

Data Mining
Prof. Pabitra Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 08
Decision Tree – I

Welcome to the lecture in the data-mining course on decision trees. Decision trees are one of the algorithm to perform what is known as a classification task. So, what is a classification task, a classification task is given and an instance an object, you have to put in one of the number of categories. So, you have a predefined set of categories, you have to put that instance or the object into one of these categories and these categories are called the class. For example, let say an object is an email; and I have two categories or classes spam or non spam email. So, this is a very common problem all of us face that we get lot of spam emails and we want to automatically put an email into one of these two categories. So, this problem is known as a classification problem.

In this particular case is known as a two class or a binary classification problem. In general, you can have k categories and you would call it a k class problem. Another example, suppose an customer has applied for a bank loan. And in this case, the object is the customer and the object the customer is described by a number of attributes for example, the income of the customer, the age of the customer, the marital status of the customer the household income of the customer. So, excuse me. So, you have a set of attributes describing an instance, and I want to put this customer the bank wants to put this customer into say into two categories again whether the customer will repay the loan or will not repay the loan. So, basically like a fraud customer or a real customer. This is another example, again a binary classification problem.

There may be other instances for example, maybe a patient comes to a doctor and the patient has one of possible say five diseases. And the symptoms of the patient would be the attributes of the patient. And the classification system suppose automated diagnosis system, we will use this attribute values and tell whether which of these five diseases this patient has. So, this is a five class classification problem.

So, what we will do in the next few lectures is to study a number of algorithms which will take as input a set of attributes attribute values describing an instance or an object

and we will produce or predict a class or a category to which that instance belongs. The way to do this is the following. So, we have something called a training set. So, this classification algorithms are also sometimes that is I called supervised learning algorithms supervised algorithms.

So, we have a training set where lot of instances along with they are categories are mentioned. So, in some attribute values and the category to which they are known to belong, they already evidence is that they belong to this. For example, the bank will have a list of previous customers with certain attribute values, we have repaid the loan; and with certain attribute values, you have not repaid the loan. Similarly a medical diagnosis system will have lot of patient cases where these diseases has happened what does the symptom just like a doctor does. And these another diseases happen what has the symptoms. So, this existing experience will be called a training set.

And what the classification algorithm is supposed to do is that it is used to take this training set learn from this training set. And when a future instance comes which is not part of the training set a new instance and unknown instance, a new customer or a new patient you have to predict the class level, the class we call it a class level. So, these algorithms that is why are also called predictive algorithms, they consider historical data in the use historical data and predict on new data. So, we will see a number of algorithms, we will start with the simplest one which we call a decision tree algorithm.

(Refer Slide Time: 06:48)



Day	Outlook	Temp	Humidity	Wind	Tennis?
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

The slide also features the IIT Kharagpur logo and NPTEL ONLINE CERTIFICATION COURSES text at the bottom.

So, let me explain this to you with an example. So, this is what I was mentioning that this table that we see represents the training set of the training examples. Let us see what this table means. In this in this toy example the instances are objects that you are talking about and nothing, but it a day, a day of the week some day of some season. So, each row in this table describes a day. So, there are there are D 1 to D 14, there are 14 days which are previous examples. And each day belongs to one of the two categories one of the two categories. If you look at the table in the slide, there are two categories whether people prefer to play tennis and outdoor sports on that day or does not prefer to play tennis on that day. So, each day belongs to two categories whether people play tennis or do not.

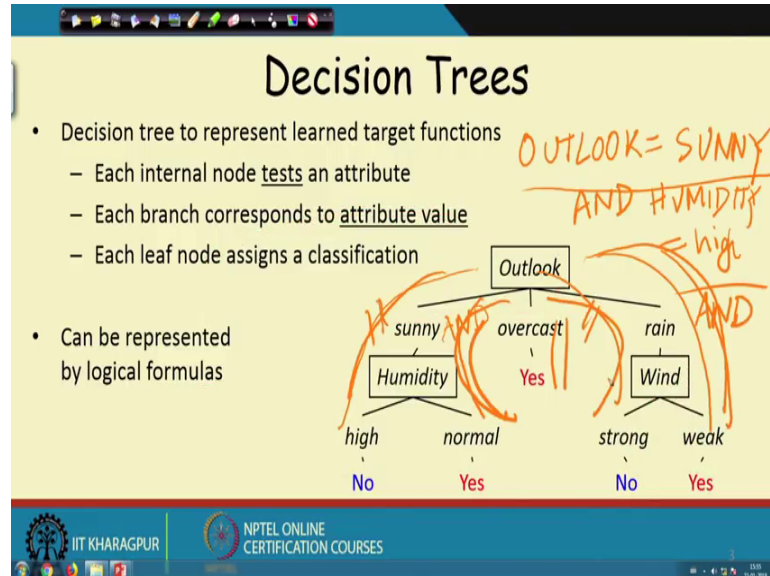
So, the last column that you see tennis question mark is actually the class level of the day. So, I have two possible class levels yes and no binary classification. Besides the final last column, each day is described by four attributes represented by these columns. So, these attributes are how does the day look like outlook I call, it is sunny or overcast or raining what is the temperature of the day, again I do not write numerical values, but I just qualitatively described it as hot or cold or mild. The humidity of the day which again takes some value higher normal or low maybe; and wind condition which is weak or strong.

So, as if I can think of a day as a four-dimensional vector, consisting of four values - the outlook, the temperature, the humidity, the wind. So, for example, day one - D 1, I describe it by this vector sunny hot high; and weekday 10 would be described as rain mild normal and weak. So, and of course, the final column that we see is the class level. So, this matrix is a very common form of representing the data, it is called a attribute instance matrix. The columns are attribute and there is one particular attribute which is the class level usually in the last column and the rows are the instances.

So, I have these fourteen instances in my training example. What my intent is that after looking at this 14 instances, if I get a new instance another D i, let us say for which I know the attribute values, but I do not know the class level. So, I know this attribute values of outlook temperature, humidity and wind, but I actually do not know what the class level is whether people will play tennis or not play tennis. So, I want to write some kind of rule or some kind of function or rule which will take this four attribute values and return a value of the tennis attribute that last level that would be my classification rule or

classification law or function. And this particular rule or function I would like to derive from my training example.

(Refer Slide Time: 12:24)



So, one particular type of rule that we are currently interested in is called a decision tree. It takes this particular form; it takes the form of a tree. So, how does it do it? First, it checks the value of one attribute in there, which is the root of the tree, the root node of the tree. So, in the diagram that you see, the outlook value is a root node. And for each possible value of this root node attribute, I have three children - three branches. So, for example, here sunny, overcast, rainy, these are the three branches. And now for each of these three branches, I can do one of the following: one of the two possibilities; either I can ask a new question, test a value of another attribute. For example, in the branch sunny, I again ask the question: what is the value of the humidity attribute. And similarly, in the branch rain, I again ask: what is the value of the wind attribute.

So, basically, this child can either lead to another internal node of the tree, or I can directly say, for example, in this overcast, that if the value of outlook is overcast, it is definitely the class level is yes. So, I directly tell the class level. So, I do not ask any further question. In other words, I form a leaf of the tree; I do not transfer further; whereas, in the sunny branch, further using the humidity attribute into two branches, and then these branches actually they do not branch further, so where they directly tell the class attribute. So, a decision tree is a tree where the nodes are attribute values and the

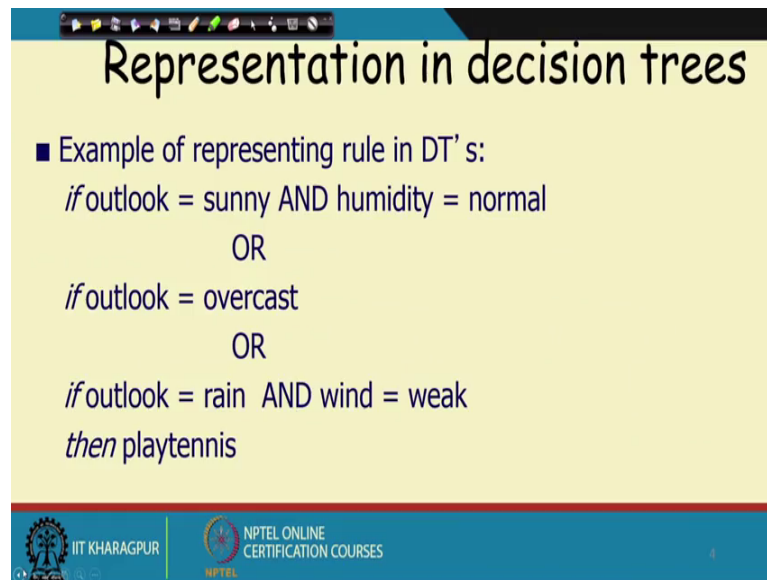
branches are the different values that attribute actual takes and the leaf nodes are some class level.

How is this tree gives me a classification rule. So, once I have this tree, once a new instance comes I just push it down that tree. I ask first asks what is the value of outlook, depending on the value I take a branch. If it is overcast I directly say the class is yes; if it is not overcast, I take sunny or rain, suppose sunny I again check the value of humidity. If humidity is high, I say class level is no; if humidity is normal I say class level is yes; similarly for the wind attribute further elements.

So, you see I can use this tree to infer the class level of an instance new instance by for which the attribute values I know I just push it down the see which leaf it leads to following which branch like internal statement. And each branch each leaf has a class level whatever leaf it leads to that is the last level of this instance. So, this can act as a classification rule. So, everybody of you please note down this tree in your exercise books. Note down this tree - simple tree.

Now you see look at each of the rows and see what is the class level you get using this tree. So, for D 1 what I get outlook is sunny I check humidity is high, so class level is no. So, you check class level is no. Of course, I will not do it for an existing training example I do it for an unknown example only. Please note that. So, this is a decision tree, this kind of tree with attributes and leafs is a decision tree. Note that we human beings use this kind of logic, we often use this kind of logic.

(Refer Slide Time: 18:40)



Representation in decision trees

- Example of representing rule in DT' s:
if outlook = sunny AND humidity = normal
OR
if outlook = overcast
OR
if outlook = rain AND wind = weak
then playtennis

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

If you look carefully this tree can be represented by some simple Boolean logic function of this particular form. So, if you look at this tree, and try to write it as a set of if then else rule you see this is what we get. If outlook is sunny and humidity is normal or if outlook is overcast or if outlook is rain and wind is weak, we have play tennis is equal to yes. do you see a pattern in this first thing you verify whether this if then else matches with this trees inference. You just check once. You will easily see that this if then else matches it. Second is you just try to see if there is a pattern in this tree in this rule that I have written you. See, if there is actually a pattern or not and think on it, how I am writing this if then rules. Just look at the rule, look at the nature of the rule, you see if you have any pattern or not.

If you have seen carefully you will see that in each of this if statement I have some AND and between this if statement I have some OR I have some ANDd and I have some OR. And what are each of these if you see each of the branches each of the branches of the tree, they correspond to an if they correspond to an if. So, all you need to do is follow a branch connect this by AND connect this by AND sorry connect this by AND. Similarly, connect. So, I have this branch another branch and so on. So, I have this as a branch, this as a branch, this as a branch, so this branch corresponds to this rule.

So, what I do is that I write down such AND rules for each branch and connect them by OR connect them by OR. So, basically let me take the all the yes branches. So, this is a

yes branch, this is a yes branch, and let me take all the no branches this is a no branch this is a no branch. So, what I will do I will say if this branch AND statement or if this branch AND statement yes or this branch actually the third branch is there yes. And then I will say if this branch or this branch then no. So, I can write equivalent logical rules for this. So, let me sorry ok.

(Refer Slide Time: 25:20)

Representation in decision trees

- Example of representing rule in DT' s:
if outlook = sunny AND humidity = normal
 OR
if outlook = overcast
 OR
if outlook = rain AND wind = weak
then playtennis

The diagram shows a decision tree with a root node 'Temp'. The left branch is labeled 'high' and leads to a node 'Humidity = high'. From 'Humidity = high', the left branch is 'Wind = low' leading to 'Y', and the right branch is 'Wind = high' leading to 'N'. The right branch from the root is labeled 'low' and leads to 'N'.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Let me give you one more small example, suppose I have this. (Refer Time: 26:03) So, suppose I have a tree like this. So, what would be the rule corresponding to this tree. So, temperature is the root node, I check high and low then I check humidity. If humidity is high, I check wind; if humidity is low maybe I check something else; if wind is low, yes play tennis, wind is high no. So, I can write it as a rule temperature equal to high and humidity equal to high and wind equal to low this path then yes or maybe some other part temperature equal to low and humidity equal to low and say wind equal to high then maybe yes again. So, what that, so this way. So, basically what I am doing is that I am writing down a set of AND OR rules set of AND OR rules as a as a pictorial as a diagrammatic manner that is all. And you know that actually any Boolean formula can be written as AND OR rule or of some AND any Boolean formula can be written like that.

So, basically if I write down this kind of tree I can actually represent any functional in functional form that means, any type of function which takes the input produces the class level I can do that. So, also another thing, I would like to mention that if you think

carefully, we humans we actually do often use this tree. First we say suppose this happens then what I do if this happens this way then I what happens am something else if it is raining today, and the buses are crowded and there is there is lot of work then I have to go to office. If it is not raining today and there is no work then I will not go to office. So, we use this kind of hierarchical this kind of tree light decision making often in our practice. So, the good thing is that you can write it as a logical formula you can express any function by this kind of decision tree. So, this is a popular classifier that people use in this kind of this kind of rule this kind of this decision tree.

What we will do next is that. So, now, I think what you can do from this is then you have understood the structure of decision tree, you can use it to predict; that means, when a new example comes whose attribute value I know, you can push it to the tree and predict the class level. So, decision using a decision tree that much you can do now, but what data mining is actually supposed to do what I will do in the next lecture is that find out the best decision tree which fits the training set. So, I have my I have this 14 days already examples in all say credit card fraud already I have seen previous examples find out or learn a decision tree from this previous training example which will help me give the best prediction. So, using some statistical properties of this table of this training set I want to learn the decision tree that is my next task.

So, please practice a few examples. What I would supposed to tell you is that you can look at this training examples this 14 that this the previous slide that I have shown you sorry this slide this table and try to think of yourself. What is the decision tree I should make to best fit this data fit this data means it for every row it gives the correct output the final column. You try to draw a decision tree from your human intuition; we try to draw it. And the hope is that if it is doing good on this fourteen examples it will probably do good on the new unknown example also that is the underlying assumption often that assumption is correct. So, you try to and you will see that if you think on this intuition of this I will tell an algorithm to actually quantify this intuition and build that. So, in the next class, we are studying the algorithms for building or construction of decision trees from training set.

Thank you for today.