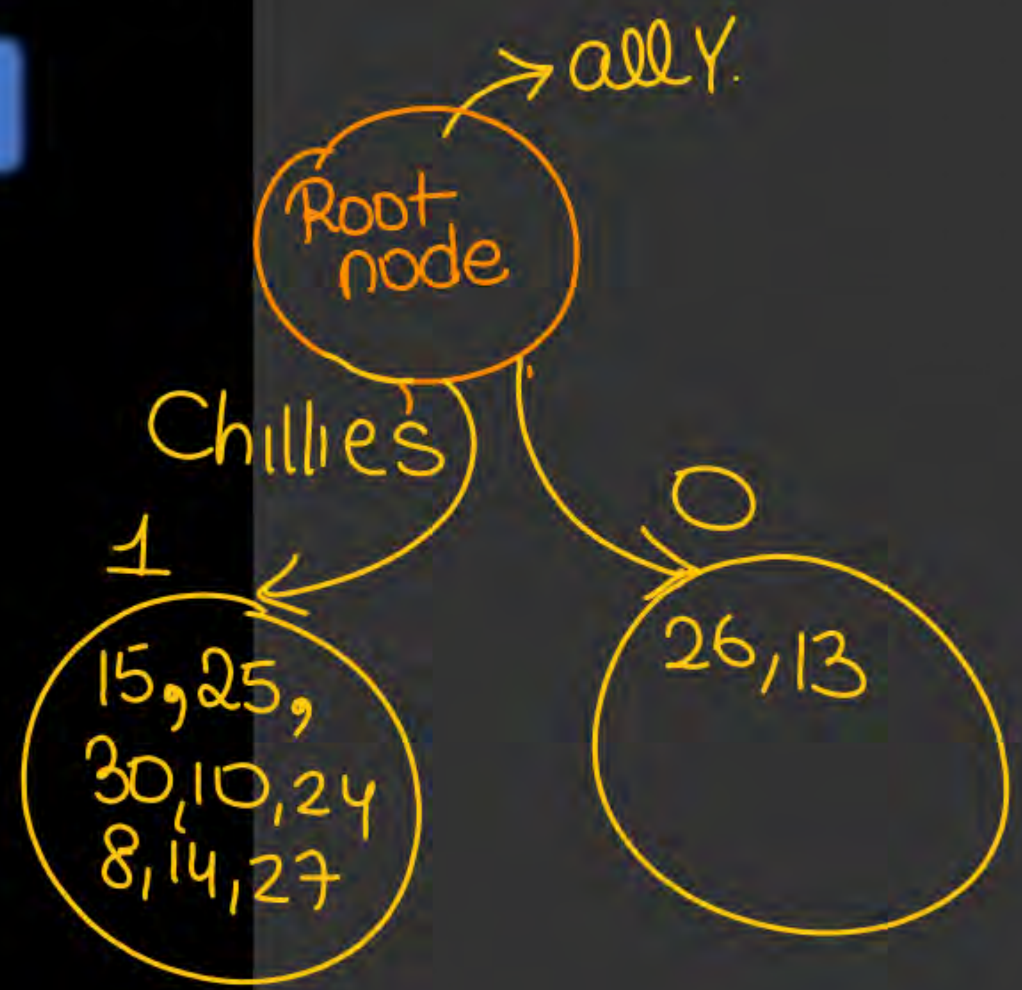
$$\Rightarrow \begin{cases} IG = 57.36 - 48.346 \\ \quad\;\; = 9.014 \end{cases}$$

Decision Tree

Variance as measure of impurity (Regression case)

label

Type of Cuisine	Chilies	Cooked for Kids	Base Ingredient	Quantity of Dish	Quantity of Chili Powder	
Indian	0	1	Rice	1300	26	1
Indian	1	1	Rice	800	15	2
Chinese	1	0	Vegetables	300	25	3
Thai	1	0	Rice	1500	30	4
Thai	1	0	Vegetables	980	10	5
Chinese	1	1	Noodles	1350	24	6
Indian	0	1	Rice	500	13	7
Indian	1	0	Noodles	200	8	8
Indian	1	0	Vegetables	450	14	9
Thai	1	0	Rice	1250	27	10



Decision Tree

Variance as measure of impurity (Regression case)

label

Type of Cuisine	Chilies	Cooked for Kids	Base Ingredient	Quantity of Dish	Quantity of Chili Powder	
Indian	0	1	Rice	1300	26	1
Indian	1	1	Rice	800	15	2
Chinese	1	0	Vegetables	300	25	3
Thai	1	0	Rice	1500	30	4
Thai	1	0	Vegetables	980	10	5
Chinese	1	1	Noodles	1350	24	6
Indian	0	1	Rice	500	13	7
Indian	1	0	Noodles	200	8	8
Indian	1	0	Vegetables	450	14	9
Thai	1	0	Rice	1250	27	10

Root node

all Y.

Quantity of dish

tree will be given

median= 800+980/2

(200 300 450 500 800 980 1250 1300 1350 1500)

avg

So in case of numerical dimension we arrange dimension into increasing order values and take mid value as threshold

or median

Decision Tree

Variance as measure of impurity (Regression case)

Type of Cuisine	Chilies	Cooked for Kids	Base Ingredient	Quantity of Dish	Quantity of Chili Powder
Indian	0	1	Rice	1300	26
Indian	1	1	Rice	800	15
Chinese	1	0	Vegetables	300	25
Thai	1	0	Rice	1500	30
Thai	1	0	Vegetables	980	10
Chinese	1	1	Noodles	1350	24
Indian	0	1	Rice	500	13
Indian	1	0	Noodles	200	8
Indian	1	0	Vegetables	450	14
Thai	1	0	Rice	1250	27

Decision Tree

CART - Classification and Regression Tree Algorithms

- Start with complete training dataset : Root Node of Tree
- Calculate Node Impurity.
- Select the feature for split that results in highest information gain (impurity reduction): ASM
- Split and continue the same process for each node until Stopping Criterion is met
- Majority Class Label : Classification
- Mean Value of target class: Regression

- Splitting help in reducing the bias, it add complexity to the model
- If we keep on splitting it may lead to overfitting

Stopping Criteria

Split till we get homogeneous nodes...

Stopping Criteria

1. Split only when Information Gain > some threshold
2. Number of nodes in split nodes > some minimum value
3. A certain threshold on depth of node
4. Some threshold on Node impurity

Stopping Criteria

Why we need some stopping criteria ?

Decision Tree

What is Pruning in Decision Tree

Remove the branches of the Decision tree

Pruning

- Removing branches from tree.
- It involves simplifying the tree structure, and in effect **regularizes** the model.

Pre-Pruning: this approach involves stopping the tree before it has completed fitting the training set. Pre-Pruning involves setting the model hyperparameters that control how large the tree can grow.

Post-Pruning: here the tree is allowed to fit the training data perfectly, and subsequently it is truncated according to some criteria. The truncated tree is a simplified version of the original, with the least relevant branches having been removed.

Pruning

⇐ prevents overfitting

- Removing branches from tree.
- It involves simplifying the tree structure, and in effect **regularizes** the model.

delete it

Pre-Pruning: this approach involves stopping the tree before it has completed fitting the training set. Pre-Pruning involves setting the model hyperparameters that control how large the tree can grow.

thresholds

Post-Pruning: here the tree is allowed to fit the training data perfectly, and subsequently it is truncated according to some criteria. The truncated tree is a simplified version of the original, with the least relevant branches having been removed.

pre-pruning ≡ stopping criterion

GO CLASSES

* Pre-pruning is simple to implement than post pruning ; also it saves the training time

* But post-pruning is better in terms of building a model because we are exploring all possibilities first & then making a conscious choice to cut some branches

* post pruning is difficult to implement as it involves heavy computation

What is Pruning in Decision Tree

Which is better Pre or Post pruning

Advantage of decision tree

Advantages of Decision Tree

- **Interpretability:** Itis simple to understand, interpret and visualize as the idea is mostly used in our daily lives. Output of a Decision Tree can be easily interpreted by humans.
- Used for both Classification and Regression
- Can handle both categorical and continuous variables.
- No Feature Scaling is required
- Handles non-linear parameters efficiently
- can handle missing values.
- Insensitive to Outliers: extreme values or outliers, never cause much reduction in RSS, they are never involved in split.

Disadvantage of decision tree

- High computation
- The decision tree is non linear and more prone to variance and less bias (Linear algo is has more bias and less variance)

Advantage of decision tree

Dis-advantages of Decision Tree

- Overfitting and High Variance
- **Unstable:** Adding a new data point can lead to re-generation of the overall tree and all nodes need to be recalculated and recreated.
- Not suitable for large datasets: If data size is large, then one single tree may grow complex and lead to overfitting. We should try ensemble model here.

We always have to create new tree if we have new data

Is linear regression also unstable ?

Practise

7. Which of the following statements is not true about Information Gain?

a) It is the addition in entropy by transforming a dataset

b) It is calculated by comparing the entropy of the dataset before and after a transformation

c) It is often used in training decision trees

d) It is also known as Kullback-Leibler divergence

Practise

8. Which of the following statements is not true about Information Gain?

a) It is the amount of information gained about a random variable or signal from observing another random variable

b) It tells us how important a given attribute of the feature vectors is

c) It implies how much entropy we removed

d) Higher Information Gain implies less entropy removed

Practise

9. Given the entropy for a split, $E_{split} = 0.39$ and the entropy before the split, $E_{before} = 1$. What is the Information Gain for the split?

a) 1

b) 0.39

c) 0.61

d) 2.56

Practise

10. Which of the following statements is not an objective of Information Gain?

a) It tries to determine which attribute in a given set of training feature vectors is most useful for discriminating between the classes to be learned

b) Decision Trees algorithm will always tries to minimize Information Gain

c) It is used to decide the ordering of attributes in the nodes of a decision tree

d) Information Gain of certain event is the discrepancy of the amount ofinformation before someone observes that event and the amount after observation

Practise

14. Given entropy of parent = 1, weights averages = $(\frac{3}{4}, \frac{1}{4})$ and entropy of children = (0.9, 0). What is the information gain?

a) 0.675

b) 0.75

c) 0.325

d) 0.1

Practise

Question: 1

Which of the following is a common method for splitting nodes in a decision tree?

(A) Gini impurity

(B) Cross-validation

(C) Gradient descent

(D) Principal component analysis

Practise

Question: 2

What is the main disadvantage of decision trees in machine learning?

(A) They are prone to overfitting

(B) They cannot handle categorical variables

(C) They cannot model non-linear relationships

(D) They are computationally expensive

Practise

Question: 3

What is the purpose of pruning in decision trees?

(A) To reduce the depth of the tree and prevent overfitting

(B) To optimize the tree's parameters

(C) To handle missing data

(D) To improve the tree's interpretability

Practise

Question: 4

Which of the following is a popular algorithm for constructing decision trees?

(A) ID3

(B) k-Nearest Neighbors

(C) Support Vector Machines

(D) Naive Bayes

Decision Tree

Practise

What is the main difference between classification and regression trees (CART)?

(A) Classification trees predict categorical variables, while regression trees predict continuous variables

(B) Classification trees use Gini impurity as the splitting criterion, while regression trees use information gain

(C) Classification trees can handle missing data, while regression trees cannot

(D) Classification trees are computationally expensive, while regression trees are computationally inexpensive

Practise

What is the primary purpose of the Random Forest algorithm?

(A) To combine multiple decision trees to improve prediction performance

(B) To optimize the parameters of a single decision tree

(C) To handle missing data in decision trees

(D) To visualize the decision boundaries of a decision tree

Practise

Which of the following is a popular method for splitting nodes in a regression tree?

(A) Gini impurity

(B) Information gain

(C) Mean squared error

(D) Cross-validation

Practise

What is entropy in the context of decision trees?

(A) A measure of disorder or impurity in a node

(B) A measure of the complexity of a decision tree

(C) The difference between the predicted and actual values in a node

(D) The rate at which information is gained in a decision tree

Practise

Which of the following is a common stopping criterion for growing a decision tree?

(A) Reaching a maximum depth

(B) Achieving a minimum information gain

(C) Achieving a minimum Gini impurity

(D) Both A and B

Practise

What is the main disadvantage of using a large maximum depth for a decision tree?

(A) It leads to overfitting

(B) It reduces the interpretability of the tree

(C) It increases the computational complexity of the tree

(D) It causes the tree to underfit the data

Practise

Which of the following techniques can be used to reduce overfitting in decision trees?

(A) Pruning

(B) Bagging

(C) Boosting

(D) All of the above

Practise

Which of the following is a disadvantage of using decision trees for regression tasks?

(A) Decision trees cannot handle continuous variables

(B) Decision trees are prone to overfitting

(C) Decision trees are sensitive to small changes in the data

(D) Both B and C

Practise

Which of the following is a disadvantage of using decision trees for classification tasks?

(A) Decision trees cannot handle categorical variables

(B) Decision trees are prone to overfitting

(C) Decision trees cannot model non-linear relationships

(D) Decision trees are computationally expensive

Practise

Which of the following is an ensemble learning technique that uses decision trees as base learners?

(A) Random Forest

(B) k-Nearest Neighbors

(C) Support Vector Machines

(D) Naive Bayes

Practise

How can decision trees be made more robust to noise in the data?

(A) By increasing the maximum depth of the tree

(B) By using a smaller minimum samples per leaf

(C) By using ensemble techniques like bagging or boosting

(D) By removing features with low importance

Practise

In a decision tree, what is the purpose of the leaf nodes?

(A) To represent the class label or value to be predicted

(B) To store the conditions for splitting the data

(C) To indicate the importance of a feature

(D) To represent the depth of the tree

Practise

What is the primary advantage of using decision trees in machine learning?

(A) They are computationally inexpensive

(B) They are easy to interpret and visualize

(C) They can handle missing data

(D) They have high predictive accuracy

THANK - YOU