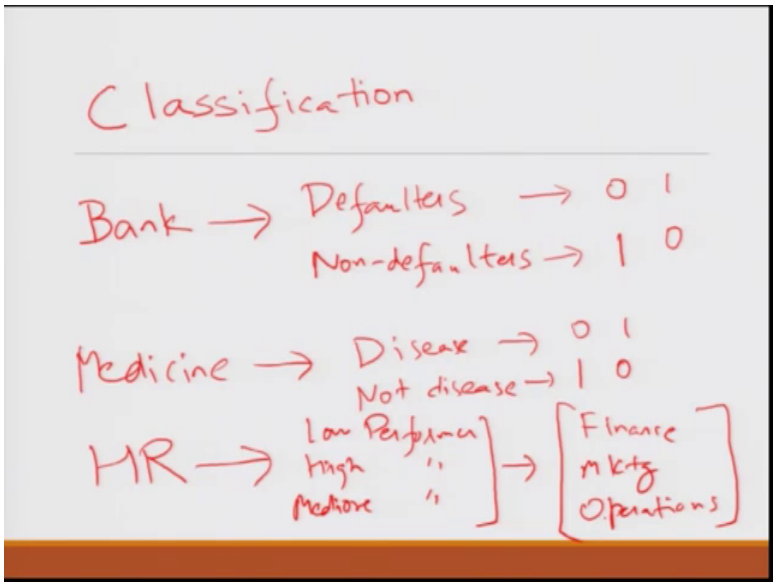
Practitioners Course in Descriptive, Predictive and Prescriptive Analytics

Prof. Deepu Philip
Dr. Amandeep Singh Oberoi
Mr. Sanjeev Newar
Department of Industrial and Management Engineering
Indian Institute of Technology, Kanpur
National Institute of Technology, Jalandhar

Lecture – 27
Machine Learning – (Part 4)

Welcome you all to another session on machine learning in our practitioner's course on analytics. We talked about different types of machine learning algorithms the philosophy the trips the traps. Let us now get into a quick crash course of the most important form of machine learning or rather the most widely used form of machine learning which is called the classification based algorithms.

(Refer Slide Time: 00:44)



So, classification just to quickly remind you is all about classifying as the word says. So, in a normal regression, our goal is to predict a particular value of the output in this case we want to categorize the output. So, for example, let us take case of a bank. So, this is 1 of the most popular uses of classification based machine learning algorithm. So, a bank wants to classify its customers into defaulters and non defaulters. So, bank is in the business of giving loans, it wants to predict which customers can default which customers have less likelihood to default. So, this is a classic or very popular case of

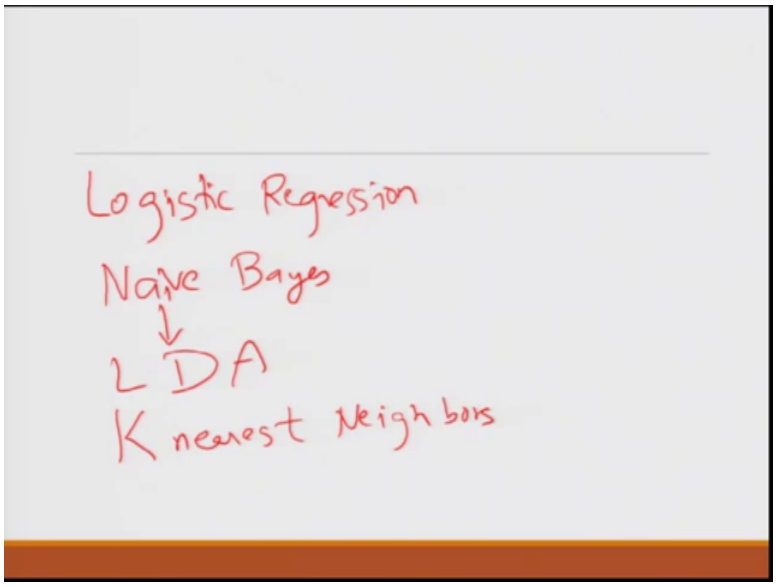
classification based algorithm, there can be similar for example, if you talk about medicine.

So, based on the different parameters, different vital statistics of a patient you want to predict whether the patient has a particular disease or not disease. So, these are typically binary classification, you label them as 1 0 or 0 1 depending upon your convenience, there can also be cases where you want to break them into multiple categories.

So, for example, you may want to based upon the say in case of hr based upon the different characteristics, different kinds of scores, you want to see to predict whether a person would be a low performer, high performer or a mediocre performer. And this particular case you can also use regression, but often this categorization can give you slightly better results or based upon different kinds of characteristics it wants to see whether a person is a better fit for finance, marketing or operations. So, what fits is profile better in terms of this potential performance. So, these kind of problems where the output variable is not a continuous variable it is a category.

So, these are the problems, which come in to the classification set of algorithms.

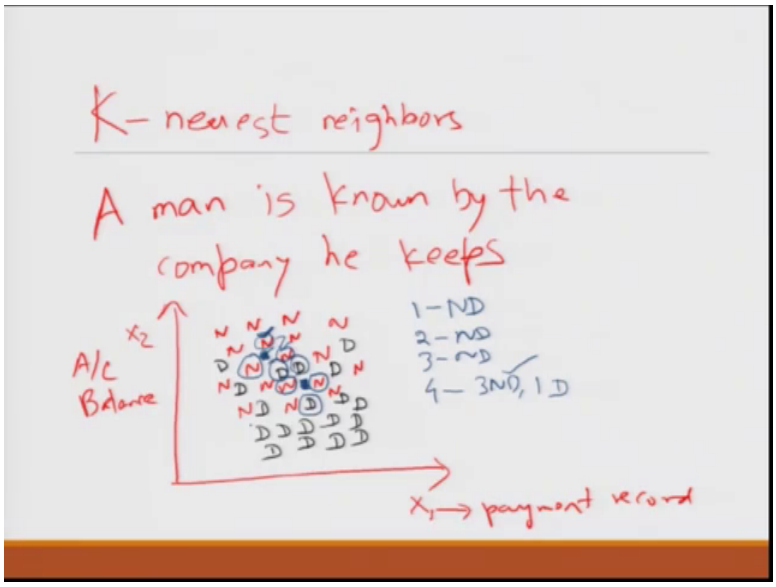
(Refer Slide Time: 03:34)



There are several methods, some of the most popular used methods are something called logistic regression, which is extension of the same principles as of normal regression or linear regression, multivariate regression or bivariate into problems where you have to

get them into different categories. Then there are these simpler, but yet to very powerful approaches Naïve Bayes and an extension of Naïve Bayes is what is generally called a linear discriminant analysis based algorithms and then you have the very classic or very intuitive K-nearest neighbors based methods.

(Refer Slide Time: 04:34)



So, let us have a quick look on each of these methods and then also understand what are the pit falls get into a bit detail of its let us start with what I mentioned at very last K-nearest neighbors. Now this is the most I would say intuitive or most natural way in which we also try to classify people around us or various things around us. So, what it says is you know the popular proverb that a man is known by the company he keeps.

So, basically what you do is, let us say draw 2 parameters let this 1 be x 1 this 1 be x 2 let this be any 2 parameters for example, if you are talking of a default that may be x 1 is a parameter like his past payment record whether he pays on time or not; x 2 may be his account balance.

So, let us not bother about what these exactly are because depending on the problem this will vary and then for example, we know certain people defaulted certain did not default. So, let us put the defaults as let us classify them as n, let us mark them as n. So, for example, I am just drawing.

So, let these are the some point some people who did not default and let us assume that these are people who defaulted. So, now, for example, assume that you get an new data point. So, you got a data point here for example, let this be a data point. So, you want to predict whether this fellow is he a defaulter or a non defaulter. So, what K-nearest neighbor says is that look at the people who are around him who are closest people around him.

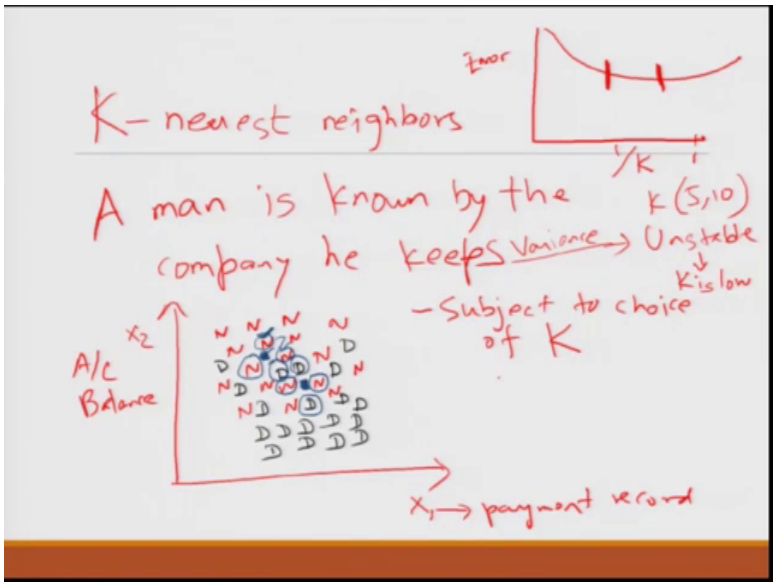
So, you can very easily say that, if I choose his closest neighbor probably this fellow is the closest neighbor if I choose two closest neighbor then probably this and these two are closest neighbor. So, this may be a neighbor, this may be a neighbor if I have to choose three neighbors these three may be neighbors, if I have to choose four neighbors then this fellow also becomes also a neighbor.

So, if I decide that I have I make a rule that, I will classify a person as per his closest neighbor. In that case I will choose this fellow and say that this record is a non-defaulter; however, if I say I will choose 2 defaults 2 closest neighbors again for one he is a non-defaulter, for two he is a non-defaulter if I choose three again here is a non-defaulter if I choose four points then three points say that he is a non-defaulter, one point says he is a defaulter you do a majority vote and then you still you say that he is a non-defaulter.

Similarly if I had another point which was somewhere here in this case probably if this was the closest fellow, then we would say he is a defaulter. If next closest was this and among the two its kind of equal you are in a tie, if you choose three points then probably again this fellow breaks the tie, if you choose four points then again its a tie. So, what we are doing in a way is that, you are pooling how many of the nearest neighbors are defaulters or non defaulters and we go by the majority vote.

So, typically very intuitive and it works perfectly fine in many simple cases. In fact, a lot of the machine learning algorithms, which are used in many recommendation based systems in shopping cards and even some sentiment based analysis they do this K-nearest neighbor methods and if you look at the results also surprisingly the results come out to be really good in many cases. In fact, they beat some of the most advanced algorithms, but there are there are few riders, the first rider is that this is very very subject to this very. So, it is a function of the it is subject to choice of K. K means the number of neighbors to choose to make the decision.

(Refer Slide Time: 10:27)

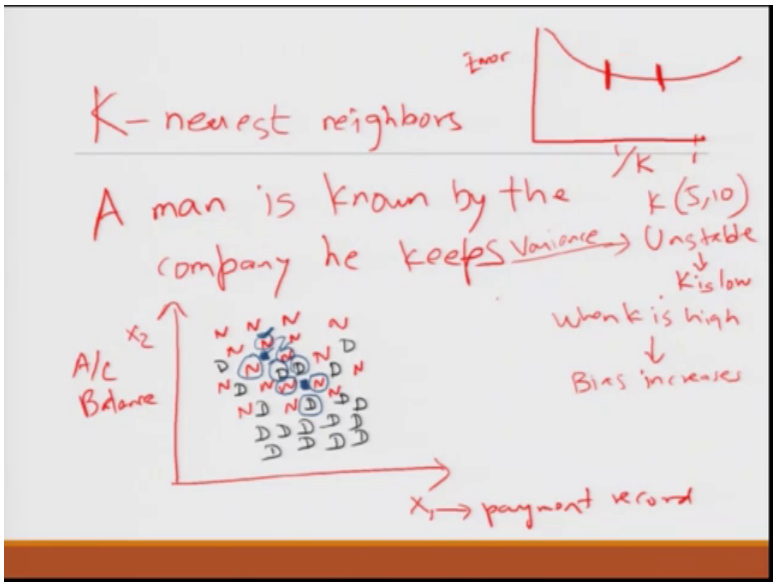


And if you look at a typical profile, if I have to make a graph out here; so, let me plot error out here, the mistakes I make if I classify using this let me put 1 by K here. So, 1 by K here means that as K increases am going close to the origin. So, the further most points are those the high the lowest can be one. So, may be this extreme is 1 and other values are beneath it and if we plot it, we see the plot is something like that.

So, as we increase the number of ks if you increase. So, somewhere between this to this, you find that you get the least error beyond it if you take more number of neighbors, then what happens is typically it is almost sampling the entire dataset and hence your accuracy reduces. In the initial part if you choose the less number of K, then what happens is the nearest neighbor becomes a very strong influences.

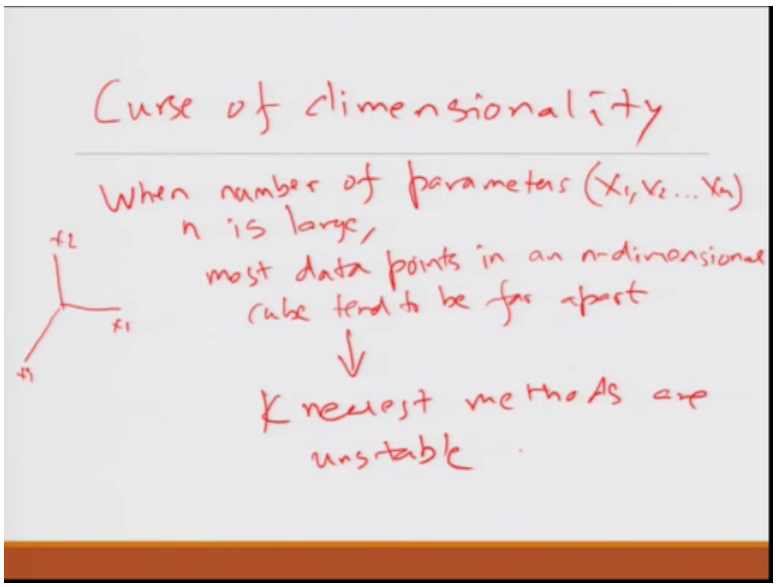
So, if there have been some odd points here and there, in that case you may still get error; so, what happens is there is an optimal typically the range between which it works beautiful in most cases. So, typically it is a thumb rule that if K is between 5 and 10, then it works good. If it is below this then the model becomes slightly unstable the variance is high if I have to talk in terms of the bias variance discussion we had last time, it is too much adapted to the data set could a new data it may not work fine if K is this is when K is low.

(Refer Slide Time: 12:56)



When K is when K is high then bias increases. So, it means you are sampling the entire dataset and hence your accuracy reduces. So, keep in mind that it is it is very much subject to and something dataset is slightly more complex the K-nearest neighbor may not also work very great. So, you have to keep in mind the between 5 to 10 years typically it works good.

(Refer Slide Time: 13:43)



But there is another rider along with it and that is the curse of dimensionality. So, when the number of parameters in this case we had worked with only 2 parameters, but if when

number of parameters I mean x 1 x 2 xn. So, if n is large, then it can be shown that most data points in an n dimensional cube. So, if you assume that all data points are part of the cube.

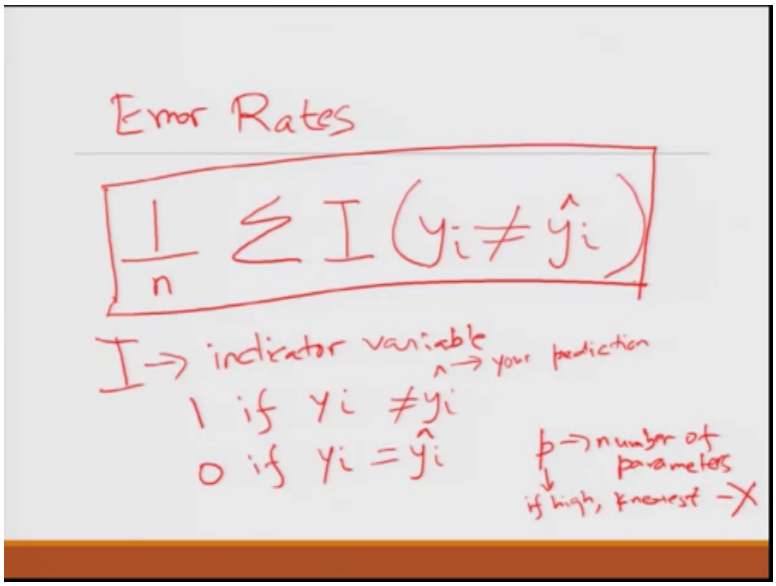
So, a three dimension is like this is x 1, x 2, x 3. So, if I make an n dimensional and I try to plot and if that be considered kind of a cube. So, most of most data points and n dimensional cube tend to be far apart. So, what happens, it is becomes very difficult to find a close neighbors. So, what happens in most of the points are actually quite very much far apart from each other and because they are far apart from each other, the entire notion of having K-nearest neighbors, does not make much sense.

So, it is like you are being in Kanpur and your closest neighbor is in say Norway the other closest neighbor is in Australia. So, Australia and Norway itself are so far away from Kanpur they does not make sense to even call them as neighbors, but yeah from a technical stand point, yes you can still come out with some results, but they are not actually neighbor in the strictest sense.

So, if you have unless if you have very huge number of data points typically finding this neighbors and actually classifying them as neighbor itself becomes a question mark. So, for when the number of parameters is large because they tend to be far apart, it means that K-nearest methods are unstable.

So, you do not use these methods in this case. So, typically what you do is when you use this methods we should also consider what is called error rates.

(Refer Slide Time: 16:29)



So, when you use these methods there are always misclassifications that happen. So, there will be the there will be data points, for which your model will classify them wrongly. So, what you do is, you calculate the metric which is. So, do not be scared about this complex term.

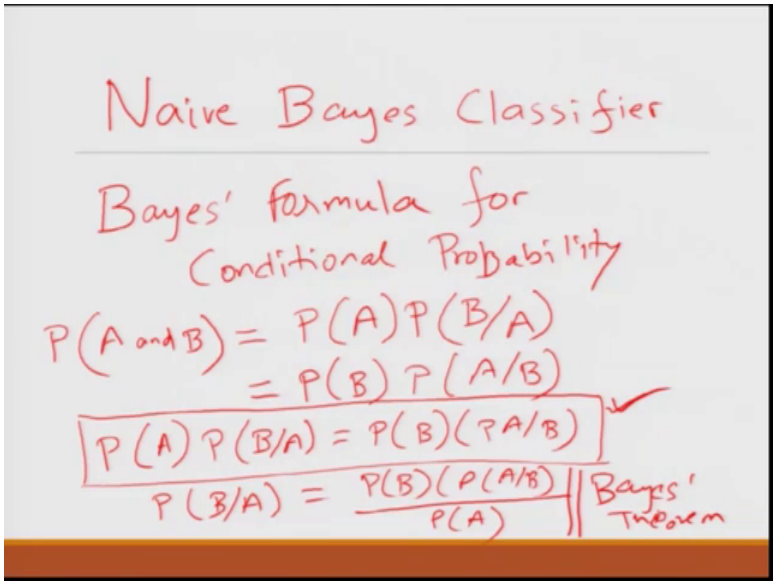
So, I is take it sides a it is a very commonly used variable in probability based methods, it is called an indicator variable. So, what happens is this I takes a value of 1, if yi not equal to yi cap. So, what happens is this yi cap. So, cap as we discussed earlier, it means your prediction. So, you predicted that something is one, but actually it was predicted as 0. So, 1 and 0 being 2 different categories; so, if it fails to predict I gets a value of 1, if it correctly predicts, I gets a value of 0.

So, in this case what happens is, basically what we are trying to do is you are trying to sum up all the number of wrong predictions you made and you divide by the total number of predictions that you made. So, n is the total number of data points. So, this gives you an average error rate and your goal is to maximize this error rate.

So, if you see that in K-nearest neighbor when the number if dimensions is very high the p number of parameters, number of parameters if high K-nearest is not recommended. So, typically in any classification algorithm that you use you calculate this error rate, you calculate this formula, your if you are using a software it will itself do the calculation, it

will give you these kind of estimates and your goal is to reduce it. So, what we find is that if K for K-nearest neighbor this does not give a very great result.

(Refer Slide Time: 19:32)



The other simple method which is again very popular is Naïve Bayes classifier. So, if you recall your study of probability in class 12, recall there was this Bayes formula. So, Bayes formula for conditional probability , actually the formula is not a very complex formula let me just give you a quick refresher you do not even have to remember many people try to remember this formula actually you do not have to remember.

So, what it says is that for example, what is the probability of P A and B both happening. So, you can say that this is the probability of P A happening and once P A has happened probability of B happening given that A has happened. So, this slash A means that given that A has happened happening of B.

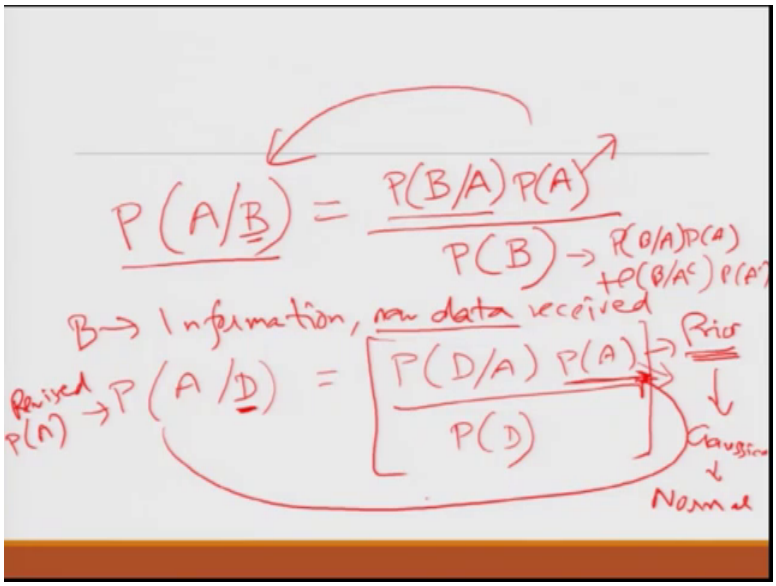
So, probability of both P A and B happening you can easily say that it means that probability of a happening and once that a has happened probability of B happening, but then you can also swap A and B right. So, this is also same as P A by B. So, if you compare these 2 formula, what we basically get is that P A into P B by A is equal to P B into P A by B. Now, Bayes theorem is nothing, but re arrangement of this, this basic equation or in other words you know simple terms P B by A is equal to P B into P A by B by P A.

So, probability of an event B happening given that A has happened is nothing, but probability of the event happening irrespective whether A has happened into probability of A given that B has happened divided by the total probability of probability of A happening.

So, if you remember this equation this becomes easy to derive at this point what is called Bayes formula or Bayes theorem and in modern analytics this very simple, but very intuitive or I would say very insightful formula is the basis of lot of machine learning algorithm.

In fact, there is an entire field of Bayesian statistics which primarily works on this formula. How it works let me just give you a brief insight and then you would understand that why Naïve Bayes classifier becomes a very very attractive tool. In fact, apart from K-nearest neighbor, this is one of the most popular tools for most of the recommendation (Refer Time: 23:30) that you would come across online. So, let us have some brief discussion on this.

(Refer Slide Time: 23:42)



So, let me write this formula again for you, basically we said that P A by B is equal to let me put it in a more ordered way divided by P B. So, so what we are doing is, if you notice this term and this term it gives the way of swapping A and B. So, moving from probability of B given that A has happened to moving to probability of A given that B has

happened. Now in case of analytics many a times what happens is, now if you look this A, A was your original estimate of probability of A happening.

So, if you consider B to be an information or new data received. So, what it says is that given that you receive a new data point or maybe you know let me call it D. So, that its easy to visualize. So, given that a new data point that you received, the probability. So, this was P A was your prior probability this is called a prior.

Now, D by A is that given that A is the situation, what is the probability that you could get this new data in the first place. So, what you find is that, and if you divide it by the probability of getting that data I respective of whether you got A or you did not get A. So, you find that if A be this, P A be the probability of an event a happening this formula gives you a way to adjust that probability given that a new data point has received. So, given that you now got new information new data point, you revise your estimate of A.

So, the probability of A now, what happens typically is that entire the region it becomes. So, powerful in analytics is that, what you do is you start with some in a typical cases you start with some assumptions about this probability of A whatever be the value and then you say you have received.

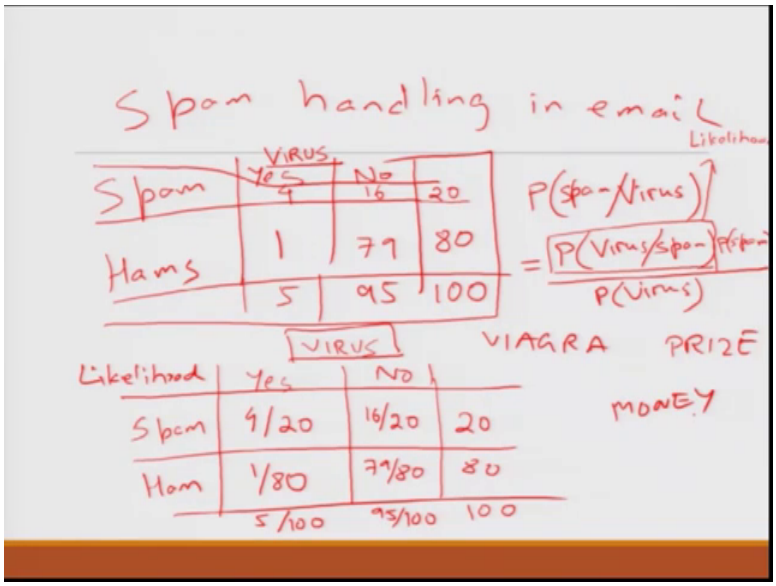
So, you take first data point, you say that what is the probability of I getting this data point given that A has happened. You multiply by this do this calculation, I will not get into the further details of in to this because it is a field in itself, but what happens is after that you after that iteration you get a revised probability of A. Now in the next step you again put this value here again get the new data and again revise your estimate. So, after some time some iterations that you do in a way simulations that you do you get a stable value of probability of a given that all the data that you have received.

So, this is the way a lot of classifications happen for example, you are deciding whether who will win an election. So, given that a new data point receives, you keep refining your probability of a winning a being say a candidate winning the probability of candidate winning. Based on new data you keep refining keep refining ultimately the model comes to a stable state and you come to your prediction. So, that is what this Bayes theorem is about. Now if you recall here this actually this P A now what happens is that this P B is actually probability of B by A into P A plus probability of P given that a

compliment into P A compliment. Now in reality a lot of this calculations become complex especially if there are multiple parameters.

So, what you do is you make certain assumptions simply those calculations and then use this Bayes theorem.

(Refer Slide Time: 28:32)



So, let us take for an example case of spam handling in email. So, say for example, you get a lot of mails, some mails are spam the mails which are not spam they are called hams, that is just lingo. Suppose you got 1 objectionable word for example, let the word be virus. So, basically you make a rule or you know that you know if a word virus comes in the mail that there is a high probability of the mail being spam. So, there can be other words also. So, for example, you do an analysis and finally, you based upon your review, you found you make a matrix of presence of word virus verses prese verses not presence of virus and a mail being spam or not spam.

So, for example, you make this table and. So, for example, there were four cases when the word spam was present and word virus was presented in it was a spam mail and there were 16 cases when the word virus were not present yet it was a spam. So, and similarly there was one case when the word virus was present, but it was not a spam and there were 79 cases when word virus was not present, and it was also not a spam it was a ham. So, let this be this.

So, we have 100 points, we have around 95 points here and we have 5 points here. So, this is what it looks like. So, what you do is you calculate the likelihood based upon this of something being a spam or not spam. So, you try to calculate likelihood out of this how you do is very simple, you just divide these by. So, if I have to convert this into a likelihood table. So, the table would look something like this.

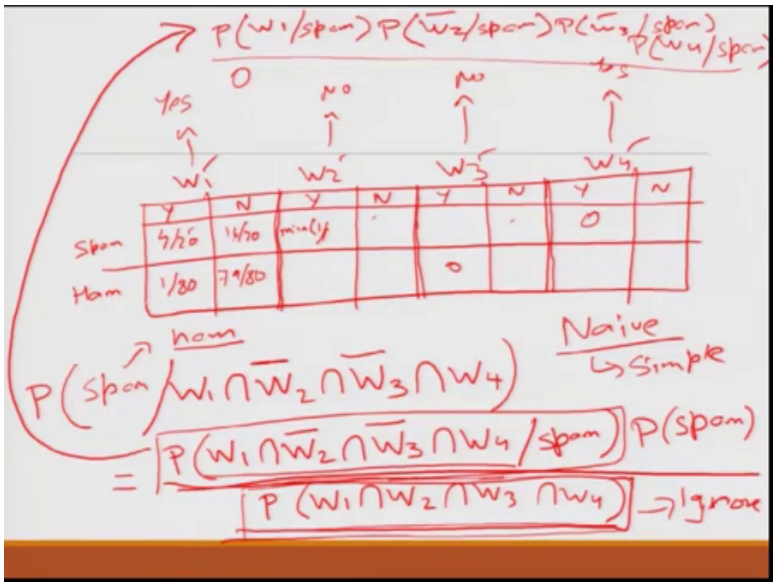
So, you had 20 points here, you had 80 here this was 100. So, you divide each of them by this 20. So, this becomes 4 by 20, this becomes 16 by 20, this becomes 1 by 80, this becomes 79 by 80, this becomes 5 by 100, this becomes 95 by 100.

So, this becomes your likelihood table. Now, what you do is the trick that you do is you say probability of something being a spam given that, the word virus came is equal to probability of word virus given that it is a spam into probability of it being spam by probability of the word virus.

So, note this term this is your likelihood. So, what we are calculating here is that probability of some being virus given that its spam. So, if you know that from here this is the prior probability of something being a spam in this case you know that the prior probability was 20 by 100. So, that is the prior probability and then you know the probability of something being virus was 5 by 100. So, in a way you calculate this you can easily calculate this probability.

So, what Naïve Bayes theorem does is something more. So, here it was a simple case of just one word virus. Now assume that there are many other words that you do. So, for example, apart from virus may be there is an another word common word viagra for example, then there may be some other word prize. So, you may have these kind of tables or you can have another word like money.

(Refer Slide Time: 34:53)



So, in a way you have a bigger table, let me just try to draw that table for you and then we will understand slightly better. So, let us just have four words to begin with let this be word 1 whatever be the word it may be Viagra, it may be grocery, it may be prize, it may be money, it may be virus. So, these are the four words and may be let me for sake of clarity, make the divider with a double line; So, this is yes no, yes no, yes no, yes no and this being a spam this being a ham. So, for each of them you have some values here in you have some in this case for example, for this we knew it was like 4 by 20, 1 by 80, 16 by 20 and it was 79 by 80.

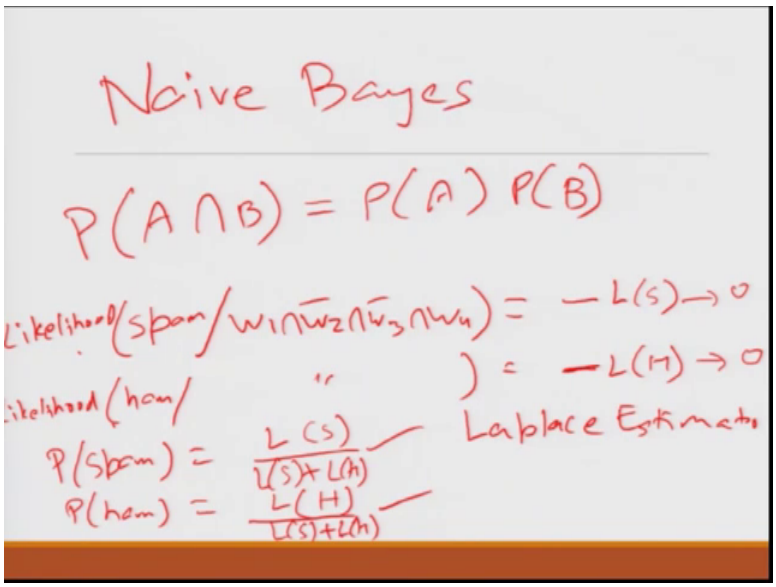
So, similarly based upon the example that you have taken you have this field. Now the probability and for example, a case came where this was yes, this was no you got a you got a mail where this particular first word was present, the second word was not present, the third word was not present, fourth word was present. So, the probability of it being a spam given that first word was present and second word was not present let me just put it by a compliment, which is the an horizontal line over this and word three was also not present and word four was present. This by Bayes theorem is nothing, but probability of given that it is a spam in to probability of something being a spam, divided by probability of.

Now, notice from this particular table you can actually calculate all the permutations combinations and then because you know in each of these. So, P w 1 and w 2 which

means you know this part and then this part and this part you know. So, you can actually calculate and do this its simple algebraic exercise that you would have done in your class 12. So, this is the way the Naïve Bayes is used; however, what happens is that this calculation of this actually in this case we has just four examples. So, maybe you can just do some kind of a manipulation, but this wherever you will do if you do by you know and that is why it is called there is this notice word naïve means very simple.

So, what happens is we unconsciously make an assumption that these this w 1, w 2, w 3, w 4 are independent in reality this independence may not be there. So, what happens is in reality this particular formula this particular term this and this. So, they are very very difficult to compute because there may be relations between w 1 and w 2. So, it is like saying that if a mail has w 1 the probability of w 2 also happening is in influenced by this. So, there are not actually independent. So, what happens is in Naïve Bayes theorem you simply assume that they are independent.

(Refer Slide Time: 40:26)



The moment you assume they are independent then you know that P A intersection B, for independent event is nothing, but P A into P B. So, in this case for example, moment you assume them to be independent all you do is you replace this by something P w 1 given it is a spam into P w not 2 given it is a spam multiplied by w not 3 given its a spam multiplied by w four given it is not a spam. So, you simply replace that by a product

similarly you replace the denominator also by a product and you just then it is a matter of simple calculation.

So, what happens is you do this calculation and for each of these combinations for example, you start filling numbers and you find out that the probability of it being a spam given that, w 1 happened and w 2 did not happen and w 3 did not happen and w 4. You calculate this value and it comes to some number similarly you calculate the value of ham you do the same exercise again that what is the probability that it was not a spam for the same thing you do a calculation of this whichever is greater. If this is greater than this then you classify it as a ham if this is greater than this you classify it as a spam. So, in reality what you do is this value actually does not strictly. So, what happens is in a smart case, you notice whether you are calculating spam or you are calculating ham.

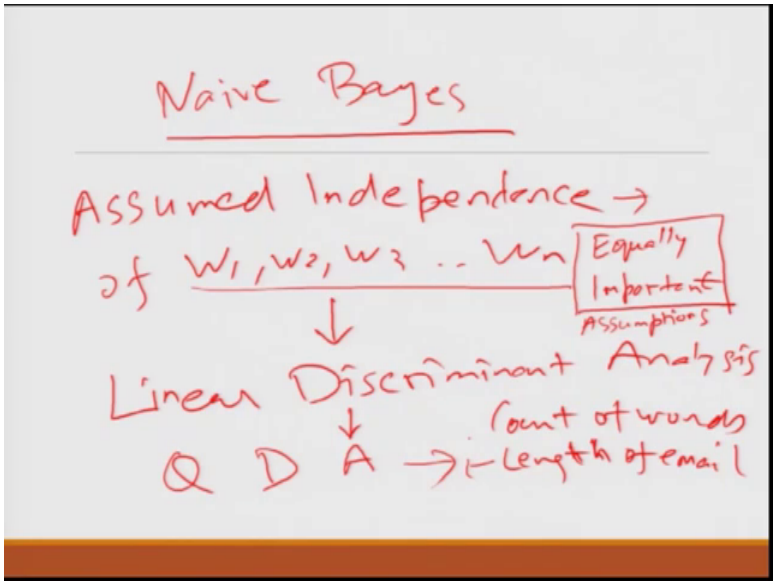
This value actually does not matter, this is constant it does not have any term of spam or ham. So, in general what happens is you simply ignore this simply ignore. The moment you ignore all you have to do is you have to just calculate this numerator. So, what happens in a calculation is that instead of probability you in a way calculate the likelihood just the numerator, similarly likelihood is you simply calculate the numerator and to convert it. So, if this P P s or L s if this be L s and this be L H then what you do is probability of spam is equal to L s by L s plus L h probability of ham is equal to L h by L s plus L h.

So, you calculate this and then you are done. Now one of the issues that comes here is that you have assumed that these are all independent then reality this independence does not hold good. The other concern that happens is you took 4 words suppose tomorrow 5 new words got added and then you got a and these are like completely new words for which you do not have any data points. So, in those cases what happens is that some of these terms may actually have 0.

So, if these are 0 and you are multiplying. So, this term actually becomes 0. So, you get actually you after doing all these calculations, these values simply because the term did not exist become 0. So, what you do is there is something called Laplace estimator which is nothing, but you say that I will not allow 0 to happen, you make sure that it is never 0 it is minimum of 1 and any value 1 by 1 divided by total values and something else. So,

you make sure its always non zero. So, if it is non-zero then calculations becomes simple. So, this is in summary the Naïve Bayes.

(Refer Slide Time: 46:04)



Now, in Naïve Bayes we assumed let us be very clear assumed independence of w 1, w 2, w 3 w n. Now there is this is the part of more generic algorithm set called linear discriminant analysis. In linear discriminant analysis what you do is that, you do not necessarily assume independence and what you do is this prior you assume it to be a Gaussian or a Norman distribution and then it is complicated formula.

So, what happens is the output is also comes out to be a Gaussian and an comes out to be a linear equations. So, that is something called linear discriminant analysis you make more complex assumptions about the nature of probability distribution, you have something called quadratic discriminant analysis. So, these are different methods which are used; however, the Naïve Bayes is computationally very efficient, very simple and for most cases it kind of suffices.

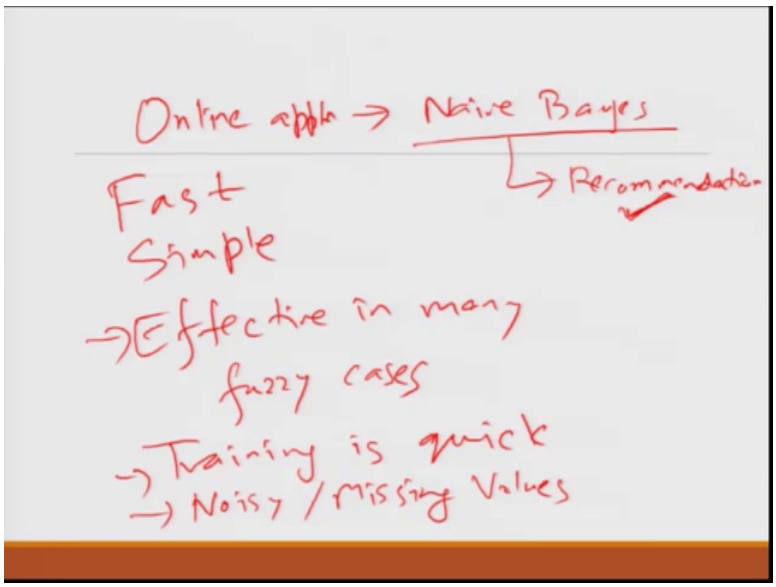
But let us be very clear that because it assumes independence, it also assumes at every factor is equally important. So, what happens is what happens is that if you take too many factors and your problem designer has been kind of too finicky about it if your models tend to get very very unstable, because there are too many factors suppose instead of these words I would have instead of phrases you would have broken phrases into words, and they would have chosen every hundreds of words then it would have

become very very complex. And also these are like good if the data is largely categorical in case of presence or non-presence of words.

If it was numeric variables then it would have problems, then you would have to break those numerical variables into different. So, for example, instead of presence or non presence of word we had something like count of words or length of email. In that case you would have to split these lengths and counts in to low medium high different categories and then try to use and you find that the algorithm then does not work very very great.

So, but it is you need to be very clear that it is very fast.

(Refer Slide Time: 49:06)

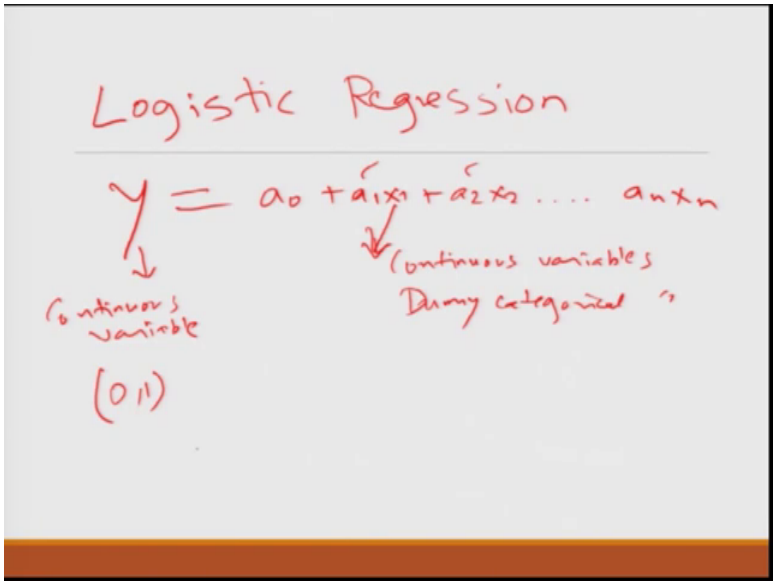


It is simple and I would say also effective in many fuzzy cases fuzzy in the sense when the data is also not very clear accurate and it turns out to be very accurate, also training is quick you do a 1 basic sampling of your emails and it can smartly train itself and also if data is noisy or there are lot of missing values with Laplace estimation it tends to work pretty very well.

So, and that is the reason that for many online applications we use Naïve Bayes or one of its modification for example, your recommendation engines in most of these shopping. So, E-commerce sites etcetera all likelihood are going to be Naïve Bayes based methods ok. So, now, let us move to a more serious method Naïve Bayes is also serious am not

saying it is not serious, but something which is used a lot in enterprise based situations very popular.

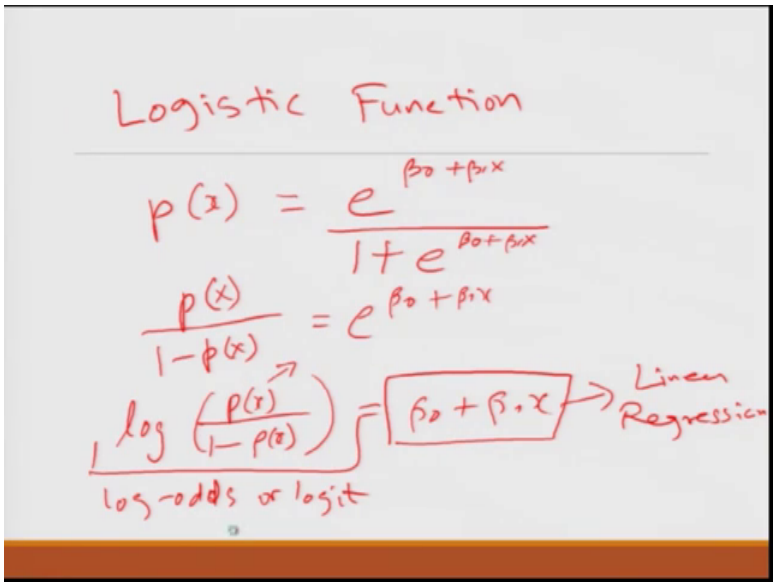
(Refer Slide Time: 50:36)



In fact, if you talk in terms of its usage, then it is probably more applied or more widely used then even linear regression this is the logistic regression. So, in logistic regression is nothing, but your linear regression was nothing, but a 0 plus a 1 x 1 plus a 2 x 2 up to a n xn were these a 0, a 1, a 2 these could be either or these x ones rather. These x ones they could be either continuous variables or dummy categorical variables and y was a continuous variable what if I want y to be a discrete variable or other a categorical variable I want finally, the answer to come as like a 0 or 1.

So, mostly logistic regression is used for binary classification, there are extensions of logistic regression for non binary or multinomial classifications, but they are not very popular the most popular uses for binary classification. So, in this case what you do is there is something called logistic function.

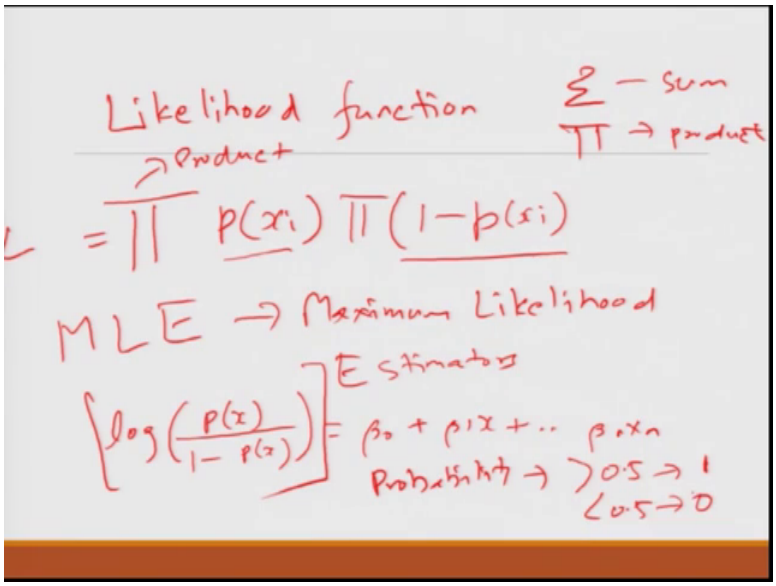
(Refer Slide Time: 52:51)



So, logistic function; so, am assuming am starting with simple case of just one single variable logistic regression and the same thing extends to multivariable. So, this is the function. So, if you swap do some elementary algebra you will find that px by 1 minus px is equal to e to the power e 0 plus B 1 x. So, if you take log of both sides natural log, you will find log of px by 1 minus px is equal to beta 0 plus beta 1 x.

So, this particular term is called a log odds or logit function. So, this term if you now note is almost same as that of linear regression. The y term is slightly different now what you try to do because it is a classification algorithm. Now this is kind of a log word like if you if you just roughly think that you know this p be a kind of probability it is kind of log of something happening verses something not happening that comparative ratio.

(Refer Slide Time: 54:36)



So, what in this case we do is we try to actually maximize something called a likelihood function.

So, this is how the mathematical form is evolved. So, which is nothing, but. So, this is basically a symbol for product. So, we all know sigma means symbolizes some this symbolizes product. So, what it is trying to do is, it is calculating the parameters for which this is called likelihood. It tries to find those parameters for different betas that maximize this particular probability. So, this is called a set of you know methods called maximum likelihood estimator method; So, maximum likelihood estimators.

So, similarly if you are doing a in this case if it was you are doing this was for a bivariate case, if you are having multiple variables the same function would look something like log of px by log of 1 minus px is equal to beta 0 plus beta 1 x plus beta n xn and the same formula that comes.

So, here what happens after you have done this thing it comes with the value, you convert that value into a probability and if the probability is greater than 0.5 you say it belongs to a class 1 if it is less than 0.5, it belongs to class 0. So, thank you very much we will continue this further and then take ahead from there.

Thank you very much.