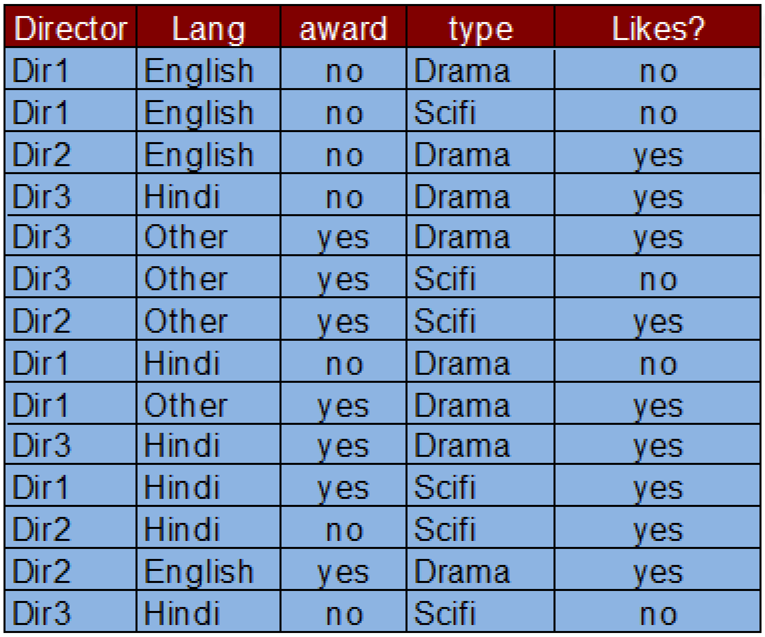
Course Name - Recommender Systems
Professor Name - Prof. Mamata Jenamani
Department Name - Industrial and Systems Engineering
Institute Name - Indian Institute of Technology Kharagpur
Week - 05
Lecture - 24

Lecture 24: Feature Engineering – III

Hello everyone. Welcome to the lecture on Feature Engineering and in this regard now today's lecture we are going to look at feature selection. Feature selection, what is this? Feature selection is about selecting the best possible subset of the available features to improve the prediction accuracy. So, this is a subset selection problem. Given a large set of features we are supposed to find out a subset. Now, there are many approaches for this like in some are information theory based approaches, Gini index, chi square statistics and so on.

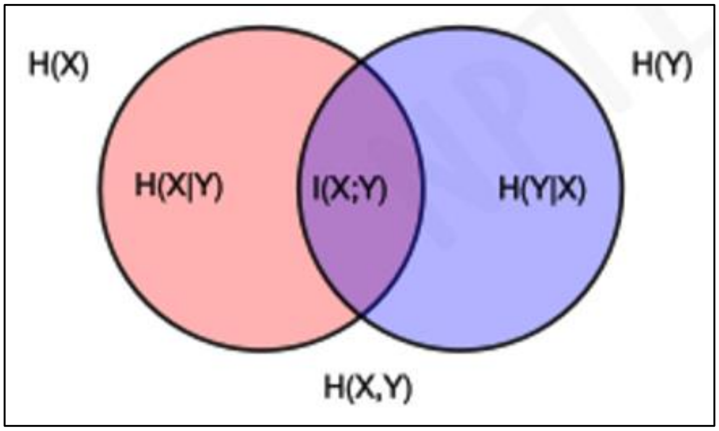
So, we start with certain information theory foundation for feature selection. So, entropy is one such measure. So, let us find out what entropy is. Our aim is to find out some variable which maxima which has maximum information content while describing the response variable.



Director	Lang	award	type	Likes?
Dir1	English	no	Drama	no
Dir1	English	no	Scifi	no
Dir2	English	no	Drama	yes
Dir3	Hindi	no	Drama	yes
Dir3	Other	yes	Drama	yes
Dir3	Other	yes	Scifi	no
Dir2	Other	yes	Scifi	yes
Dir1	Hindi	no	Drama	no
Dir1	Other	yes	Drama	yes
Dir3	Hindi	yes	Drama	yes
Dir1	Hindi	yes	Scifi	yes
Dir2	Hindi	no	Scifi	yes
Dir2	English	yes	Drama	yes
Dir3	Hindi	no	Scifi	no

Now, the question is how much information is received when you observe a specific value for a discrete random variable X. Now, what is the random variable here? For example, in this example in this example we have the details of movies ok. What are the attributes? Of course, here we have not talked about text attribute we are only considering extrinsic attributes here, but text attributes can be considered as well. So, here the attributes are director, language of the movie, whether it has got any award or not, what type, genre of the movie and like. Again for the simplicity we have kept only one column for the genre, but movie can have multiple genres as well.

So, now what is the random variable here? This all these attributes are random variables director, language. So, when a new movie comes you do not know a priori what is going to be the director, language etcetera when the movie comes then only you know. So, director in the name of director anybody's name can come here we have given 3 directors director 1, director 2 and director 3. Now when we talk about information what exactly the information? Suppose this data set would have been containing movies from only director 1. So, is there any surprising factor here? Because we know that it is always direct director 1 who has directed them directed all the movies.

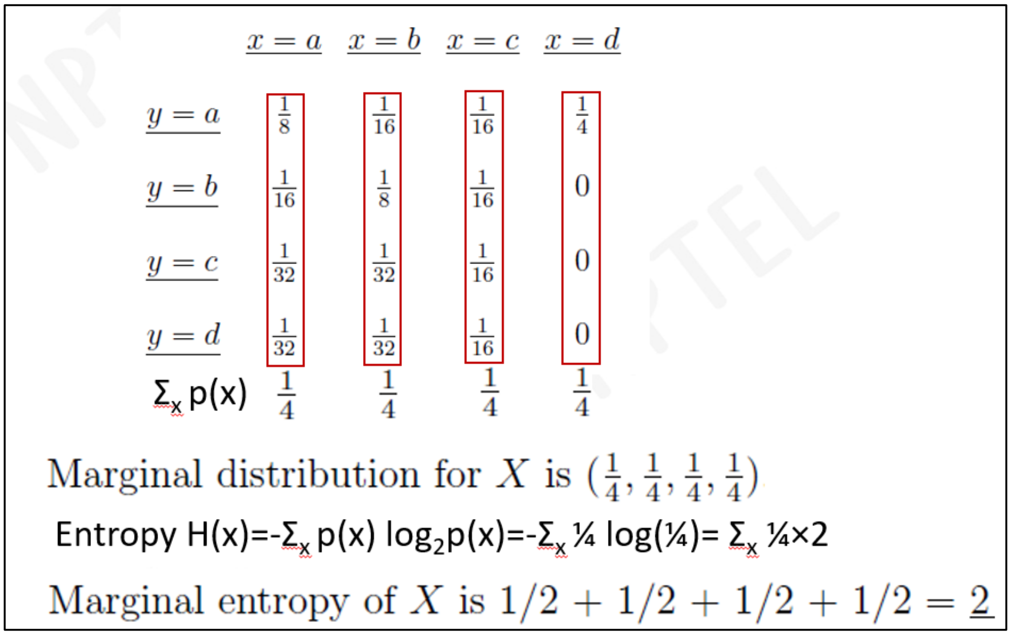


So, this particular column would have become irrelevant. So, similarly if you have let us say likes all the movies would have been no. So, which means there is no discriminatory power you know that this is all no. So, discriminatory power or the surprising thing that can come out of your analysis from one analysis of the attribute specifies what is the kind of information you have it. Now how do we get a quantification of that that is the issue.

Now depending on the probability distribution P x that is that has value between 0 to 1 we need a function H x of P x that expresses this information content. Now how do we get this P x values? So, here what is the random the random value variable here can take the term DIR 1, DIR 2 or DIR 3. So, what is the probability of having DIR 1? So, how many data we have 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14. So, you have total 14 data. So, out of 14 data in there will be a number of places the director 1 will be there and divided by.

So, number of places where director 1 divided by total records. So, that is how you can find out the probabilities of a probability distribution for P x a mass function for P x because these are discrete discrete values. Now the function H x which you are trying to discover what should it say? It should say the information content in the with respect to that ah attribute. So, when we talk about information content this H x has to have certain property. What is the property? The information content must be the sum of the information content of two resources when both are statistically independent.

So, if there are two unrelated events x and y in this case attributes and they are statistically independent then their joint probability probability of having x and y together is P x into P i. Then information content must be sum of. So, if their joint probability is this then the information content must be the sum of these two because they are independent of each other. So, they are providing some information independently. So, it has to be now added together.



	$x = a$	$x = b$	$x = c$	$x = d$
$y = a$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
$y = b$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	0
$y = c$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	0
$y = d$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	0
Σ_x p(x)	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

Marginal distribution for X is $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$

Entropy $H(x)=-\Sigma_x p(x) \log_2 p(x)=-\Sigma_x ¼ \log(¼)= \Sigma_x ¼×2$

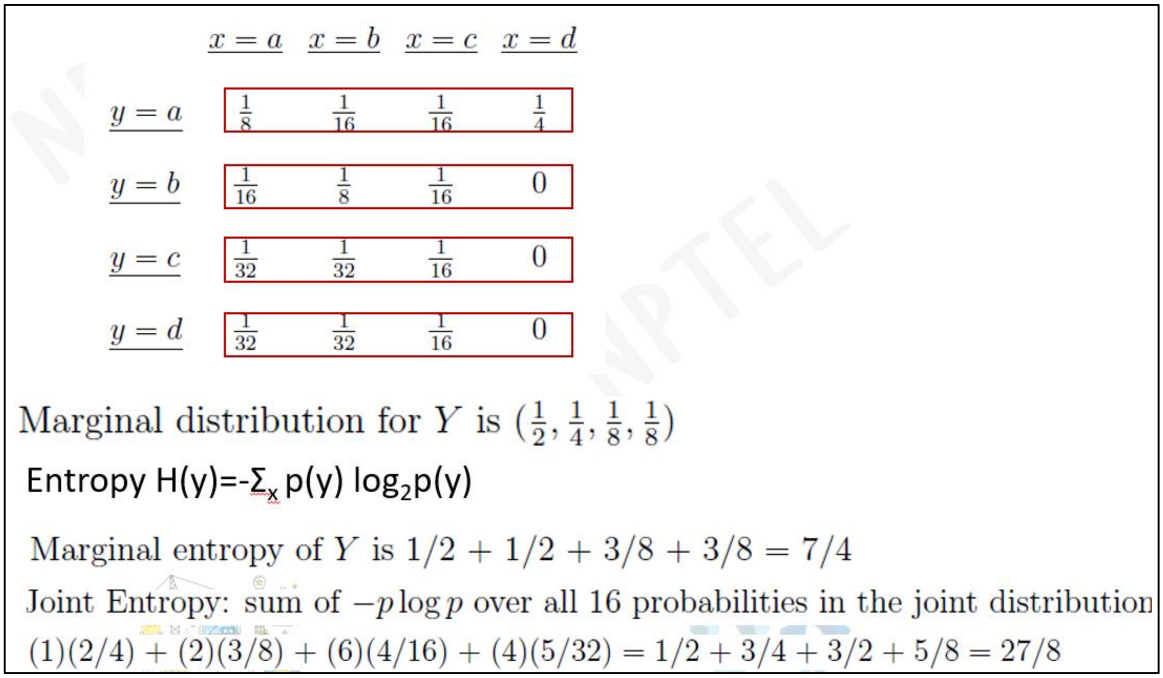
Marginal entropy of X is $1/2 + 1/2 + 1/2 + 1/2 = \underline{2}$

Now, if we are trying to derive it from probability which in case of joint probability is a multiplication and you need the function in terms of addition what is the kind of right kind of function this purpose? Log is the function this purpose if you take log of that it will be able to ah add them together ok. So, this average amount of information is the expectation with respect to P there is a blank here respect to P and is referred to as entropy. So, this is the formula for entropy. Now, why the formula for entropy is this? The reason being this one. Now, this entropy measures the uncertainty of a random variable.

How to derive this entropy? Now, let x be a discrete random variable with alphabet xi and probability mass function P x. Try to remember our last example. In last example what was x? x was the variable director. There was another variable called language. Now, what is xi? Which we call as the alphabets.

So, alphabets in case of director where 3 director 1, director 2 and director 3. So, these 3 were my alphabet set and with respect to this how to identify the probabilities that we know. So, probability of this this and this make my mass function and this is how with respect to the mass function I now get the entropy like this. Now, joint probability is another concept which is defined by this formula with respect to this joint distribution joint distribution of 2 variables x and y. Now, if both the events are independent then joint distribution is going to be the joint entropy is going to be the entropy of x plus entropy of y.

Now, look at this. In this diagram in this typical Venn diagram let us say both of them are independent. So, this will be your H x and this will be your H y. So, this there will not be any overlapping part. So, it will be the sum of that.



	$x = a$	$x = b$	$x = c$	$x = d$
$y = a$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
$y = b$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	0
$y = c$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	0
$y = d$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	0

Marginal distribution for Y is $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$

Entropy $H(y)=-\Sigma_x p(y) \log_2 p(y)$

Marginal entropy of Y is $1/2 + 1/2 + 3/8 + 3/8 = 7/4$

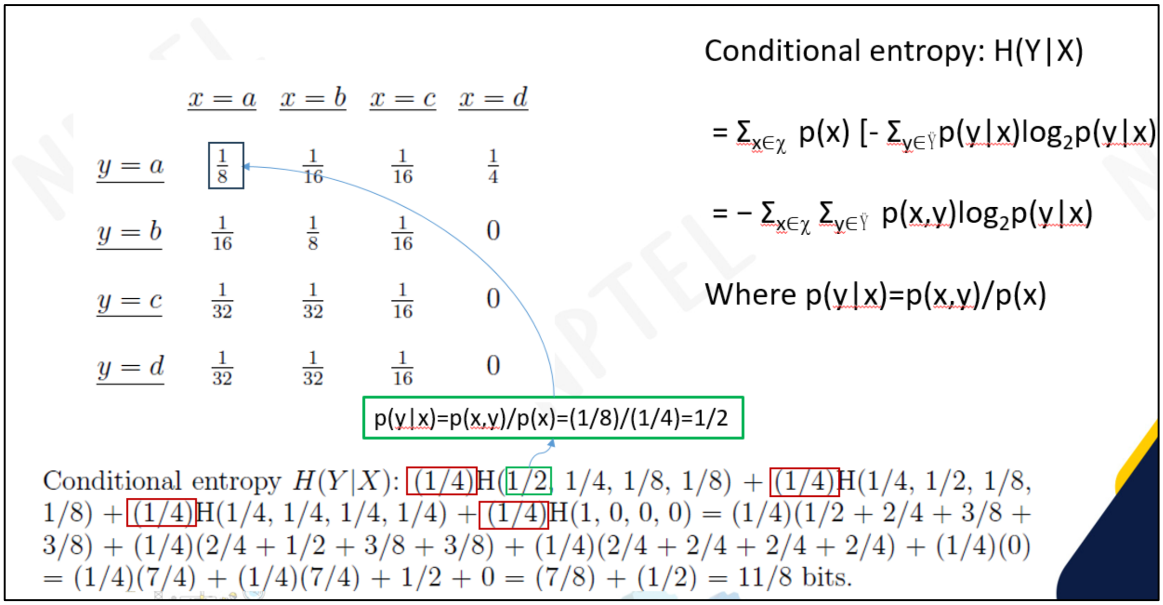
Joint Entropy: sum of $-p \log p$ over all 16 probabilities in the joint distribution

$(1)(2/4) + (2)(3/8) + (6)(4/16) + (4)(5/32) = 1/2 + 3/4 + 3/2 + 5/8 = 27/8$

Now, in case there is an overlapping part it has to be now subtracted once. So, that overlapping part is called the mutual information. So, when you have some dependency which means you have to define you can define something called conditional entropy. So, what is this conditional entropy? This is the entropy of certain random variable y giving the value of x. So, conditional entropy of y given x is this part.

So, this circle is y and this common part if you take away sorry ah yes ah this is in presence of x. So, this is your y given x. So, your by the chain rule of entropy this H this entropy of this joint distribution let us go up this was the joint entropy. This joint entropy x y is H entropy of x H of x H of x is what this full circle and H of y given x H of y y given x is what this part this together makes the joint. The otherwise you can also present it first taking the y that is the this part this circle then taking this one.

And how do you get now the mutual information? From H of x that is this circle you remove this part you get similarly from this y circle you remove this part you get this one or you you can also get it from this joint. How do you get it from this joint? This joint is basically decided by these two terms. So, oh it is just the opposite. So, you ah given this joint the fraction which is common to both these circles represent this mutual information. This is one example here we have two random variables x and y and this is the joint distribution.



	$x = a$	$x = b$	$x = c$	$x = d$
$y = a$	$\boxed{\frac{1}{8}}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
$y = b$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	0
$y = c$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	0
$y = d$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	0

Conditional entropy: H(Y|X)

$= \Sigma_{x \in \chi}\ p(x)\ [- \Sigma_{y \in \ddot{Y}} p(y|x) \log_2 p(y|x)$

$= - \Sigma_{x \in \chi} \Sigma_{y \in \ddot{Y}}\ p(x,y) \log_2 p(y|x)$

Where p(y|x)=p(x,y)/p(x)

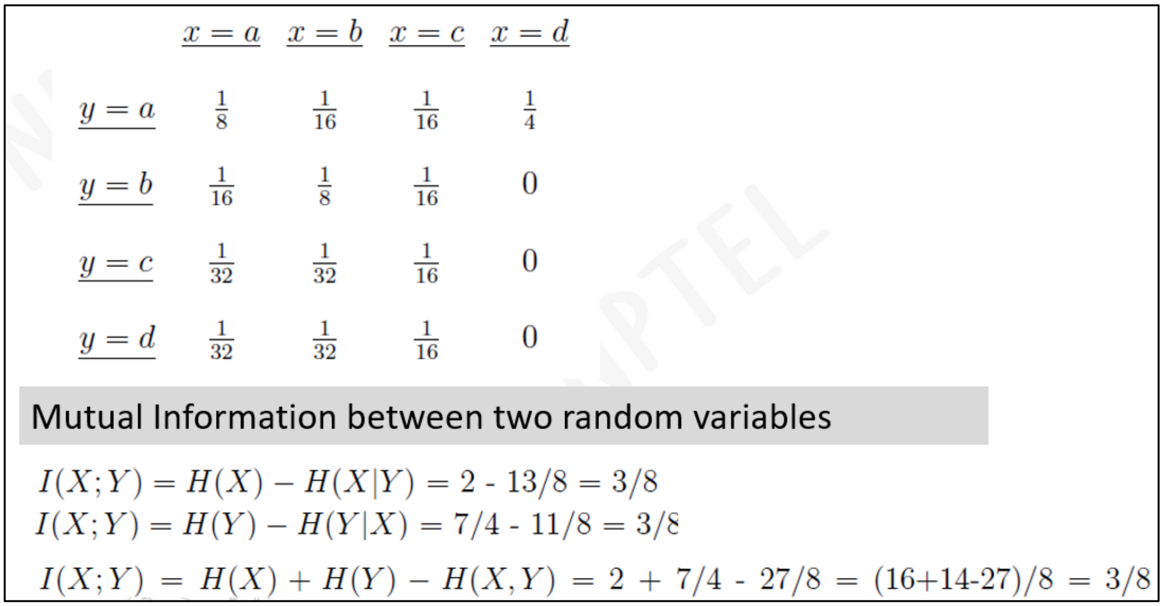
$\boxed{p(y|x)=p(x,y)/p(x)=(1/8)/(1/4)=1/2}$

Conditional entropy $H(Y|X)$: $\boxed{(1/4)}H(\boxed{1/2}, 1/4, 1/8, 1/8) + \boxed{(1/4)}H(1/4, 1/2, 1/8, 1/8) + \boxed{(1/4)}H(1/4, 1/4, 1/4, 1/4) + \boxed{(1/4)}H(1, 0, 0, 0) = (1/4)(1/2 + 2/4 + 3/8 + 3/8) + (1/4)(2/4 + 1/2 + 3/8 + 3/8) + (1/4)(2/4 + 2/4 + 2/4 + 2/4) + (1/4)(0) = (1/4)(7/4) + (1/4)(7/4) + 1/2 + 0 = (7/8) + (1/2) = 11/8$ bits.

How do you get this joint distribution? These are the alphabets this y can take four values a b c d our director was direct variable director was taking three values. What are the director 1 director director 3 language was taking language of the movie was taking three values English Hindi and others. So, consider them. So, how many times director 1 language let us say this second one language English how many times it has occurred in the entire database. So, that many times divided by number of total data points that is how you decide this joint distribution.

So, in this example this x and y are two random variables and this is the joint distribution. So, once you have this joint distribution you can find out the marginal distribution of x what is the how many times x a as x equal to a how many times b x equal to b in the data set and divided by the total number. So, that probability is the sum of this columns sum of this columns you get this probability. Now, how do you get the entropy of H x? In this case everything a for every ah column it is 1 by 4. So, you sum over all the values of x.

So, 1 by 4 into log 1 by 4 has to be added 4 times ok. So, that makes it 1 ah 1 by 2 1 by 2 that makes it 2. So, this is the entropy of x in the similar manner you can also find out the row wise sum to get marginal distribution for y and marginal entropy for y. How do we get joint entropy? To get the joint entropy individual individual elements this you have to take the sum p log p over all the 16 probabilities and you get this value you can try this by yourself. Then we can also find out conditional probability conditional entropy.



	$x = a$	$x = b$	$x = c$	$x = d$
$y = a$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
$y = b$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	0
$y = c$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	0
$y = d$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	0

Mutual Information between two random variables

$I(X;Y) = H(X) - H(X|Y) = 2 - 13/8 = 3/8$

$I(X;Y) = H(Y) - H(Y|X) = 7/4 - 11/8 = 3/8$

$I(X;Y) = H(X) + H(Y) - H(X,Y) = 2 + 7/4 - 27/8 = (16+14-27)/8 = 3/8$

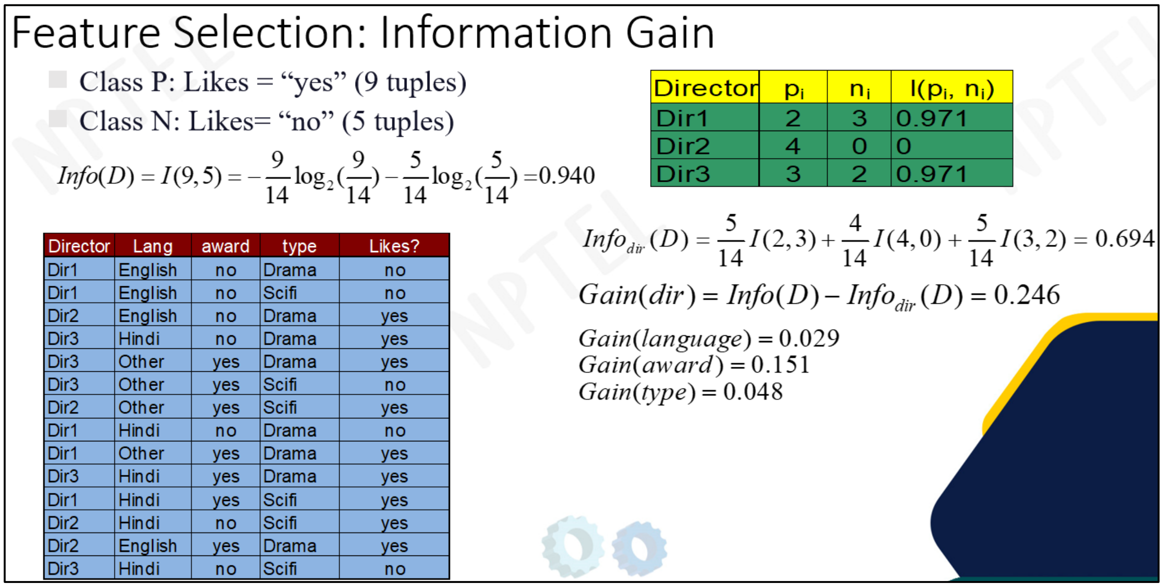
How do we find out conditional entropy? This is the formula and in this formula if you look at we have p x we have p of y given x and p of y given x. Now, in this particular it is sum of all these terms. So, this is over all the alphabets of x this is over all the alphabets of y. Now, this turns out to be 11 by 8 and because the concept of it the concept is coming from information theory. So, for historical reason it is has the unit called bits this entropy has unit called bits.

So, now, here how do we get this value look at this this formula probability of x this is p x for that particular value when x is equal to a. In fact, in our last example we saw here all these were 1 by 4 1 by 4 etcetera. So, this is 1 by 4 here 1 by 4 here 1 by 4 here and 1 by 4 here. Now, come to this this half is coming from this formula probability of y given x is joint probability of x y by p x what was the joint probability of x y this was this and what is probability of x it was 1 by 4 this 1 by 4 that makes it 2. So, that 2 1 by 2 and that 1 by 2 has to be now multiple along with others it has to be now multiplied with this.

So, so, you can carry it out and find out it in detail. Now, come to the mutual information once you find them out this mutual information can be determined from all this formula. What is this formula? This relates mutual information information content of x minus information content of x given y. So, remember that Venn diagram this was H of x and this was H of x given y.

So, this was the area. Similarly, from y side also you can determine and this was your this was your H. So, this has to be now subtracted from sum of these two. So, once you subtract you will be getting this value. So, in all the three ways you are getting the same value. Now, extending this this information theoretic measures for feature selection can have three greedy ways.

First information gain can be used as a criteria, decision tree for can also be used for decision trees can also be used for feature selection and there can be a methodology which uses mutual information for feature selection. So, now, let us look at information gain as a criteria for feature selection. So, this is our old data set that I was trying to show you. So, what we do? You find out the information content of the entire data set with respect to this last variable. In the last where total 14 number of rows how many no's how many no's there are 9 yes and 5 no's.



Feature Selection: Information Gain

- Class P: Likes = “yes” (9 tuples)
- Class N: Likes= “no” (5 tuples)

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

Director	p_i	n_i	$I(p_i, n_i)$
Dir1	2	3	0.971
Dir2	4	0	0
Dir3	3	2	0.971

Director	Lang	award	type	Likes?
Dir1	English	no	Drama	no
Dir1	English	no	Scifi	no
Dir2	English	no	Drama	yes
Dir3	Hindi	no	Drama	yes
Dir3	Other	yes	Drama	yes
Dir3	Other	yes	Scifi	no
Dir2	Other	yes	Scifi	yes
Dir1	Hindi	no	Drama	no
Dir1	Other	yes	Drama	yes
Dir3	Hindi	yes	Drama	yes
Dir1	Hindi	yes	Scifi	yes
Dir2	Hindi	no	Scifi	yes
Dir2	English	yes	Drama	yes
Dir3	Hindi	no	Scifi	no

$$Info_{dir}(D) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$$

$$Gain(dir) = Info(D) - Info_{dir}(D) = 0.246$$

$$Gain(language) = 0.029$$
$$Gain(award) = 0.151$$
$$Gain(type) = 0.048$$

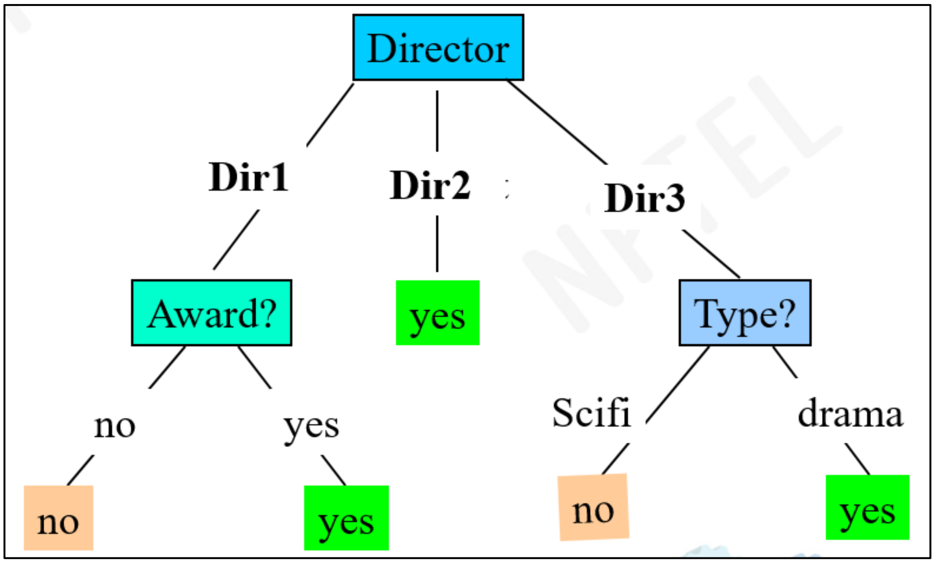
So, following the formula that P x into log P x negative of that adding over all the alphabets we get the entropy information content which we call as information content of the entropy. So, now, this is about this response variable. Our a what is our aim? Our aim is to choose out of these 4 which two will be the best maybe suppose we are considering top two which two. So, for that purpose what we do? We find out information content of each of this attribute. For example, if we consider that case of director as the attribute how many directors? 3 directors.

How many P? How many positive? How many negative? Positive in the sense yes and no. So, total 5 negative and total 9 positives. So, they are now distributed over these 3 directors here 2 4 3 are the positive here 3 0 2 are the negative among 3. Now, in individual case now the way you found it out you find out the information content. Once you find out the information content then information content of director is turns out to be this one where these probabilities were multiplied with this individual entropy values.

So, where from this 5 is coming? This 5 is coming by adding this 2 plus 3. Where from this 4 is coming? Adding this 4 plus 0. This 5 is coming by adding and total number of data sets size of the data set is 14. So, you got this value. Now, what is the gain? This is the formula for the gain.

This value is coming from here and this value is coming from here. So, both this when you subtract you get this one. So, similarly for other variables for language and our and type you get these values. Out of this wherever the information gain is highest you choose those. So, if we decide to choose 2 then this will be our first choice, this will be our second choice.

Now, decision tree itself can be considered as an attribute selection method. Now, how to construct a decision tree also depends on this concept of entropy and consider entropy that we are going to see shortly maybe in next week in the next module. So, suppose we make the decision tree. In this decision tree out of those 4 variables director, language, award and type we see only our director and type are appearing language is not appearing. So, which means language do not have much information content to decide this.



So, therefore, in our final data set we will be keeping director, award and type. But at the same time I would like to mention the decision tree itself can be used as the tool for classification. Classification when for generating recommendation. So, in that context when we talk about talk about the decision tree as a attribute selection measure, we can use it as an attribute selection measure and for better accuracy we can use another model and maybe after selecting this few attributes you can get a better decision tree, better discriminatory power of this decision tree. Then the last approach is mutual information based feature selection for which some algorithm is used.

So, in this algorithm what you do we are supposed to find the best subset of best k ah k features a subset of best k features out of total let us say n number of features. So, how do you move about it? So, first of all create 2 sets f and s, s is supposed to be 5 does not have any value and f contains all the features n. Then what you do in an iterative manner take consider all the features here and in the first step find out the one which has the highest ah information gain is highest resting. So, you you consider this n and there is a class variable class variable will have different levels. In our example it was he likes or he does not like yes or no, but it can be multiple level as well consider case of ah ah 5 Likert scale data.

So, now individually each of them here there are how many features 1, 2 up to n individually all of them you find out the mutual information with respect to this class and the best one you keep in this set. Let us say in this set now you have some feature called i ok. So, now, i is no more here i has been removed from here. Now, whatever is the remaining features you our aim is to see that if you take the I mean the ah it it this should not be predictable from this one. So, what do you do you compute the mutual information between the features which are already existing here and which is here this feature i.

Now, out of these let us choose some feature G that maximizes this mutual information which you are getting from this C and considering C and F together and considering ah C and F I mean considering this one and class variable together and this is considering ah this is there is some confusion in terms of ah symbols. So, consider this as S which is S and this S is subset of capital S ok. So, with this process you take this difference and whatever is the highest that G comes here ok. Now, originally it was I now G comes. So, for all and all this I and G all of them we call them as S.

Now, now take one more vector one more ah feature and continue the process you keep on continuing this process till total k features are here ok. So, with this we stop our discussion over here and next class we will be considering about other measures such as Gini index and all. So, these are some of my references which I have used and specifically for mutual information based feature selection I have used this feature in this paper and ah methods etcetera are mostly referred from this Char-Vagarwal's book these are my concluding remark. So, if the number of features are very large few features can be selected using methods such as entropy, information gain, mutual information etcetera. The information theory builds the foundation for methods such as entropy, information gain and mutual information.

Three greedy methods for information theoretic measures for feature selection are information gain decision tree and mutual information based selection. Gini index ah which will be discussed next class is another such approach. Thank you.