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 - 25

Lecture 25: Feature Engineering – IV

Hello everyone. Today is the last lecture in the series Feature Engineering which we are discussing as a part of content-based recommender system. So we are basically focusing right now on feature selection. In this context we have seen many methods for feature selection and today we are going to see few more. So to start with, after last class we were discussing about various information theoretic measures. So today which is pretty aligned with this probability theory is Gini index.

$$\text{Gini}(w) = 1 - \sum_{i=1}^{t} p_i(w)^2$$

This Gini index is suited for binary ratings, ordinal ratings and ratings which are distributed into small number of intervals. In case you have numeric ratings you can also discretize and get it intervalized. Now the ordering among the ratings is ignored in this scheme and each possible value of rating is treated as an instance of a categorical value. Now this formula is expressed in this way where w represents the feature and within that feature, that feature in a particular feature there will be a number of labels.

So basically over all the labels or total number of possible values you will be taking the probability. And it is to be understood that this value will always lie in the range 0 to 1 minus 1 upon t which is a very small value and smaller the value greater it indicates the discriminating power. So with respect to this response of a variable like let us find out the Gini index. So this Gini index turns out to be 1 minus 9 by 14 and 5 by 14 and taking the square of each of these. Now where from this value is coming? So if you look at this variable like in how many places it is no.

In 5 places it is no. So 5 no and we have 9 yes. So probability of getting yes is 9 by 14 and probability of getting no is 5 by 14. So this is yes, this is no. Now let us try finding out how do we make feature selection. So now look at this. Our whole aim of talking about this Gini index is finding out what is the best set of features that we should be considering along with this response variable. As I have told you with this Gini index we will be making the binary split and find out the best binary split. So if we are considering this attribute called language, this language has 3 different levels English Hindi and author. Now if we make the split making Hindi and author together and English

separately, how do we compute the Gini value for this particular split? Now how many English are here? There are 4 English 1, 2, 3 and 4.

$$Gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459.$$

$$Gini_{lang \in Hindi, other} =$$

$$= \frac{10}{14} Gini(D_1) + \frac{4}{14} Gini(D_2)$$

$$= \frac{10}{14}\left(1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right) + \frac{4}{14}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right)$$

$$= 0.443$$

$$= Gini_{lang \in English}$$

And out of these 4 we have 2 no's, 1 yes over here and another yes here. So this part is 2 by 4 for yes, 2 by 4 for no and total 4 out of this 14 number of data. Now coming to the other partition which takes together Hindi and authors, how many yes? 1, 2, 3, 4, 5, 6, 7. So 7 yes and 3 no. So this is taking English in one side and this is taking Hindi and authors in other side.

So the total for this particular partition it turns out to be this value. So similarly how many partitions, how many binary partitions you can make taking Hindi authors English 2 partitions, this is one and English other, then taking Hindi English together and authors is another. So similarly many partitions we can make 3C2 number of combinations. So for all combinations we find and the best split we consider. Similarly, for other attributes also we find out the best split.

Now this Gini value wherever this Gini value is lowest we choose it as the first feature to be considered along with this response variable. Then the next best with the next I mean the value which is just above the lowest that we choose. So that is how we select. Let us say we have to select only 3 features, then top lower 3 will be selecting. Next one is chi-square statistics.

This chi-square statistic can be computed by treating the co-occurrence between the feature and a class by making a contingency table. Now this chi-square measures, this chi-square statistic measures the normalized deviation between observed and expected value across the cells in this contingency table. Larger the chi-square value the more

likely the variables are related. Since one of them can be dropped. So basically for categorical variables we will be able to now find out the correlation in some sense.

Now if the attribute is independent of the class attribute that is low chi-square value it may be dropped. This is the formula and now we will see how to realize this formula. Look at this suppose our example has total 1500 data points. Of course now in the small part that I have shown we have only 14, but assume that there are total 15. And out of that taking the variable award and finding its chi-square contingency table with respect to the response variable like whether the person likes that movie or not like can take value yes and no, award can take value yes and no.

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

So, how many total yes? This many number of this is the actual observation. And you take the row sum and you take the column sum here this is the row sum 250, 200, 450. Similarly, column sum 250, 50, 300. So, that is how you take the total across rows and columns of the actual observation. Then this actual observation gives this value O.

Now what is this expected observation? Expected observation you will be getting by calculating by looking at this sum and calculating the probabilities. Now if we have total 1500 observations and 300 are like like where like is equal to yes the probability of getting this is 1 by 5. So, similarly probability of getting no is 4 by 5. Now look at this 1 by 5, 4 by 5 it becoming probabilities 1. Similarly, column wise also you take 450 by 450 how do you get 250 and 200, 450, 450 this is 3 by 10 and this is 7 by 10.

So, now to find out the expected value assuming that they I mean the their joint probability we will consider. So, 1 by 5 which is coming from here 1 by 5 and 3 by 10, 3

by 10 is coming from here 3 by 10. So, this 1 by 5 and 3 by 10 multiplied together with 15 will give us the expected value. So, this expected value actual value minus expected value square divided by the expected value and this is summed over all the fields all the cells in this contingency table. Now here we have only binary variables.

So, therefore, we have total 4 number of cells, but it is not limited to only binary variables we can have more number of values. For example, in case of in case of director there were 3 director in case of language there were 3 options. So, in that case it can be if we are computing between these 2 then it will be a 3 cross 3 situation. So, now once we find it out chi square value then we look for the best attributes which if an attribute is independent of the class chi square value is going to be low.

$$\chi^2 = \sum_{i=1}^{p} \frac{(O_i - E_i)^2}{E_i}$$

So, it may be dropped. So, with high chi square value whichever for whichever attribute for each of the attribute will be making one such table and for which so ever attribute these values are very high. So, which which means it in turn it says that there is a good relationship between this particular variable, particular feature and the class variable we choose that. So, next one is your wrapper based approach. This wrapper based approaches are widely adopted in the sense this particular while all others are kind of heuristics here you actually integrate this approach with the algorithm with the learning algorithm that you are using. So, what do you do? You first choose the learning algorithm.

So, learning algorithm or the classification algorithm in our case becomes the input to this wrapper class. Then what you do? You try taking subset of features and use it along with the class variable in the learning algorithm. So, so basically you consider all combination of the features taking 1 at a time, 2 at a time, 3 at a time and so on. So, if you have total n number of features then n c 1, then plus n c 2, plus n c 3.

So, this way you continue. So, which means it is becoming computationally very heavy and once you continue how do you measure? There has to be some kind of metric using which you can decide which set of features can give you the best value. So, now, this can happen. This computational ah complexity can be managed into ways. So, forward subset selection and backward elimination. In case of forward selection, you start you start with an empty set.

Let us say you have some features f 1, f 2 up to f p number of features. So, you start with an empty set. This is the total feature and out of that you will be moving few to this. This will be finally, build your model. This you will be finally, using build your model.

	Like=Yes	Like=No	Sum (row)
Award = Yes	250 (**1500×1/5×3/10=90**)	200 **(360)**	450 **(prob=450/1500=3/10)**
Award = No	50 **(210)**	1000 **(840)**	1050 **(Prob=1050/1500=7/10)**
Sum(col.)	300 **(Prob=300/1500=1/5)**	1200 **(Prob=1200/1500=4/5)**	1500

(numbers in parenthesis are expected counts calculated based on the data distribution in the two attributes)

$$\chi^2 = \sum_{i=1}^{p} \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

Now what you do? First make this set empty, then one of this you choose. So, while choosing one of this you can choose any one, but typically if you would like to have some kind of prior knowledge on this probably out of this you will be choosing the one which is highly correlated with another. So, that the other one can be dropped automatically. So, let us say you choose some f i and you have your response variable. So, with respect to each of this when you did with with respect to your criteria you got the best combination with f i and r.

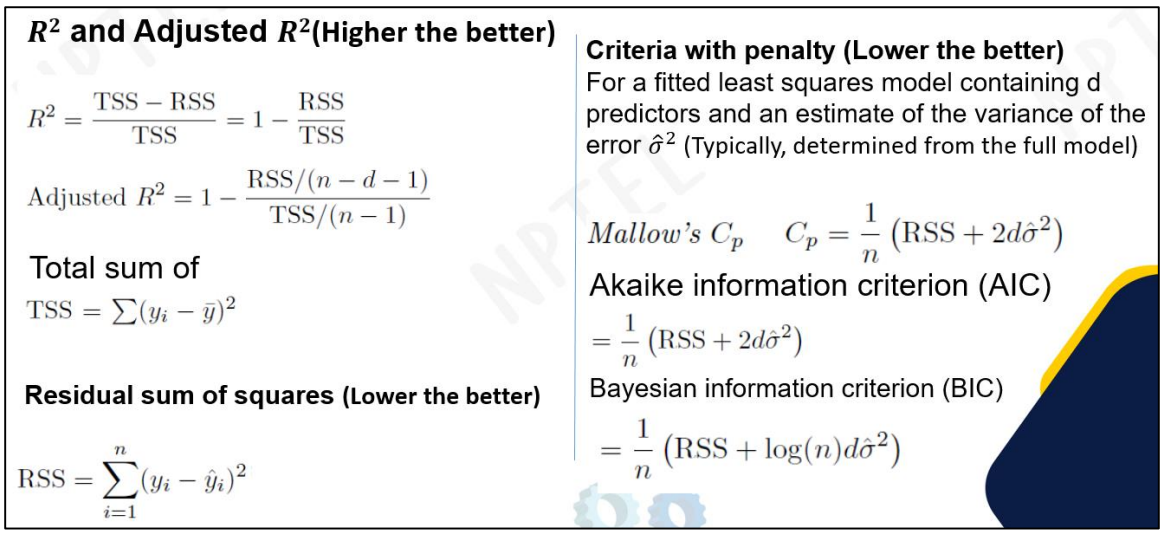
So, you decide on this. Now how many are remaining here? Here you have total p minus 1 number of remaining. So, with f i keeping, so rest p minus 1 you choose. Let us say you now choose f j. So, with all the j for 1 to p minus 1 you will be now finding taking these two together and finding the and building the model using r as the response variable. Now you will be getting certain accuracy or r square in case of regression you will be getting some such a sort of accuracy criteria or you can go for cross validation.

So, somehow you will be that metric you will try to find out with respect to each one. So, then what happens? You have found out the best combination and you have put it here. So, how many remains here? Now p minus 2. So, out of p minus 2 now let us say choose some f k. So, the k can take this p minus 2 values and with respect to r r r you find out the model ah.

I mean the model whatever you have used with respect to r as the response variable you run the model. Find out let us say some accuracy measure. Now that accuracy measure to which so ever f k that accuracy measure is the best choose that and you continue till the desired number of features are selected. Suppose you decide to keep 3 features.

So, now, at here itself you stop. In case of backward elimination just the opposite happen you take the full set of features and selectively eliminate keep on eliminating. Eliminate one test, eliminate second one test. So, that first for first elimination with each individualp you have tested p features. For second element a second elimination one once

you eliminate one for the second elimination p minus 1 remains, third elimination p minus 3 remains and so on.



So, you continue like this. A very well known wrapper based approach is your stepwise regression. Regression is something which we have already discussed in our earlier lecture when we are trying to introduce these concepts on machine learning. So, here also if you choose that your in the wrapper based approach your method is regression then what you perform for this process of wrapper based approach is called stepwise regression. So, in stepwise regression as we have discussed first you have to choose from the p number of predictors or the features. Now whichever gives you the best value in the next time you take for the two combinations.

So, this can again happen it is one wrapper based approach. So, there can be forward selection or backward elimination. So, in case of forward stepwise selection what you do for take all the p features now consider this has to be done in an iterative manner. So, out of all the I mean taking all the variables in the first iteration you build the model then check for the value where the highest the value of highest r square or smallest r s s or there can be other criteria what are these criteria we are going to see shortly. Now this will be first done with k for k equal for one variable then with that you will be having adding two variables a second variable to that you will be adding third variable and so on.

So, if you continue like this taking one feature then two features then three features and so on you will have many models and precisely p number of models and now with respect to certain validation criteria like this c p a i c b i c or adjusted r square or cross validation method you decide which is the best subset. So, in backward elimination again you start from p, p features then take p minus 1 last one feature. So, start from the entire set keep on reducing. So, assessing the accuracy of the model we can use many metrics specifically this we are talking about regression models and kind of metric that you

decide with other models as the input to this wrapper class can be may be different. So, here in this case we can have r square or adjusted r square then you can have a residual sum of square you can have certain criteria with certain penalty.

So, these you already are aware of this we discussed probably at the time of talking about the regression. So, r square is computed in this way and the value of this total sum of squares is this where this y cap is the average value of all the y and in case of residual sum of square this y i hat is the predicted value this is the average value of the training set this is the predicted value. So, now, we can think of adding some penalty as well for example, here some penalty term is added here some penalty is added here also some penalty is added. So, there are three criteria Mallows CP, Akai Kei information criteria or AIC or Bayesian information criteria or BIC. So, both all these three methods they have certain penalty term in this penalty term D is the total number of predictors considered out of P and this sigma square kr is the estimated variance of error typically it is determined from the full model, but usually it is assumed certain value and is given as input.

So, giving these inputs you find out these values. So, lower this value better is your model lower this value better is your model, but in this case higher the value better is the model. So, which means using them you can always test with different sets of features how these criteria are changing and you stop when you get the best such criteria among all the models. So, with this we stop and these are our references for Gini index we have considered this for stepwise regression we have followed this and this is our overall reference. These are our conclusions in this lecture we saw few new approaches for feature selection Gini index chi square statistics are two of them which are again kind of heuristics, but in the upper best approach you actually use the models as one of the input.

So, with respect to the model that you are finally, going to use you select the features in an iterative manner. Both Gini index and chi square statistics suits to binary ratings ordinal ratings and ratings which are distributed into small number of intervals and we can if in case of continuous we have to discretize the values. The basic strategy in wrapper based model is to iteratively refine a current set of features by success successively adding features to it. Thank you.