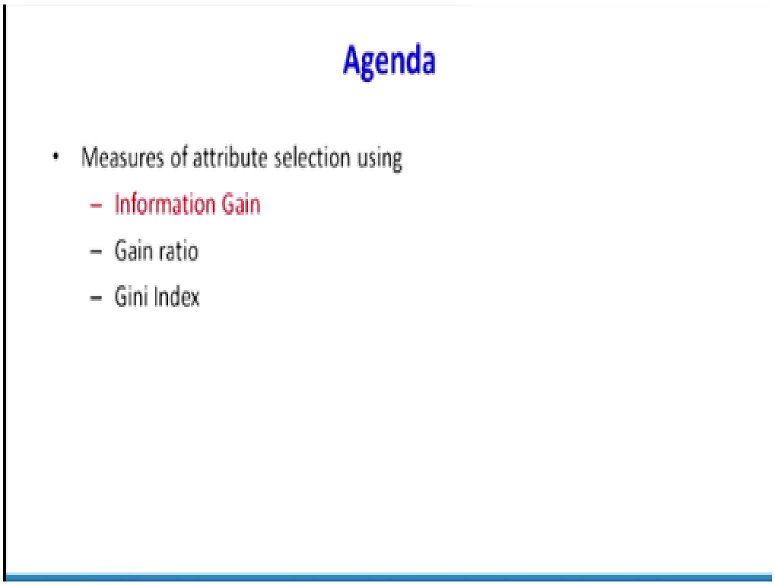
Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology - Roorkee

Lecture – 58
Measures of Attribute Selection

In our previous class, I have given introduction to classification regression tree, in this lecture I am going to take some example, some numerical examples; with the help of numerical examples, I am going to explain how to select attribute for the CART model.

(Refer Slide Time: 00:47)

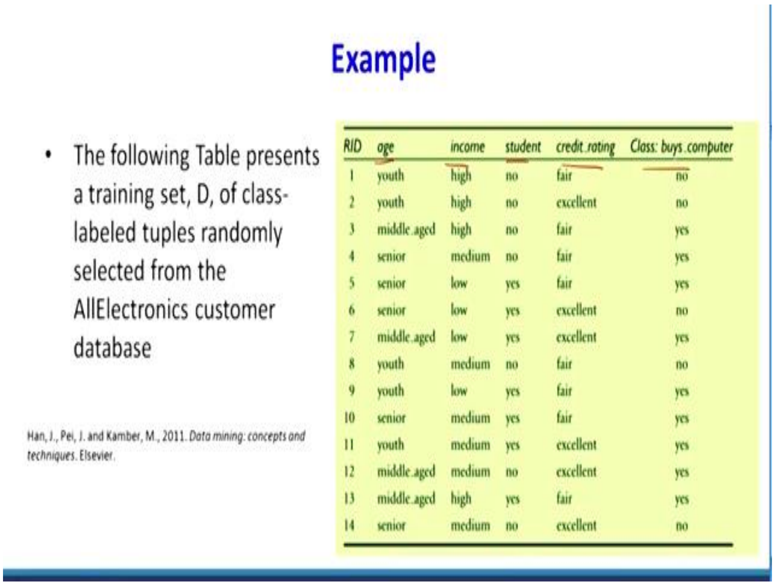


Agenda

- Measures of attribute selection using
  - Information Gain
  - Gain ratio
  - Gini Index

The agenda for this lecture is measures of attribute selection using; there are 3 measures for selecting attributes. Here, what is the meaning of attribute is choosing an independent variables for making classification, so there are 3 criteria; one is we can choose attribute with the help of information gain, another measure is gain ratio, the third one is Gini index. In this lecture, I am going to take the first criteria that is information gain.

(Refer Slide Time: 01:24)



Example

- The following Table presents a training set, D, of class-labeled tuples randomly selected from the AllElectronics customer database

Han, J., Pei, J. and Kamber, M., 2011. *Data mining: concepts and techniques*. Elsevier.

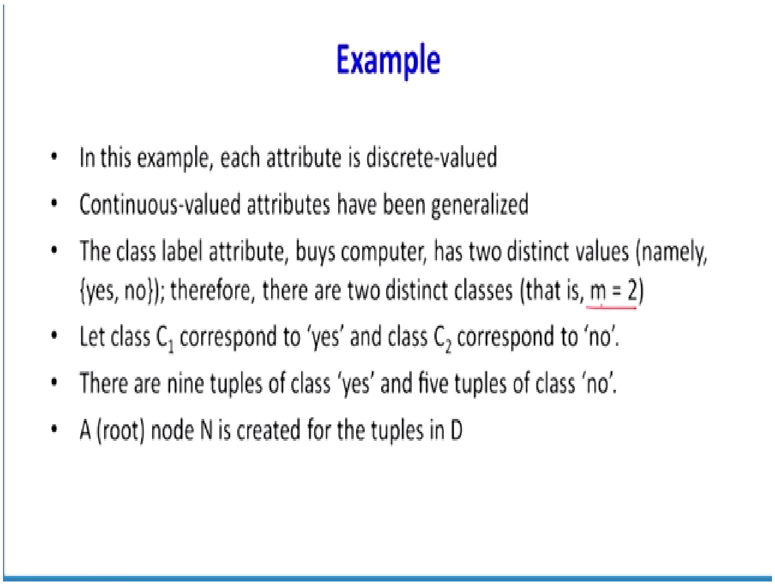
RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

With the help of this attribute, I am going to explain how to choose the attribute, now, taken one sample example, this example is taken from this book, Han and Kamber, the book title is Data mining, concepts and techniques. The problem is there are 1, 2, 3, 4, 5, there are 5 column is there, these 5 columns is called attributes, the last column is class that is a buys computer.

So, what this database says that the company; the database called all electronics customer database, they are going to see what kind of customers, they are going to buy the computer, they have 1 attributes or variable called age in that they have in different levels; youth, middle aged, senior. In income there are 3 levels; high, medium, low. In student; yes or no, in credit rating whether their credit rating is fair or excellent.

So, the final objective is we have to make a decision tree or classification tree for this dependent variable that is buy say computer, for choosing that one out of these 4 variables, we want to know from which variable we have to started. For that purpose, the information gain criteria is taken as a measure, let us see how it is working. The following table represents a training set, the data set called D of class labelled tuples randomly selected from all electronics customer database.

(Refer Slide Time: 03:13)



Example

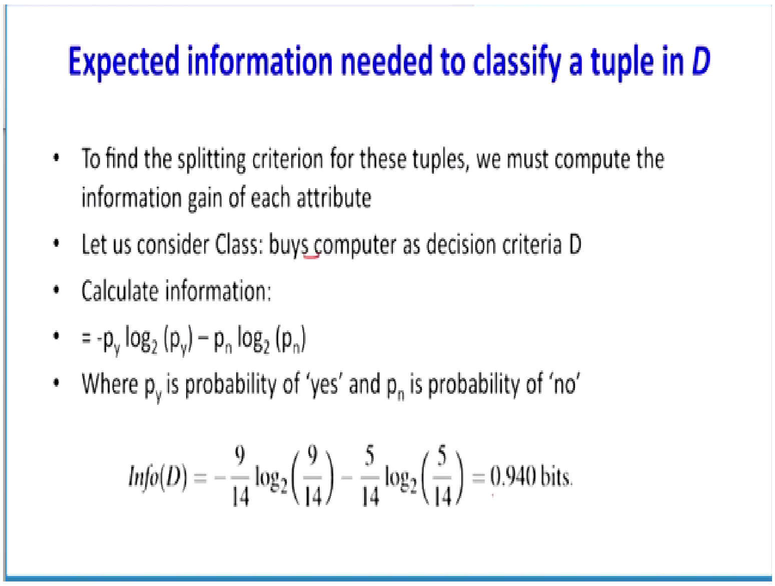
- In this example, each attribute is discrete-valued
- Continuous-valued attributes have been generalized
- The class label attribute, buys computer, has two distinct values (namely, {yes, no}); therefore, there are two distinct classes (that is, $m = 2$)
- Let class C_1 correspond to 'yes' and class C_2 correspond to 'no'.
- There are nine tuples of class 'yes' and five tuples of class 'no'.
- A (root) node N is created for the tuples in D

The tuples is nothing but the full database called tuples, in this example each attribute is discrete valued, what are the attributes here; age is an attribute, it is a discrete because there is no continuous value, income is an another attribute, student is another attribute, credit rating is another attribute and buy say computer also an attribute, all are categorical variable there is no continuous variable here.

The class labels attributes that is in the last column the variable called buys computer has 2 distinct value namely yes, no therefore, there are 2 distinct classes, this m equal to 2, so this value m equal to 2 I am going to use in coming slides, please remember this m equal to; because there are 2 levels the person is going to buy the computer or not, let the class C1 corresponds to yes, class C2 corresponds to no.

There are 9 tuples of class yes and 5 tuples of class no, a root node N is created for the tuples in D, so for making this root node we have to find out which variable, from which variable you have to start this root node, 1, 2, 3, 4 out of this 4 variables; age, income, student and credit rating, we are going to find out which variable is going to be in the root node.

(Refer Slide Time: 04:36)



Expected information needed to classify a tuple in *D*

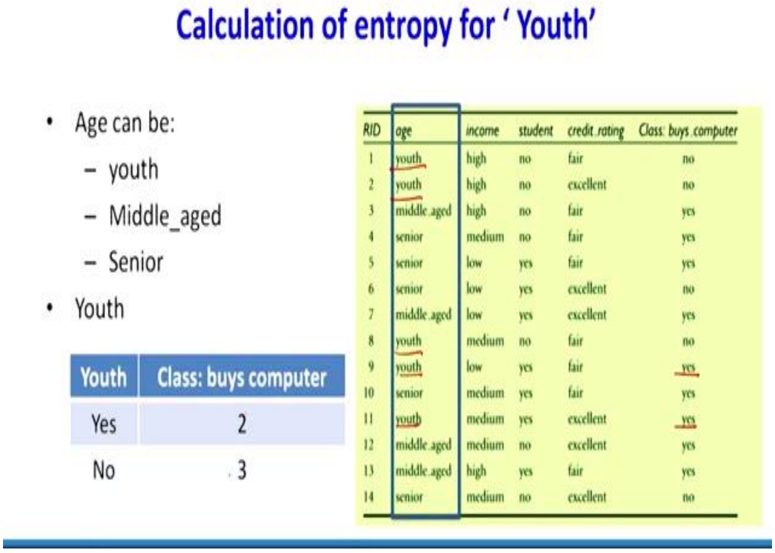
- To find the splitting criterion for these tuples, we must compute the information gain of each attribute
- Let us consider Class: buys computer as decision criteria D
- Calculate information:
- $= -p_y \log_2 (p_y) - p_n \log_2 (p_n)$
- Where p_y is probability of 'yes' and p_n is probability of 'no'

$$Info(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940 \text{ bits.}$$

Here what we are going to do; expected information needed to classify a tuples D, to find the splitting criterion for these tuples, we must compute the information gain of each attributes. Let us consider the class, buys computer actually it is variable, buys underscore as a decision criteria D, so calculate informations that is - py log py to the base 2 - pn log pn to the base 2, where the py is a probability of yes and pn is probability of no.

So, the another name for this equation is entropy, so Info D py, let us see what is a py; in the last column when you look at this how many yes is there; 1, 2, 3, 4, 5, 6, 7, 8, 9 yes is there, so 9 yes is there out of 14, - 9 by 14 log again, py 9 by 14; 9 divided by 14 to the base 2 minus; then how many no's is there because out of 14, 9 is yes, so remaining 5 is no, 5 divided by 14 log 2, 5 divided by 14.

(Refer Slide Time: 06:10)



Calculation of entropy for ' Youth'

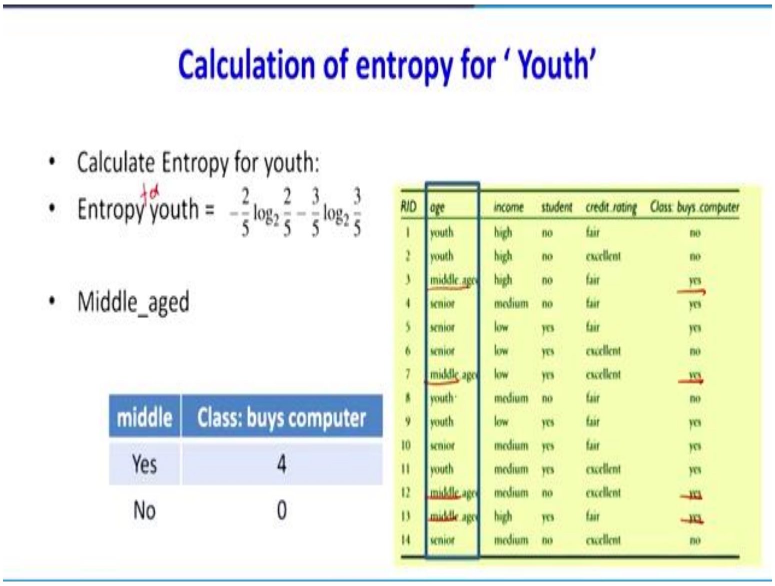
- Age can be:
  - youth
  - Middle_aged
  - Senior
- Youth

Youth	Class: buys computer
Yes	2
No	3

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

So, this is 0.940 bits for our dependent variable, now let us calculate entropy for the variable age, so age can be; see there are 3 levels is there in age, one is youth, second one is middle-aged, third one is senior. If you take youth how many people has given yes is their option, so here youth there is 1, youth because here also yes, yes there are 2 people, so youth how many people are yes; 2. So, how many people say no; this one, 1, 2, 3, so out of 5 youth, 2 youth they told their yes; yes means yes for buying computer, 3 youth told no for buying computer.

(Refer Slide Time: 07:08)



Calculation of entropy for ' Youth'

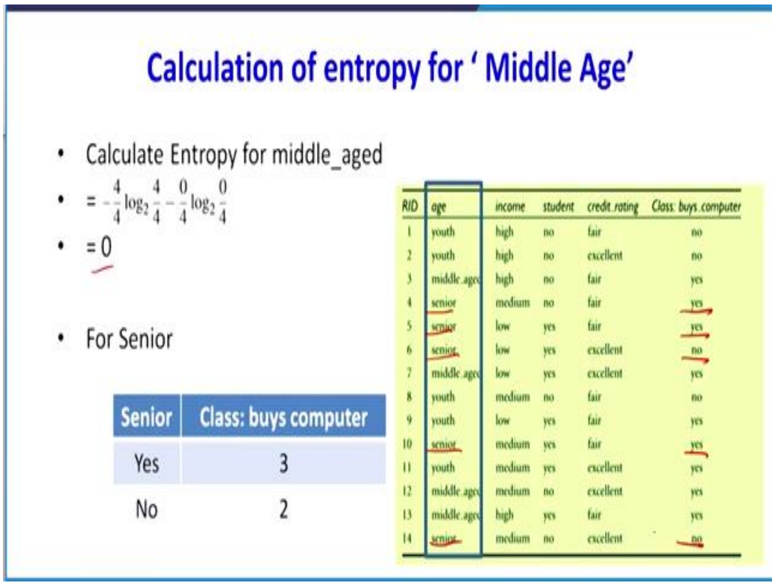
- Calculate Entropy for youth:
- Entropy fd youth = $-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}$
- Middle_aged

middle	Class: buys computer
Yes	4
No	0

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Let us calculate entropy for youth, so out of 5 entropy for youth is, this entropy for youth equal to - 2 divided by 5 log 2 divided by 5 to the base 2 - 3 people have told no, so - 3 by 5 log 3 by 5 to the base 2, so this is entropy for youth. Then we will go to the next level, the next level is middle aged; in this middle aged, how many people told yes? 1, this 2, middle aged yes, middle aged yes, so there are 4, 4 people told yes, so no is 0.

(Refer Slide Time: 08:07)



Calculation of entropy for ' Middle Age'

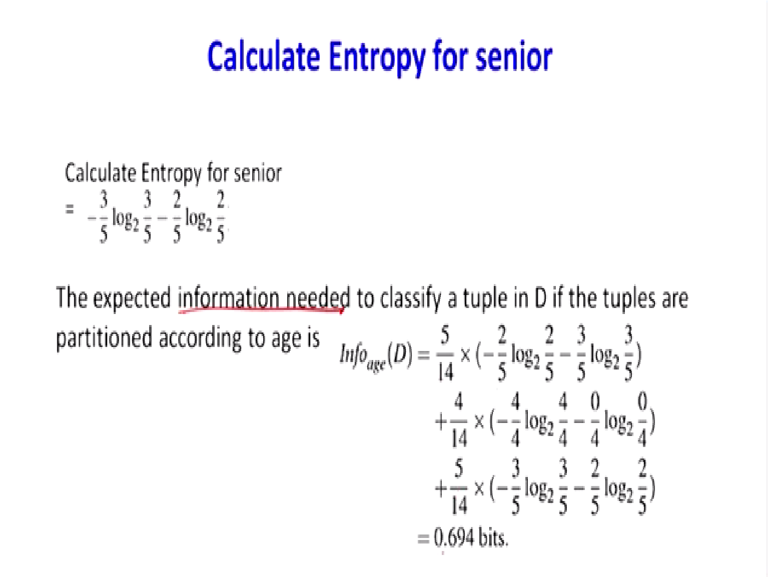
- Calculate Entropy for middle_aged
- $= -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}$
- $= 0$
- For Senior

Senior	Class: buys computer
Yes	3
No	2

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Because only there are 4; 1, 2, 3, 4 middle aged people, so calculate entropy for a middle aged here, entropy - 4 divided by 4 log 4 by 4 to the base 2 - 0 by 4 log 2 0 divided by 4, so here entropy is 0. Similarly, the next level is senior, when it is senior how many people told yes, so this senior is yes, this senior is yes, this senior is yes, so there are 3 people told yes, so how many senior level they told no; this they told no here and this senior also told no, there are 2 people, so out of 5, 3 people told yes for buying computer, 2 people have told no for buying computer.

(Refer Slide Time: 08:59)



Calculate Entropy for senior

Calculate Entropy for senior

$= -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

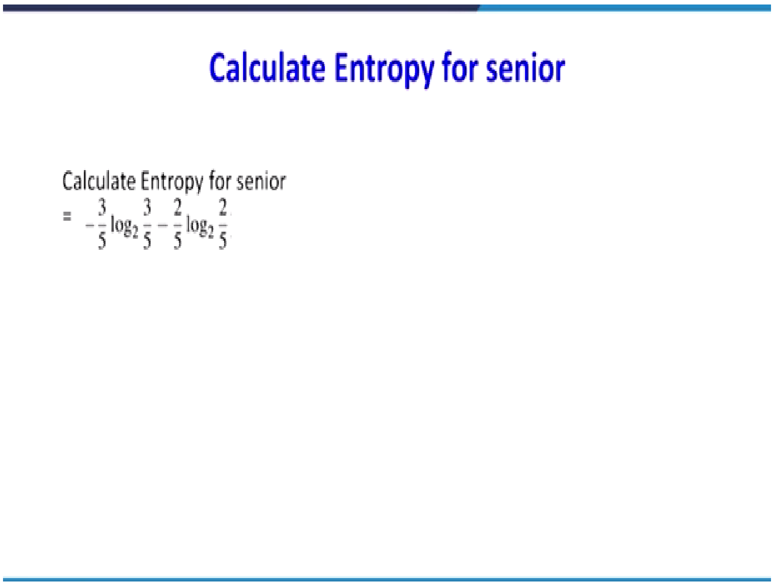
The expected information needed to classify a tuple in D if the tuples are partitioned according to age is

$$Info_{age}(D) = \frac{5}{14} \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right) + \frac{5}{14} \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) = 0.694 \text{ bits.}$$

Let us find out entropy for senior, so - 3 by 5 log 2 log 3 divided by 5 to the base 2 - 2 by 5, remaining 2 people told no - 2 divided by 5 log 2 the base 5 sorry, log 2 divided by 5 to the base 2. So, now what you are going to see, the expected information needed to classify a tuple in D, if the tuples are partitioned according to age is; so what we are going to do, we are going to find out the expected information needed.

There are 5 element right, there are 5, how we got this 5; for example, youth is 5, so it is 5 by 14, then middle aged 4 divided by 14, then senior 5 out of 14, so now we are finding the expected information needed that is 0.694 bits.

(Refer Slide Time: 10:05)



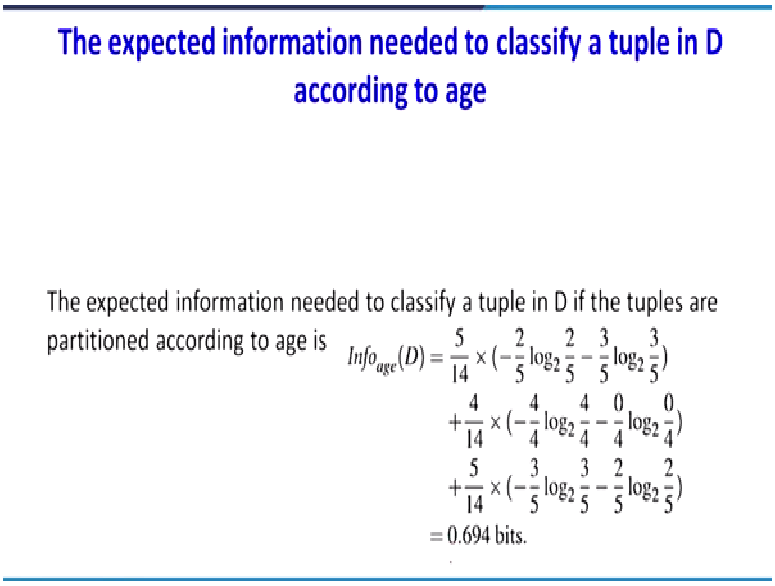
Calculate Entropy for senior

Calculate Entropy for senior

$$= -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

Calculation of entropy for senior category; so we have seen in our previous slide out of 5, there are 5 senior score is there, out of 5, 3 people have answered yes, 2 people have answered no, so the entropy is – 3 divided by 5 log 3 divided by 5 to the base 2 - 2 divided by 5 log 2 by 5 to the base 2. So, now we have got the entropy for all the levels.

(Refer Slide Time: 10:36)



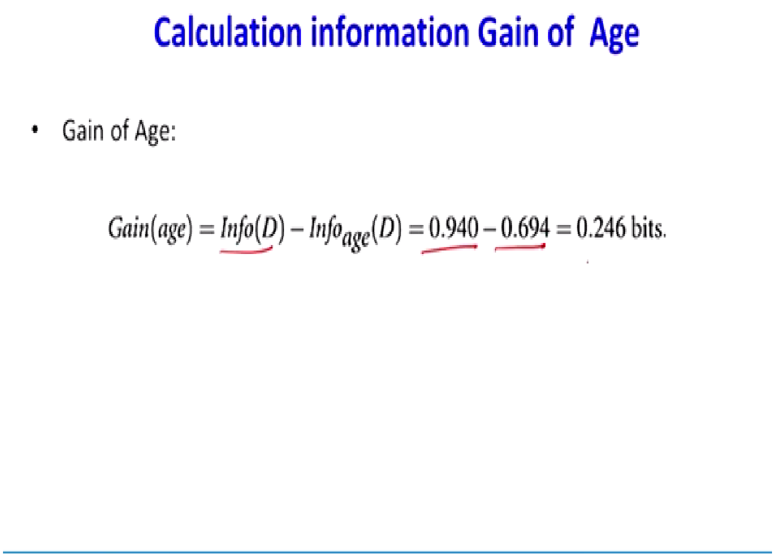
The expected information needed to classify a tuple in D according to age

The expected information needed to classify a tuple in D if the tuples are partitioned according to age is

$$Info_{age}(D) = \frac{5}{14} \times (-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5})$$
$$+\frac{4}{14} \times (-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4})$$
$$+\frac{5}{14} \times (-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5})$$
$$= 0.694 \text{ bits.}$$

Now, let us find the expected information needed to classify a tuple in D, according to age so, the expected information needed to classify a tuple in D, if the tuples are partitioned according to age is, so here nothing but we are finding weighted, so because there was a 5 youth out of 14, there are 4 middle-aged people out of 14, there are 5 senior out of 14, so 5 divided by 14, then corresponding entropy, 4 divided by 14 corresponding entropy, 5 divided by 14 corresponding entropy, so it is 0.694 bits,.

(Refer Slide Time: 11:24)



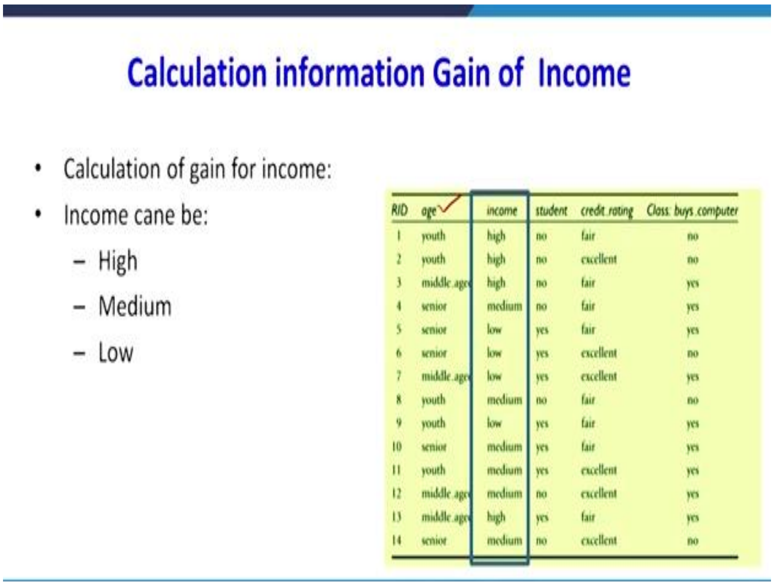
Calculation information Gain of Age

- Gain of Age:

$$Gain(age) = \underline{Info(D)} - Info_{age}(D) = \underline{0.940} - \underline{0.694} = 0.246 \text{ bits.}$$

Now, calculation of information gain of age, so the gain of age is; see that gain of age Info D – Info D for only 1 variable age, so this Info D 0.940 which we have found for the class attributes that is the our dependent variable, this is only for the info variable age. So, the difference is 0.246 bits.

(Refer Slide Time: 11:58)



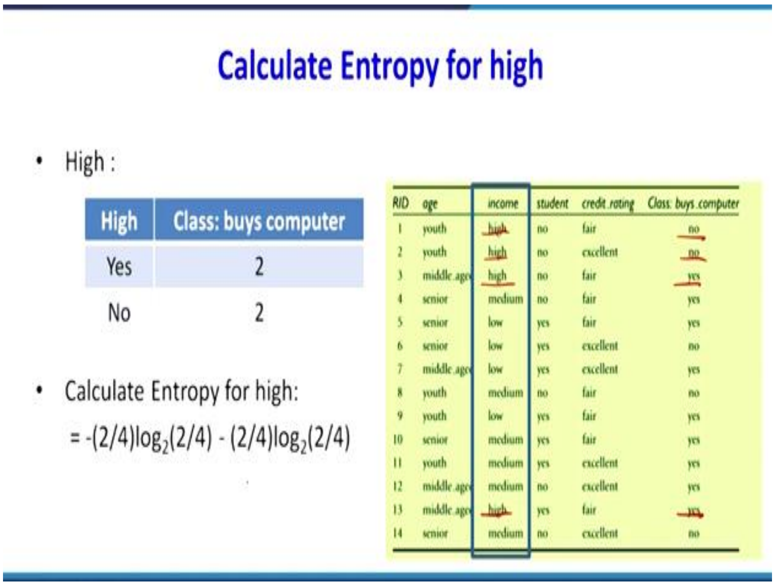
Calculation information Gain of Income

- Calculation of gain for income:
- Income cane be:
  - High
  - Medium
  - Low

RID	age ✓	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_age	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_age	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_age	medium	no	excellent	yes
13	middle_age	high	yes	fair	yes
14	senior	medium	no	excellent	no

Now, we will go to next variable, we have seen for age, now we will go for variable income. In the income, the same way we will repeat the procedure, in the income there are 3 level is there; high, medium, low, so we will find the entropy for high, medium and low, then we will find the expected information required, then we will find the information gain, how will you find the information gain? So, it is the gain from our dependent variable minus this income variable, let us find out that one.

(Refer Slide Time: 12:31)



Calculate Entropy for high

- High :

High	Class: buys computer
Yes	2
No	2

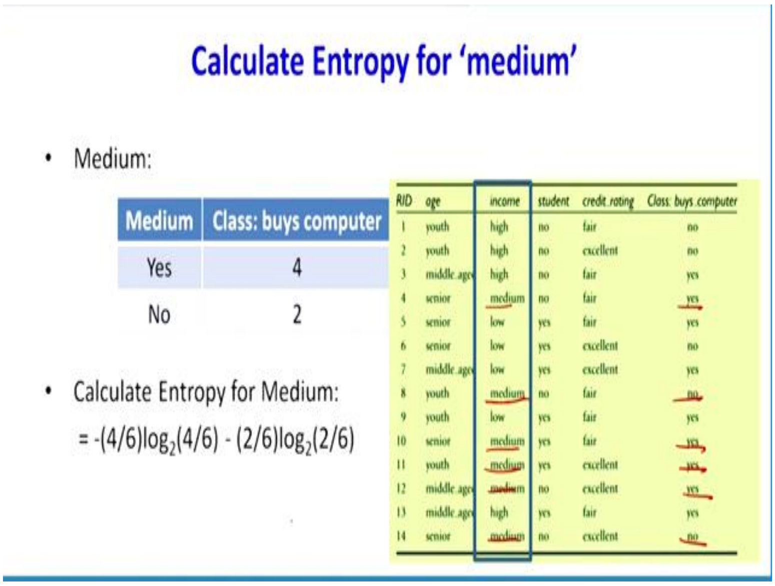
- Calculate Entropy for high:

$= -(2/4)\log_2(2/4) - (2/4)\log_2(2/4)$

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

So, when you go for high income, how many people have told yes? Here that is 1, here yes, then how many people told no, when the level is this 1, 2 that is all, so out of there are 4 values under the income level, out of 4, 2 have answered yes for buying computer, 2 have responded no for buying computer. So, we will find out entropy for high, so – 2 divided by 4 log 2 divided by 4 to the base 2 – 2 divided by 4 log 2 divided by 4 to the base 2.

(Refer Slide Time: 13:20)



Calculate Entropy for 'medium'

- Medium:

Medium	Class: buys computer
Yes	4
No	2

- Calculate Entropy for Medium:

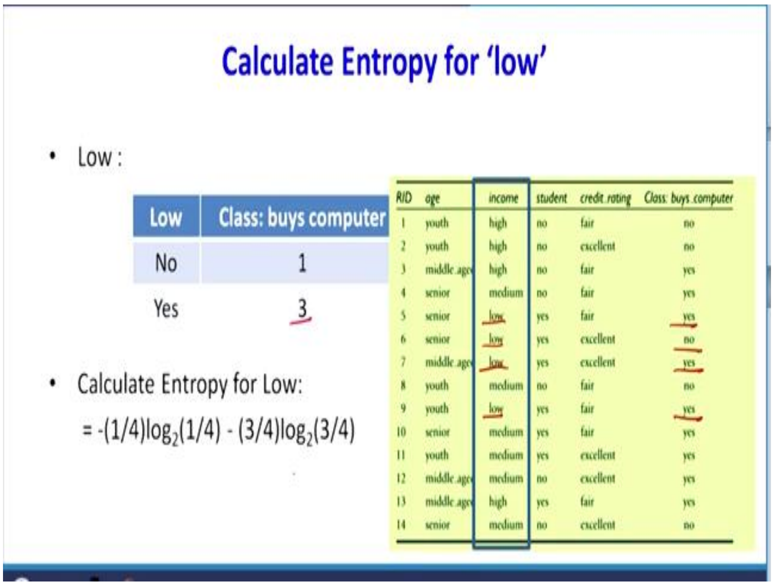
$= -(4/6)\log_2(4/6) - (2/6)\log_2(2/6)$

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_age	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_age	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_age	medium	no	excellent	yes
13	middle_age	high	yes	fair	yes
14	senior	medium	no	excellent	no

So, this value we got now, we will go to the next category level. The next category level is medium; in the medium, we are going to find out how many people are answered yes, medium yes, medium no, this is no, then medium yes, medium yes, medium yes, 1, 2, 3, 4, so there are 4 people have answered yes for buying computer. So, let us see in medium how many people answered no.

So, medium no, then medium this one, medium no, so 2 people have answered no for buying computed, so we will find the entropy, so – 4 divided by 6 log 4 divided by 6 to the base 2 – 2 divided by 6 log 2 divided by 6 to the base 2, so we will get an entropy for medium.

(Refer Slide Time: 14:11)



Calculate Entropy for 'low'

- Low :

Low	Class: buys computer
No	1
Yes	3

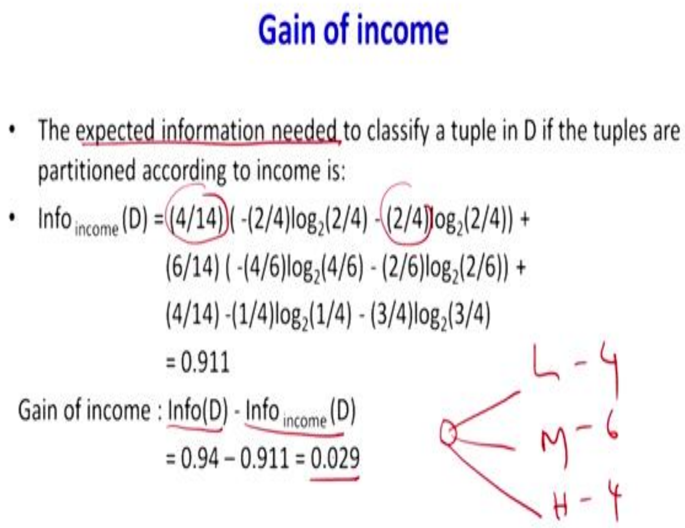
- Calculate Entropy for Low:

$= -(1/4)\log_2(1/4) - (3/4)\log_2(3/4)$

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_age	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_age	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_age	medium	no	excellent	yes
13	middle_age	high	yes	fair	yes
14	senior	medium	no	excellent	no

Then, we will find out the entropy for low; in low, how many yes is there; low yes, then low yes, then here low yes, so 3 people have answered yes for buying computer, how many people are answered no, when they are low yeah, here it is there, this low no, so only 1 people answered no. So, out of 4, 3 answered yes for buying computer, 1 answered no for buying computer. Now, we will find out the entropy; - 1 divided by 4 log 1 divided by 4 to the base 2 - 3 divided by 4 log 3 divided by 4 to the base 2.

(Refer Slide Time: 15:03)



Gain of income

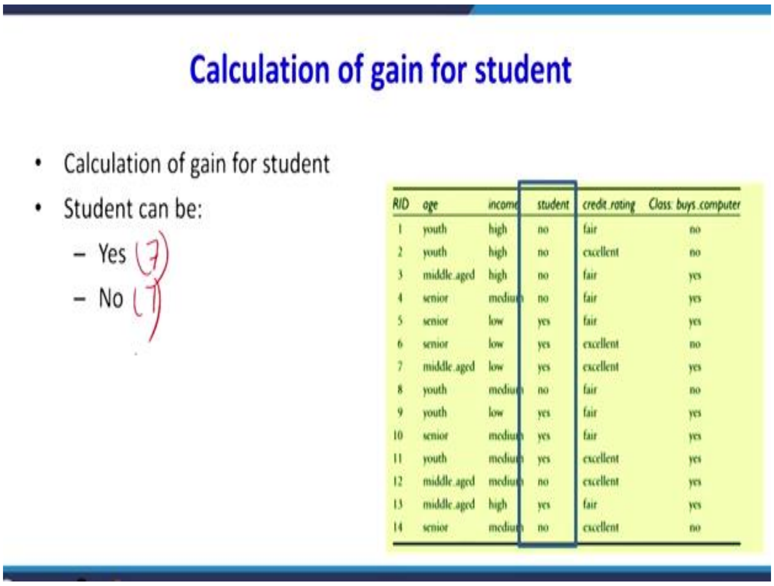
- The expected information needed to classify a tuple in D if the tuples are partitioned according to income is:
- $\text{Info}_{income}(D) = (4/14)(-(2/4)\log_2(2/4) - (2/4)\log_2(2/4)) +$
  $(6/14) (-(4/6)\log_2(4/6) - (2/6)\log_2(2/6)) +$
  $(4/14) -(1/4)\log_2(1/4) - (3/4)\log_2(3/4)$
  $= 0.911$

Gain of income : $\text{Info}(D) - \text{Info}_{income}(D)$
$= 0.94 - 0.911 = 0.029$

The expected information needed to classify a tuple in D if the tuples are partitioned according to income is; so we are finding this weightage, it is nothing but the weighted mean of the entropy, so 4 divided by 14, then it is 6 divided by 14, what is this 6; we will go back, medium will be 1, 2, 3, 4, 5, 6, so there are 6 medium. So, what is happening, it is like this; low, medium, high.

In low, how many variable is there; 1, 2, 3, 4; 4, in medium 1, 2, 3, 4, 5, 6; 6, in high 1, 2, 3, 4, so there are total 14, so 4 divided by 14 the entropy for low, then 6 divided 14, this 6 divided by 14, then the entropy plus 4 divided by 14 that entropy, the weighted entropy is 0.911. So, the information gain of income variable is; so Info D that is for our dependent variable, Info income D is this 0.91, so the difference is 0.029. So, this represents expected information needed to classify a tuple D, if the tuples are partitioned according to income.

(Refer Slide Time: 16:35)



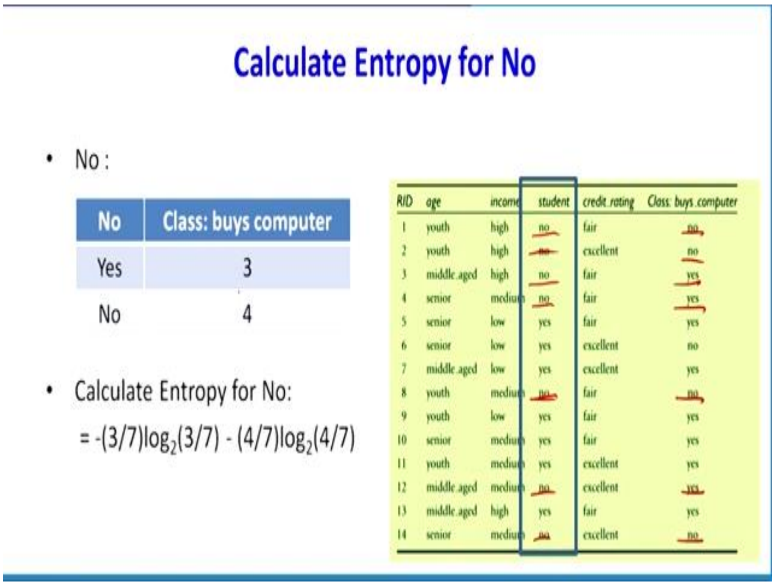
Calculation of gain for student

- Calculation of gain for student
- Student can be:
  - Yes (7)
  - No (7)

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

We will go to the next variable is a student, in student there are 2 level is there; one is yes and no. So, how many yes is there? 1, 2, 3, 4, 5, 6, 7; 7 yes is there. How many no is there? 1, 2, 3, 4, 5, 6, 7, so 7 no is there. Now, we will find out the entropy when it is yes, we will find out the entropy when it is no.

(Refer Slide Time: 17:08)



Calculate Entropy for No

- No :

No	Class: buys computer
Yes	3
No	4

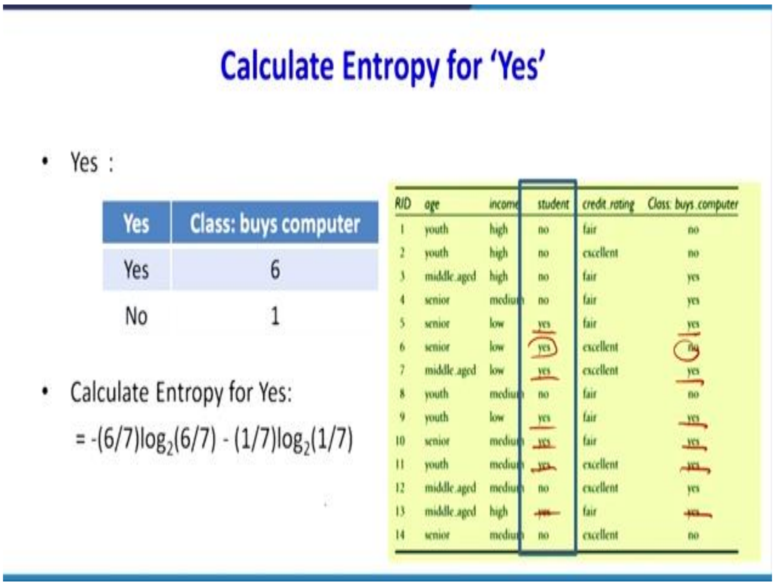
- Calculate Entropy for No:

  $= -(3/7)\log_2(3/7) - (4/7)\log_2(4/7)$

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

When it is no, how many yes, so here one is there, no, yes, no, no, no, yes, so out of 7, 3 people have answered yes to buy the computer. So, how many people answered no for buying computer when they it is no, so this is 1, 2, 3, here one more is there 4, there is a 4. So, entropy for no is - 3 divided by 7 log 3 divided by 7 to the base 2 - 4 divided by 7 log 4 divided by 7 to the base 2.

(Refer Slide Time: 18:03)



Calculate Entropy for 'Yes'

- Yes :

Yes	Class: buys computer
Yes	6
No	1

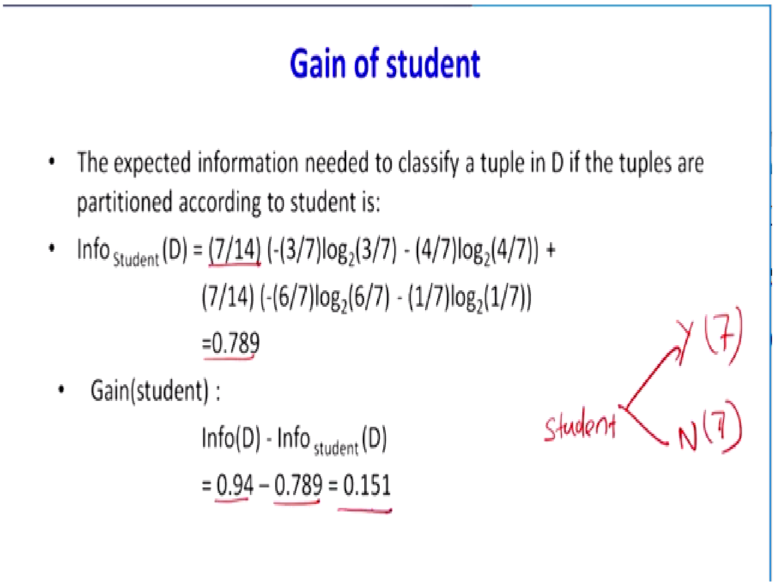
- Calculate Entropy for Yes:

  $= -(6/7)\log_2(6/7) - (1/7)\log_2(1/7)$

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	~~yes~~	fair	~~yes~~
14	senior	medium	no	excellent	no

Now, we will find out entropy for yes, so when it is yes, how many people are answered yes for buys a computer; 1, 2, 3, 4, 5, 6, so there are 6 yes is there. Now, how many no; when there is yes, how many people have answered no, here this one, there is only 1 no is there, so the entropy for yes is - 6 divided by 7 log 6 divided by 7 to the base 2 - 1 divided by 7 log 1 divided by 7 to the base 2.

(Refer Slide Time: 18:44)



Gain of student

- The expected information needed to classify a tuple in D if the tuples are partitioned according to student is:
- $\text{Info}_{\text{Student}}(D) = \underline{(7/14)}\,(-(3/7)\log_2(3/7) - (4/7)\log_2(4/7)) +$

  $(7/14)\,(-(6/7)\log_2(6/7) - (1/7)\log_2(1/7))$

  $= \underline{0.789}$
- Gain(student) :

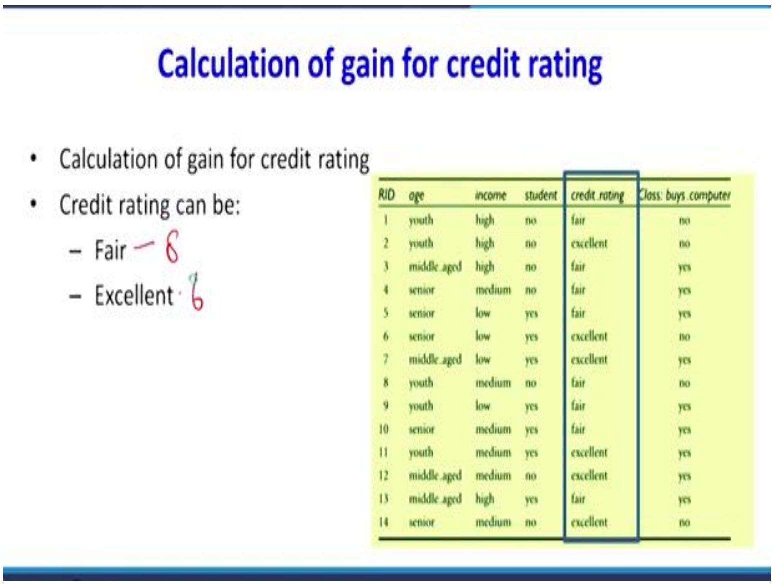
  $\text{Info}(D) - \text{Info}_{\text{student}}(D)$

  $= \underline{0.94} - \underline{0.789} = \underline{0.151}$

Now, we will find out expected information, so for the expected informations, so what we have to do; we have to find out the weighted entropy. So, the weighted entropy is 7 divided by 14 because already we have seen when there was a student, how many yes is there, how many no is there? There are 7 yes, there are 7 no, so out of 14, 7 divided by 14 and that is corresponding entropy plus for no, there is 7 divided by 14 corresponding entropy.

So, this was the expected information needed, so the gain is; so 0.94 for our dependent variable and for this student variable, it is 0.789, so we got this one the gain; the gain is 0.151.

(Refer Slide Time: 19:41)



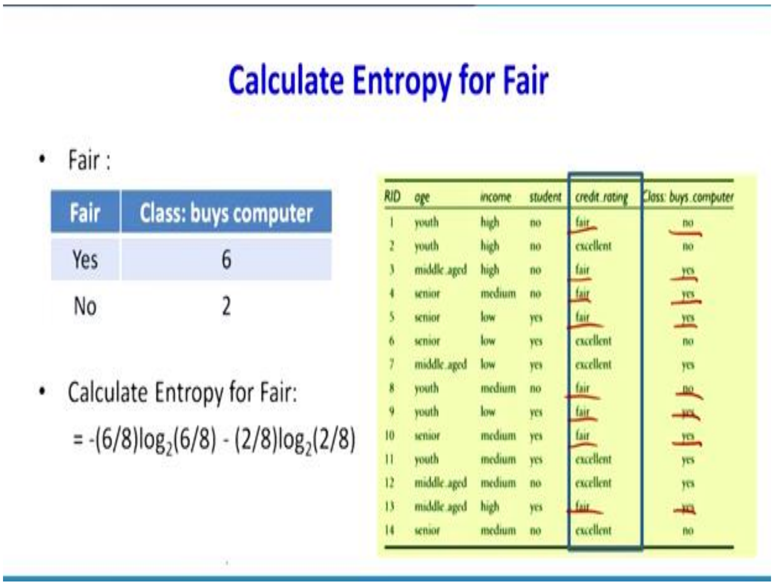
Calculation of gain for credit rating

- Calculation of gain for credit rating
- Credit rating can be:
  - Fair – 8
  - Excellent 6

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Next, we will go to the next variable; credit rating. In credit rating, how many levels is there? Fair, excellent, there are 2 level is there, fair and excellent. So, how many fair is there? 1, 2, 3, 4, 5, 6, 7, 8. How many excellent is there; 1, 2, 3, 4, 5, 6, out of 14, 6 is there. So, now we will find out the entropy, when it is a fair, we will find out entropy when it is excellent for the credit rating.

(Refer Slide Time: 20:14)



Calculate Entropy for Fair

- Fair :

Fair	Class: buys computer
Yes	6
No	2

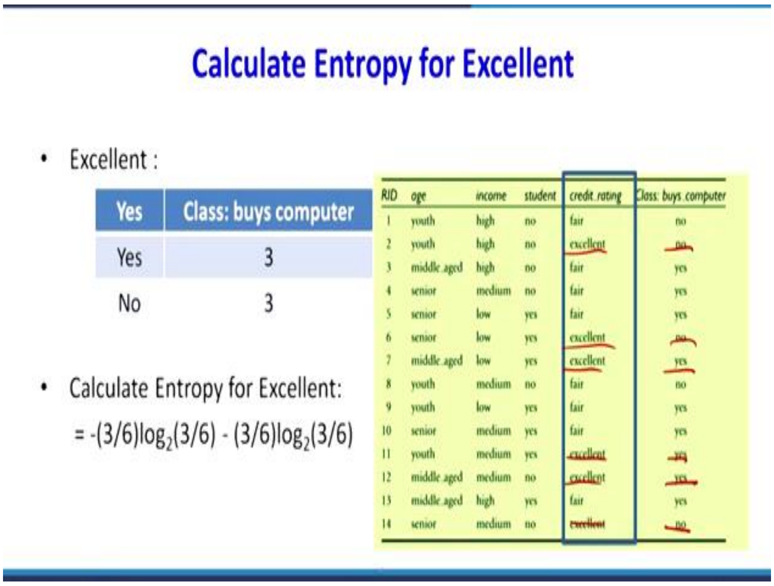
- Calculate Entropy for Fair:

$= -(6/8)\log_2(6/8) - (2/8)\log_2(2/8)$

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

So, for fair how many people are answered yes; fair, fair, fair yes, fair yes, fair yes, fair yes, fair yes, fair yes, 1, 2, 3, 4, 5, 6. So, how many people are fair and same time it is no, so this fair no, this fair no, there are 2. So, how to find out the entropy for fair; - 6 divided by 8 log of 6 divided by 8 to the base 2 - 2 divided by 8 log 2 divided by 8 to the base 2.

(Refer Slide Time: 21:00)



Calculate Entropy for Excellent

- Excellent :

Yes	Class: buys computer
Yes	3
No	3

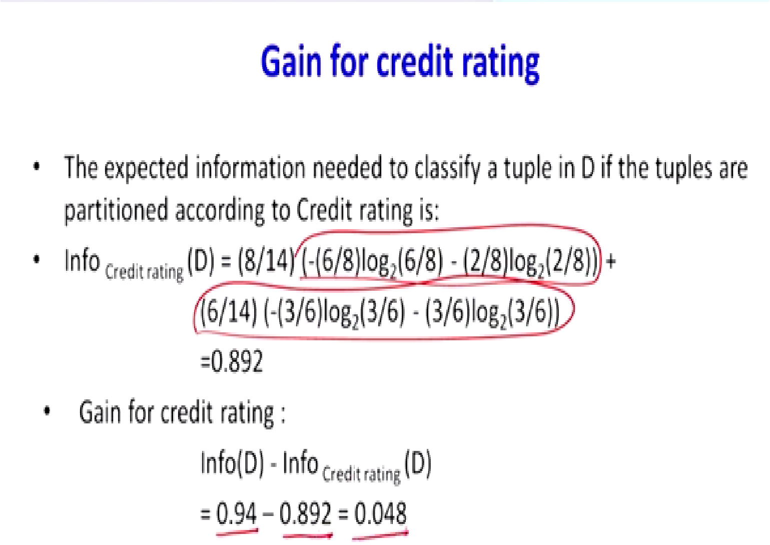
- Calculate Entropy for Excellent:

$= -(3/6)\log_2(3/6) - (3/6)\log_2(3/6)$

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Now, let us find the entropy for excellent, so how many people are yes for buying compute, when they are excellent; excellent 1, excellent 2, excellent 3, there are 3 people. How many people told no; excellent no, excellent no, excellent no, so 3 people are answered yes for buying computer when their level is excellent, 3 people are answered no for buying computer when their level is excellent. So, the entropy for excellent is - 3 divided by 6 log 3 divided by 6 to the base 2 - 3 divided by 6 log 3 divided by 6 to the base 2.

(Refer Slide Time: 21:44)



Gain for credit rating

- The expected information needed to classify a tuple in D if the tuples are partitioned according to Credit rating is:
- $\text{Info}_{\text{Credit rating}}(D) = (8/14)(-(6/8)\log_2(6/8) - (2/8)\log_2(2/8)) + (6/14)(-(3/6)\log_2(3/6) - (3/6)\log_2(3/6))$

  $=0.892$
- Gain for credit rating :

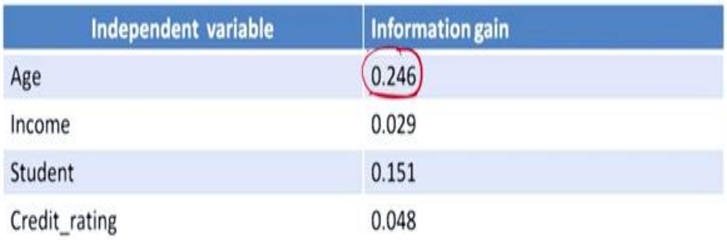
  $\text{Info}(D) - \text{Info}_{\text{Credit rating}}(D)$

  $= \underline{0.94} - \underline{0.892} = \underline{0.048}$

Now, we will find out the expected information needed to classify in a tuple D, if the tuples are partitioned according to credit rating. So, here in the credit rating there are 2 levels; one is as I told you one is fair, another one is excellent. So, for fair there are 8 is there; 8 items, so 8 divided by 14 and corresponding their entropy, the remaining 6 divided by 14 and their corresponding their entropy.

So, the expected information needed is 0.892, now we will find out gain for credit rating. What is the meaning of gain for credit rating? If we use credit rating as the root variable, how much information is required, so this variable we got it when it is 0.94 for the dependent variable, this is for credit rating variable just now we got it 0.892, so the difference is 0.048.

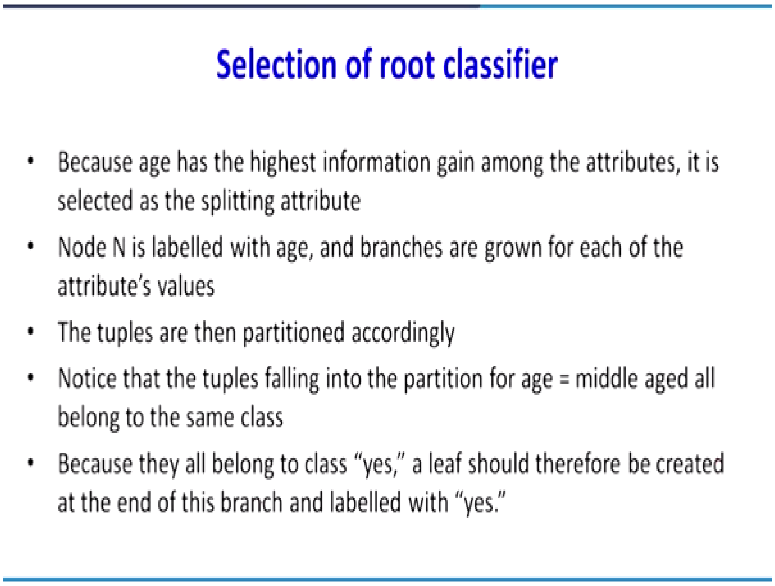
(Refer Slide Time: 22:52)



Independent variable	Information gain
Age	0.246
Income	0.029
Student	0.151
Credit_rating	0.048

Now, I have summarized; if we use age is the classifier, the information gain is 0.246, if we use income as the classifier, the information gain is 0.029, if we use student as a classifier the information gain is 0.151, if we use credit rating as a classifier the information gain is 0.048. So, out of this 4, the highest value is 0.246, so now we will start keeping age as a classifier variable, so that is application of this.

(Refer Slide Time: 23:36)



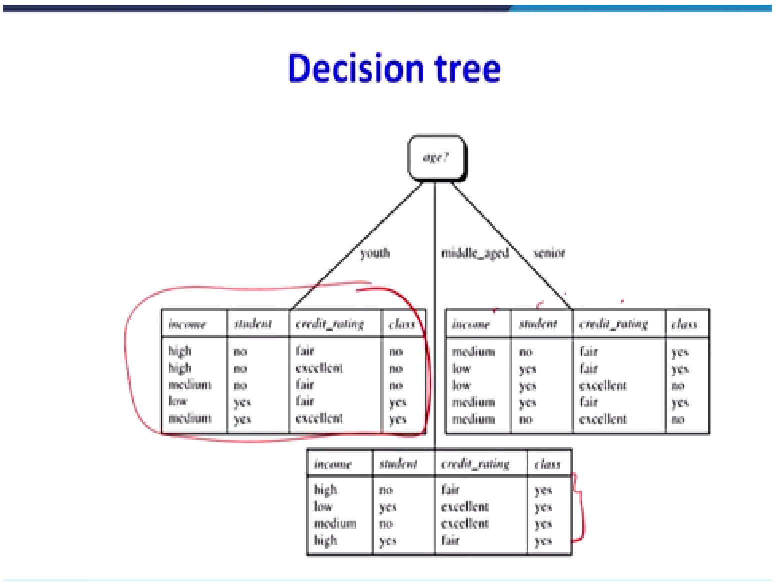
Selection of root classifier

- Because age has the highest information gain among the attributes, it is selected as the splitting attribute
- Node N is labelled with age, and branches are grown for each of the attribute’s values
- The tuples are then partitioned accordingly
- Notice that the tuples falling into the partition for age = middle aged all belong to the same class
- Because they all belong to class “yes,” a leaf should therefore be created at the end of this branch and labelled with “yes.”

So, what happened because age has the highest information gain among the attributes, it is selected as the splitting attribute, node N is labelled with age and the branches are grown for each of the attributes values. The tuples are then partitioned accordingly, notice that the tuples falling into the partition for age, when there is a middle aged all belongs to the same class.

So, we need not go for further classification because all are belongs to; all middle aged people are answered yes for buying computer, so further classification is not required because they are belongs to class yes, a leave should therefore be created at the end of this branch and labelled with yes.

(Refer Slide Time: 24:23)



Decision tree

income	*student*	*credit_rating*	*class*
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

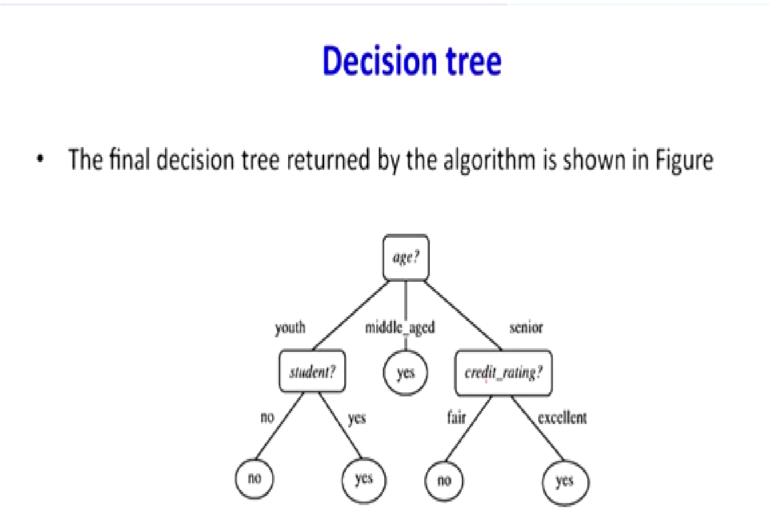
income	*student*	*credit_rating*	*class*
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

income	*student*	*credit_rating*	*class*
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

This was the decision tree, so what happened; the age is the classifier, if there is age; youth, middle aged and senior, there are 3 level is there. So, for middle aged all are answering yes, so further classification is not required. So, if it is youth there are some other variables is there, income is there, student is there, credit rating is there, so the information gain algorithm methodology which we have used can be used this group also.

Again, we can find out, out of these 3 variables; income, student, credit rating, which variable should appear here, same way the one classifier is a senior, when it is a senior there are 1, 2, 3 variable is there, you can find out this information criteria, so out of these 3 variable which variable should come into the this node, so this way we can continue our classification procedure.

(Refer Slide Time: 25:25)



Decision tree

- The final decision tree returned by the algorithm is shown in Figure

This was our final decision tree returned by algorithm I shown in the figure, we started with the age; when you started with the age there are 3 level was there in the age; youth, middle aged, senior, so we are stopping because all are yes there, we do further classification required. Then you see that as I told you previously there are 3 options income, student, credit rating, so out of this, the student yes appeared one classifier on the left hand side.

So, when there is a student we can there is a 2 possibility; yes or no and the right hand side when it is seen here, you see that there are 3 possibilities there, we can classify with respect to income, student, credit rating. So, what happened is student already we have done that one, so now we go for credit rating, this also got by using that information gain measures, so this was our final decision tree.

We have seen different measures for selecting the attributes, there are 3 measure is there; one measure is information gain, another one is gain ratio, another one is gain index. In this lecture, what I have done using information gain as a measure by taking one numerical example, I have explained how to choose an attribute. In the next class, I will take another measures for choosing the attribute that is gain ratio and Gini index with that example, we will continue in my next lecture, thank you.