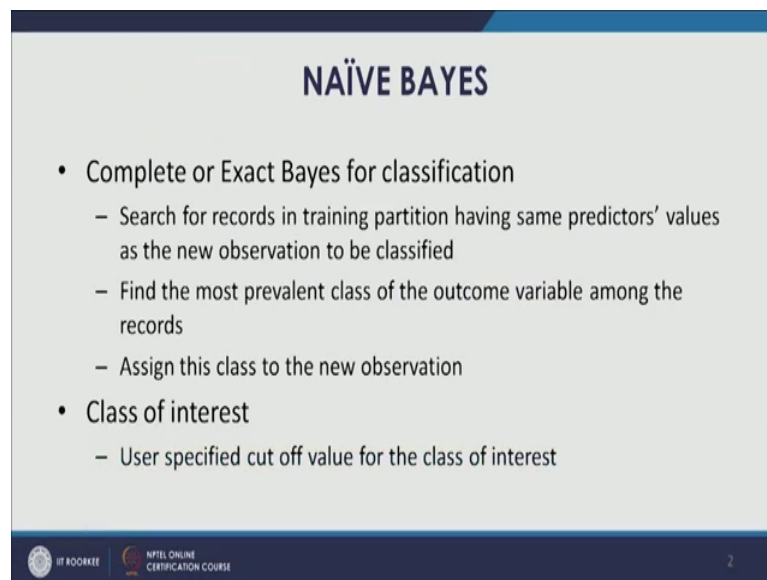


Business Analytics & Data Mining Modeling Using R
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture- 31
Naive-Bayes – Part I

Welcome to the course Business Analytics and Data Mining Modeling using R. So, in the previous lecture, we concluded our discussion on KNN, K nearest neighbors. So, in this particular lecture, we are going to start our discussion on Naive Bayes. So, let us start our discussion.

(Refer Slide Time: 00:41)



NAÏVE BAYES

- Complete or Exact Bayes for classification
 - Search for records in training partition having same predictors' values as the new observation to be classified
 - Find the most prevalent class of the outcome variable among the records
 - Assign this class to the new observation
- Class of interest
 - User specified cut off value for the class of interest

IFT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE

So, before we go into the Naive Bayes, let us discuss the complete or exact Bayes, so which is the basis for a Naive Bayes modelling. So, a complete or exact Bayes especially a specifically for classification, so let us get down the steps that would be required. So, you would see there is some similarity in these steps with a KNN algorithm. So, in the complete or exact Bayes for classification tasks, so first step is search for records in the training partition having same predictors values as the new observation to be classified.

So, in this case depending on the number of predictors that we have in our dataset 3, 4 predictors depending on the number of predictors for all those predictors the value that a new observation is having of those predators, in the training partition we have to find out all such record all such the records which are having the same value as the new

observation to be classified. So, the values should be same. So, there has to be there you know exact matches. So, the values a predictors values for the new observations right, so we have to find records in the training partition which are having exact values for all those predictors. So, that is the first steps with. So, we need to find out all such records in the training partition which are having the same predictors values, values as the naïve observation to be classified. So, this being the first step.

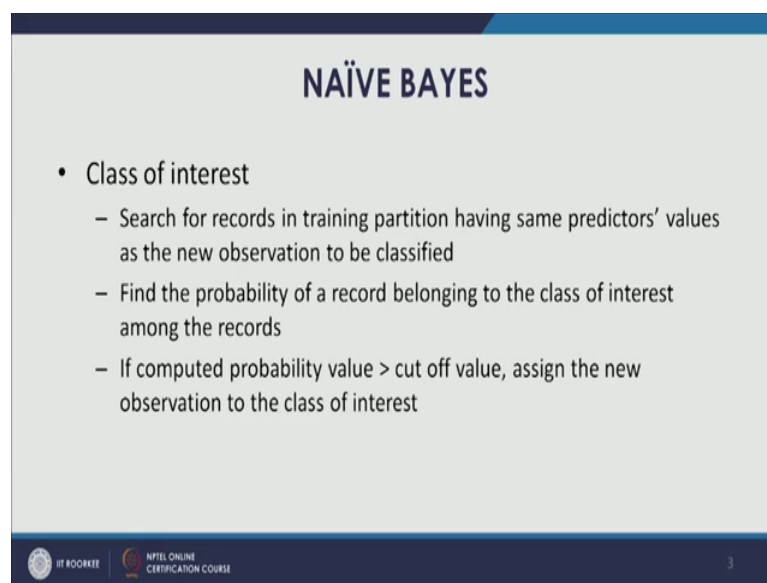
So, once we have the list of all such records, then we can find the most prevalent class we are discussing this for the classification task. So, therefore, the outcome variable of interest would actually have the classes. If it is two class scenario, we will have class one, class zero; otherwise m class scenario we have more than two classes. So, we will have to find the most prevalent class of the outcome variable among the records. So, the records that we have listed down in the step number one, first this step. So, among those records, we will have to find out the most prevalent class of the outcome variable.

Now, this particular class is in the third step as you can see this particular class is going to be assigned as the class of new observation. So, you can see there is a similarity between in terms of steps. So, the KNN approach. So, KNN approach also we talked about very simple step there we had. Finding you know doing the computations between the new observation and the observation in the training partition right then the searching for the k nearest neighbours and then looking for the most prevalent class and then assign it to the new observation. Similarly, here we look for the within the training partition, we look for the records which have the same values for the predictors values, there we used to compute the distance.

Now, here we are interested in finding the cause having the same values as the new observations. And again the next step is same and then finding most prevalent you know class within those identified in this step one and then assign that class with the new observation. Now, as we talked about the class of interest scenario which is different from the usual or typical scenario, where we focus on the overall misclassification error, overall error we would like to minimize that. And we do not have any preference for a particular class, so that is the typical scenario, but sometimes we might be interested in a particular class we would like to identify the members belonging to that particular class. So, we have a class of interest.

Then in that case as we have been talking about in the previous techniques in previous lectures as well that the user specified cut off value for the class of interest has to be established first. So, first we need to establish this particular cut off value, so that will depend on our expertise and the level of misclassification error that we would like to cut tolerate for other classes and still be able to identify more number of records belonging to the class of interest. And if the class of interest also happens to be a bit rare class then a the situation for the cut off value would also be more you know more expertise would actually be required to and get that value. So, we need to establish this value first.

(Refer Slide Time: 05:48)



The slide is titled "NAÏVE BAYES" in a bold, dark blue font. Below the title, there is a bulleted list of steps for classifying a new observation. The first bullet point is "Class of interest", followed by three sub-bullets: "Search for records in training partition having same predictors' values as the new observation to be classified", "Find the probability of a record belonging to the class of interest among the records", and "If computed probability value > cut off value, assign the new observation to the class of interest". At the bottom of the slide, there are logos for "IT ROOKIE" and "NPTEL ONLINE CERTIFICATION COURSE", and a small number "3" in the bottom right corner.

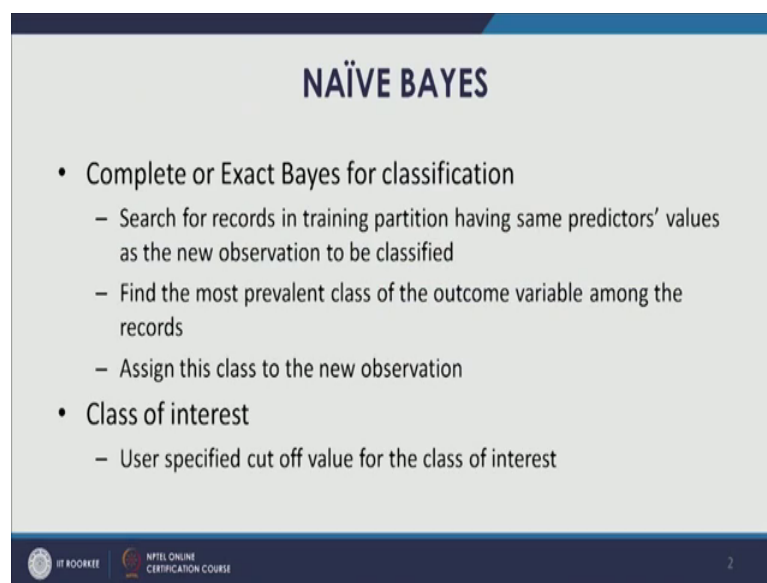
- Class of interest
 - Search for records in training partition having same predictors' values as the new observation to be classified
 - Find the probability of a record belonging to the class of interest among the records
 - If computed probability value > cut off value, assign the new observation to the class of interest

Then the next step is going to be search for records in training for partition having same predictor's values as the new observation to be classified. So, this is same as the in the previous approach where we go by the majority decision rule or the most prevalent class. So, this is step is same. So, we have to again find out the records in the training partition which are having the same predictor's values as the new observation. So, once this list is known to us, then we can find the probability of a record belonging to the class of interest among these records we have among the list that we have identified. So, within a list identified in the previous step, we can look for the probability of belonging of a record belonging to the class of interest.

Then as we a discussed in the KNN and even before that, this probability value computed probability value for the identified records, this will be compared with the cut

off value a user specified cut off value right from the step one. And if it is greater than that then the new observation is going to be assigned to the class of interest. So, these are the two approaches. One approach is where we have the equal class and or we do not have the class of interest then simply the most prevalent class. If we have a class of interest, then we have to specify the cut off value then we will have to also focus on computing the probability value of belonging to that for a class of interest. And then compare these two numbers and then do our assignment of a classification, assignment for classification. So, these are the steps in.

(Refer Slide Time: 07:37)



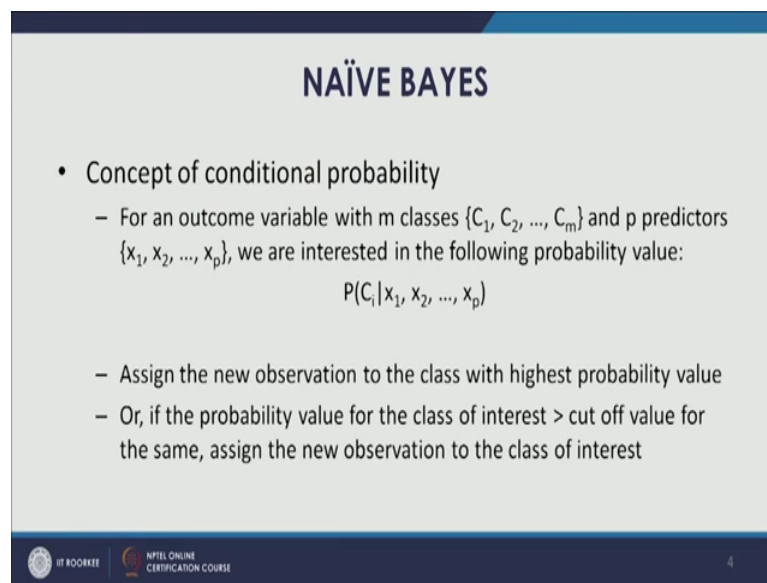
NAÏVE BAYES

- Complete or Exact Bayes for classification
 - Search for records in training partition having same predictors' values as the new observation to be classified
 - Find the most prevalent class of the outcome variable among the records
 - Assign this class to the new observation
- Class of interest
 - User specified cut off value for the class of interest

IT ROOKIE | NPTEL ONLINE CERTIFICATION COURSE | 2

So, these steps that we have talked about they are for the complete or exact Bayes. So, till now we have not started our discussion on Naive Bayes. So, we have started our discussion on the main principle main Bayes principle that is for complete or exact Bayes. Now, for the complete or exact Bayes the two scenarios that we did discussed the regular scenario and the class of interest scenario in both those scenario we have been the underlying probability related concept is the concept of conditional probability.

(Refer Slide Time: 08:11)



NAÏVE BAYES

- Concept of conditional probability
 - For an outcome variable with m classes $\{C_1, C_2, \dots, C_m\}$ and p predictors $\{x_1, x_2, \dots, x_p\}$, we are interested in the following probability value:
$$P(C_i | x_1, x_2, \dots, x_p)$$
 - Assign the new observation to the class with highest probability value
 - Or, if the probability value for the class of interest $>$ cut off value for the same, assign the new observation to the class of interest

III ROOKIEE | NPTEL ONLINE CERTIFICATION COURSE | 4

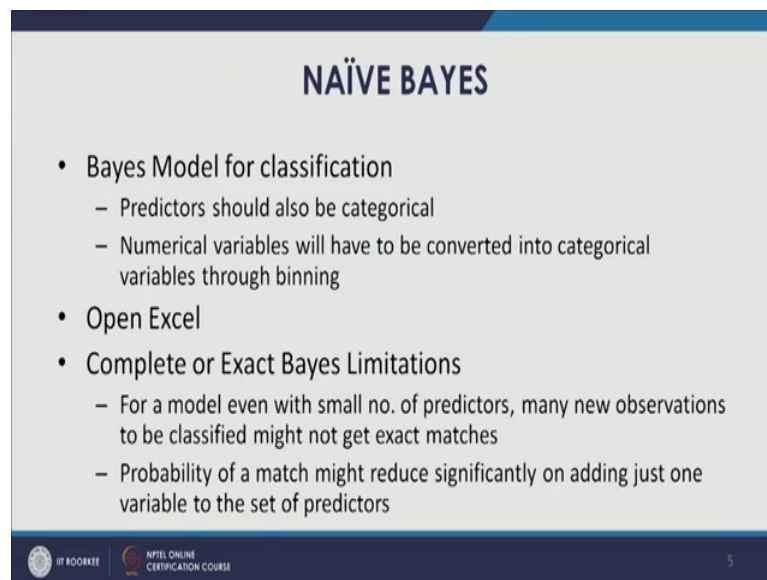
So, for an outcome variable with m classes, so C_1 to C_2 to up to C_m , and p predictors from x_1, x_2 to x_p , so we are interested in the following probability value. So, probability of a particular observation belonging to class i , given the predictors p predictors values x_1, x_2, x_p given these values. So, conditional on these values x_1, x_2, x_p , we would like find the probability of a record belonging to class C_i . So, we are interested in these conditional probabilities value in both the scenarios.

Now, in the scenario number one, which is the regular scenario typical scenario, where we are interested in minimizing the overall error all right. Then in that scenario, we would like to assign the new observation to the class with highest probability value right. In the other scenario, so the value would be computed using this conditional probabilities right. So, what these steps that we talked in there that we talked about you know before, so that has been expressed in this probability value format. So, we need to compute the probabilities value using this conditional probability for all the classes and given the predicted information. And once that is done new observation can be assigned to the class with the highest probability value in the regular scenario.

And we have the a class of interest scenario then if the probability value for the class of interest is greater than the cut off value for the for the same class, then assign the new observation to the class of interest. So, main idea being that the conditional probability when we talk about finding the records in the training partition which are having the

same values as the new observation, then essentially we are looking to compute this probability, because in the next step we would be finding the most prevalent class or the class having the highest probability value. So, actually the probability value would be the conditional probability and using this particular expression.

(Refer Slide Time: 10:42)



NAÏVE BAYES

- Bayes Model for classification
 - Predictors should also be categorical
 - Numerical variables will have to be converted into categorical variables through binning
- Open Excel
- Complete or Exact Bayes Limitations
 - For a model even with small no. of predictors, many new observations to be classified might not get exact matches
 - Probability of a match might reduce significantly on adding just one variable to the set of predictors

IT ROOKIE NPTEL ONLINE CERTIFICATION COURSE 5

Now, as you might have understood by now that the predictors that we have to that we are going to include in our Bayes modeling they have to be categorical. So, in the for example, in KNN or other classification other techniques other techniques which could be used for the classification and tasks, generally the predictors could also be the continuous variable. But in this particular algorithm in the Bayes model, the predictor should also be categorical it is not for the classification task task, not says the outcome variable has to be categorical even the predictors should also be categorical. If they are numerical variables then they will have to be converted into categorical variable through binning.

So, when if we are you going to use Naive Bayes modeling or Bayes modeling in general then all the variables whether the predictors or the outcome variable all of them they should be categorical, if they are continuous and they will have to be converted into categorical variable you through binning. So, this is an important difference. So, the techniques that we are going to cover in this course this is a one particular technique

which requires which had such condition which requires all the predictors to be categorical.

So, therefore, from this you will also understand that Bayes modeling cannot be used for prediction tasks. So, this is mainly it can be used it is used for the classification. So, even though theoretically we can apply Bayes model for prediction tasks, but it would be very difficult for us to find records in the training partition which would be having the same values as the new observation right especially for the prediction tasks. It is going to be very difficult to match the numeric values of a new observation with and records that are there in the training partition, so very difficult to get those matches.

And therefore, and this particular difficulty can increase multi fold as the number of predictors increase right. So, even if there is one mismatch that particular observation cannot be selected for further steps. So, therefore, it becomes impractical to apply Bayes model for prediction tasks. So, most of the discussion that we are going to do in this particular topic Naive Bayes and overall Bayes as well both, right now we are discussing the a complete or exact Bayes, so would be around the classification task. And therefore, the results the predictors or the outcome variable both have to be categorical variable or converted to categorical variables through binning. So, what will do we will do an exercise using excel. So, we will apply the complete or exact Bayes, and see what are the issues that we might encounter and how it can actually be applied on a real problem.

(Refer Slide Time: 14:03)

Complete or Exact Bayes				Prior Legal Trouble	Company Size	Status	Complete or Exact Bayes calculations			
	Prior Legal (K+1)	No Prior Legal (K+0)	Total	yes	small	T	P(F yes,small)	F	yes	small
Fraudulent (C1)	50	50	100	no	small	T	P(F yes,large)	F	yes	large
Truthful (C2)	180	720	900	no	large	T	P(F no,small)	F	no	small
Total	230	770	1000	no	large	T	P(F no,large)	F	no	large
P(fraudulent prior legal)	0.22			no	small	T	Naive Bayes calculations			
P(truthful prior legal)	0.78			yes	small	F	P(F yes,small)	F	yes	small
Most Probable class	truthful			yes	large	F	P(F yes,large)	F	yes	large
Cut off probability	fraudulent			no	large	F	P(F no,small)	F	no	small
method	fraudulent			yes	large	F	P(F no,large)	F	no	large
cutoff+2				yes	large	F	P(F no,large)	F	no	large
(siden)				yes	large	F	P(F no,large)	F	no	large

So, let us open this excel file. So, this particular example is about this is audit data. So, this is for the financial source on so many some funds are required to submit some of the financial document, some of these financial statements to regulatory bodies for further inspection all right. So, before that they are required to get their audit done with the accounting firms. So, for the accounting firms they have to apply use a lot of their human resources and other resources, systems and analytic solutions and software to analyze the financial statements, the financial reports submitted by their clients. And because they have a big incentive to find out the fraudulent statements, fraudulent reports; otherwise they are going to be penalized by the regulatory body. So, they are required to certify that those firms have all the legal reports, legalized reports. So, the responsibility lies with them. So, therefore, it is very important for them do find out the fraudulent reporting any erroneous reporting.

So, therefore, if a firm has some information because if they have loyal customers, loyal clients, so they might also have information about the previous such reporting previous such you know previous troubles that the legal troubles that might have occurred for different clients. So, the accounting firm might be interested in knowing that whether that information, whether the historical information about the auditing of these financial reports whether that can be used to find out the potential you know fraudulent reports, potential fraud fraudulent statements. And that can make their task of auditing much easier, because they can do more intense inquiry on a few identified records that few identified reports or statements right. So, where the fraudulent, the chances of the reports being fraudulent or more likely, so they would like to identify those reports where the chances of those reports being fraudulent are on the higher side.

So, let us say we have a 1000 such reports. So, this is sample size is 1000. And the variable is from the historical information for their clients the accounting firms clients that we have is whether that particular whether that particular client had a prior legal trouble or did not have prior legal trouble. Whether in previous years the reports submitted by those clients, whether and that was found to be problematic, and therefore, they had some legal trouble or they did not have any legal trouble. So, with that information, can we identify you know some of the fraudulent reports and all identify the truthful reports.

So, this classification matrix or this based on the data set also, we can produce this kind of summary table using a pivot table option that we have in excel. So, this kind of summary we can easily summarily we can easily generate using our data. So, you can see this side we have prior legal, this is the predictor, this is just one predictor that we are using here. And as we talked about if there are more number of predictors, as we talked about the chances of finding exact matches always go down right. So, here this is $x = 1$, so that means, prior legal trouble; x value a 0 this is no prior legal trouble. Then we have whether the report was fraudulent or truthful.

So, 50 fraudulent reports they had prior legal trouble, 50 fraudulent reports they did not have prior legal trouble, 180 truthful report they did not have a prior legal trouble, and 720 truthful reports they had no prior legal trouble. So, this is the hypothetical example that we have and then totals for all these scenarios are have also been done. So, now, from this, we can actually compute some of the conditional probability values that we talked about.

For example, probability of a particular record belonging to fraudulent class that is $c = 1$ and given the prior legal trouble that just one predictor that we have in this case prior legal trouble. Given that information probability of a particular record belonging to the fraudulent class would be 0.22 computed period by using these numbers. 50 or 50 times 50 reports have been fraudulent reports I also had prior legal trouble, so this is the number 50 and then divided by the 230 total reports which had. So, out of total reports that is 230 which had prior legal trouble 50 of those were also identified as fraudulent. So, conditional probability would be can be computed using these two numbers, this comes out to be 0.22.

If we want to compute the probability of a particular report being truthful given the information, we have a related to prior legal trouble. So, this also can be computed. So, out of all the records which had all the all the reports which had a prior legal trouble that is 230, 180 were a found to be truthful. So, the conditional probability would be come to computed using these numbers 180 divided by 230. So, this will give us the probability of a particular require belonging to truthful class given they had the prior legal trouble, this number comes out to be 0.78.

So, if we apply the most probable class method here right. So, most probable class method would say that among these so once those records as we talked about these steps, once those records having exact match matches have been identified, then we will have to compute the most prevalent class among those records among. So, we already have the information related to probability of different classes. So, whether put fraudulent or truthful, so we have 0.22 and 0.78. If we look at the higher number it is at being 0.78. So, if we apply the most probable last method then based on that 0.78 being the higher value, so the record new observation would be classified as the truthful report. So, most probable class method would lead to this conclusion.

So, this whole example is quite simple in this case because we are dealing with just one predictor information right. Now, if we apply a different approach, the cut off probability method so suppose we are so interested in one particular class. So, we have a class of interest. So, in this particular case, the class of interest is typically going to be the as we discussed about the problem typically is it is going to be the fraudulent report. So, we are more interested in identifying the fraudulent financial reports, because we like to do more intense, scrutiny of such reports as an accounting firm. So, therefore, we are more interested in identifying such reports. So, the class interest is the fraudulent class.

So, for this we will have to specify the cut off probability for this class of interest. So, let us say the cut off probability is 0.2 as indicated here. So, if this is the established you know usually specified value. So, any probability value that we compute following those steps finding the records from the training partition which are having the exact matches. So, this matches we do not have to perform in this example. And then the second one being identifying the most prevalent class right, so that we can do the probability values we have using the information that we have here.

For this example, so if the cut off value is 0.2, and if we compare it to compare this one with the computed probability value which is 0.22. So, you can see that the probability value is greater 0.22 is greater than 0.2. So, in that case, the new observation is going to be classified as the fraudulent class all right. So, you can see depending on the method, our answer might change, because the cut off value for a class of interest can actually be specified by the user because we are interested in one particular class right. In the typical regular case where we do not have, we would like to minimize the overall error. So, there

we would just go with the majority rule or the most prevalent class in that case as you can see the observation would be identified as the truthful.

For m class scenario the formula that we talked about the conditional probability formula that we talked about can be expressed using this particular expression. So, p probability of particular observation belonging to class C_i given the predicted information x_1 to x_p . This conditional probability can be computed using this. In the numerator, we have probability of x_1, x_2, \dots, x_p probability of those predictors value given particular class C_i and this multiplied by the probability of a particular record belonging to C_i . Then in the denominator we have a for every class for example, C_1 we have the same expression probability of a record belonging to this, but you know predictors values and given it belongs to class 1 and then multiplied by the probability of belonging into class one. And for all the classes that is how the denominator would be there.

So, this is the typical conditional probability formula that you might have studied in your previous degrees, in 10th or 12th or during your graduation. So, using this particular conditional probability, we can find out the exact probability value, and then that can be used to find out the prevalent class and then all can be used to compare with the cut off user specified cut off value in case of class of interest. So, you can see when we have just one predictor it is quite easy for us to perform these computations and then do your comparisons, and follow the steps, and therefore do your classification.

But what we are talking about is the finding exact matches right. For the new observation, we will have to find out the records, which are having the same predictors value. Now, this becomes quite difficult if the number of predictors increase right. So, for an example, if you know in a university we have to find out a person who is doing a B. Tech in a particular engineering let us say computer science and engineering. And then he might have also taken an elective on data analytics and then his CGPA would be some you know 7.8 or something or more than that a 7.8 it exactly if you know all that at that grade in that particular subject, because we are talking about all the predictors being categorical. So, and then whether that particular student being male or female, all belonging to a particular city or state.

So, as we keep on increasing the number of predictors and the values, so it would become more difficult for us to find exact matches as the number of a predictors even

though these predictors are categorical. So, they will have now classes they will have the labels right. So, even for a limited number of labels for a particular predictors as the number of predictors increase it will be very hard to find the exact matches. So, therefore, that becomes the practical limitation of applying complete or exact Bayes.

So, as you can see in this particular slide as well complete or exact Bayes limitation for a model even with small number of predictors many new observations to be classified might not get exact matches right. So, as we talked about with an example that even if we have a small number of predictors, we talked about only four or five predictors, the gender, the city or the hometown, gender, hometown, the program the undergraduate program that he might have he or she might have might be involved to. Then the particular course that he or she might have taken then the grades, grade in that particular course. So, four, five, six variables and even with these variables you would get the idea that it becomes quite difficult out of a you know thousands of students that might be in the campus with the specific details that I have given.

If I give the name of the city, if I give the engineering that he or she might be doing student might be doing, and then if I take an name of an elective course that he might be a studying, and a particular grade within that the for that course, the gender information. Then you would see we will end up with the you know out of so many students ten thousand students, we will end up with just a 10 or 5 students or 10 not even 10, because they have to come from the same city hardly two or three students.

So, the number of exact matches with just 5 or 6 variables they come down to such a small number. So, hard to find exact matches. And therefore, many probability value will not be to compute. Therefore, how are we are going to classify the new observation if there are no exact matches, we would not be able to classify because we need to find the exact matches and then within those matches then we need to find out the most prevalent class right. And even then if we are able to find the exact matches, and there are very few exact matches, then the classes might not be all the classes might not be covered all the probability values that could be there because of the few exact matches they might not remain meaningful right.

So, the idea is there should be a lot more matches, so that the values that we get they are slightly stable given the sample given the problem and so that our classification might be

more accurate, and will not be dependent on the data. And those results should be more stable, our classification results should be more stable. So, we need more matches fewer matches even if we find matches, if there are very few that could also be problematic.

The next point in the limitation is the probability of a match might reduce significantly on adding just one variable to the set of predictor. So, if we add just a one additional predictor in these set of predictors, and that one predictor even if it has just two classes right, so that and the those classes are you know the frequency of those classes that is equal in that case equally occurring you know with equal frequency classes occurring with equal frequency, then that would reduce the number of exact matches by the factor of two right, two classes by factor of two it will reduce the probability of finding a match. So, therefore, if we include a predictor and it has more than you know four or five classes and that would reduce the probability of finding exact matches Bayes you know factor of five right.

So, therefore, the complete or exact Bayes, so because of these limitations it is quite difficult to apply complete or exact Bayes for our modeling exercise. So, what is the solution, solution is the Naive-Bayes some assumption, some simplification of complete exact Bayes that we perform that allows us to apply our Bayes modeling on for different classification problems.

So, we will stop here and we will continue our discussion on Naive Bayes.

Thank you.