Data Science and Artificial Intelligence

Machine Learning

Decision Tree

Lecture No. 3

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- Topic
- Topic
- Topic
- Topic
- Topic

Decision tree → Variance

Topics to be Covered

- Topic
- Topic
- Topic
- Topic
- Topic

Questions

Stopping Criteria

ID3, CART

Pruning


"If you are working on something
exciting that you really care about,
you don't have to be pushed.
The vision pulls you."
- Steve Jobs
ROOPVIBES.COM

Basics of Machine Learning

Variance → Regression decision tree

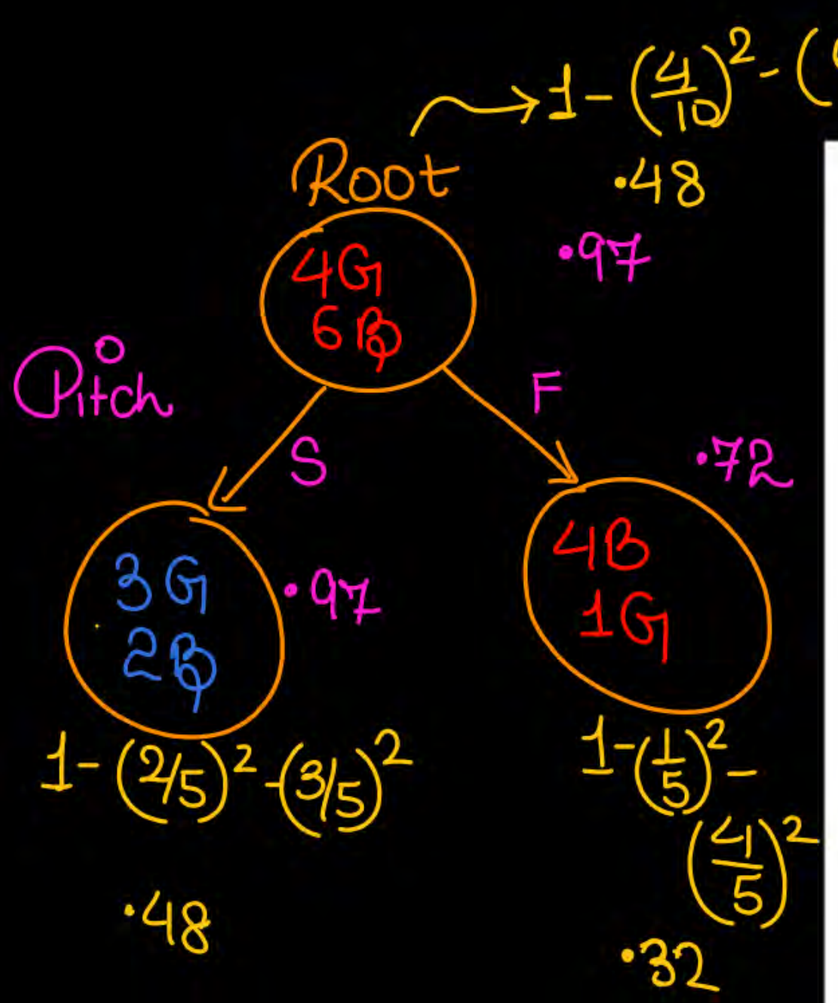
$$\frac{\sum (y_i - \bar{y})^2}{\text{Number of Points}}$$

$IG = \text{Variance}^{P} - \text{Variance}^{Child}$ → Weighted avg.

Basics of Machine Learning

Information gain

done



Q.62 Details of ten international cricket games between two teams "Green" and "Blue" are given in Table C. This table consists of matches played on different pitches, across formats along with their winners. The attribute *Pitch* can take one of two values: spin-friendly (represented as *S*) or pace-friendly (represented as *F*). The attribute *Format* can take one of two values: one-day match (represented as *O*) or test match (represented as *T*).

A cricket organization would like to use the information given in Table C to develop a decision-tree model to predict outcomes of future games between these two teams.

To develop such a model, the computed InformationGain(C, *Pitch*) with respect to the Target is _____ (rounded off to **two** decimal places).

classification

Table C

Match Number	Pitch	Format	Winner (Target)
1	S	T	Green
2	S	T	Blue
3	F	O	Blue
4	S	O	Blue
5	F	T	Green
6	F	O	Blue
7	S	O	Green
8	F	T	Blue
9	F	O	Blue
10	S	O	Green

$GI^C \Rightarrow \frac{5 \times .48 + 5 \times .32}{10}$

$\Rightarrow .4$

So $I.G = GI^P - GI^C$

$= .48 - .4$

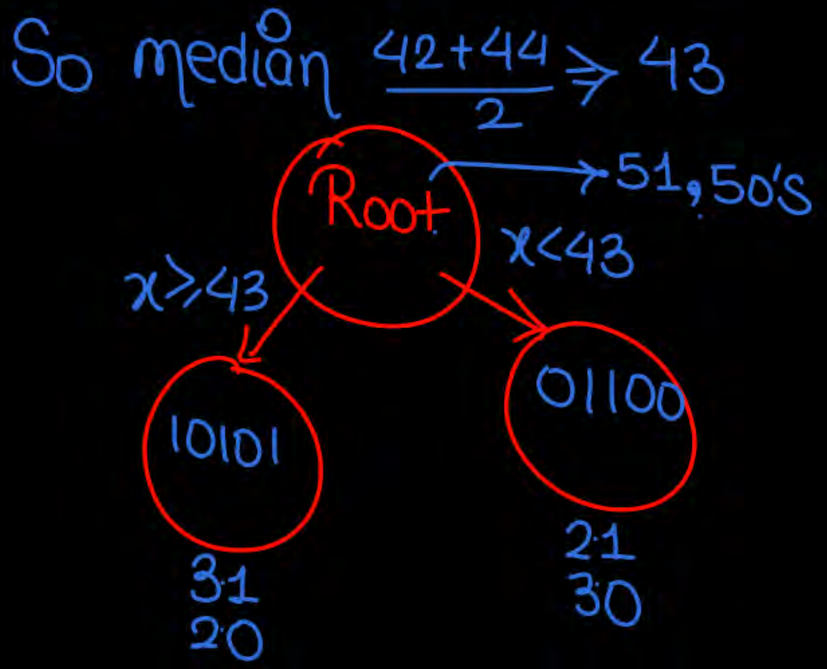
$= .08.$

Entropy $\Rightarrow$

.125

Find the Information gain if Likes gravity is used as a dimension for splitting.

- Use age as splitting criteria take median value as threshold



age	likes dogs	likes gravity	going to be an astronaut
24	0	0	0
30	1	1	1
36	0	1	1
36	0	0	0
42	0	0	0
44	1	1	1
46	1	0	0
47	1	1	1
47	0	1	0
51	1	1	1

→ Label.

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

$GI_{>43} = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = .48$

$GI_{<43} = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = .48$

$GI^C = \frac{5 \times .48 + 5 \times .48}{10} = .48$

$IG = .5 - .48 = (.02)$

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

Decision Tree Algorithms

There are many algorithms there to build a decision tree. They are

1. **CART** (Classification and Regression Trees) — This makes use of Gini impurity as the metric.
2. **ID3** (Iterative Dichotomiser 3) — This uses entropy and information gain as metric.

Decision Tree

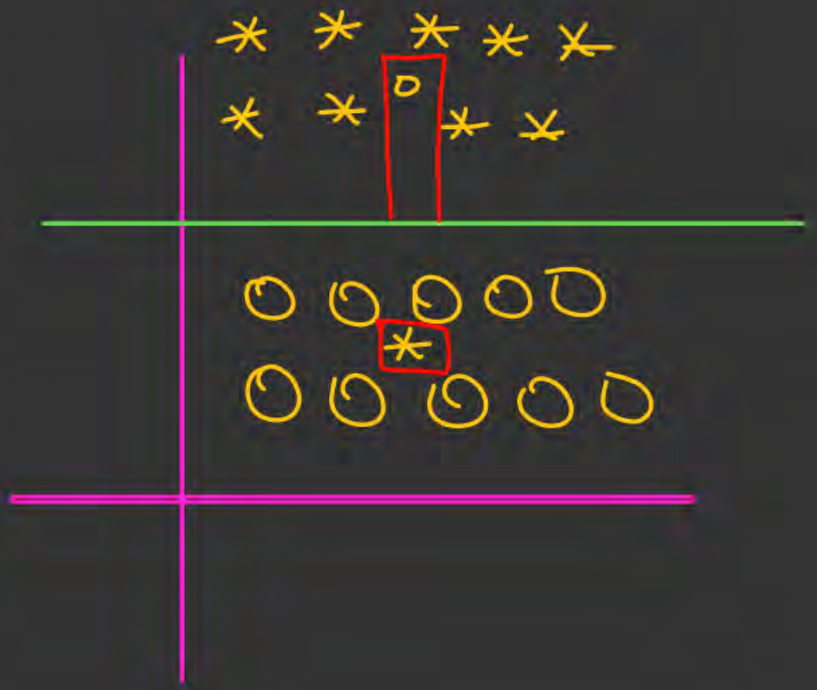
CART - Classification and Regression Tree Algorithms

- 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

} leaf node

ASM ⇒ Attribute Selection measures

More the split ⇒
- model Complex
- Bias reduce
- overfitting



Decision Tree

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

Decision Tree

Stopping Criteria

Split till we get homogeneous nodes...

- we were spitting till we get homogeneous nodes ⇒ But this lead to overfitting.

- Stopping Criteria
  - GI → 0
  - Entropy → 0

1. If GI of any node < Threshold then do not Split.
2. the Information gain < threshold, then do not Split
3. depth of tree Can be a Constraint
4. Number of point in a node < threshold, then do not Split.

Decision Tree

Stopping Criteria

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

Decision Tree

Stopping Criteria

- helpin reducing overfitting.

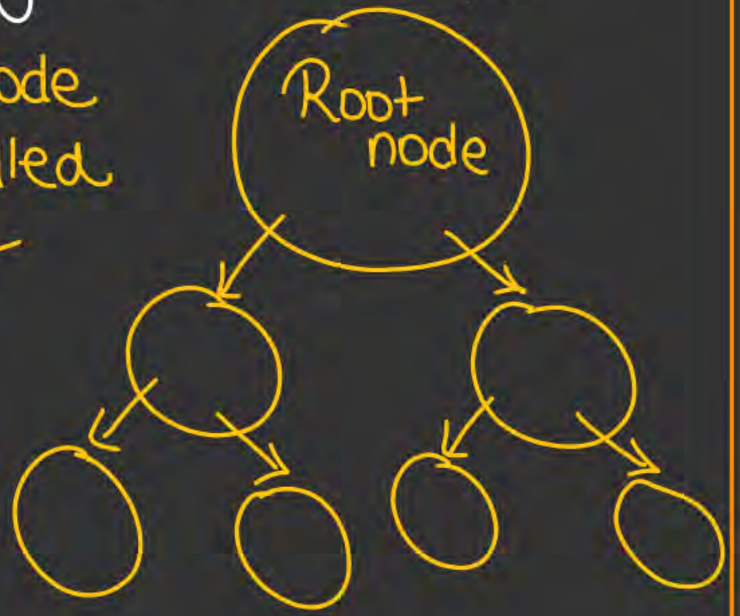
Why we need some stopping criteria ?

we have 2 methods to do this

1. Pre-Pruning → stopping
2. Post-Pruning

- Pre Pruning ⇒ Tree is created Considering the thresholds.

i.e we start with Root node and splitting is Cancelled if thresholds are not met.



Post-Pruning → First of all the DT is allowed to overfit data

→ means DT will observe the whole data and patterns in data

→ After creation of DT the branch are cut based on thresholds.

So Post Pruning is better but Computationally intensive

Decision Tree

What is Pruning in Decision Tree

Remove the branches of the Decision tree

Pruning → Stopping

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

Remove overfitting

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.

Stopping Criteria

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.

Decision Tree

What is Pruning in Decision Tree

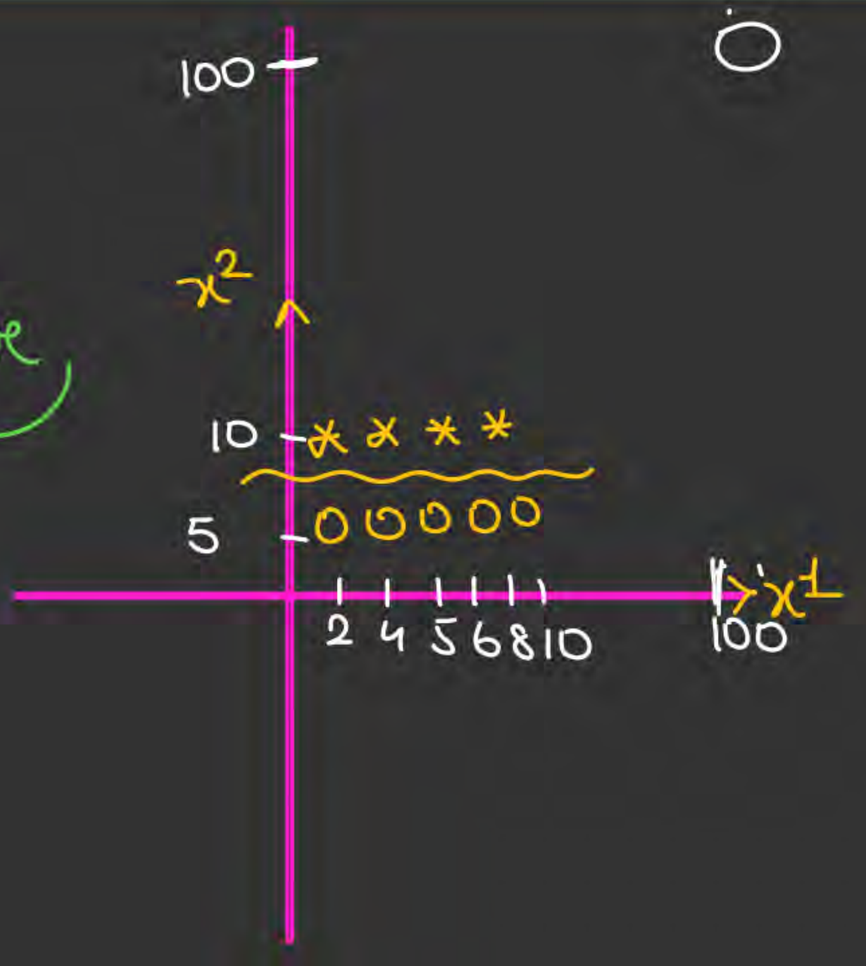
Which is better Pre or Post pruning

→ Post pruning is good but we use pre pruning


Class1
Class0
Class
Class1
DT is non linear Algo.

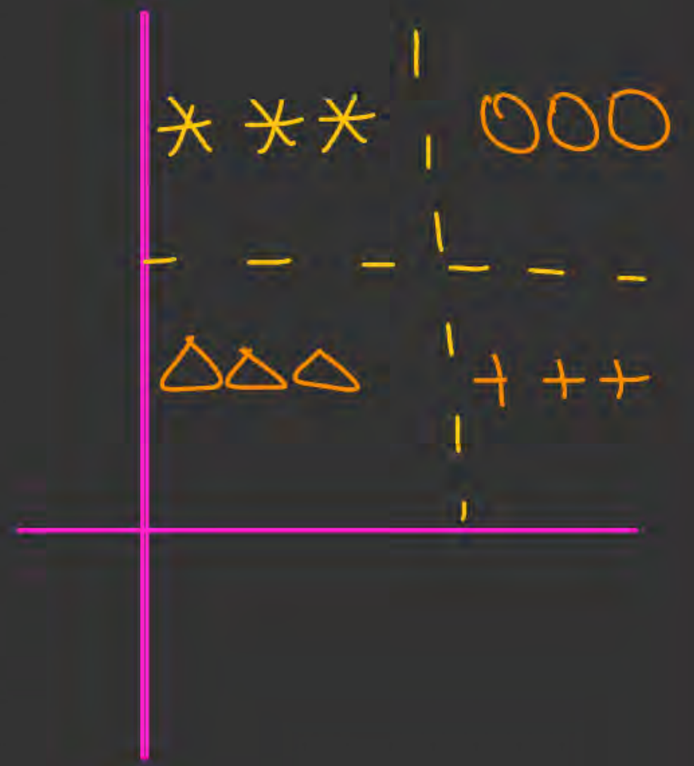
Why DT is not effected by Outliers

x^2	$x1$
5	2
5	4
5	5
5	6
10	8
10	10
10	
100	100



Advantage of DT ⇒

- → High interpnet ability
- → Can be used for R+C
- → non linear, nonparametric
- → not effected by outlier



because always the splitting is started from median value of dimension.

median is not effected by outlie points.

- and in DT the most imp thing is the bifurcation/splitting threshold.
  - If these threshold are not effected by the Outlier, DT is not effected by them.

→ Feature Scaling is not needed.

→ Space Complexity is v. low.

disadvantage

- → training time Complexity huge
- → prone to overfit
- → Computationally extensive
- → unstable Algo, If data changes we have to create whole new DT.

→ not good for large dataset,

that is why Random forest agorithm.

ensemble Technique

Decision Tree

Disadvantage of decision tree

done

- 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)

Decision Tree

Practise

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

a) It is the decrease 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

Decision Tree

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

higher

Decision Tree

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

(1 - .39)

Decision Tree

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~~ max. 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

Decision Tree

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

$E^C = .9 \times 3/4 + 0 \times 1/4$

$\Rightarrow 2.7/4$

$IG = E^P - E^C$

$= .325$

Decision Tree

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

Decision Tree

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

Decision Tree

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

Decision Tree

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

Decision Tree

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

Decision Tree

Practise

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

(A) Gini impurity

(B) Information gain → Parent $ child Impurity.

(C) Mean squared error

(D) Cross-validation

Decision Tree

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

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, C

Decision Tree

Practise

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

ABC

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

Decision Tree

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

Decision Tree

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

Decision Tree

Practise

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

B, D

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

Decision Tree

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

Removing unwanted features

- inc overfit
- reduce overfit

Decision Tree

Practise

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

↳ donot allow overfit.

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

Decision Tree

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

Decision 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