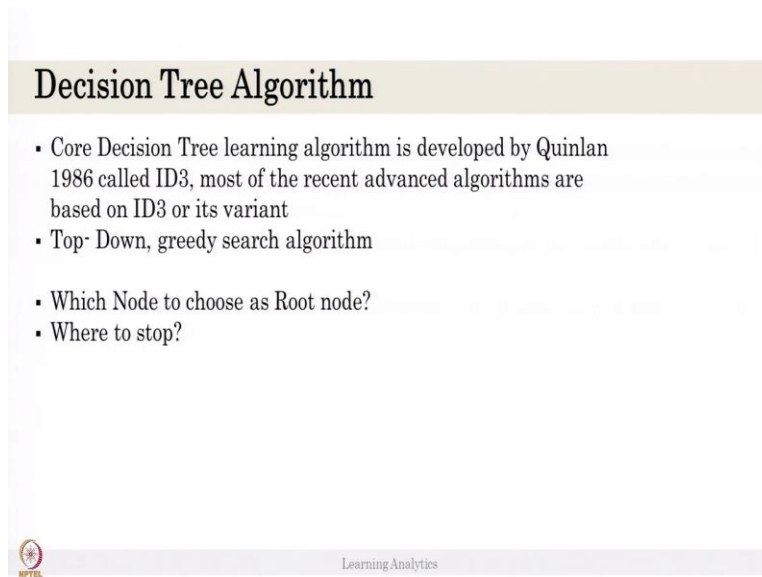



Learning Analytics Tools
Professor Ramkumar Rajendran
Department of Educational Technology
Indian Institute of Technology Bombay
Lecture 9.4
Decision Tree Classifier

(Refer Slide Time: 00:26)



Decision Tree Algorithm

- Core Decision Tree learning algorithm is developed by Quinlan 1986 called ID3, most of the recent advanced algorithms are based on ID3 or its variant
- Top-Down, greedy search algorithm
- Which Node to choose as Root node?
- Where to stop?

 Learning Analytics


In this video, we will continue on Decision Tree Classifier. So, we saw that decision tree is developed in 1986 as ID3 algorithm and lot of variance of them is used in the current modern tools. Let us look at what is ID3 algorithm. It actually answer two things, which node to choose as root node? And where to stop? This is a very key questions, you have to answer in decision tree.

(Refer Slide Time: 00:43)

Decision Tree

Stud.ID	Attendance in %	Mid Term Marks	Final Marks > 70
1	56	45	No
2	45	32	No
3	85	56	Yes
4	80	73	Yes
5	90	65	Yes
6	100	80	Yes
7	95	65	Yes

Which Node to choose as Root node?
Where to stop.

 Learning Analytics 3


So, let us see with this particular table you have attendance, mid-term marks, final marks is greater than 70 you have seen this table multiple times. Which node to choose at the root node? And where to stop? I am not going to explain solving decision tree in this particular table, but I am going to explain the mathematics behind which node to choose and where to stop and you can apply that formulas on this.


(Refer Slide Time: 01:09)

Entropy and Information Gain

- Entropy (S) = $-P(\text{yes}) \log_2 P(\text{yes}) - P(\text{no}) \log_2 P(\text{no})$
 $-0.71 \log_2 0.71 - 0.29 \log_2 0.29 = 0.87$
- To measure the amount of uncertainty in the data

$$\text{Entropy (S)} = \sum_{i=1}^c -p_i \log p_i$$

 Learning Analytics



Decision Tree

Stud.ID	Attendance in %	Mid Term Marks	Final Marks > 70
1	56	45	No
2	45	32	No
3	85	56	Yes
4	80	73	Yes
5	90	65	Yes
6	100	80	Yes
7	95	65	Yes

Which Node to choose as Root node?
Where to stop.

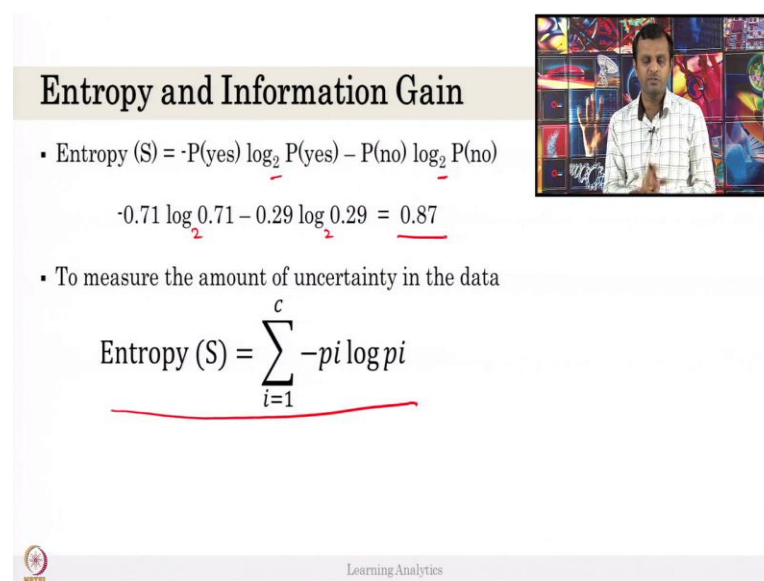
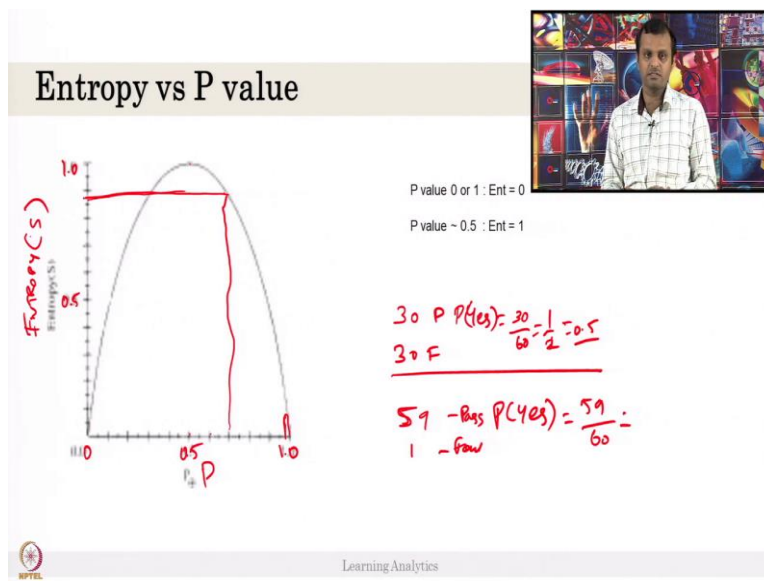


It involves only two things one is Entropy and Information gain, you all might know what is entropy, entropy is different in different fields. It is to measure the uncertainty in the data. So that is entropy formula is simple, this is the formula, $-\sum p_i \log p_i$, this i is number of classes, if it is binary that is yes or no classes, it is two. If it is four classes to classify in decision tree, it will be four something like that. So probability of p_i number classes, so p_i is probability of yes and the p_2 is probability of no.

So if I apply this to two class problem like yes or no, probability of yes log probability of yes and probability of no log probability of no, that is exactly given in this equation. We have \log_2 check this base is very very important, there are two classes that is why we should log 2. If you compute this with log 2 base value for the previous table point previous table, let us see how many yes and how many no are there.

So 5 out of 7 are yes, this 0.71 and 2 out of 7 are no, 0.29, if you compute the entropy value that then it is the value 0.87. What is this 0.87 means?

(Refer Slide Time: 03:14)



Let us understand what is entropy with the probability the simple diagram. This is the probability value along x-axis, this is the entropy along y. So, this probability value is 0.5, then this entropy value is 1 and entropy is 0 and probability value is 0.5 and. So probability value 0 or 1 the entropy will be 0, the probability value 0 or 1 entropy will be minimum for two class classification.

For probability values is 0.5, the entropy will be maximum. So let us try to understand what is entropy measures from the probability value, so that this value is maximum. So what it tells is in a previous example, consider I have 30 students passing the exams and 30 students failing the

exam. So, probability of passing the exam probability of yes that is a pass the exam is 30 by 60 total students, so 1 by 2, so 0.5 probability of passing the exams 0.5, which means opposite, that probability of failing is 0.5. If there are 60 students of anyone of them can be pass or fail there is uncertainty like, pass or fail. If you pick any student, it can be either pass or failed. So the entropy is high. Consider again for 60 students, I have say 59 students pass and only 1 student fail.

So, which means probability of yes is 59 divided by 60 it is almost 1, 0.99 for something maybe. So it is almost 1 if probability of S is almost 1, hence the entropy will be very minimum. So it tells the uncertain to classifying this thing is very very low. So if you pick that value any student from the class, you can say pass because 59 out of 60 pass only one time it will fail.

Our aim in a decision tree classifier is to bring that entropy equal to low not high because I want to make a decision of one leaf to be this student will attend the class this all students here or they will not attend the class. So my aim here in decision trees to bring that in entropy value to as minimum as possible that is it. That is a basic concept of what decision tree works off.

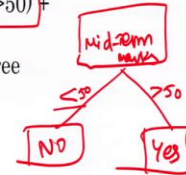
So, which node to choose, now you can understand which node to choose you have to apply entropy on all classes so there are seven classes entropy is 0.87 and you have to select one of the feature and create one of the application say the attendances is high, if you put the attendance equal to high, what happens to the entropy? If information gain entropy is reducing or not. How do you know the entropy values is reducing or not? It is can be computed by the information gained.

(Refer Slide Time: 08:11)

Entropy and Information Gain

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

- Gain(S, Mid-Term Marks) = Entropy(S) - (5/7 * Entropy(>50) + 2/7 * Ent(<50))
- Select a node with highest attribute will be used split the tree



Learning Analytics

Decision Tree

Stud.ID	Attendance in %	Mid Term Marks	Final Marks > 70
1	56	45	No
2	45	32	No
3	85	56	Yes
4	80	73	Yes
5	90	65	Yes
6	100	80	Yes
7	95	65	Yes

Which Node to choose as Root node?
Where to stop.



Learning Analytics

3

Let us look at what is information gain. What happens is, it takes the entropy of the complete set without making any decision, it takes all root. So entropy of complete set is 0.87, I am not taking any decision just given a particular student whether student will be pass or not I just make a decision out of it.

So, I will true about 71 percentage of time but we want to better classifier. Let us say I will select one feature. The feature is selected is mid-term marks.

So, I consider the mid-term marks as a node, you have to select all other features and select different values for features. Let us say mid-term marks as the node and if I consider mid-term

marks as the node so I have select a one value. So, let us see if I select a mid-term marks as a node we need to compute, what are the classification you can make from the mid-term mark.

Now you need to make a decision whether you want two branches out of it a three branch out of it, that is two child's, three child's. I make it two child, one is less than 50 and one is greater than 50 so I make a decision here. So how this number comes up it is how also the mission like you have to compute different numbers so you can make it to different categories or that is all about how do you start with it.

So, let us say I will take a mid-term marks the root node, and I computed only two decision, less than 50 and more than 50.

So let us consider the midterm marks has a two values. I want to make it two decision less than 50 more than 50 only two values. So that is exactly two values which means what happens here is entropy of S this is basically less than 50 marks there are five students who got less than 50 marks and there are two students out of seven got less than 50 marks.

Let us look at the table again once, in the mid-term marks that is exactly here. There are two students who got less than 50 marks and all other student's five students who got more than 50 marks. So let us classify it.

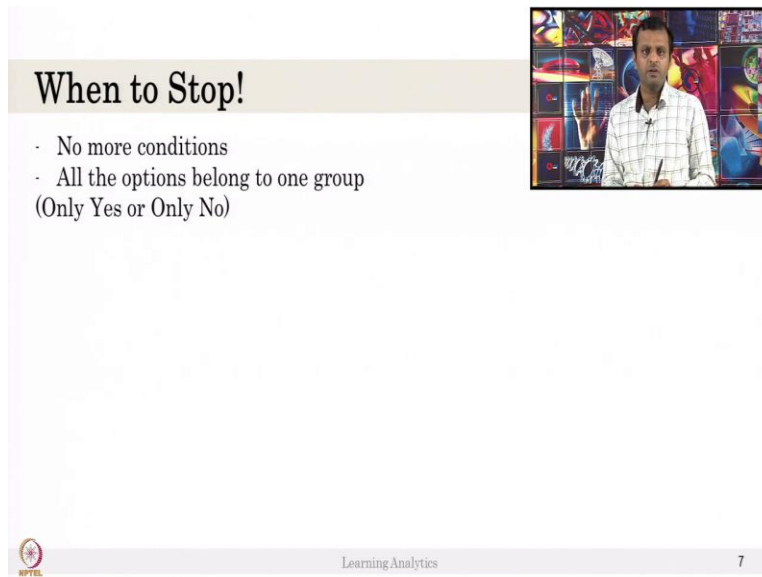
So let us come back here. When you come back here the 5 students out of 7 who got less than get 50 marks exactly this class that is what 5 by 7 here complete set how many people are particular value with decision you making 5 by 7 entropy of getting more than 50. So we saw that probably of students who are passing more than 50 is 1 if the probability of 1 which means entropy equal to 0. Similarly, probability of less than 50 were getting probability of passing the exam is 0, which means entropy is zero.

If it is 0 this value will be complete 0. So the entropy gain is the this node is 0.87. So there is no information loss. So maximum gain you select a node which has the highest gain to the to as a root node. So if you pick any other node or any other decision might have a different value, that value be less gain.

So you have to pick a root node, which gives you the maximum gain, and the gain will be computed by the current set.

So for example in this particular term if less than 50 you can simply draw that is it, it is simple, the decision tree of this particular given table is very simple yes or no, there is only one root node. Try this decision tree with much complexity for that you create your own data, with the 30 students attendance, 30 student mid-term marks and 30 students engagement in the class 30 student are submitting assignment in time and their interaction with the or something like that you create a big table and also why and compute and see how this decision tree is made that is the idea.

(Refer Slide Time: 14:41)



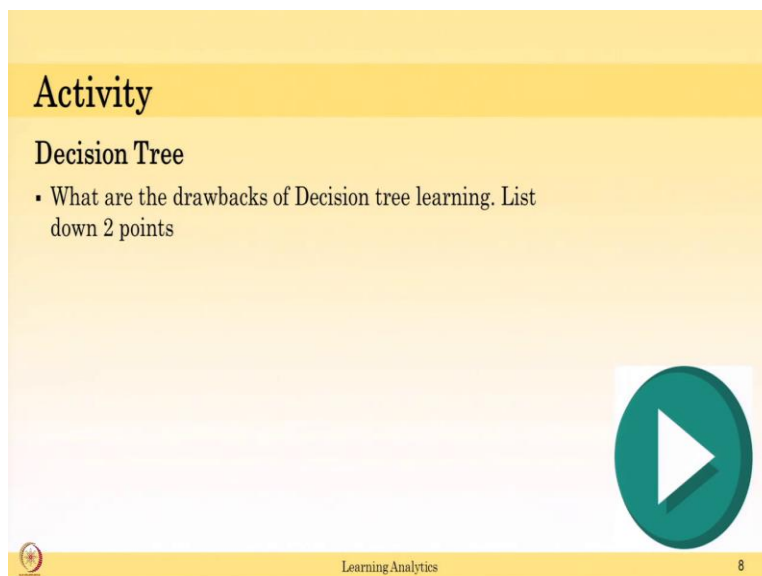
When to Stop!

- No more conditions
- All the options belong to one group
(Only Yes or Only No)

Learning Analytics 7

Let us see what is the root node and you want to know when to start. Suppose you have seen all the conditions all decision has been made and also all the options belong to one group consider you already made a decision everything to be property 1 or 0, all of all the students will be in yes or only no. If you assumed that particular reach that level you have to stop the decision tree.

(Refer Slide Time: 15:09)



Activity

Decision Tree

- What are the drawbacks of Decision tree learning. List down 2 points

Learning Analytics 8

So, you have seen decision tree, what are the drawbacks of decision tree? List down two points please list it down after listing it down resume the video to continue.

(Refer Slide Time: 15:24)

Activity

Decision Tree

- Complexity increases if number of decision increases
- Overfitting – performs well for training data set
- Greedy Algorithm – Assumption that optimization depends only on local decision node might not lead to global optimum solution
- Not applicable for continuous data

Pruning is used to overcome the first two disadvantages



So the complexity increases if number of decision increases, it looks very easy if you have say five levels or 10 decision to make the tree looks good nice. If you have say very complex problem, I have the 15 features or 20 features and each has a different values categorical more category not just two decision more decision to make than tree will be very very complex to look even it is visually good but you will not understand anything because to complex tree.

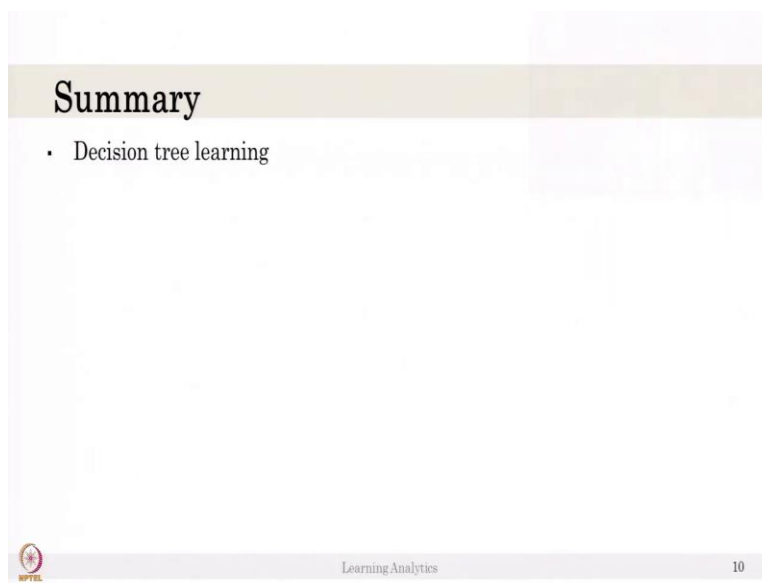
Decision tree has a very very main drawback that is over fitting, it tries to create a decision tree for the training data very well. So it performs very good on the training data, 100 percent accurate, but on a test data, it is not known to perform well just because in the training data, it takes all the small conditions it tries to get a new branch for fixed condition, but in test data is not possible, so that is a bigger problem and the greedy search algorithm I mentioned in the just starting of this video there is a greedy search tree, it is means only local decision made might lead to the next decision.

Sometimes you have to combine decision of two three features to make a good decision, but that is not possible in decision tree. Also, it is not applicable for continuous data you have to make a decision splits before you start decision tree, but latest algorithms automatically makes bins based on the probability of each bin so that is not issue.

So I said that there are two problems the other is complexity, so complexity and over fitting that can be solve by pruning. What is pruning? It is basically cutting out the branch, it is trimming the tree , as tree is very is very dense to big and to dense to cut down the branches.

So make it trim so that we can grow further and makes looks good that is also happening in the decision tree also, here what they do is they cut down the branches which makes not much decision and also then cut down the branches to reduce the complexity. Pruning helps in better performance. So in the sense it helps in improve the performance of the decision tree in the test data set also that is how the two advantages disadvantages, but that can be solved by the pruning methods and other also can be taken care by the latest machine learning algorithms, so it is not an issue.

(Refer Slide Time: 18:01)



So in this video we saw what is decision tree learning and I request you to go and try decision tree in a tools and check it out I did not give a complex problem for decision tree, but I would request you to practice decision tree using the complex set of problems with the more features more decisions to make and see how it works. Thank you.