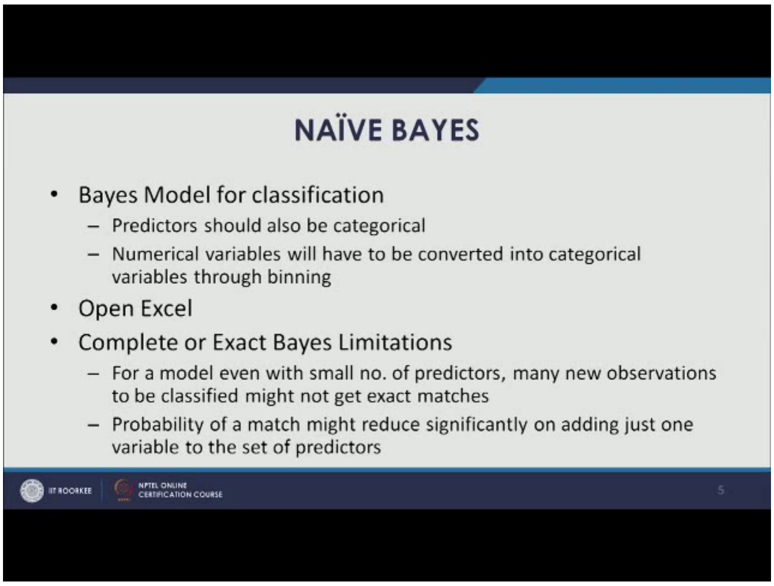
Business Analytics & Data Mining Modeling Using R
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture - 32
Naive Bayes - Part II

Welcome to the course Business Analytics and Data Mining Modeling Using R. So, in the previous lecture we started our discussion on a Naive Bayes and then we started our discussion on complete or exact Bayes. We also understood the different steps that would be required if we do our modeling following complete or exact Bayes. We also discussed if you discuss few more a points about class of interest and other things like limitation of complete or exact Bayes.

(Refer Slide Time: 01:04)

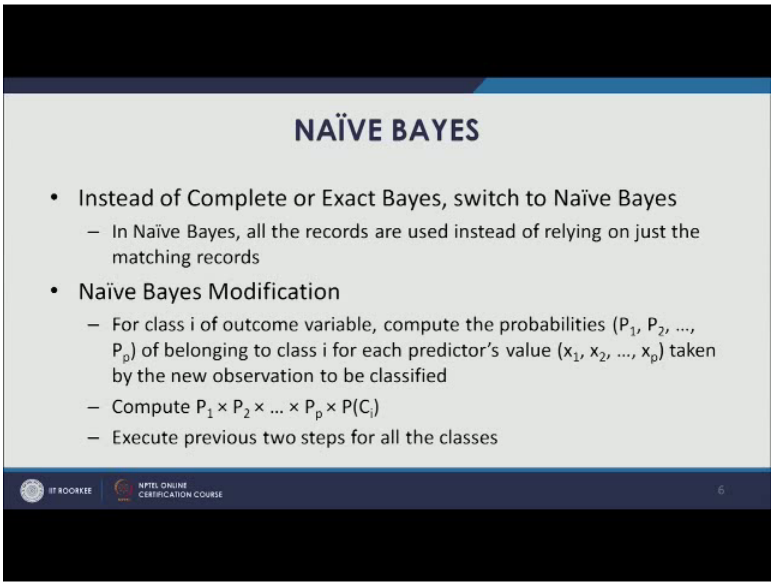


So, while discussing the limitation we talked about that incomplete or exact Bayes sometimes it is difficult to you know if we have even if we have a small number of predictors many new observations to be classified they might not get exact matches the values, for those values operators being taken by those observation those new observation which we want to classify. So, they might not get the exact matches in the data set in the training data set.

So, and also we talked about that if we include a one or two more predictors having few categories few classes, how that decreases the probability of finding a match. So, because of these limitations complete or exact Bayes app application of complete or exacts Bayesd in typical you know data mining modeling exercises that becomes difficult. So, what is the solution? So, that is the part that we are going to cover in this particular lecture.

So, instead of complete or exact Bayes we can actually switch to what is called naive Bayes the core discussion for this particular lecture, for this particular topic.

(Refer Slide Time: 02:19)



So, therefore, we should we should switch to Naïve Bayes we will talk about what happens in Naïve Bayes. So, in Naïve Bayes all the records are used instead of relying on just the matching records.

So, as we talked about in the complete or exact Bayes for a new observation that we want to classify. So, the values of different predictors that is being taken by that new observation all those values have to be matched with the training a partition records and the matching records have to be identified. But as we talked about that sometimes that may not be possible. But in Naïve Bayes, we do not actually a just rely on the matching records rather than the procedure or steps that we take they incorporate the entire dataset while performing Bayesian calculation Bayesian probability values. So, that is of course, with some relaxation, with some assumptions that we you know that we incorporate in Naïve Bayes. So, let us talk about the modification that we do on that we perform on complete our exact Bayes to convert it into a Naïve Bayes scenario.

So, let us discuss the Naïve Bayes modification. So, for example, if there are m classes m classes of outcome variable then let us take a particular class for example, per class i of outcome variable. So, these are the steps that we are talking about these steps for Naïve Bayes and then we will analyze how they are different from complete all exact Bayes. So, first step being that for class i of outcome variable there could be m classes. So, class i is one of those m classes. So, for class i of outcome variable we compute the probabilities P 1, P 2 and up to P p that is for we have we have P predictors.

So, therefore, we are computing P probabilities. So, we compute the probabilities of belonging to class i for each predictors value. So, value could be x 1, x 2, x p, taken by the new observation to be classified. So, the values for the new observation predictors values for the new observation. So, for each of those values we compute the probabilities of that per turn record and having that particular value for a particular predictor of belonging to class i. So, for class i we do compute these probability in the first step.

So, for example, if there is a predictor x 1, there is a predictor x 1 and it has three classes and one of those classes is actually the class of one of those classes is actually the value of new observation for that predictors. So, for that value we try to compute the probabilities of that new observation having that value and the probabilities are belonging to class i with respect to that value. So, that would be the P 1.

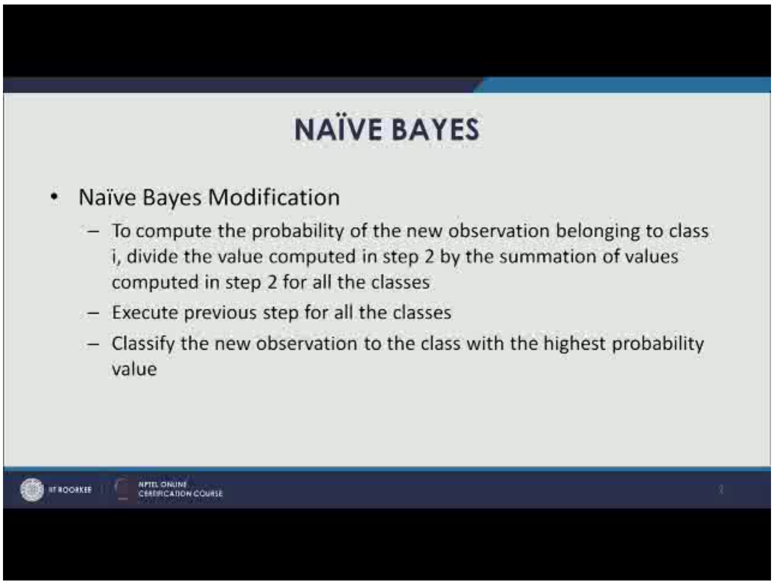
Similarly for the second predictor x 2 it will also have some value, but it being a categorical variable as we talked about in the previous lecture that all the variables are going to be whether outcome variable or the predictor wherever they are going to be the categorical variable. So, they will have many classes. So, therefore, for second predictor as well x 2 there are going to be classes different classes of this predictor and one of those would be taken by this not new observation and for that particular value for new observation will try to find out the probability of that particular value belonging to class i. So, in this fashion for each of the predictors values, each of the predictors values x 1, x 2 to x P we will try to compute the probability of these values belonging to class i.

So, once this is done once we have computed these values P 1, P 2 up to P p for the P up the predictor then we multiply these values with each other as you can see in this step next step compute P 1 multiplied by P 2 and up to the multiplication goes up to a P p and then we multiply it with the proportion of records that are belonging to this class i. So, out of all the records that we might have, how many records out of the total records actually belong to a class i. So, that proportion we will multiply using my, but we will multiply that proportion into this particular value right. So, we will have to compute P 1 into P 2 into you know multiply up to P p and then the proportion P of C i probability of C i.

So, now, these two steps as expressed in the third step. So, these two steps have to be repeated for all the classes. So, we talked about for a specific class i, but these steps. So, there could be m classes in an outcome variable. So, the m classes for each of those classes C 1 to C 2 to C 3 up to C m for each of those classes we will have to do, we will have to do this exercise step 1 and step 2 for each of these steps we will have to compute and finally, arrive at the values that we have in this step number 2 for all the classes. Once this is done, once we have done these computations for all the classes with respect to the new observations values then we come to the next step.

Now, to compute the probability of the new observation belonging to class i divide the value computed in step 2 right. So, all those probabilities multiplied by the proportion for that class. So, we divide that value by the summation of values computed in step 2 for all the classes. So, we talked about in the step 3, that this step 1 and 2 have to be repeated for all the classes. So the values that we get for different classes we will have to sum those values and this value will become the denominator and will you divide the value computed in step 2 for class i with this summed up value. So, that will give us the probability of new observation belonging to class i.

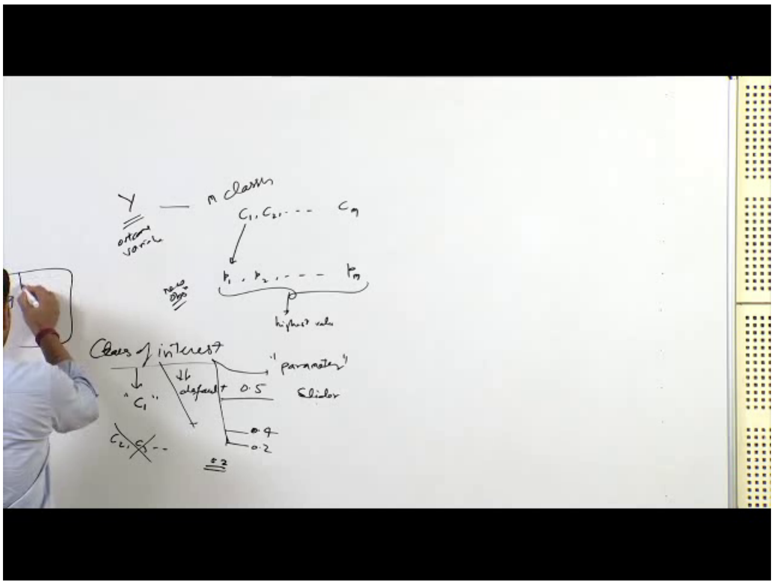
(Refer Slide Time: 09:41)



Now, in the next step we will execute what we have done for class i for the remaining classes of the outcome variable. So, for all other classes also will have to perform this. So, we from this you would see we will have to do lots of probability computation we will have to compute lots of probabilities and then followed by other multiplication or division operations and that has to be done for all the classes.

So, for all the classes once we exude this then will have the probability value, probability values of the new observation belonging to each class of the outcome variable. So, if there are m classes. So, if there are m classes in the outcome variable.

(Refer Slide Time: 10:55)

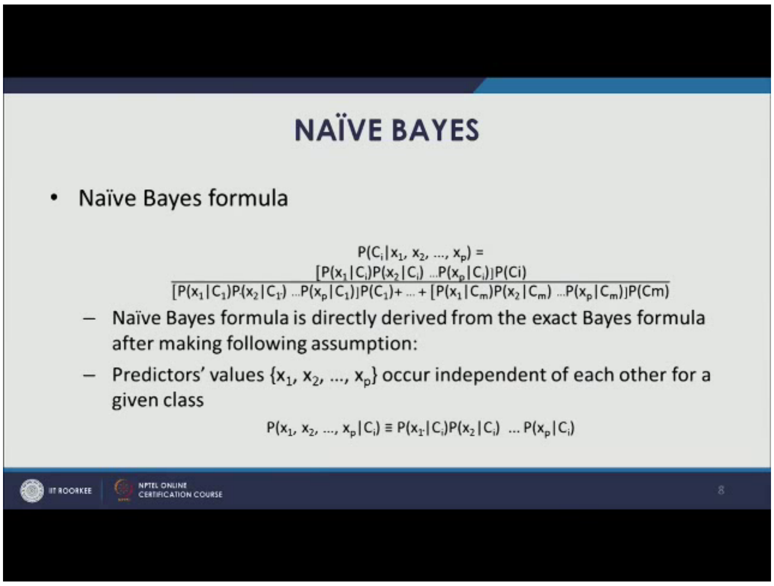


If our outcome variable is having m classes C 1 to C 2 to C m. So, once we have executed the probability value for that new observation P 1, P 2 for all these classes. So, once we have this step then we are going to classify the new observation to the class with the highest probability values. So, among these values that we have computed following 4 or 5 steps right then will find out the highest value and the class of that highest value highest probability value that would be assigned to the new observation. So, new observation would be classified to that particular class. So, you can see most probable class method that is being applied, so class having the highest probability that would be assigned.

So, let us recap some of these steps. So, what happens in Naïve Bayes modification is slightly different. Some calculations are different from complete or exact Bayes which will discuss later on. So, in the first step we compute the probabilities of belonging belonging to class i for each predictors value taken by the new observation to be classified right. So, as we discuss for each of those values x 1, x 2 we tend we try to find out the probabilities we try to compute the probabilities of those values belonging to class i. So, then we compute this expression of all these probabilities multiply then multiplied by the proportion or for that class and then this is executed for all this step. Once we have the values for all these classes for each of those class the value for that class and it would be divided by the sum of all those numbers and that is how we will get the probability for all the classes.

So, let us look at the Naïve Bayes formula. So, this is the formula we are going to compare how this is different from the complete or exact Bayes formula that we saw in the previous lecture.

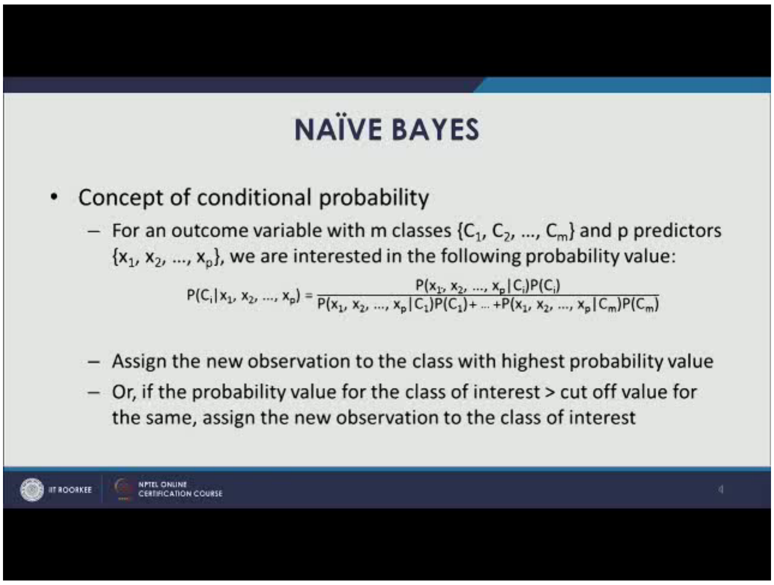
(Refer Slide Time: 13:33)



So, you can see P C i probability of a observation belonging to class C i given the predictors values that are x 1 x 2 and up to x p. So, P predictors are there for each of the predictors then observation is going to have some value. So, given those values what is the probability of that observation belonging to class i. So, this can be computed using this particular expression right.

So, here you can see that a probability of x observation belonging to x 1 given the this particular class C i and probability of probability of this predictor value x 2 belonging to probability of this particular x 2 given the class C i, then the probability conditional probability of x C x P with respect to say C i. So, all these numbers have been multiplied then the proportion is also multiplied proportion of records that belong to this particular class C i that is also multiplied. And then in the denominator you can see the same expression is there for all the class with the 5 expression is for C 1 and it is going to be followed by the same expression similar expression is going to be for the C 2 and similarly for other classes up to the last class that is the a m'th class. So, you can see the same is here.

(Refer Slide Time: 15:05)



Now, let us have a look at the compare the, let us have a look at the formula for compete or exact Bayes this was the formula for the complete exact Bayes.

So, you can see here. So, this is the conditional probability expression that we generally use. So, you can see probability of a particular observation belonging to C i given the predictors values and this is how it is expressed right. So, you can see one difference that is in the numerator as well as in the denominator the different expression that we have this particular value, this particular expression, this particular value probability of different predictors values right, probability of particular observation having these predictors values and given the class C i.

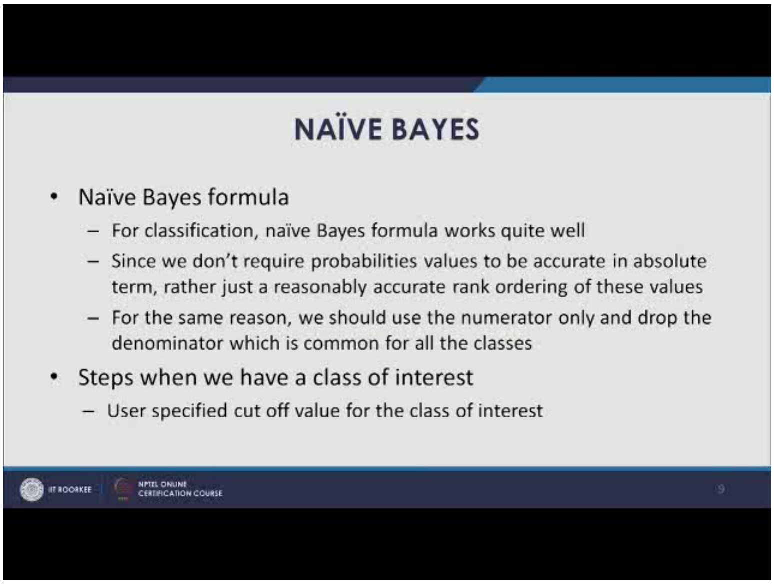
So, this particular equation has been changed to this multiplied product of different probabilities right. So, this is what is explained in the few more points. So, Naïve Bayes formula is directly derived from the exact Bayes formula after making following assumption. So, what is that assumption? That predictors values x 1 x 2 x P occur independent of each other for a given class. So, this is the approximation that we have done P, x 1 x 2 x P this was the probability that was that we were supposed to compute in the complete or exact Bayes and now this is the product of probabilities that we are now computing under Naïve Bayes right.

So, you can see for each of the observation once we say that these values occur independent of each other. So, this particular expression can be written in this for form well as you might have studied in your probability courses that if the events occur independently then the probability of those events happening R is going to be the product of the independent product of the probabilities of the independent event. So, this is the same thing that we have done because these value can occur you know independent of each other that is the assumption. So, this particular expression can be decomposed into this form.

So, from the conditional probability we have these two in way unconditional probabilities with respect to unconditional in the sense that the vectors values are going to be independent of each other. So, this is the approximation that has been done. However, these this particular assumption might not, sometimes it might not be true in practical situation a, but however, what has been seen that the results that we get by applying Naïve Bayes are quite good in comparison to other techniques despite you know some of the times when this assumption is not met.

So, let us move forward. So, as we discussed that for classification Naïve Bayes formula works quite well. Now, since another thing another important aspect is since we do not require probabilities values to be accurate and absolute term, rather just a reasonably accurate rank ordering of these values.

(Refer Slide Time: 18:46)



So, essentially as we discussed here that once we compute the probability values right then we need to find out the highest value and the particular new observation is going to be assigned to that class belonging to that highest probability value. So, therefore, we just need to find out which probability value is higher which are lower in a sense we need to just have the rank ordering of these probabilities value.

So, even if some of these probabilities values are not accurate, not accurate why we are talking about this because we have made one assumption in the Naive Bayes we have converted the conditional probability into the product of unconditional probabilities for the predictors values, right. So, because of that approximation because of that assumption that those predictors values are going to be independent of each other. So, the our probabilities value are not going to be accurate more often than not, but that is not going to create any problem for our modelling exercise reason being that we just need to find out the order of those values we need to just know the rank ordering. And that rank ordering is going to help us in finding the highest rank and, therefore, the class the new observation is going to be assigned to the sign to the class of the highest probability value just need to have the rank ordering. And that is why when we were saying that and even if that assumption of independent predictors values is not met in practice the Naïve Bayes model works quite well reason being that we just have now we just need accuracy in rank ordering and not in the actual probabilities values.

So, for the same reason that we have been discussing we should use the numerator only. So, and drop the denominator which is common for all the classes. So, let us go back to the formula that we discussed. So, this is the Naïve Bayes formula. So, if we are just arranging the rank cordoning of different probabilities values for different classes. So, you can see the numerator or this as per this expression a the Naïve Bayes formula the the denominator is going to be the same for all the classes because the it is the summation of probabilities value a computed for all the classes. So, this is going to be the same.

So, therefore, for the comparison of probabilities value to create the rank ordering this is not going to matter. So, we can just focus on the numerator and compare the numerator values and it still will end up the same rank ordering. So, that is another simplification in Naïve Bayes that can be performed. So, we will not have to compute all the values you know some of the calculation. We will not have to perform which are related to denominator and we can just focus on the numerator and the computation that are required for the same.

Now, till now what we have been discussing is mainly when we try to minimize the overall classification error or try to maximize the overall classification accuracy. So, as we have been discussing for other techniques multiple linear regression or k-NN specifically k-NN because there we discussed mainly about the classification tasks. So, when we have a class of interest right. So, how our steps will are actually going to change.

So, the earlier steps were actually the typical scenario which is that we are looking to minimize the overall error. So, when we have a class of interest, we want to identify more records belonging to the class of interest even if it comes at the expense of a miss identifying or misclassifying some of the records belonging to other classes right where the focusing is on one particular class of interest. For example, the financial reporting a example that we discussed in the previous lecture of Naïve Bayes we talked about that the an auditor and accounting form would be interested in finding the financial reports which might be fraudulent and not in the reports not he might the odd auditor or the firm might not be interested in you know identifying the truthful reports because they would like to a do more serious scrutiny of the financial repo reports which might be fraudulent.

So, therefore, they would like to you know identify more of those reports which could be the candidate for serious scrutiny which could be the fraudulent reports and therefore, it would be important for them to identify those reports those statements which could be fraudulent even if it comes at the expense of misclassifying some of the truthful reports as fraudulent reports.

So, let us discuss steps when we have a class of interest. So, a first step is typically going to be a specifying a cutoff value for the class of interest. So, this cutoff value as we have been talking about in the previous lectures as well that this cutoff value when we have a specific class of interest. So, the cutoff value, the default cutoff value is typically 0.5 when we have a class of interest. So, we are interested in identifying more of particular class members and other classes that could be there we are not too much focused on them. So, therefore, we would like to identify more records of this class as we have discussed. So, this particular a cutoff value can be treated as a parameter of the model, this can be treated as a parameter of the model or slider right. So, we can slide this value this cutoff value. So, it can sometimes be 0.2, it can sometimes be 0.4, the main idea is to identify more of to identify more up class one records right. So, if the there are many classes of course, the probability of a record belonging to class one that is class of interest is going to be slightly on the lower side.

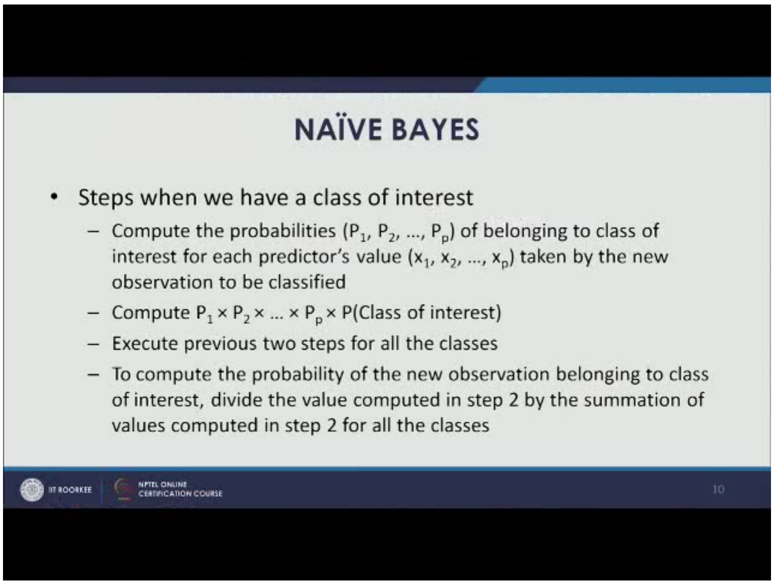
So, if we change our cutoff value that we are supposed to specify to let us say 0.2. So, any value that you know any estimated probability value for a particular new observation if it is more than 0.2 then that particular record is going to be classified as belonging to class of interest that is seem otherwise not; that means, it would be belonging to one of the other classes right. So, this is the typical idea. So, we would like to specify a cutoff value which can help us in identifying a more of the records belonging to class one.

So, once we do our modeling exercise once we have estimated the probabilities values for different observations right. So, once for different observations, once we have the probabilities value of belonging to a class one right. So, we can always slide the cutoff value and in that fashion our results of all the classification matrix they will change as we slide the cutoff value and therefore, we can find the appropriate cutoff value which can help us in identifying more of class one records more of class of interest records.

So, as a first step we have to specify user specified cutoff value. So, one of the when we discussed the performance when we were discussing performance matrix then also we had done one particular exercise using excel where we were actually changing the cutoff value and we had saw that the results were in the classification matrix they were also changing. So, that is the particular exercise I am referring to. So, that kind of exercise can help you and finding we also created a one way variable table to have a different cutoff values and the respective the classification error numbers and accuracy number. So, that kind of exercise can help us in specifying in establishing a cut off value for the class of interest.

So, once this is done we move to step number 2. So, this is quite similar to regular scenario step.

(Refer Slide Time: 28:40)



So, we compute the probabilities P 1 to P p of belonging to class of interest for each predictors value that could be x 1 to x 2 to up to x P taken by the new observation to be classified. So, as discussed in the regular scenario the new observation will have some values for different predictors. So, they could be x 1, x 2, x p. So, for each predator the value could be some class. So, for that particular class we have to find out the probability of that class belonging to class of interest.

So, once these probabilities are computed then we can move to the next step where we multiply all these probabilities P 1 to P 2 to P p. So, that is with respect to each of the predictors value for the new observation and then we multiply it by the proportion of the records a belonging to class of interest so that is I think what the probability of a recall belonging to class of interest. So, once this is done, we will have this particular value for the class of interest.

Now, will have to execute a these two steps for all the classes right. So, especially this is required if we want to compute, if we are not just focusing on the numerator and we want to compute the value, the Naïve Bayes value for the probabilities. So, in that case this term number three would be required for all the classes. And even if we are not using the denominator even then for the comparison purpose, we would be when we will be required to compute these values especially in the regular scenario not so much in the class of interest. So, let us move to the next step that is to compute the probability of the new observation belonging to class of interest we divide the value computed in step 2 by this summation of values computed in step 2 for all the classes right.

So, in this step 3 we said about computing the step 2 value for all the classes. Now, all those values would be summed up and then that would be used as the denominator to divide the value computed in step 2 for the class of observations that will give us the probability of new observation belonging to class i, so in this particular as I clarified that we have both the numerator and around and not just the numerator.

So, once this is done we can classify the new observation to the class of interest if computed probability value is greater than the cutoff value defined in step 1. So, once this probability value is known as we discussed there that for example, if the value is, if the value comes out to be 0.25 and the cutoff value is 0.2 then probably the particular new observation is going to be classified to the class of interest.

So, with this will stop here and in the next lecture will learn Naïve Bayes modeling through an exercise in excel and also will try to do an exercise in R.

Thank you.