Data Analysis and Decision Making - II

Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 19
Decision Trees

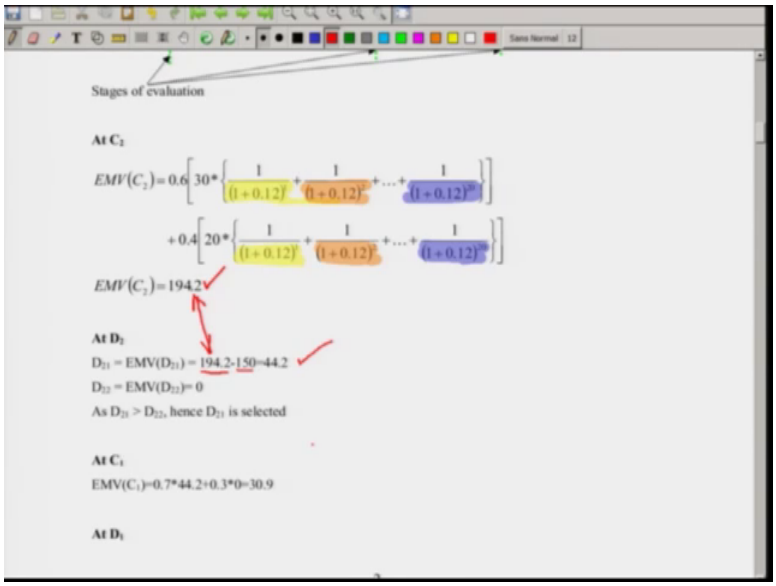
Welcome back my dear friends and students a very good morning, good afternoon, good evening to all of you. And as you know this is the DADM-2 course, under the NPTEL MOOC series. DADM means the Data Analysis and Decision Making-2. And as you know in this whole course, we will be discussing the non-parametric methods of decision making, multi-attribute utility theory, optimisation multi-criteria optimisation, not from hard core mathematics, but more from an subjective point of view, how you use AHP, ANP, topsis, electra, and all these methods.

So, initially we consider utility theory in a very simplistic sense, then we went into decision trees. And then decision trees we are doing now, then we did into the formulation of DA, we will come back to DA later on. So, let me continue with discussing the concepts of decision trees.

And as you know this would be the 19th lecture of which will be in the 4th week. And each week, there are five lectures, each lecture being for five half an hour. And after each week, we have assignments, so we have already completed. As I speak you must have completed three assignments. So, we will be going to the 4th-assignment, after the end of 20th-lecture.

So, if you remember, where we stop the last day we are considering the different types of trying to find out the present value of the amount of money, which you obtain. So, for the decision tree if you remember for high and low demand, I would not going to the details of the background again. But, for high and low demand, you are getting 30 million per year for the coming 20-years, and for low demand you got 20 millions per year for the coming 20-years.

(Refer Slide Time: 02:08)



And based on that I will just use the highlighter, based on that you found out the there were two arms. So, with a 60 percent probability the high demand, 30 was the value per year. And these were if you remember this was the first year interest rate, so if you got, you obtain 30 million after one year, then the present value all that money be 30 divided by the one, which is marked in yellow. I should not mark in this in yellow, let me remove it.

So, and then if you obtain another 30 million after 2-years, so that would be divided by wait, so that would be divided by 1 plus r to the power square means, square means 2-years. Similarly, if you obtained another 30 after 3-years, it would be divided by at amount which is not mentioned here, and if we continue here. And if you obtain 30 millions, after 20-years because that was the time frame, then the whole amount of values which you will get would be 30 million divided by 1 plus r to the power 20.

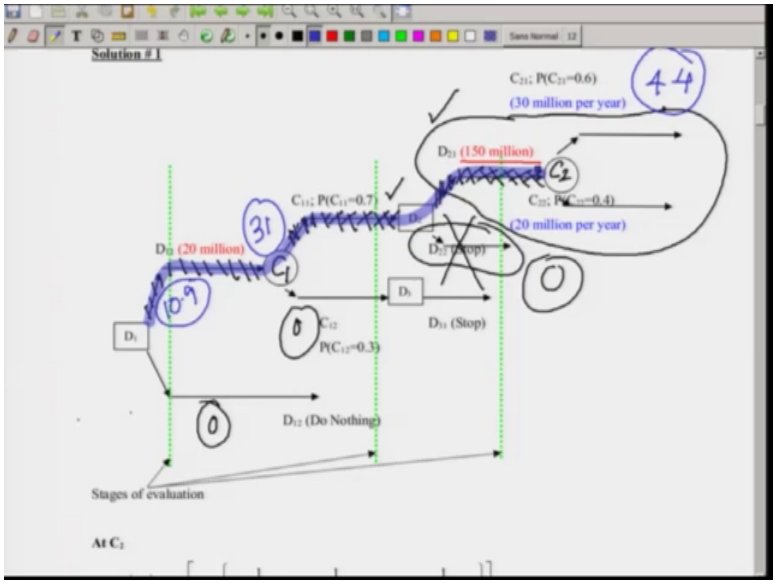
Similarly, you would use the same logic. If we go for the low demand, then the low demand probability is 40 percent, I am not marking it, I am just highlighting it with my or trying to point it out with the electronic pointer 40 percent, 20 million per year. Similarly, you will follow the same concept after 1-year, if you got 30, you will divide it by 1 plus r to the power 1.

If you got another twenty sorry 20 million after 2-years, so you will divide it by 1 plus r to the power square, and continue doing it. Till the last term which is after 20 years, if

you got 20 millions, you will divide by 1 plus r to the power 20. So, this is exactly the same concept, which you use to rank the decisions in utility theory. So, obviously you can use other method also. So, the net value comes out to be 194, I will just mark it 194.

Now, I would in this lecture, I will go back and forth with the calculation and the diagram, so it will be easy for you to understand. So, this is the first time, I am doing it. For any of the courses which we have taught, because they were all going in sequential manner one slide of the other but, in this case because the diagram is so huge, so I had to make it in this word document converted into pdf.

(Refer Slide Time: 05:04)



So, let me go back to the diagram. So, this is one what we considered 30 millions per year for 20-years, 20 million per year for 20-years, probability 60 percent, 40 percent for high demand and low demand. Now, that value if you calculated, it come came out to be about 194.

Now, 194 was the money coming back to in your pocket, but then you should ask that did you invest something to generate this revenues, the answer is yes. Because, if you note this 150 million rupees or yens or dollars whatever it is was basically for setting up the actual plant. And in other step was do nothing that means, stop the production.

So, if you consider 150 as a value, which went out of pocket that is why it is shown in red colour, and others which are coming inside your pocket coming back to you (Refer

Time: 05:55) blue in colour. So, if you consider this 150, so the expected value of the net flow both revenues, and cost being 194.2 which is coming from here. And this 150 is the value which is the cost which you incur for setting up the plant, so the total value comes out to be about 44.2, I will mention it 44.

So, now this 44 which you are which you have which is positive you should consider with respect to what. So, let me again go back to the diagram. So, this total arm, let me use the colour black, so it be easy for me to so this total decision is 44. I am not using the decimals. In order to make it little bit more clear, let me put it 44 as blue one or the so this is 44. So, this is a positive one.

Now, if you come here to the other arm, so what is the value you do nothing that means, the value would be 0. So, I will put a 0 here. So, you will compare 44 plus with respect to 0, and then consider that the actual decision which you will take in this case would be you would never go into is this direction, I am just putting a cross mean that you never go.

So, obviously your actual decision would be the one where you will set up the plant, and go for high demand, low demand values of 30 and 20 as shown. Now, at this stage when you take a decision, your net value is comparison between 44 and 0. So, let me come here. So, D21 is 44, D22 is 0. So, as D21 the arm which is giving you 44 is more than the arm, which is giving a net present value of 0, so obviously D21 will be selected.

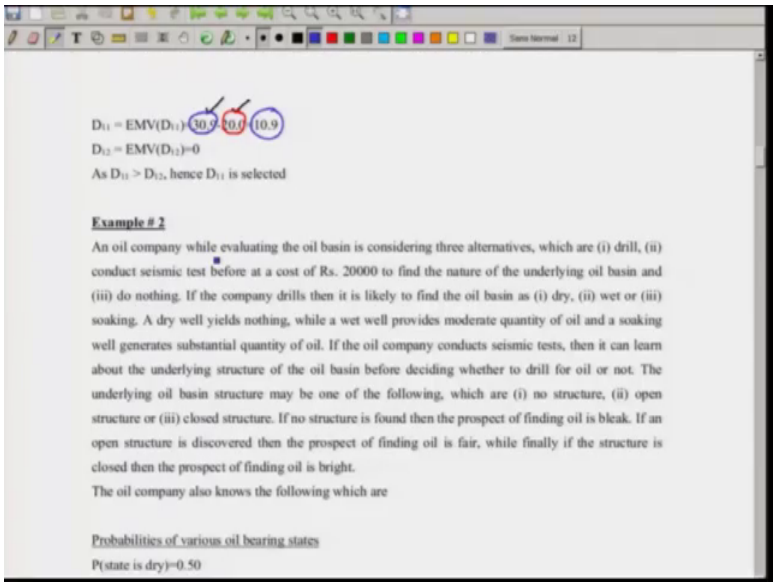
Now, when you have reached D21, you are here. So, this is 44. Now, this 44 when you consider, you will basically find out that what is the net present value or the expected value at C1. Now, you coming comparing C 1, there are two arms already. One is that 44 with the probability 0.7 70 percent, and another thing you do nothing with the probability of 0.3.

So, obviously the your actual expected value at C 1, I should mark this as C 1, and this is C 2. C 1 and C 2 would be such that the expected value of C 1 would be 44 and again ignoring the decimal 44 into 70.7 and 0 into 0.3. So, hence the expected value at this level would be like this 44 into point 7. And 0.3, which is the probability in 0, because 0 is the value for which you do nothing.

So, the expected value at C 1 is 30.9, which is 31. Now, when it is 31, then again you go back. So, you I am just going back and forth in the calculation and diagram. So, this is 31 at this stage ok, this should be positive. So, let me remove it. So, this is positive, so this is 31. And obviously, this value is 0, this value is 0.

So, till now if I use, I have taken this path, and what comes out to be high demand and low demand will come later now. So, the hashed one is the path which you will take. Now, when you come to C1 to arrive at C1, you will find out that going if I take the rout D 11. So, what is the actual cost? Cost means, what is the amount of money which you may have spent. So, if you remember for pilot production and test market, you have to invest 20 millions. So, this 31 revenue should be adjusted corresponding be to the value of 20 million already spend, when you have done the test marketing.

(Refer Slide Time: 11:08)



So, let us go back. So, at D1 31 you have already obtained as a revenue, 20 was negative. So, technically, this is a blue, this is blue, and this is red, because this is minus. So, the total value which comes out to be 10.9 which is 11. So, this is the value which is positive.

So, you will now you again compare 10.9, which is the value here. So, 10.9 considered against the value of 0. So, at D1, you will compare 10.9 and with the positive for the upper arm with respect to the value of 0 value, where you do nothing. So, obviously again you will see that you would take this decision that means, you do the test

marketing, pilot plant, then you basically with probability 70 percent, go for building up this factory spend 150 million.

And then see there are high demand or a low demand with probability 60 percent, 40 percent. So, the whole arm which you will consider let me highlight it using blue, so this will be now. And it can be either high or low there can be other arms also, but this is very simplistic one.

So, hence you will to pick take that arm. Now, I will go into second example ok. Before, I go to the second example; they would have been many quarries. So, let me and probable quarries which come up in class. One would be why do we take, I will come to the diagram, so it will be easy for me to explain. Number-1 question can be that this 30 million which you are getting, may not be same for each year for 20-years right, it may not be same. It may change from year to year 0.1. Point number-2 is that is 20 million also may not be for a low demand, may not be same from year to year, it may be on a decreasing 10. I agree, so this can also change.

Third question would be why consider 20-years as high and low demand time frame for both 60 percent, 40 percent yes that is true. High demand can be say for example 20-years, low demand can be say example 15-years that is also agreed. 4th point would be why do we take the interest rate as same for this high and low demand, yes we agree that the interest rate can be different for high demand and low demand.

The 5th why do we consider the interest rate to be fixed for each and every year, it can be r 1, r 2, r 3, r 4 correspondingly to the fact for so high demand. Similarly, can be r 7, r 8, r 9, r 10 different interest rate for the low demand yes that is possible we can consider that. 6th one can be say for example, rather than having only high and low demand, we can have high demand, medium demand, low demand yes that is possible.

So, all the arms which we have taken rather than 2, it can be more than 3, it can be 4, 5, 6, similarly for 70 percent probability setting up the plant and investing 150 million. It could have been you set up the plant of a biggest scale of 150 million spending with 55 percent probability. Another can be you spent 30 percent probability with a 30 percent probability, you spend 100 millions. And with 20 percent probability, you do not do nothing. So, these are the different type of base.

In this case also, they can be more than two arms that means, you spend 20 million with a certain probability, and another case you do nothing. So, rather than 20 millions, it can be broken three stages. Three stages means like in one case you spend 50 millions for a bigger marketing and empire plant with certain probability.

Then you spend or else you spent say for example 30 million for a lower marketing survey and lower smaller pilot plant. So, all these combinations of arms with different probabilities can be done, but obviously remember the sum of all the probabilities for anyone decision should always add up to one.

So, these assumption, we can also expand think that all the 7 or 8 different type of ways of assumptions, which we have taken which can be considered can also be done for the second problem, which we are going to consider now. This is a little bit more involved, so please bear with me. An oil company while evaluating the oil basin is considering three alternatives, it can be more than three also, which are drill you do a drilling. Number-2 will be conduct seismic test, so do some geographical test, geological test. Before, you drill, and that would cost about 20,000. So, amount can vary, but let us consider very simplistically is 20000.

And you want to find out the you conduct the seismic test, to find out what type of geo physical properties are there for the soil or the earth level, where you are going to drill for the oil basin. And third option can be you do nothing, so that means directly going to drill, conduct test, and do nothing. So, they can you more than three also. So, let us make it simple.

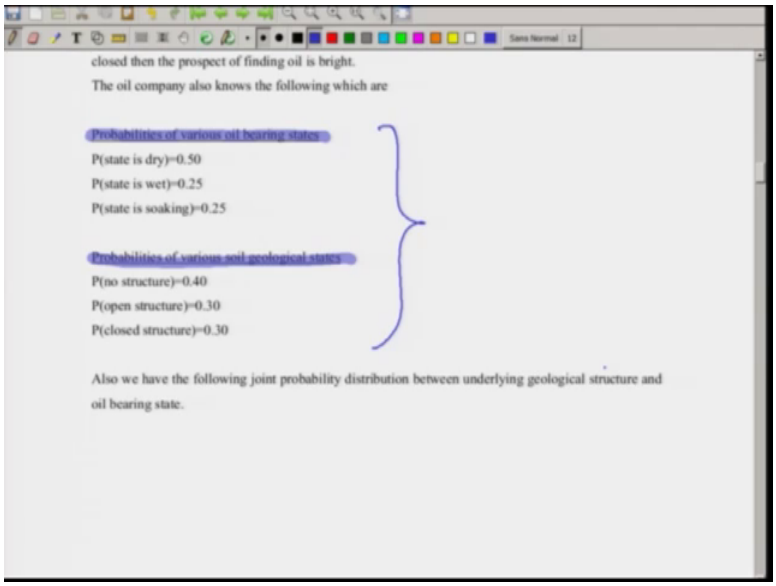
If the company drills, then it is likely to find that the oil basin. So, if it drills, it finds it is dry, it is wet, and it is basically socking, socking means you get a lot of oil. So, a dry well yields nothing, while a wet well provides moderate quantity of oil, and a soaking while generates substantial quantity of oil.

So, dry minimum quantity of oil, wet means substantial amount of oil, and soaking means quite a huge amount of oil. If the oil so let me again read it, a dry well yields nothing, wet well provides moderate quantity of oil, and soaking will generate substantial quantity of oil. If the oil company conducts seismic test that means, we are going back to the case, where the company does conduct the seismic test.

If the oil company conducts seismic test, then it can learn about the underlying structure the oil basin before deciding, whether to drill for oil or not. So, it conducts stress finds out the characteristics of the geo physical properties of the basin, and then decide whether you wants to drill or not. The underlying oil basin structure may be of one of the following type, which means that there is no structure, there is open structure, there is close structure. These are properties of the geo physical properties of the earths surface or of the basin, where you want to conduct the test.

So, if no structure is found, then the prospect of finding oil is very bleak, very low. If an open structure is discovered, then the prospect of find oil is quite fare. And if the structure is closed, the prospect of finding oil is bright, so it is quite high. Now, the oil company oily also knows the following. Now, this probability which I am going to state, can also be given in another way, I have also given that in another way, which I am I want to make it very clear to you such that they should not be any confusion, when you want to find out.

(Refer Slide Time: 19:02)



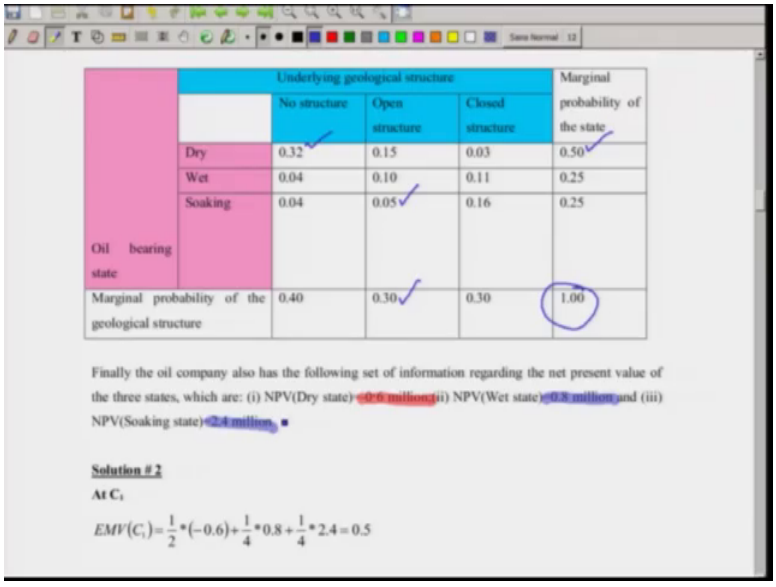
Now, look at here, what it gives? Let us consider the first case, so there are two things. So, first let me mention the probabilities of various oil bearing states. So, what are the oil bearing states, they are dry, wet, and soaking. The probability these are unconditional probability they not depend on any conditions given.

The probability of finding a dry state is 50 percent, probability of finding a wet state is 25 percent, probability of finding soaking state is 25 percent. So, unconditional on whatever it is probabilities are 50, 25, 25. Then I go to the probabilities of various soil geo logical state or the geo physical states of the soil or the earth crust are given as no structure is there 40 percent, open structure problem probability is 30 percent, closed structure is 30 percent.

So, again there are unconditional does not depend on any and condition based on which you can find it. Now, remember this 50, 25, 25, and 40, 30, 30 would be coming up later also why, and how they will come up I mention that again.

Now, rather so for this for the time being, let us consider these are not given. Even though, I will I will mentioned I will mention, but consider they are not given, what we will do. So, if this structure which would be given, the structure of the probabilities which will be given to us is like as follows. So, also let us re read it. Also we have the following joint probability distribution based on the underlying geophysical structure and the oil bearing states.

(Refer Slide Time: 20:49)



So, now we are basically finding out the joint probability of these two conditions. One is oil bearing states, and one is soil geo physical or geological state. Now, with this if I find out the probability structure, it is like this. Let us understand, it its very simple. And if

you remember the condition unconditional probabilities, which I gave will come out immediately from here.

So, on one end we have the oil bearing states which is dry, wet, soaking. And on the other end, we have the geological structure which is no structure, open structure, and closed structure. So, if I consider the probabilities combined probabilities, then they would be given in each cell. So, obviously you will have 0.32, 0.15, 0.10, then 0.04, 0.10, 0.11, 0.14, 0.15, 0.16. So, all these probabilities which are given are conditions on two factors.

So, if I consider wet, it means that it is both open structure, and wet. So, if I consider this 0.1, so obviously, it will mean that is a open structure, and wet oil bearing state. If I consider 1.6, it will be closed structure and soaking state. Now, if you remember, I mention that 50, 25, 25 and 40, 30, 30.

So, look at the marginal values that means, the extreme column, and the bottom most row extreme right column, so the marginal probabilities the states which were unconditional was given as 0.5. So, if you add up these values 0.32 point plus 0.15, 0.03 comes out to be 0.5. Similarly, if you add up the second row comes out to be 25; I am not mentioning the decimal.

If you add up the third row, it comes out to be 25 again. Now, if you add up the columns that means, the first column for no structure, it comes out to be 0.4. Add up the second column, corresponding to open structure, it comes out to be 0.3. Add up the third column corresponding to close structure, it comes out to be 0.3. And obviously, if you find out the sum of the unconditional probability, this should definitely come out to be one as it should be by addition of the right most column, and the bottom most row.

Now, with this given everything is solved. So, if I ask you a question, that what is the probability that it is no structure corresponding that of fact that is as dry oil bearing state. So, obviously he will use the values of 0.32 and 0.5. If I ask you the question, that what is the probability that it is soaking given that is an open structure, then you will obviously use the values of 0.05 and 0.3.

So, corresponding to what conditions are given, whether it is open structure, and then soaking or it is basically a wet structure or then no structure, you can basically wet state

sorry wet state, and no structure you can find out. Depending on which you direction you are looking whether it row or column, you will divide the corresponding cell. Cell means in this 3 cross 3 values, which are there you will divide it by the corresponding extreme most right most column or the bottom most row to get the conditional probabilities accordingly, we will utilise this structure accordingly.

Now, you can ask let me open the discussion before solving the problem, you can ask that what if we consider a three stage conditional probabilities that means, depending on probability of A, you have B, depending on probability B, then obviously, you have C. Then obviously, the probability of C would be depended on B, which will be depend a yes in that case that that will be possible, so obviously we will have the conditional probabilities depending on states A, B, C accordingly.

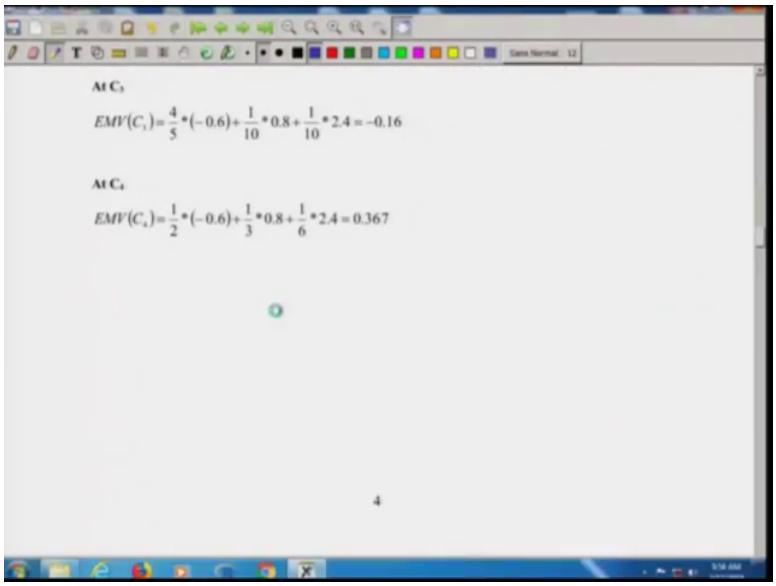
Similarly, it can be done for four stages that means, A decides B, B further more decides C, and C furthermore decides D. So, when we find out, so these are the events. So, the probability of D would be dependent on C, which will be dependent on B, which will be dependent on A. So, this we are I am reading the problem.

Finally, the oil company also has the following set of informations, informations are as follows mean regarding the net present value of the three states, which is the net present value of the dry state is given as minus 0.6. So, this should be red in colour, because this is negative, if you are following the same (Refer Time: 25:22). Net present value for the wet state is plus 0.8, so obviously it is positive. And net present value for the soaking state is plus 2.4, so obviously is positive. So, we have the net present value.

So, obviously the second question which we come up again I am making it, an open discussion that you could think that rather than having the net present value, you could have the payment happening after drilling the payment, which you are getting after each year.

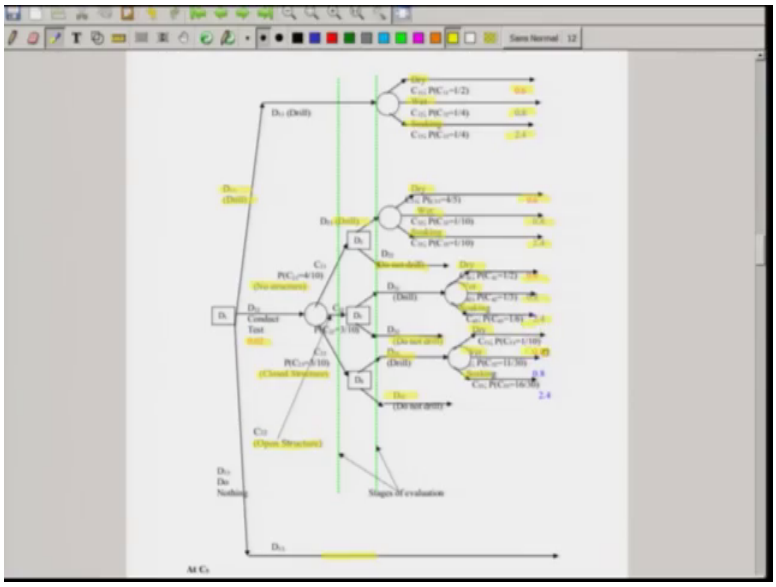
Then you have to basically find out the net present value, considering the concept which used in the first problem where you divide each and every years written by 1 plus r to the power n, where n keeps changing that year 1, year 2, year 3, so on and so forth. You can find out the net present value, where r is basically interested. Again you can assume the interest rate r is changing from year to year, and so on and so forth. So, it will also change depending on high demand, low demand, all these things are possible.

(Refer Slide Time: 26:31)



Now, before I solve the problem, let us see the diagrams. Diagram would not be cluttered, but I have tried my level best to draw it. So, let us see the diagram, so I have just added the calculations first.

(Refer Slide Time: 26:37)



So, here is the diagram. It is a little bit in word, but let me go let me try to reduce the size if possible, well sure you can make something so if required, I will zoom in as required. So, here is the let me a little bit down. So, the whole diagram is visible to us yes. So, the

overall layout of the problem is like this. Again we are drawing it from the source to the sink, but we will do the calculations from sink to the end.

So, you have D1 which is given as drill, which is D11 which is this arm. Let me use a different colour to highlight it, so D1 is still D1. So, D11, D12 is conduct I have considering for this class, I have another two minutes, I will just mention the diagram in a in a very generally, and then come back to the solution. D 12 is conduct seismic test, and for this value is given is 0.02 with a red colour which is negative. And the other arm is D13 is nothing.

If we drill obviously, they would be dry, wet, and shocking which show here, I would not go into the problems probabilities immediately, I will wait I will do it later on. The values if you remember of net present value was given as minus 0.6 plus 0.8, 2.4 minus plus plus. If you conduct seismic test, there would be no structure, they would be closed structure, and there would be open structure.

Once you find out no structure, close structure, open structure, again you will drill or do nothing. So, if you drill, again you have dry, wet, soaking, and do nothing that means, in the same way to say where as you are doing here do nothing. In the case when there is close structure, again if you drill, you have dry, wet, soaking; again I am not going to come to the probabilities immediately. And do nothing is that means, in spite of finding out close structure, you do nothing. And for the final stage for the open structure, so that was sorry for the open structure. And for the closed structure question, again you drill do nothing. And if you drill, you have the dry, wet, and soaking.

The probabilities, I will come to the next class. And the values or the net present values or net present value of minus 0.6 plus 0.8, 2.4 would be repeated for all the cases, when you do the drill. This I have kept very simplistic, and if you remember, I did mention this values can change depending on the time frame, the probability is the interest rate and so on and so forth. So, with this I will end the 19th-lecture, and continue discussing about the decision tree, and try to complete the problem in the 20th-lecture which will be the last class for 4th-week.

Have a nice day, and thank you very much for your attention.