Course Name: **Machine Learning and Deep learning - Fundamentals and Applications**

Professor Name: Prof. M. K. Bhuyan

Department Name: Electronics and Electrical Engineering

Institute Name: Indian Institute of Technology, Guwahati

Week-5

Lecture-22

Welcome to NPTEL online course on machine learning and deep learning fundamentals and applications. Up till now I discussed the concept of Bayesian decision theory. I explained the concept of parametric and non-parametric estimation techniques. Today I am going to discuss the concept of another classifier that is decision trees. It is a non-parametric supervised classification technique. Decision tree is a hierarchical tree type structure consisting of root nodes, internal nodes and the leaf nodes and the connecting branches.

So, decisions are taken at the leaf nodes. This decision tree can be used for both regression and the classification. So, let us discuss about the concept of the decision trees and how to build a decision tree. So, that concept I am going to explain today.

So, let us start this class. So, in this class I discussed one new classification technique that is the decision tree. In the first figure I have shown one decision tree and you can see the root node is available and two internal nodes and the leaf nodes. The decisions are taken at the leaf nodes and you can see the interconnecting branches.

The root and each internal node test one attribute.

Each branch from a node select one value for the attribute and the leaf node predict a particular class. So, this is the definition of the decision tree and corresponding to the first figure I have shown one example of a decision tree and in this case the example is for weather forecasting. So, outlook attribute we are considering and corresponding to this attribute I have the values are sunny, overcast, rain and corresponding to the sunny I am considering another attribute that is humidity. So, the values may be high or normal and overcast the value may be yes. For rain we are considering another attribute that is the wind and the values may be strong and weak and answers are yes or no.

And based on this the answers yes or no, I can take a classification decision. So, this is the fundamental concept of a decision tree. So, I am repeating this it consists of a root node and internal nodes and the leaf nodes and the interconnecting branches. So, decisions are taken at the leaf nodes and this root or the internal node test one attribute and these branches from a node select one value of the attribute and finally, this leaf node predict a particular class. So, this is about the fundamental concept of a decision tree.

So, corresponding to this decision tree again I am showing the same example outlook is the attribute and the values are sunny, overcast and rain and I am considering another attributes like humidity wind and the values are high normal strong weak and the answers are yes or no. So, based on this we can do predictions and the questions may be like this outlook is sunny temperature is hot humidity is high and wind is weak. So, corresponding to this I can do some classification decision or I can do the prediction. So, this is the fundamental concept of concept of a decision tree.

The decision tree is device the phasor space into hyper rectangles.

So, each region is labeled with one label. So, corresponding to the decision tree is shown here in the figure I have shown some conditions $x_2 < 3$, $x_1 < 4$, $x_1 < 3$, $x_2 < 4$. So, these conditions I have shown and these are mainly the attributes and corresponding to this I have shown the values in the leaf node 0 1 0 1 like this and here I have shown the phasor space and two phasors x_1 and x_2. So, corresponding to the first condition $x_2 < 3$. So, this is the first condition second condition is if it is yes $x_1 < 4$ if it is yes.

So, corresponding label is 0. So, if you see these zeros up to this because x_2 should be less than 3. So, this is 3 and x_1 should be less than 4. So, that means I will be getting this portion this portion I will be getting and corresponding to the second condition the condition is x_2 is less than 3 yes and x_1 is less than 4 no then corresponding to this I will be getting ones. So, that means corresponding to this I am getting this label.

So, I am getting the region corresponding to all these decisions. So, this is one example how the decision trees divide the phasor space into hyper rectangles and each region is labeled with one label. So, move to the next slide. So, what function can decision tree represent? So, decision tree can represent any function of the input attributes. So, maybe we can consider Boolean operations like n operations x or operations.

So, like this all these operations I can consider in the decision trees like this all the Boolean functions I can consider in the decision trees. So, corresponding to this I can give one example. So, move to the next slide here I have shown the logic is x or so here you can see I am showing two inputs a and b these two input variables a and b and output is the a x or b. So, corresponding to this condition false and false a is false b is false the output is false corresponding to the second condition a is false b is true the output is true corresponding to the condition a is true b is false the output is true and finally, in the last condition a is true and b is true.

So, output is false corresponding to this you can see in the feature space I have shown.

So, first condition is a is false and b is false corresponding to this I am getting this one second condition is that a is false and the b is true. So, I will be getting true. So, corresponding to this I will be getting the true is like this, this I will be getting. So, like this I will be getting and these outputs and you can see corresponding decision boundaries. So, this is one example of a Boolean decision making.

So, move to the next slide corresponding to this xor logic I can draw a decision tree. So, you can see the decision tree I have shown here. So, a is the root node and if you see the internal nodes are b and b two internal nodes. So, corresponding to a I may have two values one is false another one is true corresponding to b also I have two values one is false another one is true corresponding to b already I told you I have two values one is false another one is true and corresponding to this what is the output in the leaf node if a is false b is false then I will be getting this one the false if a is false b is true then I will be getting true if a is true b is false then I will be getting the output true and corresponding to a is true and b is true I will be getting the output false. So, this is the decision tree corresponding to the x or logic and you can see the decision boundaries in the feature space.

So, I have shown these two variables a and b that is the two features or the two inputs a and b in the feature space. So, in case of the decision tree regarding the inputs can be used for both discrete and continuous input attributes. So, the input attributes may be discrete or continuous. So, corresponding to this I have given two examples here you can see the first one is the discrete input attributes.

So, this is regarding the weather forecasting the attributes are outlook.

So, outlook has three values sunny overcast and the rain and another attribute is humidity and the another attribute is wind. So, corresponding to humidity the values are high and the normal corresponding to wind the attribute is strong and weak. So, this is about the discrete input attributes regarding the continuous I am showing the second case in case of the continuous you can see I am considering these cases $x_1 < 0$, $x_2 < 0$, $x_1 < 0.5$. So, like this I am considering these values.

So, these attributes may be continuous the input attributes may be continuous and maybe we can consider mixed attributes input attributes. So, in the last figure you can see I am showing this condition that is the mixed attributes you can see gender is the discrete attribute is you can see it is the discrete attribute, but if you see other case like this age is greater than 9.5 age is less than greater than equal to 9.5.

So, for this you can see we are considering the continuous input attributes.

So, move to the next slide. This decision trees can be used for both classification and the regression. So, in the first figure I have shown the case of classification and the output levels are yes or no. So, you can see these are these are the output levels and that is obtained in the leaf nodes and the in the second figure I have shown that the decision tree can be used for regression also. So, you can see just I am finding the mean value, the mean values are like this 8.571.

So, these values I am obtaining based on the conditions and this is nothing but considering all the samples value and I am finding the mean value. So, this 0.807.

So, these values are obtained from the decision trees.

So, that means the decision tree can be used for both classification and the regression. So, what are the outputs of a decision tree. So, already I told you. So, it can be used for classification or it can be used for regression. So, if I consider a decision tree is part from root to leaf define a region Rm of the input space.

So, here we are considering some training examples. So, corresponding to x^{m_1} the target value is t^{m_1} corresponding to the input x^{m_k} the target value is t^{m_k}. So, we are considering training examples and that fall into the region R_m and for the classification decision tree I am getting the discrete outputs. So, I am getting the discrete outputs in the leaf nodes.

So, leaf value y^m typically set to the most common value in the targets the target values are $t^{m_1}, t^{m_2}, t^{m_k}$, like this.

So, leaf values y^{m_k} typically set to the most common value in that set of target values. In case of the regression decision tree we have continuous outputs. So, leaf value y^m typically set to the mean value of the target values. So, target values are $t^{m_1}, t^{m_2}, t^{m_k}$ and we are considering the mean values from this target value. So, you can see the distinction between the classification decision tree and the regression decision tree.

So, now already we have understand the concept of the decision tree. Now, how to learn the decision trees.

So, let us discuss about the learning of the decision trees. So, move to the next slide.

So, how to train a decision tree. So, for this we can consider a greedy heuristic learning technique start from an empty decision tree. So, we are considering one empty decision tree split on next best attribute. So, we have to split the decision tree based on the attribute and after this we have to do this repeatedly and what is the best attribute. So, that I have to determine the best attribute can be determined from the information theory point of view. So, I will be explaining what is the information theory how I can determine the best attribute the best attribute I can determine from the information theory point of view.

I may consider entropy, entropy means it is a measure of uncertainty and based on the entropy I can determine the information gain and based on this I can determine the best attribute. So, let us discuss about this. So, which attribute is better to split in this example I am showing a discrete decision tree and two variables I am considering two attributes X_1 and X_2 and corresponding output I have shown. So, if X_1 is true corresponding to this output Y_{true} how many times t is equal to 4 times.

So, corresponding to X_1 is true. So, true means it is true, true, true, true.

So, how many times the y is true. So, 4 times.

So, 1, 2, 3, 4. So, 4 times it is true. So, that means the first condition we are considering corresponding to X_1 true and y is false. So, it is 0 times like this we are considering. So, the idea is to count as leaps to define probability distribution and based on this we can measure the uncertainty. So, in case of the this deterministic all are true or false. So, just one class in the leaf, but if I consider the uniform distribution all classes in the leaf equally probable.

So, what distribution I can consider because I have to determine the best attribute and based on the best attribute I can split the decision tree. So, if I consider the deterministic only we have true or false and if I consider uniform distribution all classes in the leaf equally probable. So, what about distribution in between. So, let us move to the next slide. So, considering this one, how to quantify uncertainties.

So, consider one random experiment that is the coin flipping experiment. So, the output is head or tail. So, that means I can say it is 0 and 1. So, corresponding to the sequence 1 you can see first one is tail 0 means tail 0 tail 0 tail 1 means head 0 0 0 0 all 0's again it is coming 1 0 0.

So, this is the first sequence corresponding to the experiment coin flipping and similarly corresponding to the same experiment I am considering the outputs corresponding to the sequence 2.

So, it is 0 0 means tail 1 means head 0 1 0 1 1 1 like this I have the second sequence. Now, corresponding to the sequence1 how many outcomes we are considering 18 outcomes 18 and out of 18 16 is 0 total number of outcome is 18 and out of 18 16 is 0. So, that means we are considering 16 times it is 0 and out of 18 2 times it is 1 in this experiment. So, corresponding to this in the sequence number 1 I have this one.

So, that means 0 16 times out of 18 times and 1 2 times out of 18 times and corresponding to the second experiment that is the sequence number 2 0 8 times.

So, this is corresponding to the sequence number 2 and 1, 10 times and based on this how to quantify uncertainties. So, move to the next slide. So, in the first sequence we considered out of 18 16 times that is equal to $\frac{8}{9}$. So, it is $\frac{8}{9}$ and the second one is $\frac{2}{18}$ that is 1 the occurrence of 1 how many times 2 times out of 18.

So, it is equal to $\frac{1}{9}$. So, like this I am getting. So, corresponding to the first sequence I am getting this one $\frac{8}{9}$ and that is corresponding to 0 the another one is $\frac{1}{9}$ that is the probability corresponding to 1 and corresponding to the sequence 2 the probability of 0 is $\frac{4}{9}$ and probability of 1 is $\frac{5}{9}$ and this is the expression for the entropy. So, this is mean entropy means the uncertainty. So, with the help of this formula I can determine the uncertainty corresponding to the first experiment that is the sequence number 1 and also I can determine the uncertainty corresponding to the second experiment that is the sequence number 2. So, corresponding to the first sequence I am getting this value that is that is $\frac{1}{2}$.

So, this uncertainty is 1 by 2.5 and corresponding to the sequence number 2 my uncertainty is 0.99. So, it is almost certain. So, this is regarding the experiment of the flipping of the coin. So, first we have determined the probability of occurrence.

So, corresponding to the first sequence the probability of occurrence of 0 is $\frac{8}{9}$ and corresponding to the first sequence the probability of occurrence of 1 is $\frac{1}{9}$. So, here I want to compare the uncertainties. So, from the previous example I have shown corresponding to the first sequence I have these probabilities 1 is $\frac{8}{9}$ corresponding to 0 another one is $\frac{1}{9}$ corresponding to 1 and corresponding to the sequence 2 the probability of 0 is $\frac{4}{9}$ and probability of 1 is $\frac{5}{9}$. So, what is a high entropy variables has a uniform like distribution that is the flat histogram and value sample from it are less predictable.

So, that means the high entropy corresponds to this case. So, variables has a uniform like distribution flat histogram and value sample from it are less predictable. So, that is the first case. Next one is the low entropy distribution of variable has many peaks and valleys histogram has many lows and high and value sample from it are more predictable. So, that means that corresponds to the second case. So, you can see the concept of the high entropy and the low entropy.

So, that is about the quantifying uncertainties. Now, from the information theory point of view. So, we have to determine the information gain that means from the entropy I can determine the information gain. So, what is information gain? Entropy to information gain. So, entropy $H(x)$ entropy is defined by $H(x)$ of a random variable is x. So, entropy is defined like $H(x) = -\sum_{x \in X} p(x) log_2 p(x)$.

So, this is the definition of the entropy and from this definition of the entropy, we can determine the specific conditional entropy. Specific conditional So, specific conditional entropy is $H(x|y)$, y is equal to suppose the value is v of that is x the meaning is x given y, y is equal to v. So, that is $H(x|y = v) = -\sum_{i=1}{}^{n} P(x = i|y = v) log_2 P(x = i|y = v)$. So, this is the specific conditional entropy and based on this just I can write the conditional entropy $H(x|y)$, x is a random variable and y is a random variable of x given y that is the conditional entropy.

So, $H(x|y)$ summation. So, all the values of y we can consider values of y=v that means all the values of y we are considering the probability $P(x|y = v)$, y is equal to v. This is the definition of the conditional entropy. And based on this we can determine the mutual information that is a very, very important definition mutual information. So, mutual information is mutual information of x and y that is I can represent like this $I(x, y)$ that is the mutual information the entropy of x minus entropy of x given y or I can write entropy of y minus entropy of y given x. This is the definition of the mutual information and actually this is also called the information gain.

So, based on this information gain I can determine the best attribute of the decision tree. So, how to determine the best attribute of a decision tree based on this mutual information I can explain in my next slide. So, move to the next slide. So, based on this definition that is the definition of the mutual information you can see I am giving one example and this information gain I can calculate like this entropy of the parent, parent means if I consider a decision tree. So, this is the parent and this is a decision tree and these are the children's.

So, we can determine the entropy of the parent and also we can determine the average entropy of the children. So, the difference between the entropy of the parent and the average entropy of the children that actually the information gain that is the information gain. So, corresponding to this example I am considering a decision tree and here you can see this is the parent. So, corresponding to the parent we are considering 30 examples 30 instances and I am showing the outcomes or the values some are plus and some are circle the green circle. So, if you count here out of these 30 examples 16 corresponds to the green circle and 14 corresponds to the plus labels that is the plus examples corresponding to this parent I can determine the parent entropy.

So, by using this formula I can determine the parent entropy. So, corresponding to this 14 plus examples out of 30 I can determine $\frac{14}{30} log_2 \frac{14}{30}$ minus we have to consider that the green circles how many green circles 16 green circles out of 30. So, $\frac{16}{30} log_2 \frac{16}{30}$ and we can determine the entropy of the parent. Now, let us consider the cell entropy. So, I have 2 cells corresponding to the first cell we are considering 17 examples and in this case you can see out of 17, 13 are plus and 4 are green circles. So, corresponding to the this cell you can see we can determine the entropy of the cell.

So, 13 out of 17. So, $\frac{13}{17} log_2 \frac{13}{17}$ and 4, 4 means the 4 green circles out of 17 $\frac{4}{17} log_2 \frac{4}{17}$. So, we can determine the entropy of the cell. So, it is 0.787 and similarly corresponding to the second cell.

So, how many examples 13 examples. So, out of 13 examples 12 are green circles and only one is plus and corresponding to this also we can determine the entropy the cell entropy and after this we can determine the average entropy of the children. So, average entropy of the children we can determine like this. So, it is 0.615 and after this we can determine the information gain the information gain is that is the entropy of the parent the entropy of the parent is 0.996 minus average entropy of children. So, that is 0.615 and ultimately it is 0.38. So, that is the information gain. So, by using this formula that is the formula of the mutual information we can determine the information gain corresponding to a decision tree and based on this information gain I can select the best attribute.

So, corresponding to this I can give one example. So, this example how to build a decision tree corresponding to this example I want to explain. So, this example is taken from the YouTube. So, you can see this example and here we are considering 14 examples that means 14 days we are considering D1 D2 D3 up to 14. So, here we are considering 4 attributes outlook temperature humidity and wind. So, 4 attributes we are considering and the play tennis that is actually the target variable and corresponding to the attribute outlook we have the values the sunny sunny overcast rain overcast rain.

So, corresponding to outlook I have these values sunny overcast and the rain corresponding to the temperature I have the values hot mile and the cool. So, like this I have these values corresponding to these attributes. Now how to build this decision tree based on the information theory point of view let us discuss in the next slide. So, first we are considering the attribute outlook and here we are considering 14 examples. So, first we are determining entropy of the overall set entropy of the overall set is S.

So, how many times it is positive. So, if you see this table 9 times it is positive yes yes yes yes yes yes yes yes yes yes. So, 9 times it is positive and 5 times no. So, corresponding to the complete set that corresponding to the complete set or corresponding to the entire set I am determining the entropy. So, that means 9 times it is positive and 5 times negative.

So, you can see we are determining the entropy by using this formula. So, 9 out of 14 because we have 14 examples. So, 9 out of 14 we are considering that these are the positive examples and 5 out of 14 these are negative examples. So, this is about the complete set after this we are considering the attribute that is the outlook and corresponding to this outlook I have the values sunny overcast and the rain. So, corresponding to the sunny what are the values how many positive examples corresponding to the sunny you can see in this table I have 2 positive examples yes and yes and I have 3 negative examples. So, here you can see in the table sunny this is 1 positive example this is 1 positive example.

So, that means 2 positive examples and 3 negative examples this is no no and also no. So, corresponding to the sunny I have 2 positive examples and 3 negative examples. So, that means total is 5 3 plus 2. So, out of 5 2 positive examples and out of 5 3 negative examples. So, based on this we can determine the entropy and similarly the another value is overcast that is the value corresponding to the attribute outlook. So, in this case only 4 positive examples and 0 negative examples if it is all the positive examples no negative examples then you will be getting the entropy 0 if it is all the negative examples no positive example then this case also you will be getting the entropy 0 and finally that another value rain corresponding to the attribute outlook.

So, 3 positive examples 2 negative examples and you can determine the entropy and finally we can determine the information gain corresponding to the attribute outlook with respect to the complete set the complete set is the S is the complete set. So, in entropy of the complete set that is the entire set minus we are considering entropy for the sunny entropy for the overcast entropy for the rain. So, here you can see $Entropy(S) = \sum_{v\in(Sunny,Overcast,Rain)} \frac{|S_v|}{|S|} Entropy(S_v)$. So, v maybe sunny, v maybe overcast or v maybe rain. So, corresponding to this what is the information gain of the outlook corresponding to S, S is the complete set.

So, entropy of S minus 5 divided by 14 because corresponding to sunny only I have 5 values out of 14 total examples are 14. So, if you see this table only I have in the 5 places I have the sunny values. So, 5 out of 14 next one is the overcast if you see in the table I have 4 overcast value overcast overcast overcast overcast. So, only 4 overcast out of 14. So, we can consider 4 divided by 14 entropy of the overcast and finally, another value we are considering the entropy of the rain.

So, how many times 5 times out of 14. So, we are considering the entropy of the rain and based on this we can determine the information gain of the attribute outlook corresponding to the complete set S. So, this 0.2464 it is the information gain of the attribute outlook. So, move to the next slide.

So, now we are considering the another attribute that is the temperature. So, corresponding to this temperature I have the values the hot, mild and the cool. So, we want to determine the information gain corresponding to the attribute temperature with respect to the complete set or with respect to the entire set. So, entire set is S. So, 9 positive examples and 5 negative examples. So, like in the previous slide we have determined the entropy for the complete set or the entire set and that entropy is 0.94. And let us consider these values of the temperature one is the hot. So, how many times it is hot. So, 2 positive examples corresponding to hot. So, one positive example is this. So, another positive example is this.

So, 2 positive examples of hot and also 2 negative examples of hot. So, 2 negative examples of hot these 2 negative examples of hot. So, 2 positive examples and 2 negative examples and based on this we can determine the entropy. So, $-\frac{2}{4}log_2\frac{2}{4}$ because we have 4 examples corresponding to the hot. So, $-\frac{2}{4}log_2\frac{2}{4} = 1$. And similarly for mild also 4 positive examples, 2 negative examples we can determine the entropy of the mild and also the entropy of the cool.

So, 3 positive examples and 1 negative example we can determine the entropy. After this we can determine the information gain corresponding to the attribute temperature. So, we are considering entropy of the overall set that is the complete set minus the entropy of S_v. So, the values may be hot, mild and the cool. So, $Gain(S,Temp) = Entropy(S) - \sum_{v\in[Hot,Mild,Cool]}\frac{|S_v|}{|S|}Entropy(S_v)$. So, from this formula you can see the entropy of the overall set entropy of the hot entropy of the mild entropy of the cool we can determine and we can determine the information gain corresponding to the temperature.

After this we can determine the information gain corresponding to the attribute humidity. So, the like the previous case also we are determining the information gain corresponding to the humidity. So, corresponding to the complete set the entropy is 0.94 that already I have explained and corresponding to the humidity I have 2 values one is the high another one is a normal. So, corresponding to high you can see 3 positive examples here I have shown 3 positive examples corresponding to high and 4 negative examples. So, based on this we can determine the entropy of the state high and entropy of the normal the normal humidity also you can determine 6 positive examples and 1 negative examples.

So, corresponding to normal you can see 1 negative example and this is the negative example corresponding to the normal only 1 negative example the remaining are positive examples 6. So, that is why out of 7, 6 positive examples and 1 negative examples. So, it is a $\frac{6}{7}\frac{1}{7}$ and we are getting the entropy of the state normal and again we can determine the information gain corresponding to the attribute humidity.

So, this information gain is 0.1516. So, that means we have determined the information gain corresponding to the outlook temperature humidity now we have to determine the information gain corresponding to the another attribute that is the wind. So, same principle we are applying the information gain corresponding to the attribute wind. So, corresponding to the complete set the entropy is 0.94 and in this case in case of the wind I have 2 values one is the strong another one is the weak corresponding to the strong 3 positive examples and 3 negative examples and corresponding to this you can see entropy is 1. So, that is the highest entropy you can see and corresponding to the weak 6 positive examples and 2 negative examples and corresponding to this entropy is 0.113 that you can get and corresponding to this attribute win we can also determine the information gain the information gain is 0.478 that already I have shown how to determine the information gain.

So, out of all these attributes you can see I am getting these values that is the information gain corresponding to the attribute outlook that is 0.2464 corresponding to the information gain corresponding to the attribute temperature this is 0.0289 and corresponding to the humidity information gain is 0.1516 corresponding to the win the information gain is 0.0478. So, I have this information gains now here you can see out of all these which one is maximum out of this the maximum is the outlook that means the best attribute is the outlook. So, we have to consider the outlook is the best attribute because the information gain is maximum in this case and here you can see corresponding to the outlook I have this this one d1 d2 sunny d8 is also sunny d9 is also sunny and d11 is sunny. So, corresponding to the sunny I have 5 examples d1 d2 d8 d9 and d11 corresponding to overcast I have a 4 examples d3 d7 d12 and d13 and corresponding to rain I have the examples the example is d4 d5 d6 d10 and d14.

So, I am repeating this we have determined the information gain for the outlook for temperature for humidity and for the wind and out of this the maximum information gain is gain corresponds to the attribute outlook. So, that is why we are considering the attribute outlook and corresponding to this outlook I have the values like sunny overcast and the rain 3 values and corresponding to sunny I have the example 5 examples d1 d2 d8 d9 and d11. So, like this so now move to the next slide. So, my decision tree will be like this. So, outlook I have to consider as a root node and corresponding to this outlook I have the values sunny overcast and the rain.

So, based on the information theory point of view and based on the information gain point of view I am considering the attribute outlook and corresponding to the outlook I have the values sunny overcast and the rain after this how we can build a decision tree. So, let us move to the next slide. So, you can see that means we are considering 14 examples corresponding to the outlook. So, d1 to d14 and 9 positive examples because 9 plus and 5 minus that is 5 negative examples out of 14 and corresponding to the sunny already I told you the examples are d1 d2 d8 d9 d11 and 2 positive and 3 negatives.

So, you can see d1 is negative d2 is negative d8 d8 is also negative d9 is positive d11 is positive. So, that means it is 2 positive examples and 3 negative examples. So, it is uncertain. So, that is why we cannot determine what will be the best attribute corresponding to the overcast what are the examples d3 d7 d12 d13. So, out of these 4 positive examples and 0 negative examples that means certainly we can say it is es and corresponding to the last one that is the rain.

So, I have these examples d4 d5 d6 d10 and d14 out of these 3 positive examples and 2 negative examples. So, that also we cannot decide. So, for decision we have to consider the information gain. So, based on the information gain, I have to take a decision.

So, now let us consider this sunny. So, corresponding to sunny we have to consider only the examples d1 d2 d8 d9 and d11. So, move to the next slide. So, since we are considering only the sunny. So, that means I have to consider the examples d1 d2 d8 d9 d11.

So, only these examples I have to consider. Now, let us consider the attribute temperature. So, the values are hot, mild and the cool corresponding to the sunny because corresponding to the sunny I have only 5 examples out of 5 examples 2 are positive and 3 are negative, we can determine the entropy of the sunny that entropy is 0.97. So, that means entropy of the set this set the subset I can determine and after this I can determine the entropy of the hot that will be 0 because only 2 negative examples entropy of the mild 1 positive 1 negative. So, entropy is 1 entropy of cool that is 0 because 1 is positive example and 0 is no negative no negative examples.

And based on this we can determine the information gain. So, entropy of sunny that is this set entropy of the sunny this set and minus summation of entropy of S V that is V is the hot mild and the cool. So, we can determine the information gain of the temperature corresponding to sunny. So, that information gain is 0.570 that information gain I can determine. Similarly, we can consider another attribute, the attribute is a humidity. So, corresponding to the humidity, the values are high and the normal and already I have determined the entropy of the set this set is 0.97 because we have 5 examples out of 5 examples 2 are positive and 3 are negative. So, we can determine the entropy of sunny and after this we can determine the entropy of the humidity high and the humidity normal.

So, it is 0 and 0 and we can determine the information gain corresponding to the humidity with respect to sunny. So, it is 0.97. And similarly, we can determine the information gain corresponding to the another attribute that is the wind. So, it has 2 values strong and weak and like the previous case, we have determined the information gain corresponding to sunny and corresponding to the wind strong wind and weak wind, we can determine the entropy one is 1 another one is 0.9183 and we can determine the information gain and out of all this information gain I have to determine the maximum one, which one is the maximum. So, based on this, you can see we are determining the information gain corresponding to the temperature corresponding to the humidity corresponding to the wind and out of this, this is the maximum the information gain corresponding to humidity is maximum.

So, that means I have to consider the attribute humidity. So, in the next level of the decision tree, I have to consider this attribute. So, here you can see. So, the sunny we are considering and now another attribute we are considering that is the humidity. So, this attribute we are considering based on the information theory point of view, because it has maximum information gain and corresponding to this humidity, you can see high value is d1 d2 d8.

So, these are no, if you see the table and corresponding to the normal, these two examples d9 and d11. So, it is yes, I have forecast already I told you that there is no requirement of any, any development of the decision tree, because answer is already yes, because four positive examples zero negative examples. So, answer is yes. So, there is no requirement of growing the decision tree and like this in this case also we have to grow the decision tree based on the same principle.

So, in the third node also we have to grow the decision tree. So, like this we have to build a decision tree. So, this is one example. So, you can see finally, I have this decision tree outlook is one attribute humidity is one attribute, the wind is one attribute. And finally, you can see how I can do the classification. So, corresponding to d1 d2 d8 no d9 d11 yes d3 d7 d12 d13 it is yes d6 d14 no d4 d5 d10 it is yes.

So, I will be getting the decision tree like this. So, this is the procedure of building a decision tree. So, now I will discuss the concept of overfitting in case of the decision tree. So, many kinds of noise can occur in the examples. Two examples have same attributes value pairs, but different classifications. Some values of attributes are incorrect, because of errors in the data acquisition process or the pre processing phase.

And some of the examples are labeled incorrectly. Some of the examples are labeled incorrectly plus instead of minus, and some attributes are irrelevant to the decision making process. So, for example, suppose if I consider color of a dye that is irrelevant to its outcome. So, that is that is not an attribute, the color is not an attribute.

So, we should not consider this one. So, these are the reasons of overfitting of a decision tree. So, in this case, I have shown the case of overfitting of decision trees. So, in the x axis, I have shown the size of the trees, that is a number of nodes. And in the y axis, I have shown the accuracy.

So, here you can see this dotted line is for test data. And this bold line is for the training data. So, if I consider more number of nodes in the decision tree, then the accuracy is very high corresponding to the training data. That means we are considering a very complex decision tree. And corresponding to this complex decision tree, I am getting the best accuracy, the highest accuracy I am getting.

But corresponding to the unseen data, unseen test data, the accuracy drops. And that is actually the overfitting of the decision tree. So, that means I should not consider a very complex decision tree. So, this complex decision tree may be good for the training data, but that may not be good for the unseen test data. So, that is the concept of the overfitting.

So, how to stop overfitting, stop growing when data split is not statistically significant. And we have to consider more training data. And we have to remove irrelevant attributes. And also we can consider the pruning of the decision tree. For first I have to grow the decision tree.

And after this I have to do post pruning. So, that concept I am not going to discuss how to do the pruning. So, how to select the best tree, measure performance over training data, measure performance over separate validation data set, and also add complexity penalty to performance measure. That means heuristic simpler is better. So, we can select the best decision tree based on this. And also we can avoid overfitting by considering these cases that is, we have to consider more number of training data, remove irrelevant attributes.

And also we can do post pruning. Also we can consider other metrics other than entropy. Entropy we considered to determine the information gain. So, maybe we can consider Gini index, the Gini index is something like this $G = \sum_{i=1}{}^{C} p(i) * (1 - p(i))$. So, this is a very simple measure. And with the help of this Gini index, we can also take decisions.

So, Gini criterion is must faster because it is less computationally expensive. But this entropy criterion gives slightly better results. So, decision trees are widely used as they are simple to understand and interpret. So, advanced topic that you can explore that I will not cover in this course.

So, one is the pruning methods. So, how to deal with the overfitting. So, that concept I am not going to explain in this course. So, there are different pruning methods to consider the overfitting. And another one is the random forest for better accuracy. So, the concept of the random forest I will be discussing in my next classes. Whenever I discuss the concept of the ensemble classification, I will be discussing the concept of the random forest.

So, briefly I can explain the concept of the random forest, you can see the training set. This training set is divided into number of subsets training data set one training dataset to training dataset tree. Like this we are considering the subdivision of the training dataset. After this we are considering multiple decision trees. decision tree 1 decision tree 2. Like this we are considering multiple decision trees, and we are taking the outputs from all the decision trees. And we are doing the voting that is the averaging we are taking and based on this we are doing the prediction. So, this is the fundamental concept of the random forest. So, anyhow I am going to explain the concept of the random forest in my next classes. So, this is the fundamental concept of the random forest to improve the accuracy.

So, in this class I explained the concept of the decision tree, I also explained the concept of the entropy and information gain. So, from the information gain, I can determine the best attribute. After determining the best attribute, I can develop or I can design the decision tree. And also I explained the concept of the overfitting in case of the decision tree.

And finally, I explained the concept of the random forest. So, briefly I explained the concept of the random forest. And this concept I will be explaining again in my next classes. So, let me stop here today. Thank you. .