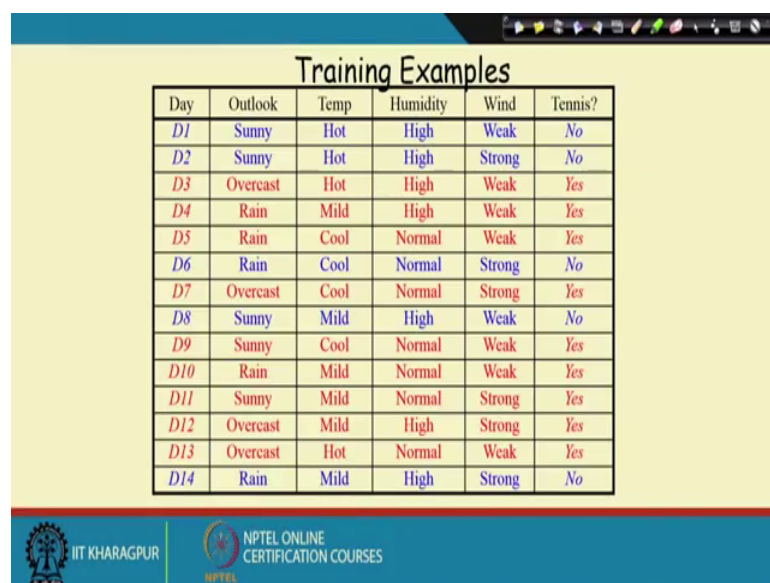


**Data Mining**  
**Prof. Pabitra Mitra**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 11**  
**Decision Tree - IV**

Let me explain the process of Decision Tree construction with the training examples we consider.

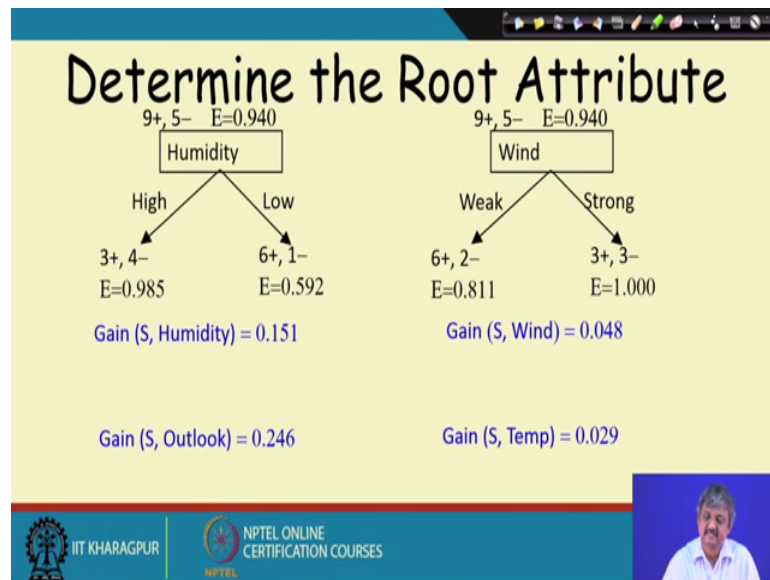
(Refer Slide Time: 00:29)



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

So, we come back to our table of examples that we had seen earlier, I have 14 examples with belonging to 2 classes. So, what I do is that in order to consider I mean construct the decision tree initially I consider all the 14 examples ok.

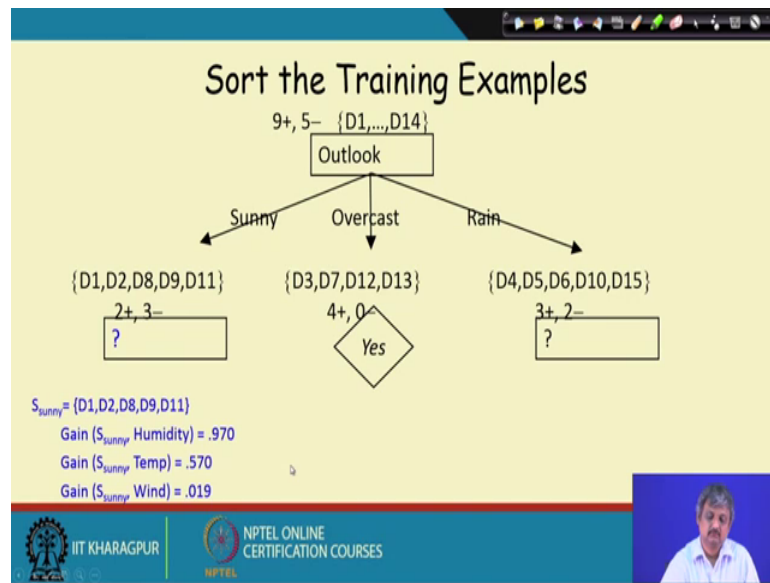
(Refer Slide Time: 00:52)



And then for each of the attributes I perform a split for example, if you see. So, there are 14 examples of the out of the 14 examples 9 belong to plus class S class 5 belongs to no class. And if we split on humidity let us say there are 2 groups high and normal and if you count 7 examples have humidity high, and 8 have humidity low. So, if we split along that and using these values of n plus and n minus, if you can compute the entropy these are the values we get.

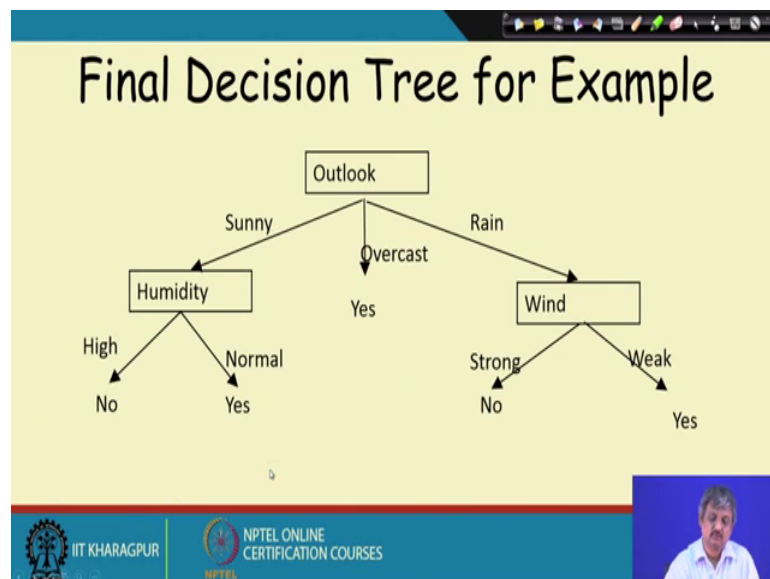
So, the gain is this much 0.985 plus 0.592 subtracted from that 0.940, so what do you do similarly for the wind we calculate, this is the gain similarly for the outlook this is the for the outlook, this is the gain, and temperature this is again. So, we see that outlook has the highest gain.

(Refer Slide Time: 02:27)



In order to construct the decision tree we first split on outlook. So, outlook equal to sunny I have these 5 examples overcast this, and 2 of them are plus t minus you see overcast is pure only plus class so it is a leaf. So this is my new S these 5 examples, on this I recursively split again. So, gain on humid is this gain on S sunny not the full set only on S sunny wind is this, humidity is highest.

(Refer Slide Time: 03:20)



So, I split on humidity and I repeat the exercise for this and I get pure leaf nodes and this is what I get.

(Refer Slide Time: 03:35)

### Overfitting the Data

- Learning a tree that classifies the training data perfectly may not lead to the tree with the best generalization performance.
  - There may be noise in the training data the tree is fitting
  - The algorithm might be making decisions based on very little data
- A hypothesis  $h$  is said to overfit the training data if there is another hypothesis,  $h'$ , such that  $h$  has smaller error than  $h'$  on the training data but  $h$  has larger error on the test data than  $h'$ .

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, one problem that happens in this kind of classification algorithm is you get, over fitting. So, if you make the training set quite small and if there is some noise in the training set if you get a deeper and deeper tree, your training set accuracy will keep on increasing, but test set accuracy may decrease after some time. So, x axis is height of the tree.

(Refer Slide Time: 04:16)

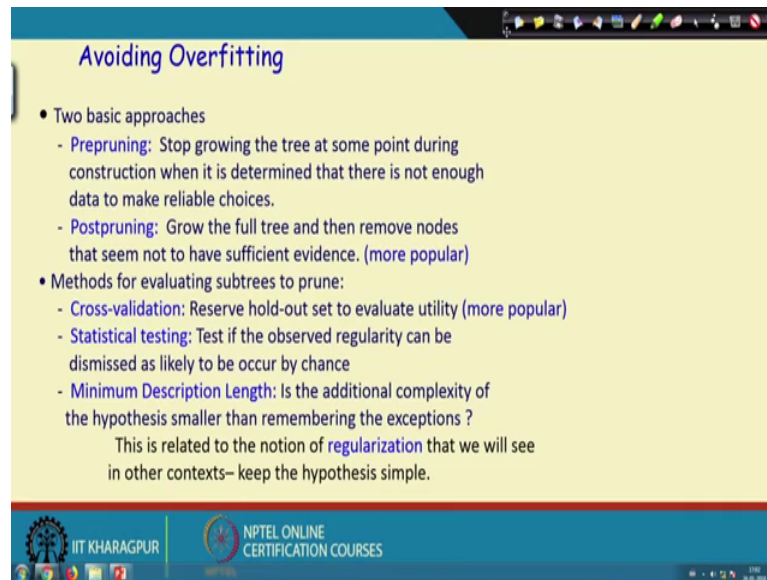
### When to stop splitting further?

A very deep tree required  
To fit just one odd training  
example

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

I explain this suppose there is a noise this point is a noise. If I keep on splitting it if I keep on splitting it to fit that, I have to draw few more lines you correctly get a pure leaf and I will have a deeper tree.

(Refer Slide Time: 04:50)



**Avoiding Overfitting**

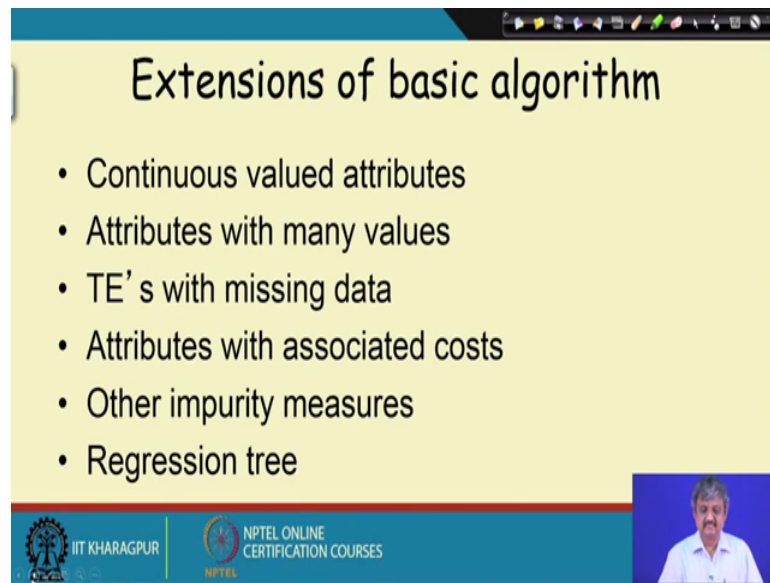
- Two basic approaches
  - **Prepruning:** Stop growing the tree at some point during construction when it is determined that there is not enough data to make reliable choices.
  - **Postpruning:** Grow the full tree and then remove nodes that seem not to have sufficient evidence. (more popular)
- Methods for evaluating subtrees to prune:
  - **Cross-validation:** Reserve hold-out set to evaluate utility (more popular)
  - **Statistical testing:** Test if the observed regularity can be dismissed as likely to be occur by chance
  - **Minimum Description Length:** Is the additional complexity of the hypothesis smaller than remembering the exceptions ?  
This is related to the notion of **regularization** that we will see in other contexts– keep the hypothesis simple.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, to solve this problem what one does is that you build a tree, then sometimes instead of pure node 100 percent pure node maybe 95 percent pure node is enough like maybe this much impurity you can allow and not split further. So, this means I am clipping the tree and not building it further this process is known as pruning. There are 2 approaches pre pruning while growing that tree you stop or post pruning you grow fully, then cut some branches which branches to cut you do something called a cross validation; that means, you take the training set error test consider sorry the test set error.

And see how much error you are getting. If it is high you stop do not you prune that, but do not build it further grow that branch. Further otherwise there is something called a description length; that means, description length of a tree is the height of the tree plus the number of examples it misclassifies, I want to build a tree which has the minimum description length ok.

(Refer Slide Time: 06:41)



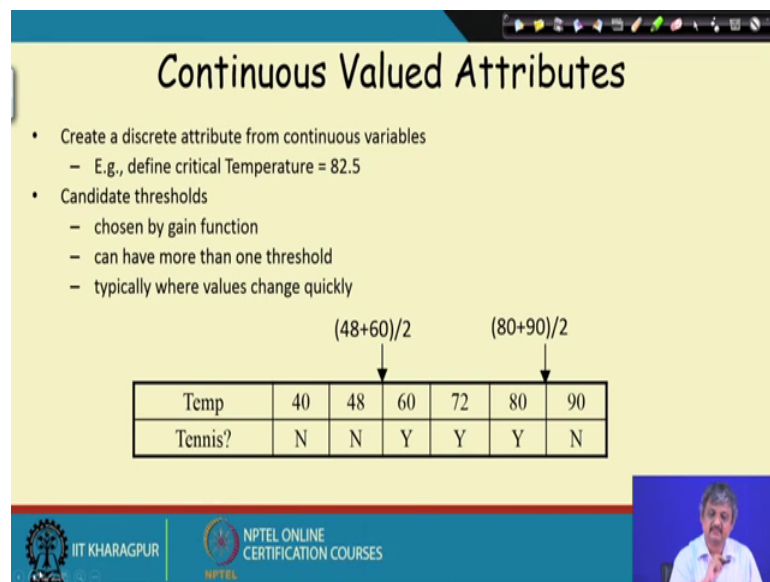
### Extensions of basic algorithm

- Continuous valued attributes
- Attributes with many values
- TE' s with missing data
- Attributes with associated costs
- Other impurity measures
- Regression tree

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this is about tree pruning I will few extensions of the basic tree algorithm what about continuous values.

(Refer Slide Time: 06:53)



### Continuous Valued Attributes

- Create a discrete attribute from continuous variables
  - E.g., define critical Temperature = 82.5
- Candidate thresholds
  - chosen by gain function
  - can have more than one threshold
  - typically where values change quickly

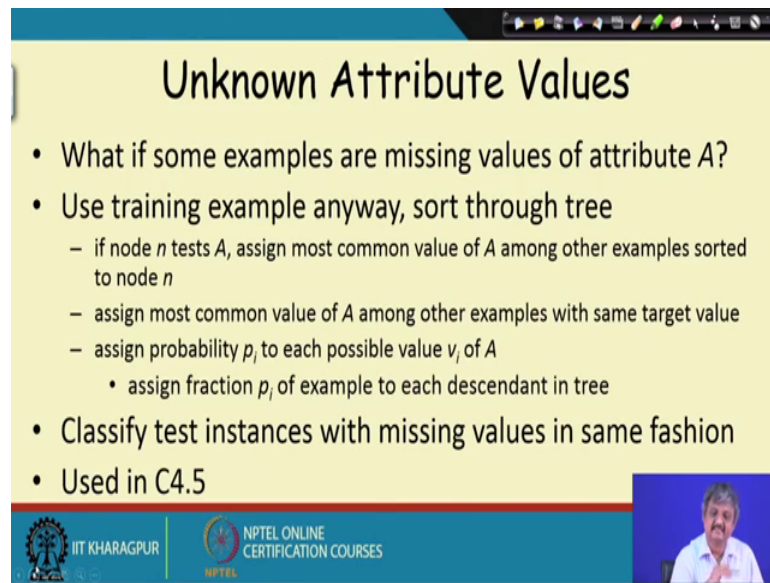
Temp	40	48	60	72	80	90
Tennis?	N	N	Y	Y	Y	N

$(48+60)/2$        $(80+90)/2$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Often what we do you first discretize the continuous value into intervals by some method; some method you discretize, and then consider design that is the most common.

(Refer Slide Time: 07:18)



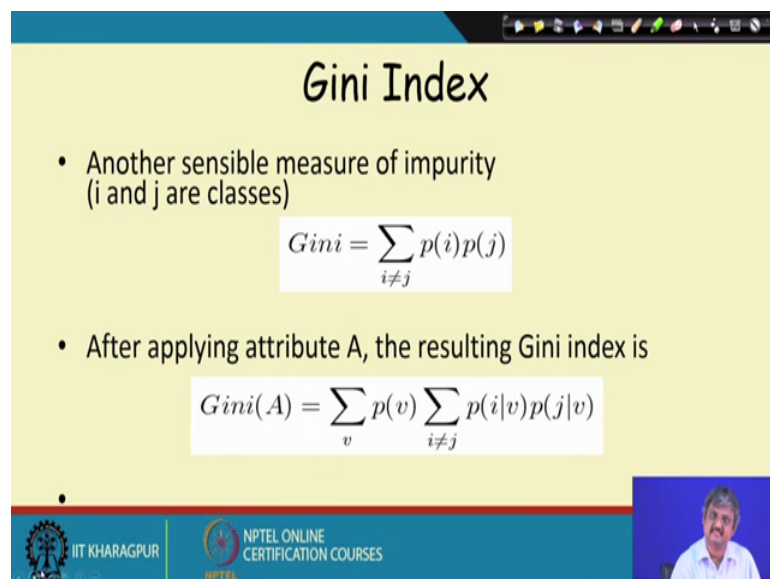
### Unknown Attribute Values

- What if some examples are missing values of attribute A?
- Use training example anyway, sort through tree
  - if node  $n$  tests  $A$ , assign most common value of  $A$  among other examples sorted to node  $n$
  - assign most common value of  $A$  among other examples with same target value
  - assign probability  $p_i$  to each possible value  $v_i$  of  $A$ 
    - assign fraction  $p_i$  of example to each descendant in tree
- Classify test instances with missing values in same fashion
- Used in C4.5

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

I will let us discuss some properties of discretize some algorithms for discretization. Similarly missing attributes you take often these are the strategies. So, C4.5 is popular software for decision tree which uses all these strategy.

(Refer Slide Time: 07:41)



### Gini Index

- Another sensible measure of impurity (i and j are classes)

$$Gini = \sum_{i \neq j} p(i)p(j)$$

- After applying attribute A, the resulting Gini index is

$$Gini(A) = \sum_v p(v) \sum_{i \neq j} p(i|v)p(j|v)$$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES


Sometimes instead of the entropy an alternate index is used called the Gini index which is defined like this.




(Refer Slide Time: 07:54)

## Regression Tree


- Similar to classification
- Use a set of attributes to predict the value (instead of a class label)
- Instead of computing information gain, compute the sum of squared errors
- Partition the attribute space into a set of rectangular subspaces, each with its own predictor
  - The simplest predictor is a constant value



IIT KHARAGPUR



NPTEL ONLINE CERTIFICATION COURSES


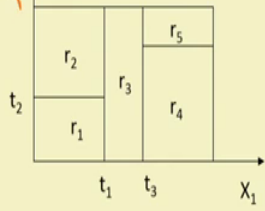
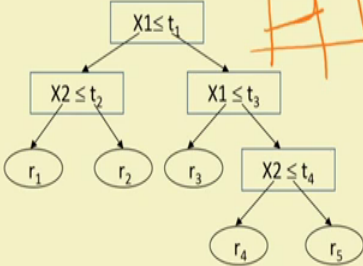


One more extension is a regression tree, what we do here is that instead of leafs being classes each of leafs are some regression line.


(Refer Slide Time: 08:07)

## Rectilinear Division


- A regression tree is a piecewise constant function of the input attributes



IIT KHARAGPUR



NPTEL ONLINE CERTIFICATION COURSES

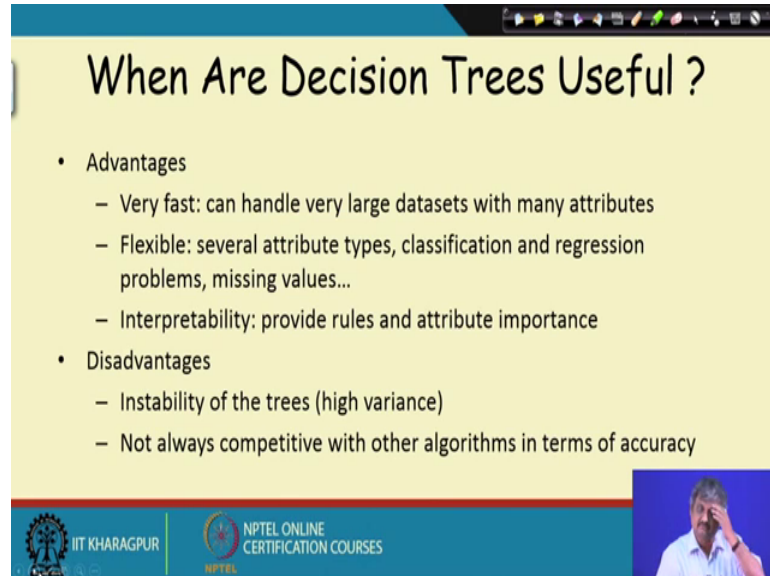


Some constant value or some constant piecewise, constant function some values or 1 or 2 instead of class their values. So, how do I use this? So suppose I want to fit a regression to this kind of a function I will piecewise fit constant values, and these boundaries will be decided by the decision tree. So, this algorithm is a popular algorithm called the cart



algorithm classification and regression tree, when I covered regression later I will cover more details of this.

(Refer Slide Time: 09:11)



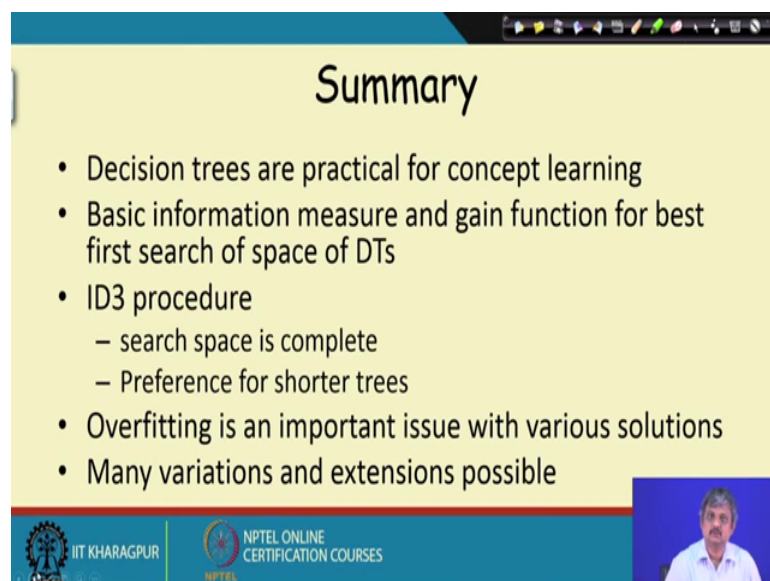
**When Are Decision Trees Useful ?**

- Advantages
  - Very fast: can handle very large datasets with many attributes
  - Flexible: several attribute types, classification and regression problems, missing values...
  - Interpretability: provide rules and attribute importance
- Disadvantages
  - Instability of the trees (high variance)
  - Not always competitive with other algorithms in terms of accuracy

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

These are some of the advantages of decision tree they are very fast flexible interpretable you can find rules. Disadvantage it depends very much on the training set you use some and also sometimes get less accuracy to the algorithms. So, we will consider in our next lectures here.

(Refer Slide Time: 09:36)



**Summary**

- Decision trees are practical for concept learning
- Basic information measure and gain function for best first search of space of DTs
- ID3 procedure
  - search space is complete
  - Preference for shorter trees
- Overfitting is an important issue with various solutions
- Many variations and extensions possible

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Is the summary of the topic of decision tree note that again, ID3 is all this method I discussed is they are usually clubbed together in an algorithm called ID3, which is a popular decision tree algorithm, so it is a method of learning classes classification uses information gain it prefer shorter tree over fitting is an important issue, and people have used various variation extensions to different cases.

So, if we are face to it and classification problem with say discrete attribute or few values. The first thing to try is a decision tree if it does not give good accuracy you go for other algorithms and decision is implemented in most of the software's.

(Refer Slide Time: 10:35)

**Software**

- In R:
  - Packages tree and rpart
- C4.5:
  - <http://www.cse.unwe.edu.au/~quinlan>
- Weka
  - <http://www.cs.waikato.ac.nz/ml/weka>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, with this these are some of the softwares, I which you can use to implement a decision tree C4.5 is a popular Weka is a machine learning toolbox which you can use and R has all the decision tree algorithms with this, I close my discussion on decision tree. In our next lecture we will go to other algorithms by classification.

Thank you.