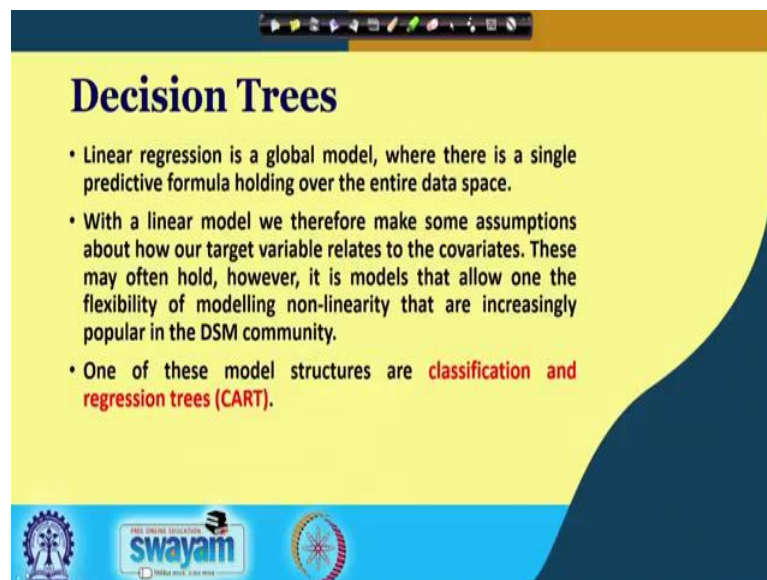


Soil Science and Technology
Prof. Somsubhra Chakraborty
Department of Agriculture and Food Engineering
Indian Institute of Technology, Kharagpur

Lecture – 58
Modeling Continuous Variables (Contd.)

Welcome friends to this 3rd lecture of week 12 of Soil Science and Technology. And in this lecture we will try to cover several quantitative models for digital soil mapping and let us start with a decision tree.

(Refer Slide Time: 00:28)



Decision Trees

- Linear regression is a global model, where there is a single predictive formula holding over the entire data space.
- With a linear model we therefore make some assumptions about how our target variable relates to the covariates. These may often hold, however, it is models that allow one the flexibility of modelling non-linearity that are increasingly popular in the DSM community.
- One of these model structures are **classification and regression trees (CART)**.

The slide features a yellow background with a dark blue wavy shape on the right side. At the bottom, there are logos for IIT Kharagpur, Swayam, and a circular emblem.

So, in the in the last lecture we have covered this linear regression which is a global model where there is a single predictive formula holding over the entire data space. However, sometime the soil data are nonlinearly related, so we may be interested to explain or may be model them through non-linear methods. So, one of these models structure non-linear model structure is called the classification regression trees.

(Refer Slide Time: 00:58)



CART

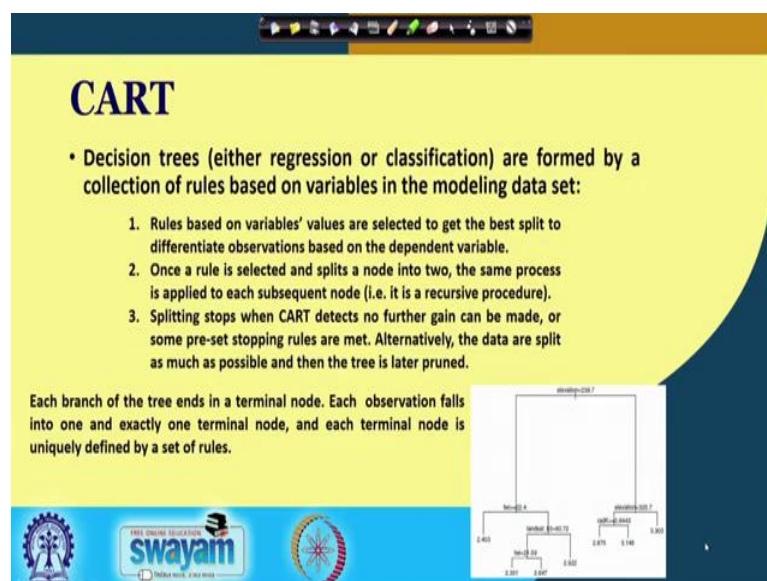
- These models are a non-parametric decision tree learning technique that produces either classification or regression trees.
- Regression trees when our target is numeric: a continuous variable
- **Classification trees for categorical variables**

Logos: IIT Bombay, swayam, and a circular emblem.

A small video inset shows a man in a white shirt speaking.

And this classification regression tree or CART are a non parametric decision tree learning techniques that produces either classification or regression trees. And regression trees; we use the term regression tree where our target is numeric when or in other words it is a continuous variable. And we use the classification trees where we use a, we use to model a categorical variable. Now categorical variable are not continuous variable; categorical variable meaning any name nominal variable are considered as categorical variables.

(Refer Slide Time: 01:35)



CART

- Decision trees (either regression or classification) are formed by a collection of rules based on variables in the modeling data set:

1. Rules based on variables' values are selected to get the best split to differentiate observations based on the dependent variable.
2. Once a rule is selected and splits a node into two, the same process is applied to each subsequent node (i.e. it is a recursive procedure).
3. Splitting stops when CART detects no further gain can be made, or some pre-set stopping rules are met. Alternatively, the data are split as much as possible and then the tree is later pruned.

Each branch of the tree ends in a terminal node. Each observation falls into one and exactly one terminal node, and each terminal node is uniquely defined by a set of rules.

Diagram of a decision tree structure:

```
graph TD
    Root["splitting on X1"] --> Node1["Node 1"]
    Root --> Node2["Node 2"]
    Node1 --> Node3["Node 3"]
    Node1 --> Node4["Node 4"]
    Node2 --> Node5["Node 5"]
    Node2 --> Node6["Node 6"]
```

Logos: IIT Bombay, swayam, and a circular emblem.

So, decision trees are basically either regression or classification are formed by a collection of rule based on variables in the modeling data set. So, remember this classification tree or regression tree is basically a rule based algorithm when the rule based on variable values are selected to get the best split to differentiate observation based on the dependent variable. And once a rule is selected and splits a node into two, the same process is applied to each subsequent node. And the splitting stops when CART detects no further gain can be made or some pre set stopping rules are met; alternatively the data are split as much as possible and then tree is later pruned.

So, here let me explain; so here you can see this is a class, this is a decision tree or classification regression tree. Here we are trying to, you know, we are trying to model: for example, say for a particular soil property based on our soil organic carbon based on several soil covariates. For example, elevation, then topographic wetness index, then Landsat band three reflectance values and then radiometric potassium from gamma radiation and so on so forth.

So, basically our target in this case is soil organic carbon and we want to differentiate or want to divide or want to make a rule so that we can predict the soil organic carbon based on this rules. So, in the first step we will select which among them we have four here; four to five different predictor variables.

So, we will select the based split and or best variable values to get the best split. So, that it can give you the maximum separation between the observation based on the dependent variable in our case it is organic carbon. So, here it might we have selected this elevation value of 238.7 because it gives the maximum clear separation between two groups of data.

And once this rule is selected and split is node and this split the data into two nodes; into two parts the same process is applied to the subsequent node. So, in the subsequent node; so this is a first node and this is a subsequent node again we select which among these five or four different variables are important and what are their values.

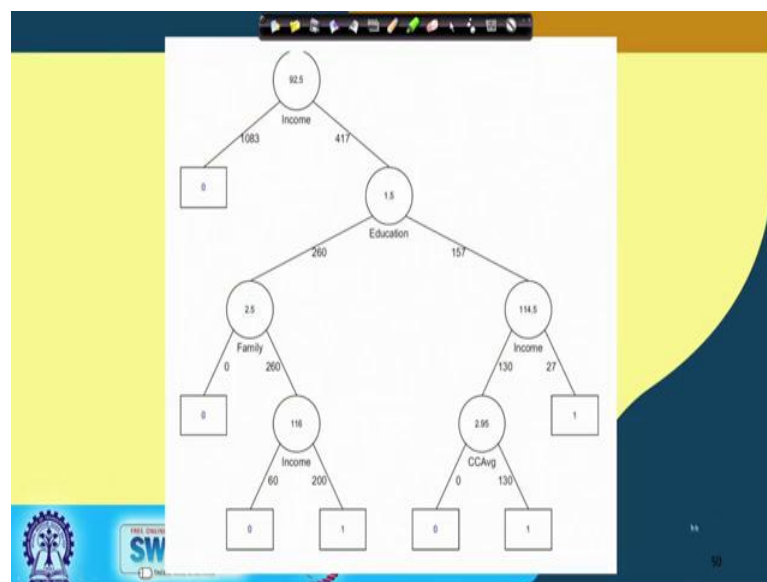
So, we have seen that in the subsequent node this t w i; value of 22.4 is a most important splitting criteria which can split the data into two more groups and so on so forth; you can see this process grows on grows on. So, that is why we call it recursive splitting process. And when this splitting stops? This splitting stops when CART detects no

further gain can be made or some pre set stopping rules. We will discuss this details what are these gains and what are the stopping rules; alternatively the data are split as much as possible and then tree is later pruned.

So, other way we will go we will allow this tree to grow continuously and ultimately once these trees are grown, these are individual leaves and terminal nodes these terminal nodes will go for pruning. You know the cutting of the branches is called pruning. So, similarly here if we cut these branches cut this individual branch so; obviously, this will be further clubbed together; so, this will call pruning we will discuss this later on.

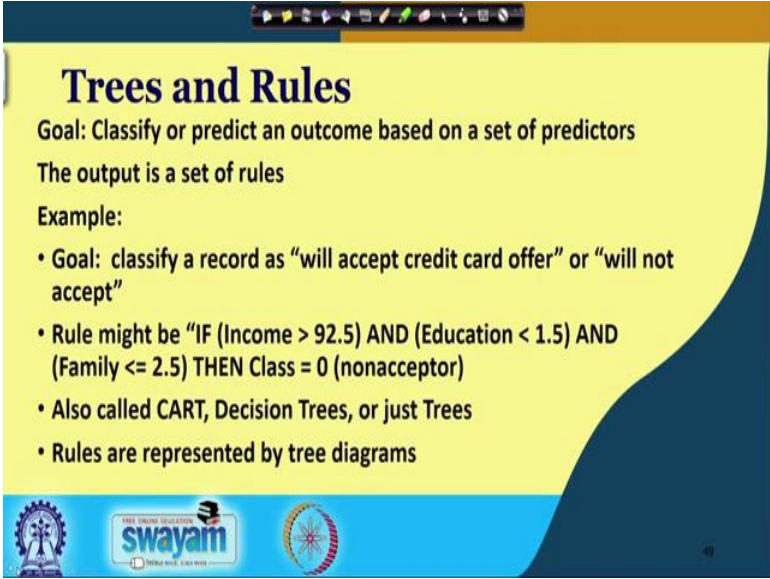
So, each branch of the tree ends in a terminal node and each observation falls into exactly one terminal. So, you can see based on this rule based thing you can; you can you can segregate each of the observation into each of this any of this rule. For example any observation can come in this rule where which is defined by an elevation values of less than 238.7 and t w i value of 22.4. So, this is the rule for classifying a particular observation into this terminal node. So, you can see each of the observation can be, you know, generalized in terms of these rule based theme.

(Refer Slide Time: 06:17)



So, let us see good example I have found this example.

(Refer Slide Time: 06:21)



Trees and Rules

Goal: Classify or predict an outcome based on a set of predictors

The output is a set of rules

Example:

- Goal: classify a record as "will accept credit card offer" or "will not accept"
- Rule might be "IF (Income > 92.5) AND (Education < 1.5) AND (Family <= 2.5) THEN Class = 0 (nonacceptor)"
- Also called CART, Decision Trees, or just Trees
- Rules are represented by tree diagrams

Logos at the bottom: UGC, swayam, and a circular emblem.

So, I want to show you so that it will be very much easy to understand. So, let us see in, you know, we have to classify or predict an outcome based on a set of predictors and the output is a set of rules. So, example here our goal is to classify record whether a person will accept a credit card offer or will not accept a credit card offer I am just giving an example so that it will be easier for you to understand what is going on in this classification regression tree.

So, our goal is to classify a record as either it he will accept credit card offer or he will not accept a credit card offer. So, the rule might be if income is greater than 92.5 and education is less than 1.5 and family is less than equal to 2.5, then class equal to 0 that is non acceptor; so they will not accept. So, this is the individual rule we generate using this classification regression tree.

So, also called CART decision trees are just trees. So, rules are represented by this tree diagram as you have seen in the last slide. So, as you can see here we have classified all the observation based on some rules. So, again here as I have told you if the income is than less than, you know some you know, this 92.5; then there are 417 observations.

And then based on that if the education is greater than 1.5, some index then it will further classify into 157 observation and 260 observation and so on so forth; ultimately you can see ultimately the end product is, you know, either it will accept or reject and reject is 0 and accept is 1. So, this is how we can differentiate any value. So, you can see all the

observations are differentiated into this individual rules. So, let us see how we create this thing.

(Refer Slide Time: 08:24)

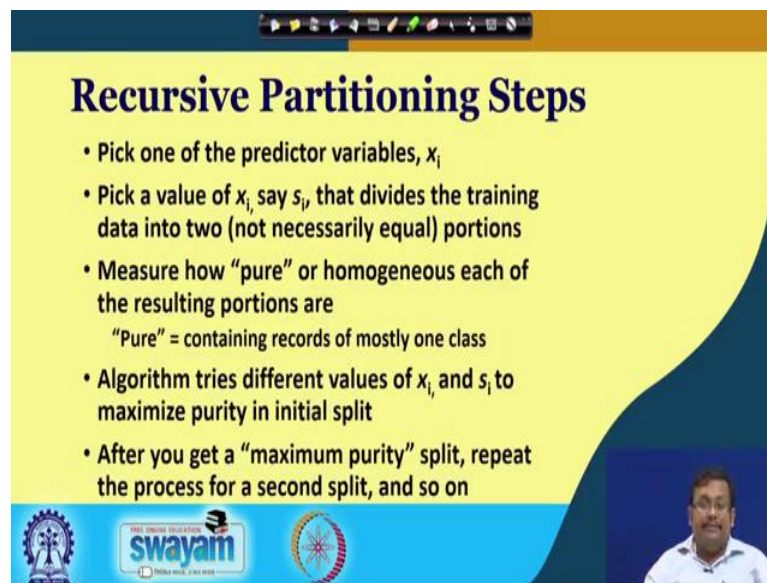
Key Ideas

- Recursive partitioning:** Repeatedly split the records into two parts so as to achieve maximum homogeneity within the new parts
- Pruning the tree:** Simplify the tree by pruning peripheral branches to avoid overfitting

So, key ideas in this case is recursive partitioning and recursive partitioning as you have known that is a repeatedly split the records into two parts so as to achieve maximum homogeneity within the new parts. So, this splitting criteria says that we will select that particular value of the variable as a splitting node, where we will get the two split and in this two split the variables will be maximum, you know, will feel we will see maximum homogeneity within these variables within this observation within a single split,

Again we will select the splitting node or the particular variable and their respective value in the splitting node in a such a fashion so that we can differentiate the groups of observation and in this individual outcome groups those observation in this groups will be maximally homogeneous. So, once we get that another term is pruning which is simplify that tree by pruning peripheral branches to avoid over fitting.

(Refer Slide Time: 09:41)



Recursive Partitioning Steps

- Pick one of the predictor variables, x_i
- Pick a value of x_i , say s_i , that divides the training data into two (not necessarily equal) portions
- Measure how “pure” or homogeneous each of the resulting portions are
“Pure” = containing records of mostly one class
- Algorithm tries different values of x_i and s_i to maximize purity in initial split
- After you get a “maximum purity” split, repeat the process for a second split, and so on

swayam
INDIA RISE, AS THE WORLD RISES





We will discuss that later on. So, let us see what is how we will do this recursive partitioning steps. So, we have to pick up one predictor variables set as x_i and pick up a value x_i say pick up a value of x_i say s_i that divides the training data into two not necessarily equal portions then we have to measure how pure or homogeneous each of the resulting portions are; pure means containing records of mostly one class.

An algorithm tries to different values of x_i and s_i to maximize purity of the initial split. So, there is, this algorithm tries several combination of x_i and s_i to find the best split criteria. After you get the maximum purity split repeat the process for a second split and so on. So, that is why the tree grows on and you can see why we call it as tree? Because a tree has several branches.

(Refer Slide Time: 10:40)


Example: Riding Mowers

- Goal: Classify 24 households as owning or not owning riding mowers
- Predictors = Income, Lot Size



So, as you can see, you know, this classification regression tree is are also having several branches and then sub branches and also so on so forth and then this branches again split. Let us see one example our goal here I am just taking one example from, you know, which I have seen lately so that it will be easier for you to understand. Let us see our goal is to classify 24 household as owning or not owning riding mowers.

(Refer Slide Time: 11:10)

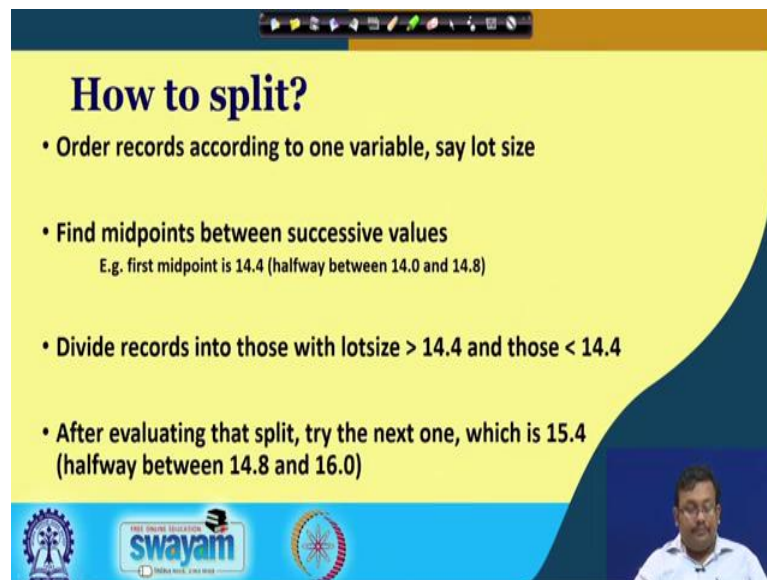


Income	Lot Size	Ownership
60.0	18.4	owner
85.5	16.8	owner
64.8	21.6	owner
61.5	20.8	owner
87.0	23.6	owner
110.1	19.2	owner
108.0	17.6	owner
82.8	22.4	owner
69.0	20.0	owner
93.0	20.8	owner
51.0	22.0	owner
81.0	20.0	owner
75.0	19.6	non-owner
52.8	20.8	non-owner
64.8	17.2	non-owner
43.2	20.4	non-owner
84.0	17.6	non-owner
49.2	17.6	non-owner
59.4	16.0	non-owner
66.0	18.4	non-owner
47.4	16.4	non-owner
33.0	18.8	non-owner
51.0	14.0	non-owner
63.0	14.8	non-owner

So, our predictors here is income and lot size. So, you can see these are 24 observations; the first column shows that the first column is basically the income. And the second

column is the lot size and finally, third column is the, you know, ownership either it is own owner or either owner of that land mower or non owner. So, based on that we have to predict using a classification regression; then here our target will be these two classes owner and non owner.

(Refer Slide Time: 11:43)



How to split?

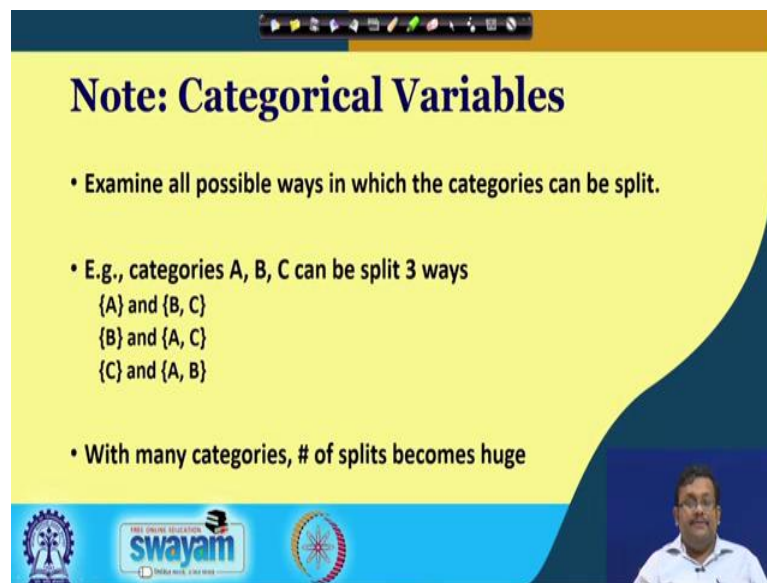
- Order records according to one variable, say lot size
- Find midpoints between successive values
E.g. first midpoint is 14.4 (halfway between 14.0 and 14.8)
- Divide records into those with lotsize > 14.4 and those < 14.4
- After evaluating that split, try the next one, which is 15.4 (halfway between 14.8 and 16.0)

swayam
INDIA RISES WITH EDUCATION

So, how to split? So, order record according to the one variable, say, lot size and then let us find the midpoint between successive variables. For example, first midpoint is 14.4 which is halfway between 14 and 14.8, then divide the records into these two those with a lot size for example, greater than 14.4 and those less than 14.4. And after evaluating that split try that next one which is 15.4 which is halfway between 14.8 and 16.

So, you can see we are trying different combination of values to find out the maximum, you know; you know you know, to find out the best split which will give us the outputs which will be homogeneous among themselves.

(Refer Slide Time: 12:34)



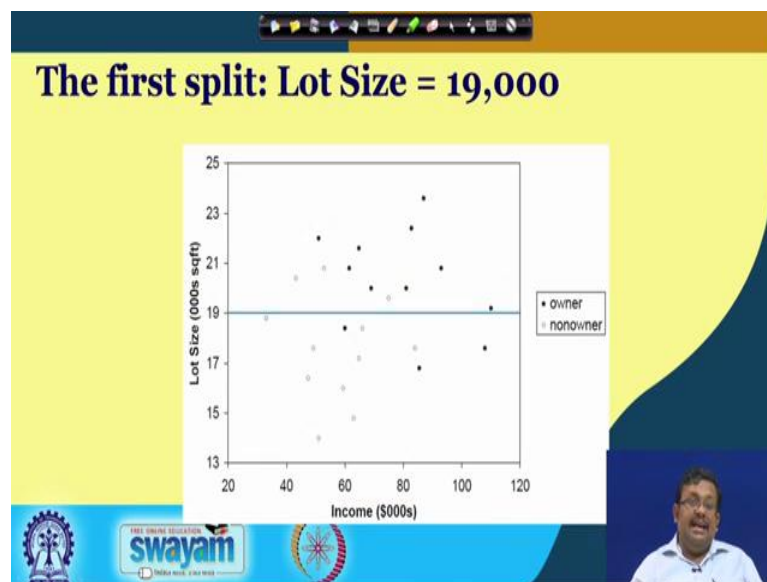
Note: Categorical Variables

- Examine all possible ways in which the categories can be split.
- E.g., categories A, B, C can be split 3 ways
 - {A} and {B, C}
 - {B} and {A, C}
 - {C} and {A, B}
- With many categories, # of splits becomes huge

The slide features a yellow background with a dark blue wavy shape on the right. At the bottom, there are logos for 'swayam' and 'INDIA RITE, A PRAJN' along with a small video of a presenter.

So, examine let us see examine all possible ways in which categories can be split in case of categorical variables let us see categories A B C can be split in 2 to 3 ways either A and B C then B and A C and C and A B.

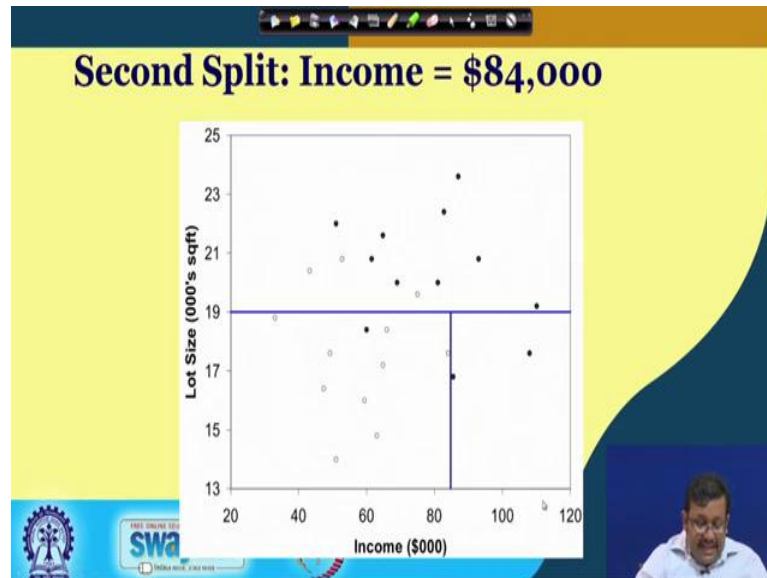
(Refer Slide Time: 12:49)



So, so with many categories the number of split becomes; obviously, huge. So, let us see in our case in case of in this more example; the first split in our case the lot size is 19000. So, our first splitting criteria is the lot size. So, you can see we are dividing the whole space into two groups this black dots are owner and the, you know, and the other dot,

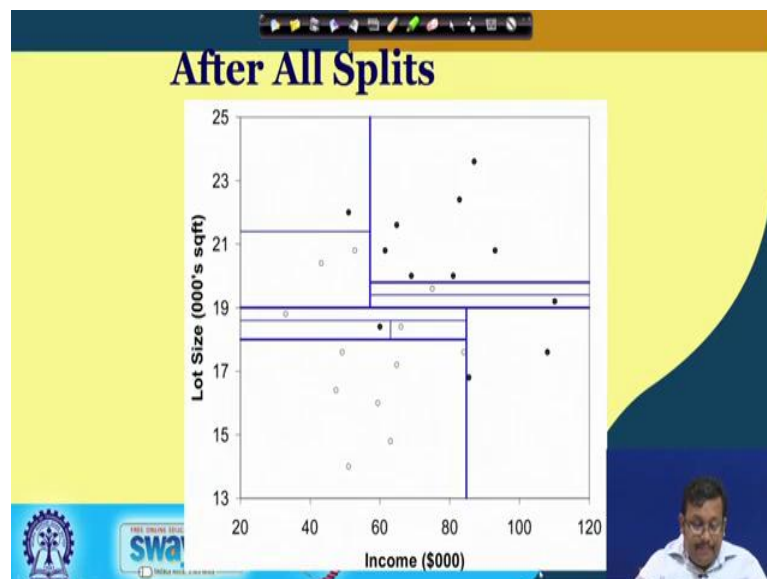
you know, the solid dots are owner and the holdouts are non owner. So, you can see the maximum homogeneity between these two split within this individual split.

(Refer Slide Time: 13:32)



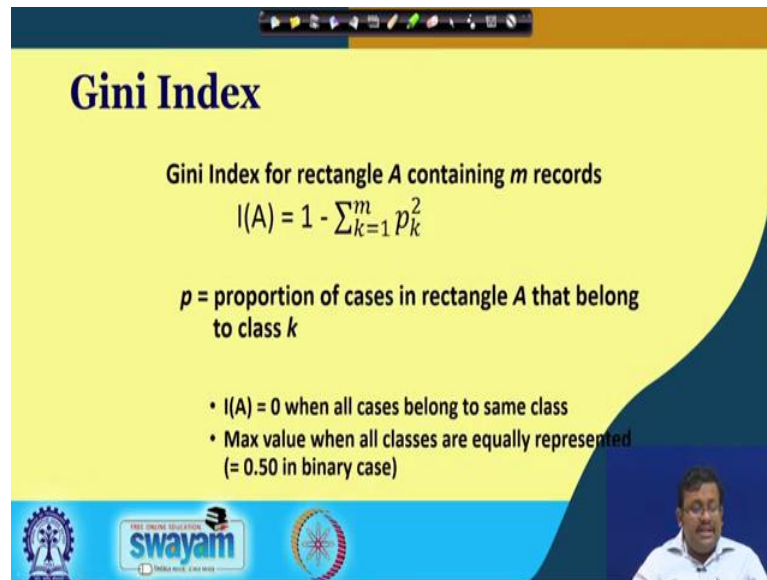
The next step we will split based on the income value of 84000 dollars. So, you can see there are three more split there are two splits again within this within this we there are two more splits.

(Refer Slide Time: 13:49)



And after all the split we will get this individual split; the final results. So, these within this individual blocks or individual areas the observations are homogeneous or maximum homogeneous.

(Refer Slide Time: 14:09)



Gini Index

Gini Index for rectangle A containing m records

$$I(A) = 1 - \sum_{k=1}^m p_k^2$$

p = proportion of cases in rectangle A that belong to class k

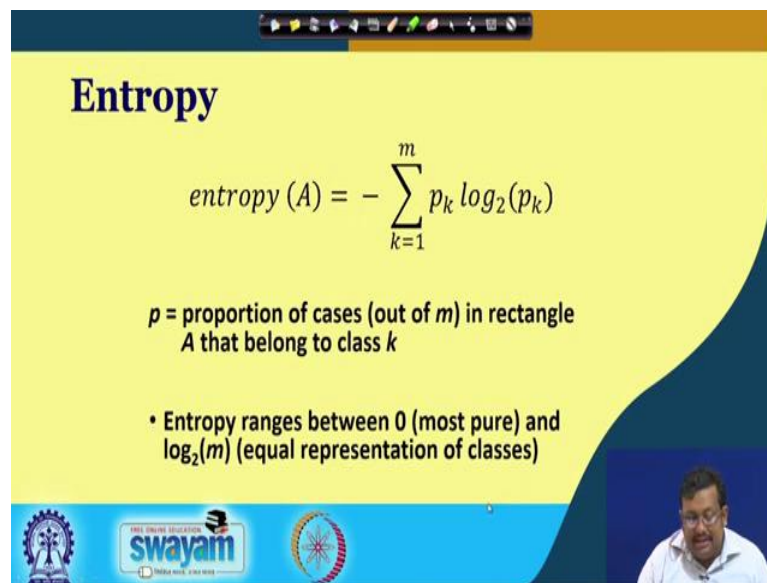
- $I(A) = 0$ when all cases belong to same class
- Max value when all classes are equally represented (= 0.50 in binary case)

swayam

So, this is how these rules are basically formed. Another important term is Gini index and Gini index for a rectangular A containing m records can be calculated by using this formula where p is the proportion of cases in rectangle A ; that belong to this class k and $I(A)$ is 0, when all classes belong to the same class and maximum values when all class are equally represented where 0.50 in case of binary case.

So, based on this Gini index criteria we select the best split criteria again this Gini index or rectangular A containing m records is calculated by this formula where p is the proportion of case in rectangular A that belongs to class A , where $I(A)$ equal to 0 when all cases belong to the same class and, you know, maximum values when all classes are equally represented.

(Refer Slide Time: 15:06)



Entropy

$$\text{entropy}(A) = - \sum_{k=1}^m p_k \log_2(p_k)$$

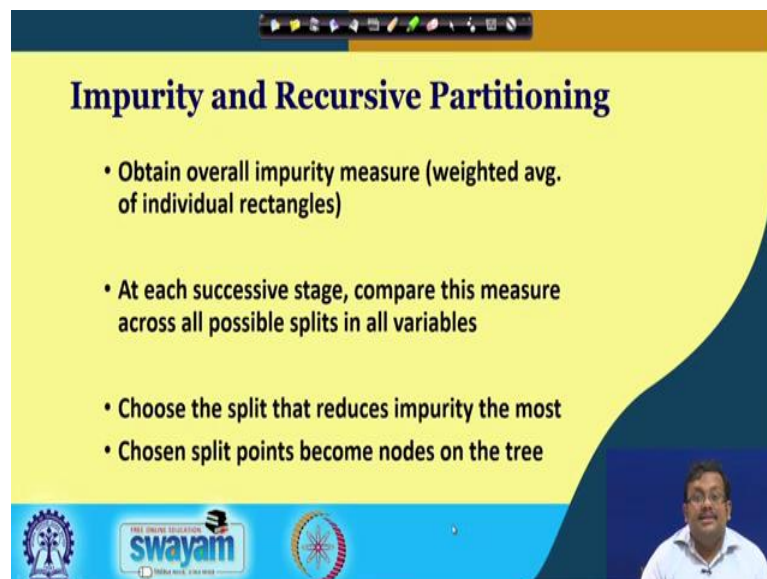
p = proportion of cases (out of m) in rectangle A that belong to class k

- Entropy ranges between 0 (most pure) and $\log_2(m)$ (equal representation of classes)

The slide features a yellow background with a blue wavy border on the right. At the bottom, there are logos for 'swayam' and other educational institutions, along with a small video inset of a man in a white shirt.

Another important term is entropy; an entropy can be calculated by using this formula where p is the proportion of cases out of m in rectangular A that belongs to class k and entropy ranges between 0 which is representing most pure and $\log_2 m$ where equal representation of all classes.

(Refer Slide Time: 15:26)



Impurity and Recursive Partitioning

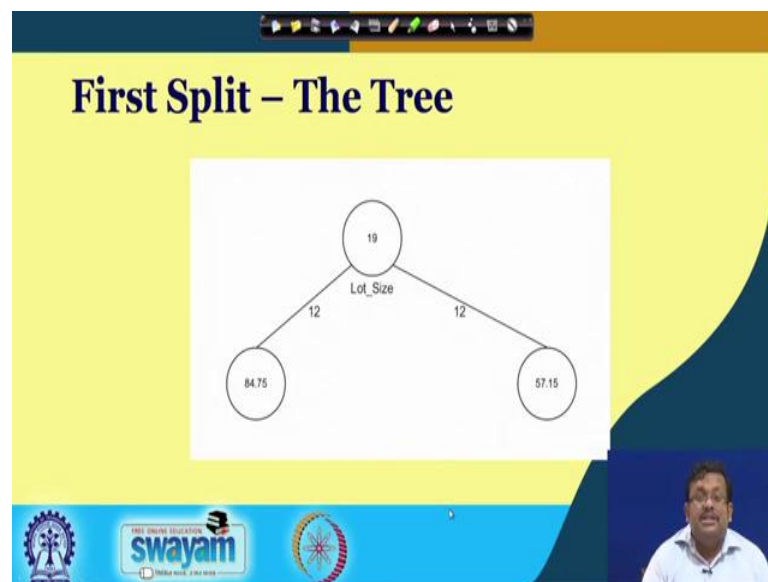
- Obtain overall impurity measure (weighted avg. of individual rectangles)
- At each successive stage, compare this measure across all possible splits in all variables
- Choose the split that reduces impurity the most
- Chosen split points become nodes on the tree

The slide features a yellow background with a blue wavy border on the right. At the bottom, there are logos for 'swayam' and other educational institutions, along with a small video inset of a man in a white shirt.

So, what is impurity and recursive partitioning? Obviously, we have to obtain overall impurity measures weighted average of individual rectangles. At each successive stage we have to compare this measure across all possible split in all variables. So, we told you

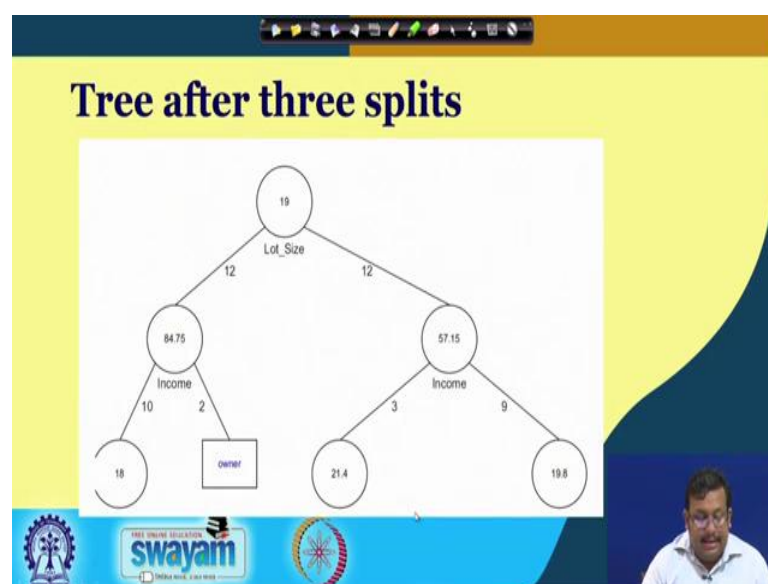
that for each step we are selecting the best combination of variable as well as its value like x_i and s_i . And this variable and their value combination is selected by measuring the overall impurity measures. So, we have to choose the split that reduces the impurity the most and choose the split point become nodes of the tree.

(Refer Slide Time: 16:12)



So, in the first split; obviously, this is the tree; obviously, we are selecting it based on the lot size and; we are selecting 12 12 observations.

(Refer Slide Time: 16:24)



And tree after the three splits you can see here looks like this.

(Refer Slide Time: 16:30)

Tree Structure

- Split points become nodes on tree (circles with split value in center)
- Rectangles represent “leaves” (terminal points, no further splits, classification value noted)
- Numbers on lines between nodes indicate # cases
- Read down tree to derive rule
E.g., If lot size < 19, and if income > 84.75, then class = “owner”

swamyam

And the tree structure; obviously, split point become nodes on the tree, circle which split values in the center, if you see this is the split point these are the circle these are the split individual split points from which further split is occurred. So, split points become nodes of the tree and from these nodes further split occurs and rectangle represents a leaves.

(Refer Slide Time: 16:54)

Determining Leaf Node Label

- Each leaf node label is determined by “voting” of the records within it, and by the cutoff value
- Records within each leaf node are from the training data
- Default cutoff=0.5 means that the leaf node’s label is the majority class.
- Cutoff = 0.75: requires majority of 75% or more “1” records in the leaf to label it a “1” node

swamyam

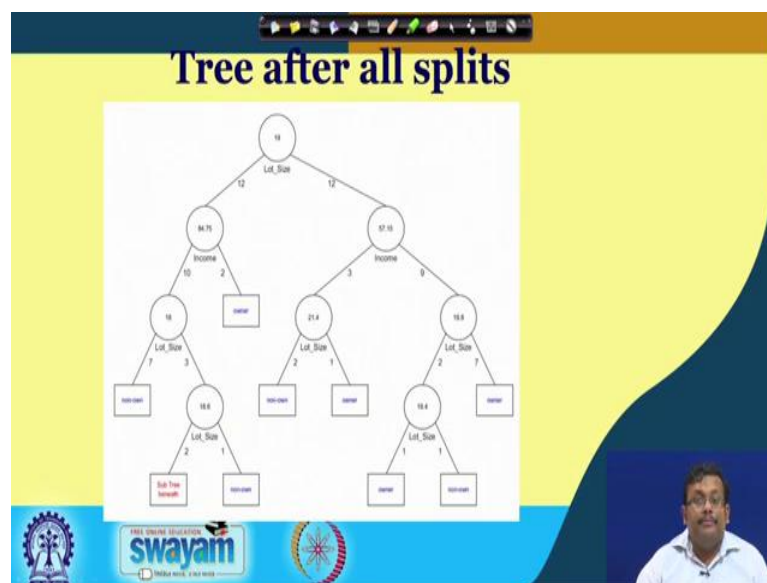
Here the rectangle leaves these; rectangle actually denotes the individual leaves and these leaves terminal points no further split classification value noted. So, basically here in this terminal this is called the terminal node or leaf and from here there will be no

further division. So, here it this branch stops; so, this is an example of how this tree grows.

And number of lines between nodes indicates the number of cases; obviously, you can see here they are the number of you know here the 3; that means, the number of cases and 9 here is the number of cases obviously. So, if the lot size read down the tree to derive rule if the lot size is less than 19 and if income is 84.75, then class is owner and so on so forth.

Determining the leaf node level; obviously, each leaf node level is determining determine by voting of the record within it and by the cut off values. And records with each leaf node are formed by training data and default cut off value is 0.5 means the leaf node level is the majority class and cutoff of 0.75 requires majority of 75 percent or more one records in the leaf to be label as one node.

(Refer Slide Time: 18:21)



So, after all the splits you can see these are the, these circles are the splitting nodes and in the splitting nodes you see we have selected a particular variable and the particular value of that variable which gives the lowest impurity. And then and ultimately the tree grows and these are the branches and ultimately you see this rectangular boxes are the terminal nodes from where there will be no further splitting. And ultimately you will see that there is ultimately we are specifying whether they are owning their owner or non owner which will our which is our target classes.

(Refer Slide Time: 19:09)



Stopping Tree Growth

- Natural end of process is 100% purity in each leaf
- This **overfits** the data, which end up fitting noise in the data
- Overfitting leads to low predictive accuracy of new data
- Past a certain point, the error rate for the validation data starts to increase

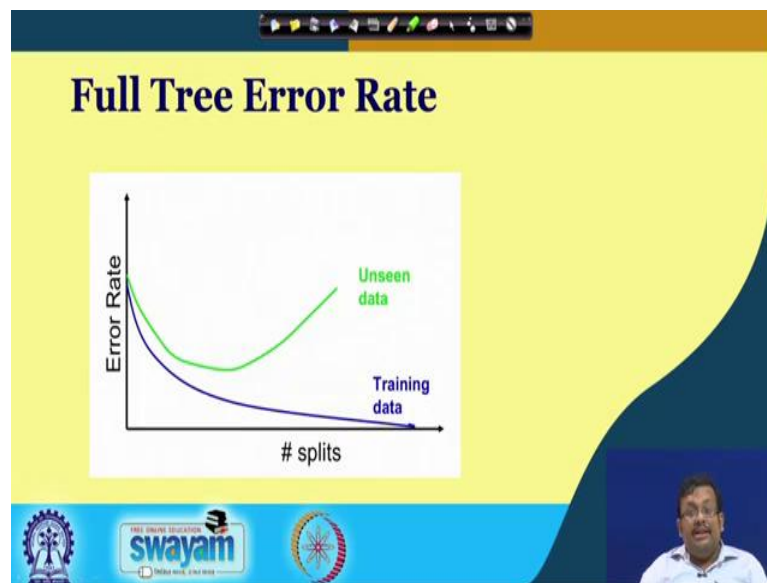
The slide features a yellow background with a dark blue curved shape on the right side. At the bottom, there is a blue banner with logos for 'swayam' and 'INDIA RISE, SKILL RISE'. A small video inset in the bottom right corner shows a man speaking.

So, what is the over fitting problem? Obviously, remember that if you continue go on and create the trees and create this branches and all these things, ultimately a time will come that you will perfectly fit all the observation because each an observation each of the observation we will have their own set of rules; so you can create any set of rules to fit any particular observation.

However, that would create over fitting. Now what is over fitting? If any calibration model performs poor in case of new set of data or independent data set that is called over fitting. For example, you have created very good calibration model and you try that calibration model to predict some unknown samples and that performs miserably.

So, that is called over fitting; so natural end process of these CART or classification regression trees is that you will get 100 percent purity of the leaf. Because you can create any rule to put any sample, you know, you can you can you can create any rule to define any sample and this over fits the data which ends up fitting noise in the data. So, over fitting leads into low predictive accuracy of the new data as I have told you; so past a certain points you see the error rate for validation data sets start to increase.

(Refer Slide Time: 20:31)



You can see here as the number of splits increases up to a certain point, you can see that if you increase the number of split a time will come that that training data error will go down and ultimately reaches 0; as you can see in this green in this blue line. However the problem is as although this training data set produces very lowest error as we continuously increasing the number of split.

However, you will see if we if we test this training or calibration model using the unseen data or independently drawn data, their error rate will increase after a certain split. So, this is basically saying that this model is over fitted. So, we need to do some kind of treatment to reduce this over fitting because our ultimately aim is to get a representative model which can predict reasonably the unknown or unseen data. How would you do that?

(Refer Slide Time: 21:37)

A presentation slide titled "CHAID" in a large, bold, dark blue font. The slide has a yellow background with a dark blue curved shape on the right side. The text on the slide is in black. At the bottom, there is a blue banner with logos for "swayam" and other educational institutions. A small video inset of a man is in the bottom right corner.

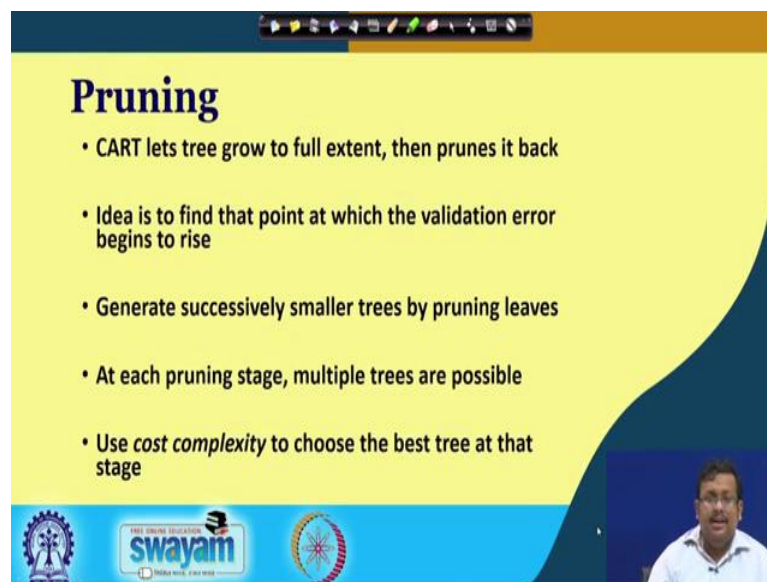
CHAID

CHAID, older than CART, uses chi-square statistical test to limit tree growth

Splitting stops when purity improvement is not statistically significant

Another one method is called chain method older than which is older than CART method which is using chi square statistical test to limit that tree growth and this splitting stops when purity improvement is not statistically significant.

(Refer Slide Time: 21:50)

A presentation slide titled "Pruning" in a large, bold, dark blue font. The slide has a yellow background with a dark blue curved shape on the right side. The text on the slide is in black. At the bottom, there is a blue banner with logos for "swayam" and other educational institutions. A small video inset of a man is in the bottom right corner.

Pruning

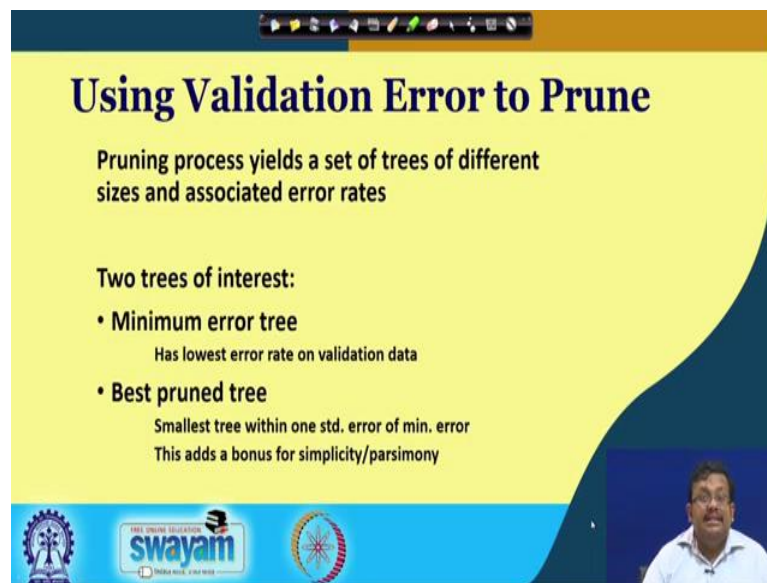
- CART lets tree grow to full extent, then prunes it back
- Idea is to find that point at which the validation error begins to rise
- Generate successively smaller trees by pruning leaves
- At each pruning stage, multiple trees are possible
- Use *cost complexity* to choose the best tree at that stage

The most important way for reducing these over fitting or, you know, to remove this over fitting is the pruning and CART. Let us first of all in this pruning method; the CART lets tree grow to full extent then prunes it back; that means, cutting the individual branches.

And when we are cutting the individual branches; obviously, we are reduce we are eliminating that particular rule.

So, idea is to find that point at which the validation error begins to rise and generally successively smaller trees by pruning leafs when we are, you know, pruning the leaves; obviously, it will get the trees will get smaller. And at each pruning stage multiple trees are possible and use the cost complexity to choose the best tree at that stage.

(Refer Slide Time: 22:38)



Using Validation Error to Prune

Pruning process yields a set of trees of different sizes and associated error rates

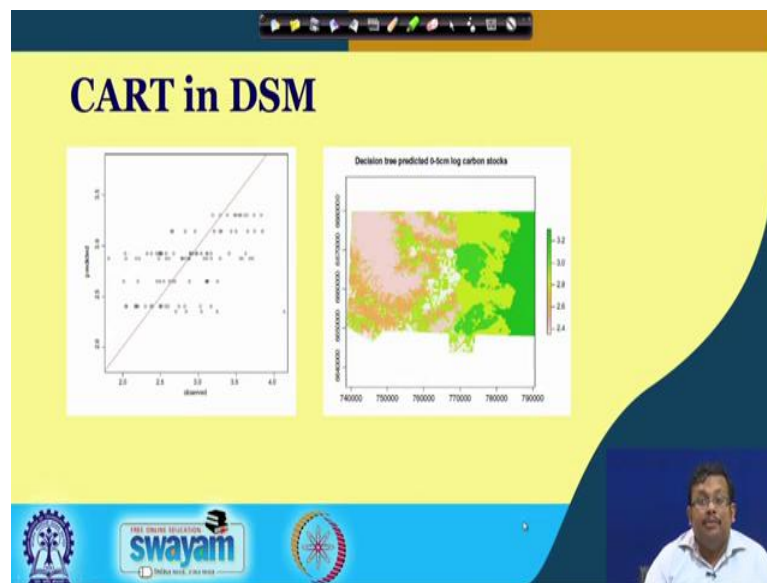
Two trees of interest:

- **Minimum error tree**
Has lowest error rate on validation data
- **Best pruned tree**
Smallest tree within one std. error of min. error
This adds a bonus for simplicity/parsimony

Logos: IIT Bombay, Swayam, and a circular logo with a tree.

Obviously, we have to use the pruning across yields a pruning process yields a set of trees of different sizes and associated error rates. And two rates of interest are there for us for while we are doing the pruning minimum error rate that is has the lowest error rate on validation data and best pruned tree; obviously, shows smallest tree with one standard error of minimum error and this adds to bonus for simplicity or parsimony.

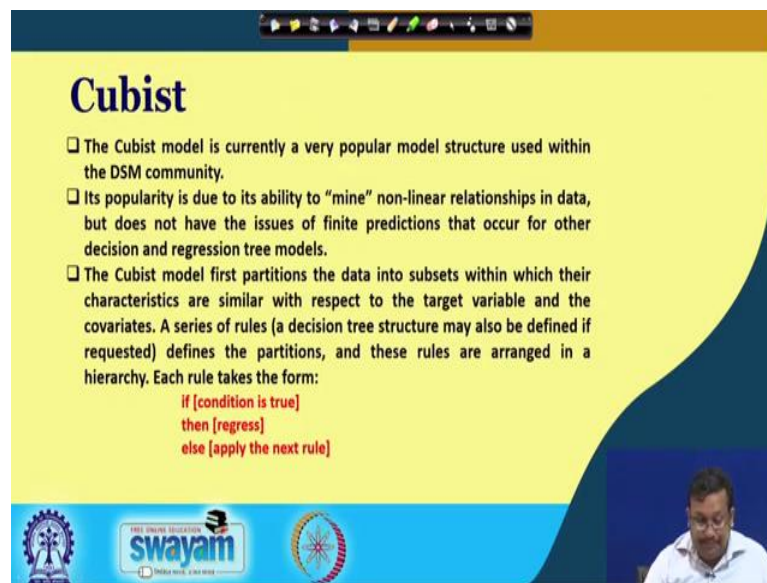
(Refer Slide Time: 23:09)



So, let us see what is the actual example of use of classification and regression tree in DSM? As you can see here we are predicting the soil organic carbon to 0 to 5 centimeter using this classification and regression tree. And this classification regression tree you can see here using that we are predicting; this is actually the observe verses predicted values. The problem in classification regression tree is that you can see some specified levels of predictions. So, this is a problem in case of classification regression tree again; this classification regression tree you will see some specified levels of predictions.

And this is a problem in classification regression tree which we generally address by another tree method, we will discuss them. And, but here is an example of use of classification regression tree or we call it finite, you know, finite predictions. So, this is an example of decision tree predicted 0.5 centimeter long log carbon stocks. So, again this classification regression tree is a very important method, non-linear method to generalize the; to generalize the to produce the non-linear prediction of soil properties.

(Refer Slide Time: 24:49)



Cubist

- ❑ The Cubist model is currently a very popular model structure used within the DSM community.
- ❑ Its popularity is due to its ability to “mine” non-linear relationships in data, but does not have the issues of finite predictions that occur for other decision and regression tree models.
- ❑ The Cubist model first partitions the data into subsets within which their characteristics are similar with respect to the target variable and the covariates. A series of rules (a decision tree structure may also be defined if requested) defines the partitions, and these rules are arranged in a hierarchy. Each rule takes the form:

```
if [condition is true]
then [regress]
else [apply the next rule]
```

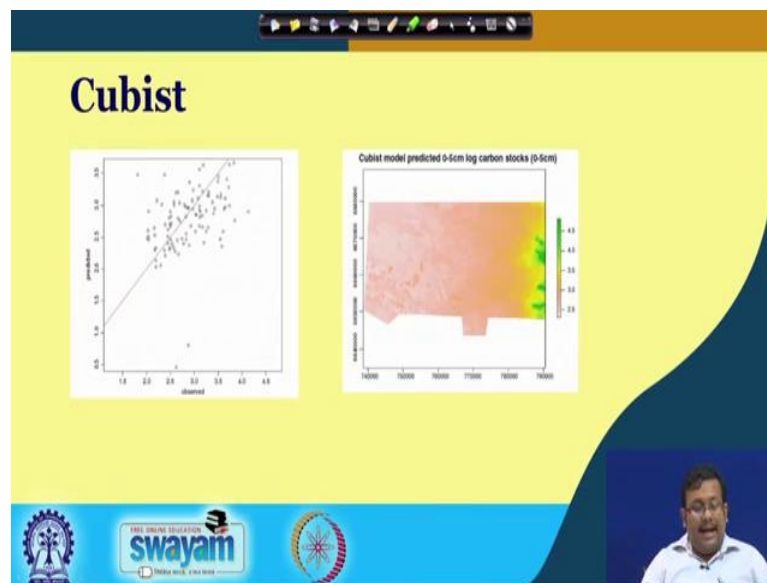
swayam
INDIA RITE, A NEW WAVE

The cubist model another model is cubist model; the cubist model is currently the very popular models structure, you know you know, model structure used within the DSM community and its popularity is due to its ability to mine non-linear relationship in the data.

But does not have the issue of finite prediction as we have seen in case of classification regression tree and the cubist model first partitions the data into subset within which their characteristics are similar with respect to the target variable and the covariates a series of rules. For example, a decision tree is a structure may also be defined if requested; defines this partition. And these rules are arranged in a hierarchy in each rule has the form either if with; it is basically shows if the condition is true; then will fit a regression model and if the condition is not true then else we will apply the next rule.

So, this is how it will fit, you know, the whole data and ultimately based on a condition is met or not it will create a regression; in this particular ultimate terminal node. So, this is how a cubist model works and cubist model is now a days considered as the most you know advance method as far as the digital soil mapping of soil is concerned.

(Refer Slide Time: 26:13)



And this is an example of cubist produce soil organic carbon map for a particular area of Australia. And you can see this is the cubist predicted 0 point 0 to 5 centimeter logarithm carbon stocks and this is the predicted verses observed values and you can see there is no finite predictions problem as we have seen in case of classification regression tree.

(Refer Slide Time: 26:40)

The slide titled "Random forests" contains three bullet points:

- ❑ Random Forests are a boosted decision tree model.
- ❑ Random Forests are an ensemble learning method for **classification (and regression)** that operate by constructing a multitude of decision trees at training time, which are later aggregated to give one single prediction for each observation in a data set.
- ❑ For regression the prediction is the average of the individual tree outputs, whereas in classification the trees vote by majority on the correct classification (mode).

The bottom of the slide features logos for "swayam" and "INDIA RITE, YOUNG RITE" along with a small video feed of a presenter.

Another important aspect another important method is random forest and random forest is a boosted regression tree method. Remember that random forest is an ensemble learning method for classification and regression that operates by constructing multitude

of decision trees and training time which are later aggregated to give one single prediction for each observation in a data set. And for regression the prediction is the average of the individual tree output whereas, in case of classification tree votes by majority of the correct classification mode.

So, what happens? We have known: what is a tree in case of classification regression tree. So, in case of random forest we create thousands and thousands of individual tree by randomly selecting the variables as well as randomly selecting the samples. And once we create these thousands and thousands of trees based on their prediction results and; once we create these thousands and thousands of trees we call it random forest. And what is the output value? For regression the prediction is an average of individual tree outputs.

So, we are getting average we are getting the individual tree output and once we are taking the average of this individual tree output that will be the overall output or average output of this random forest. And in case of classification, we take the majority of the votes from these individual trees. So, that is how a random forest basically operates; a random forest is more advanced method, a random forest is a widely used method in case of digital soil mapping.

(Refer Slide Time: 28:24)



Random forests

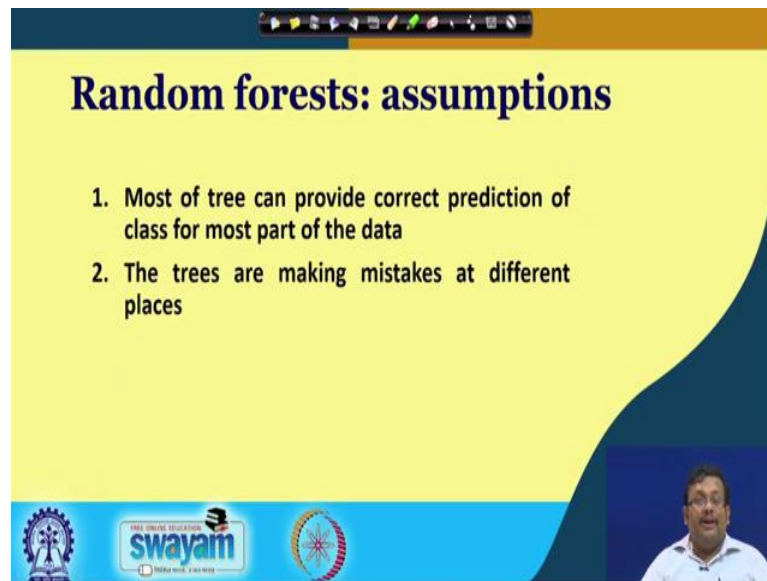
- ❑ Forms lots of decision trees with random selection of samples and random selection of features.
- ❑ Provides the class of dependent variable based on many trees
- ❑ Random trees
- ❑ Many random trees=Random forest

swayam

So, random forest forms lots of decision trees with random selection of samples random selection of features that is why it is called random forest.

First of all it is called forest because we are using thousands and thousands of trees. And it is called random because we are using random selection of samples and random selection of features for creating one particular tree. So, that is why it is called random forest and it provides the class of dependent variable based on many trees that is why it is called random trees; many random trees random forest.

(Refer Slide Time: 28:59)



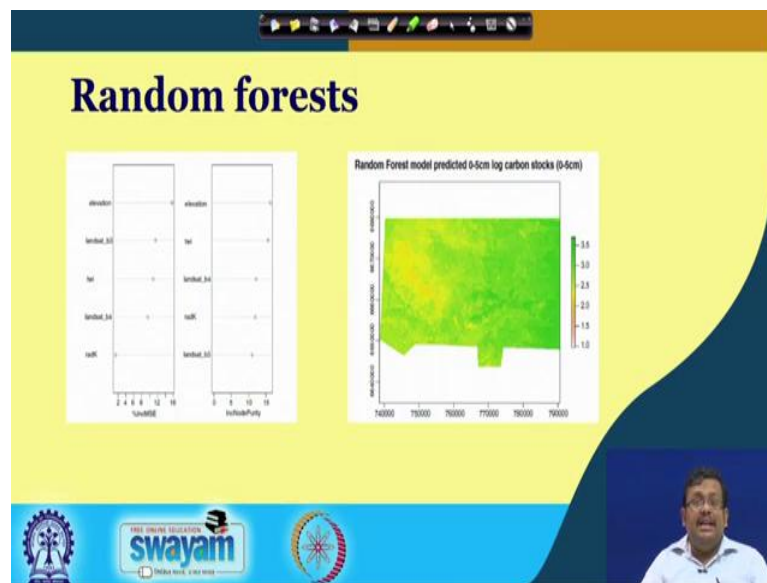
Random forests: assumptions

1. Most of tree can provide correct prediction of class for most part of the data
2. The trees are making mistakes at different places

The slide features a yellow background with a dark blue curved shape on the right side. At the bottom, there is a blue banner with logos for 'swayam' and 'MOE, Government of India'. A small video inset in the bottom right corner shows a man in a white shirt speaking.

The random forest basically assumes that most of the trees can provide correct prediction or class for most part of the data and the trees are making mistake at different places. So, based on these observation, it calculates the final output; in case of regression it calculates the average of the output and in case of classification it calculates the majority of the vote.

(Refer Slide Time: 29:24)



So, here you can see the example of random forest and random forest; obviously, you can see these random forest predicted 0 to 5 centimeter logarithmic carbon stocks. And another beauty of random forest is based on certain criteria you can get the ranking of the important variables. For example, here in this random forest example we use elevation, land sat B 3 band, TWI that is Topographic Wetness Index, land sat B 4 band and radiometric potassium all this five covariates as predictors.

And you can see here the; they are arranged based on their certain properties like increase in mean squared error and increase in node purity. So, these are, you know, arrange based on their importance. So, that is why random forest has the capability of arranging the variable based on the relative importance also. So, this is an example of random forest application for predicting, you know, this log carbons.

(Refer Slide Time: 30:36)

Regression Kriging

- The Best Linear Unbiased Predictor of spatial data
- Matheron (1969) proposed that a value of a target variable at some location can be modelled as a sum of the deterministic and stochastic components:

$$Z(s) = m(s) + \varepsilon'(s) + \varepsilon''$$

We know that both deterministic and stochastic components of spatial variation can be modelled separately. By combining the two approaches, we obtain:

$$\hat{Z}(s_0) = \hat{m}(s_0) + \hat{\varepsilon}(s_0)$$
$$= \sum_{k=0}^p \hat{\beta}_k \cdot q_k(s_0) + \sum_{i=1}^n \lambda_i \cdot e(s_i)$$

where

- $\hat{m}(s_0)$ = fitted deterministic part
- $\hat{\varepsilon}(s_0)$ = interpolated residual
- $\hat{\beta}_k$ = estimated deterministic model coefficients
- λ_i = kriging weights determined by the spatial dependence structure of the residual and where $e(s_i)$ is the residual at location s_i

The slide also features a Swamy logo and a small video inset of a speaker in the bottom right corner.

Another important, you know, hybrid method is called regression kriging. Now regression kriging is basically combination of both regression and kriging and, you know, kriging is basically the best linear unbiased predictors of spatial data.

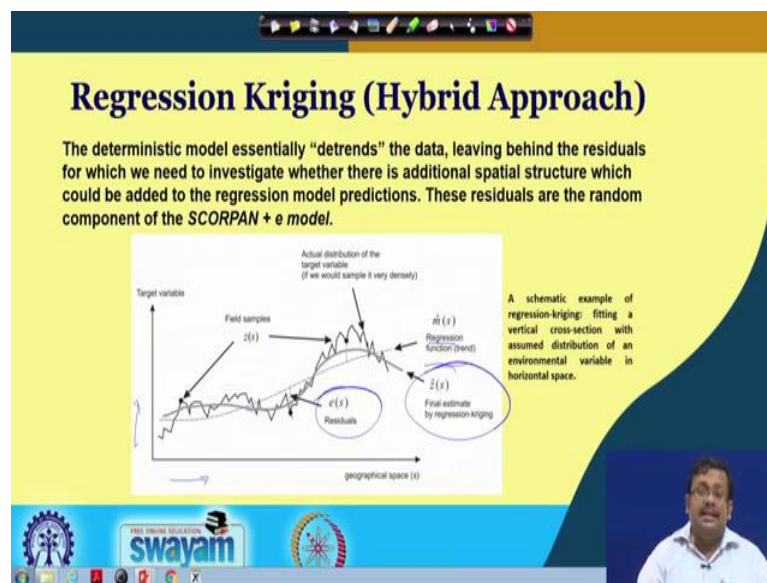
We know from our geo statistical discussion that a value of a target variable at some location can be model as a sum of the deterministic and stochastic components. So, it is a value of a particular variable at some location can be divided into, you know, it is a deterministic part and it is a stochastic part and; obviously, there are some errors. So, we know that both deterministic and stochastic components of spatial variation can be modeled separately. So, you can see; so by combining these two approaches we obtain a final output where ultimately this is the output of this stochastic; you know, deterministic component and this is an output of stochastic component.

So, this is basically the deterministic component and this is the stochastic component. So, in the deterministic component you can see this is basically linear regression equation, where these are basically the, you know, individual coefficient and here this is basically you can see it is a kriging; you know, kriging output where $m(s_0)$ is a fitted deterministic part and it is a interpolated residual this is an interpolated residual and this $\hat{\beta}_k$ is estimated deterministic model coefficient.

So, if we are using the linear regression model, these are the model coefficient values and λ_i is basically kriging weights determined by the spatial dependence structures

of the residual where $e(s_i)$ is the residual at location $z(s_i)$. So, basically you can see here we are using two components first of all we are trying to produce a model output based on the deterministic model; the deterministic model may be linear or non-linear. And using the residual we are interpolating there and we after calculating the residual from this model; we are basically interpolating them. So, this regression kriging is basically the combination of deterministic model output and the interpolated residuals. So, this is why it is called a hybrid method.

(Refer Slide Time: 33:22)



So, how this works? So, you can see this regression kriging deterministic the deterministic model essentially detrends the data living behind the residuals for which we need to investigate whether there is an additional spatial structure; which could be added to the regression model prediction. These residuals are the random components of the SCORPAN plus e model. As you can see if these are the field samples these are the field samples, so these are basically the distribution of actually distribution of the target variable; if we it sample it very densely. So, once we collect this samples these are the field samples.

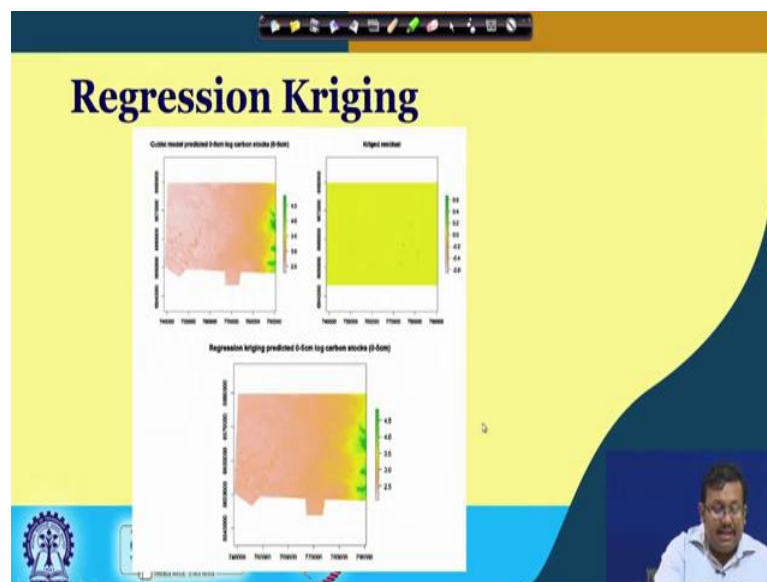
Our next step in the deterministic model you know this in the x axis we it is a geographic space, in the y axis this is the target variable. So, basically we are detrending the data by fix first fixing a regression function or deterministic part. So, this dotted line is basically the regression function and by using this regression function we are basically detrending

the data. So, once we are detrending the data this is the residual this difference between the actual point and the point on the regression function or regression line. So, these residual the next step we will basically kriging the residuals and ultimately you see these line is basically the final estimate by regression kriging.

So, it is basically the combination of kriging residuals as well as the output from deterministic regression function. So, it basically shows the schematic example of regression kriging fitting a vertical cross section with assumed distribution of an environmental variable in a horizontal space. So, I hope now this regression kriging is clear to you.

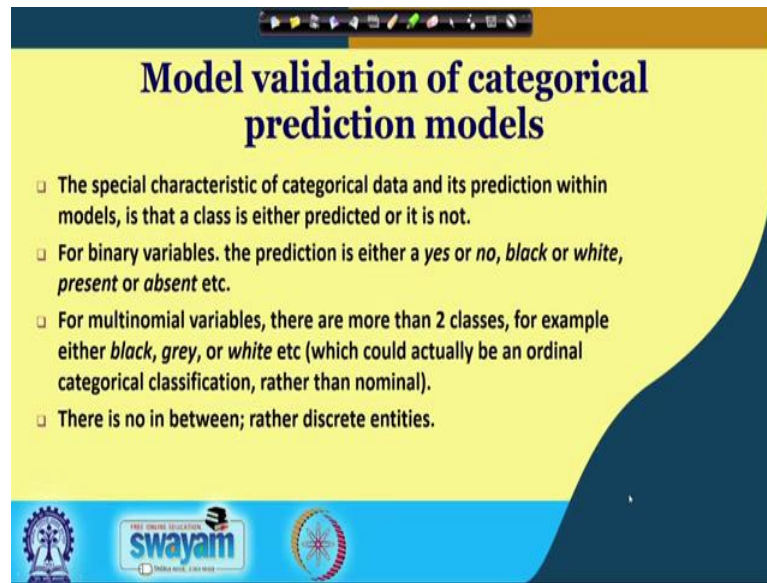
So, that is why now a days we not only rely on kriging we also rely on different types of advanced models; either it is a linear model either it is a non-linear model. So, we basically combine the output together both from kriging and deterministic model to give the most unbiased estimator of the spatial variability and that is called the regression kriging.

(Refer Slide Time: 35:41)



And these are the, you know, example of some regression kriging method; you can see the first one is cubist predicted 0.5 centimeter logarithmic carbon stock of 0 to 5 centimeter. And then this as with respective kriged residuals and finally, this is a regression kriging predictors 0.5 centimeter 0 to 5 centimeter log carbon stocks by combining these two output.

(Refer Slide Time: 36:17)



Model validation of categorical prediction models

- ❑ The special characteristic of categorical data and its prediction within models, is that a class is either predicted or it is not.
- ❑ For binary variables, the prediction is either a *yes* or *no*, *black* or *white*, *present* or *absent* etc.
- ❑ For multinomial variables, there are more than 2 classes, for example either *black*, *grey*, or *white* etc (which could actually be an ordinal categorical classification, rather than nominal).
- ❑ There is no in between; rather discrete entities.

swayam

So, guys let us start another important thing that is modeling and mapping of categorical variables. Remember that we have we have covered the continuous variables and now another important term is categorical variables and sometime it is very very important to model the categorical variable in case of digital soil mapping also.

And remember that the spatial characteristics of categorical data and its prediction within the model is a class is either predicted or it is not. And for binary variables the prediction is either a *yes* or *no*, you know, *black* or *white* *present* or *absent* etcetera. For multinomial variables there are more than two cases for example, either *black*, *grey* or *white* etcetera which could actually be in ordinal category ordinal categorical classification rather than nominal and there is no between rather discrete entities.

(Refer Slide Time: 37:07)



Important quality measures

1. Overall accuracy
2. User's accuracy
3. Producer's accuracy
4. Kappa coefficient of agreement

swayam

THINKING WITH A DIFFERENT MIND

And important quality measures are overall accuracy, user's accuracy, producer's accuracy and kappa coefficient. Let us wrap up here and in the next lecture we will discuss these different accuracy measures and different types of categorical models which we use for digital soil mapping.

Thank you.