

Marketing Analytics
Professor Swagato Chatterjee
Vinod Gupta School of Management,
Indian Institute of Technology, Kharagpur
Lecture 54
Text Mining and Sentiment Analytics (Contd.)

Hello everybody, welcome to Marketing Analytics course this is Doctor Swagato Chatterjee from VGSOM, IIT, Kharagpur who is taking this course for you. We are in week 10, session 4 and we will be discussing about Text Mining.

So in the last class, we have discussed about text mining in the context of reading a data, Martin Luther King's data we have read and then we have done some amount of pre-processing of the data and in this particular class, I will discuss about a new thing which is called spam detection and we will be using Naive Bayes algorithm to do that.

So, before I start, we should discuss a little bit about, little bit about what is Bayes theorem and we have done this in probably, in our statics class, very early days but I will just repeat once more. So, if you remember that there was a thing called probability. What is probability?

(Refer Slide Time: 01:21)

$P(E) = \frac{N(E)}{N}$

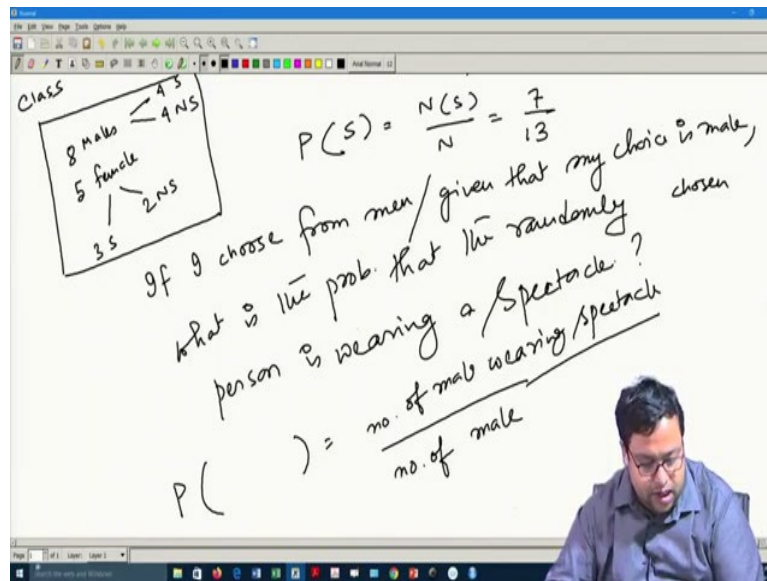
$P(M) = \frac{N(M)}{N} = \frac{8}{13}$

$P(S) = \frac{N(S)}{N} = \frac{7}{13}$

CLASS

8 Males → 4 S
5 female → 3 S, 2 NS

If I choose from men / given that my choice is male,
what is the prob. that the randomly chosen
person is wearing a Spectacle?



Probability of anything, any event is basically, number of times that event is occurring / the total number of observations. That is something called probability of a particular event. So then, then let us say, in a particular class, is a particular class, there are 8 male are there, and there are 5 female are there. And out of these 8 males, 4 are wearing spectacle and 4 are not wearing spectacle.

And out of these 5 females, 3 are wearing spectacle and 2 are not wearing spectacle. Now, my question is that, what is the probability that if I blindly choose some person in this, from this particular class that person will be a male or is the probability of male? So, what is that answer? Number of male by total number of observations, fair enough, so what is the number of males here?

Number of males is 8, total number of observations is 13, so, 8/13. If I say, what is the probability if I blindly chose somebody, he will be wearing a spectacle? So, if you see that, in this thing, there are 7 people who are wearing spectacle and there are 13 number of people, so this is how we generally try to find out probability.

So, now my question is that, listen to this question carefully. So, if I choose from men or given that my choice is male, whatever be the language, what is the probability that the randomly chosen person is wearing a spectacle? That is the question, that if I chose from men, what is the probability that a randomly chosen person is wearing a spectacle? So, how will I answer that, what is given here, so if this is something that I am asking you.

So then probability of something I will write it later. So, how many, so if I choose from men, that means the total number of people, total number of possible cases, unique possible cases

is basically the number of male, number of male and he has to wear spectacle means. Number of male wearing spectacle, that is something that we are trying to find out. Number of male and number of male wearing spectacle.

(Refer Slide Time: 04:52)

The whiteboard contains the following content:

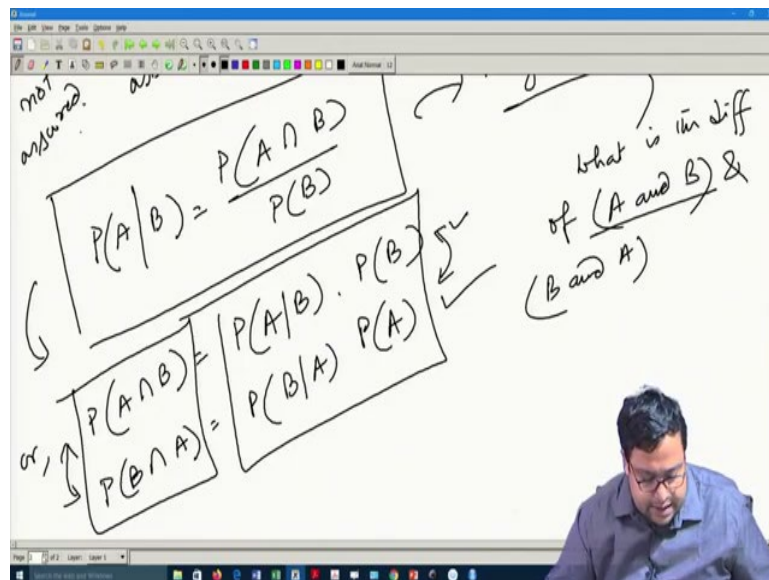
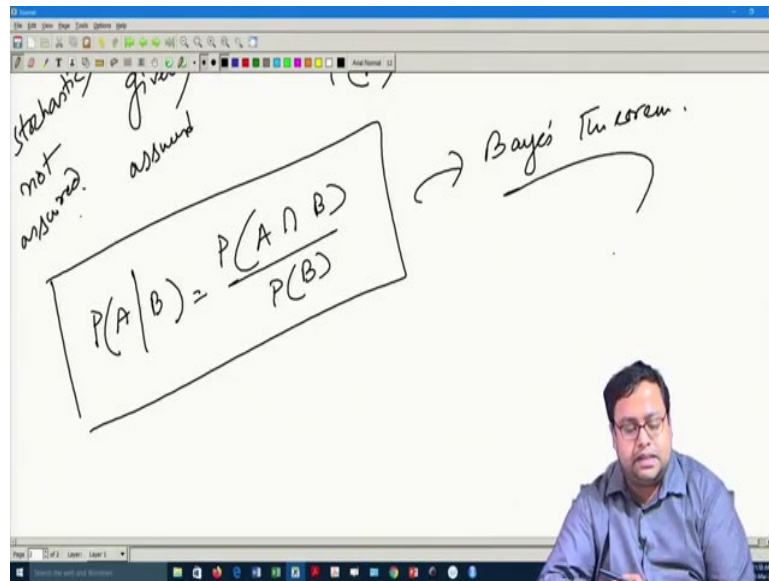
- On the left, the expression $P(S|M)$ is written. A bracket under 'S' is labeled 'stochastic, not assumed'. An arrow points from 'M' to the text 'given, assumed'.
- In the center, the derivation is shown:

$$P(S|M) = \frac{N(M \cap S) / N}{N(M) / N} = \frac{P(M \cap S)}{P(M)}$$
- On the right, the text 'male and spectacle' is written with checkmarks under each word.

So, if I try to write this in mathematical term, whatever the probability that we are trying to find out is basically $P(S|M) = \frac{N(M \cap S)}{N(M)}$. What is and? And means both are thing are happening. Male is happening, that the person is male and spectacle is happening that means, he is wearing spectacle, both are happening. Now, if I just divide it both numerator and denominator with 13 which is the total number of persons then do I get this? $= \frac{P(M \cap S)}{P(M)}$

From here, if I divide the both numerator and denominator with the same value, then it remain same. So then then this numerator becomes $P(M \cap S)$ and denominator becomes $P(M)$. And this particular part we call it S given M. Why S given M? Because male is given, this is given, this is will happen, this is assured and this is a stochastic part. This is something which is not assured. Okay? So that is why S given M.

(Refer Slide Time: 06:11)



So, using this similar thing, there is something called Bayes Theorem which is $P(A|B) = \frac{P(A \cap B)}{P(B)}$ this is called Bayes theorem and this is something that we will use here.

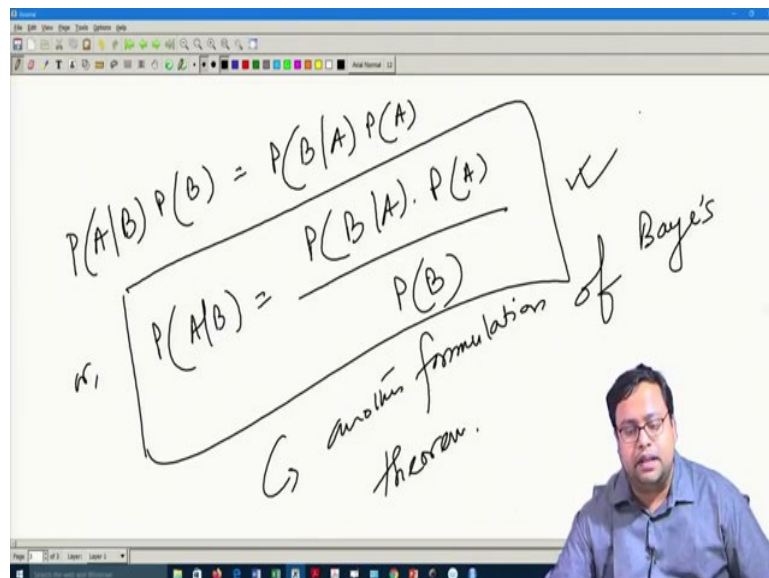
Before I go ahead with Bayes theorem, we will do a little bit of maths on this particular formula. Can I just write? $P(A \cap B) = P(A|B) \cdot P(B)$

So, from here can I just write this? See the numerator remains same and this denominator which has been PB is pushed here. So, this denominator which is PB this one are pushing it to this part. So, this one I am pushing it to this part. This guy I am pushing it to here. So, that is what I am doing. And while doing that, I am getting this formula, kind of. Now if this formula is true, what is a difference between A and B and B and A?

Is there any difference? A and B and B and A, is there any difference? So, if you say that you and I are going to the school versus I and you are going to the school, the persons were going to the school are they different? Probably not. So, if this two are not different, can I write this one, just replacing A and B. If this is true then this is also true, no. Just replaced A and B to, in their spaces.

And if I can replace A and B in their spaces and put A in this side and B in that side, just replacing the space, then if these two guys are same then these two guys are also same, if the right hand side of these two equations are same then the left hand side of these two equations are also same, in fact the other, the left hand side of this two cases are same then the right hand sides are also same.

(Refer Slide Time: 08:41)



That means, $P(A|B)P(B)=P(B|A)P(A)$ So that was the formula. So, $P(A|B) = P(B|A).P(A)/P(B)$. So, this is another formulation, another formulation of Bayes theorem. So, these two formulations, just remember these two formulations, we will be requiring to do something called Naive Bayes algorithm. We are doing spam detection.

So, what is spam detection? So, just tell me that what exactly do you understand as spam? So, the spam term, if I am not wrong came from, from a burger, burger was not good or something like that. So, not good burger was named as spam and good burger was named as hams or Something like that. So, you just read in Google, there is a history of spam and ham by these two terms.

Basically, in your email, whatever mail comes or in your SMS box, whatever SMS comes is, what about SMS is useful to you that is something which is ham and whatever which is not useful to you is something called spam. So, this is how we name spam and ham. Ham is something which is useful and spam is something which is not so useful.

(Refer Slide Time: 10:33)

Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

	Viagra		Total
spam	4	16	20
ham	1	79	80
Total	5	95	100

	Viagra		Total
spam	4/20	16/20	20
ham	1/80	79/80	80
Total	5/100	95/100	100

$\frac{4}{20} \times \frac{20}{100} = \frac{4}{5} = 0.8$
 If viagra word is occurring, the prob. of the msg being spam = 0.8

$P(\text{spam} | \text{Viagra}) = \frac{P(\text{Viagra} | \text{spam}) P(\text{spam})}{P(\text{Viagra})}$

$\frac{4/20 \times 20/100}{5/100}$

Spam

Naïve Bayes Classification

Swagato Chatterjee
VGSOM IITKGP

- supervised
- unsupervised

spam not useful
ham useful

IBA

→ supervised learning

So, spam is not useful, spam is not useful and ham is something which is useful. So, we have to find out, now if you remember at one point of time Google was not doing good job or you have, was not used to give, doing good job in spam detection. But now it is they do very good job in spam detection. So, there are some ways that they follow, one of the basic ways to do that is this Naive Bayes Classification. So, any classification is basically a supervised learning.

So, if you have, I hope that people who are in this class have done introduction to business analytics and I hope that they are, what is supervised and what is unsupervised learning has been discussed, if not, search Google. So, this classification is a supervised learning where you have to be pre-give to this particular classification algorithm that which of your observations are spam and which of your observations are ham.

There will be past data which are pre-labeled for the, for the learning purpose. So, will it say, there is certain 100 emails that you got. And out of them, 20 were spam, 80 were ham. You manually check that and you are also checked about the word, there is a word called Viagra, that is very common in US based spam mail and you will see that, out of these hundred SMSs or emails that you checked, 5 had the word and 95 did not had the word.

And if I just convert this table to the probability table, it looks like this. So 4/20, 1/80, 5/100 and 5/100. Now, then what is the probability that SMS will be a spam given that there is a word Viagra? If the word is present, what is the probability that it will be a spam? That is probability of $V.P = \frac{P(V|S)P(S)}{P(V)}$ Is it not? It is V given S by probability of, so what is I would say V given S? So I will just delete this probably.


So, what exactly is this V given S? So, given that it is spam, what is a probability, it is Viagra? If you check the thing is, 4/20. What will be probability of spam? It is 20/100. What is the probability that Viagra will occur? 5/100, so the net probability is $4/20 \times 20/100 / (5/100) = 4/5 = 0.8$.

That means that if Viagra word is occurring, that means whatever is there in my message, the probability of the message being spam is 0.8. If this value is high cut of that you set on your own, if this particular value is high than cut off that you set on your own then that particular message will be a spam message. So that is how we try to deal with spams.

Refer Slide Time: 14:53)

Typically, Bayesian classifiers are best applied to problems in which the information from numerous attributes should be considered simultaneously in order to estimate the probability of an outcome. While many algorithms ignore features that have weak effects, Bayesian methods utilize all available evidence to subtly change the predictions. If a large number of features have relatively minor effects, taken together their combined impact could be quite large.

Strengths	Weaknesses
<ul style="list-style-type: none">• Simple, fast, and very effective• Does well with noisy and missing data• Requires relatively few examples for training, but also works well with very large numbers of examples• Easy to obtain the estimated probability for a prediction	<ul style="list-style-type: none">• Relies on an often-faulty assumption of equally important and independent features• Not ideal for datasets with large numbers of numeric features• Estimated probabilities are less reliable than the predicted classes



Now, this is very simplistic this is Naïve Bayes because it is naive, it is very simplistic to find out the probability and it is based out of Bayes theorem that is why this is Naive Bayes. What are the strengths? It is simple, it is fast and very effective. Does well with noisy and missing data, so even if some of the data is missing or noisy, it can handle that.

It records relatively few samples for trainings, so the data set, that is recorded the training data, the past data that is required, might be small, no issues. And also works well with very large number of examples, so even with if it is large or small, it can handle it. And easy to obtain the estimated probability for a prediction. So, that is something which are strengths.

What are the weakness? The weakness are it relies on an often faulty assumption of equally important and independent features. So, we will discuss about that, this particular thing, we will discuss at the moment I will go to the next line. They are not ideal for the data sets with large number of numeric features, so generally it cannot handle all the categorical variables like Viagra word yes or no, it is a categorical thing.

So, that kind of situations, it can handle better than numeric variables. And estimated probability, so that is why in text mining it is used more than other classification problems. An estimated probabilities are less reliable than predicted class. So, it is better to predict whether it is a spam or ham using these. Not what is the probability because remember this particular thing is not good with numeric data.

So, if you want to do let us say credit scoring, credit scoring tries to find out the probability of somebody will default or not. Not, so they are good in finding out the probability, they are

not good in finding out exactly classification and saying that okay, this, this guy will be defaulter and this guy will not be defaulter. But that gives probability.

Similarly here, if we try to get the probability of defaulting or not defaulting, that kind of a problem, it will not do good job. But if you try to find out whether it will be a spam or ham, a classification problem instead of a probability as an outcome, then it will go to be the do a good job. So that is something which are the weaknesses of this particular thing.

(Refer Slide Time: 17:19)

P(A, B) = P(A ∩ B) = P(A)P(B) if A and B are independent

Naïve Bayes Algorithm


Likelihood	Viagra (W _i)		Money (W _j)		Groceries (W _k)		Unsubscribe (W _l)		Total
	Yes	No	Yes	No	Yes	No	Yes	No	
spam	4/20	16/20	10/20	10/20	0/20	20/20	12/20	8/20	20
ham	1/80	79/80	14/80	66/80	8/80	71/80	23/80	57/80	80
Total	5/100	95/100	24/100	76/100	8/100	91/100	35/100	65/100	100

Assumption is the words are independent occurring.

$$P(s|v, u, \bar{m}, \bar{g}) = \frac{P(v|s)P(u|s)P(\bar{m}|s)P(\bar{g}|s)P(s)}{P(v)P(u)P(\bar{m})P(\bar{g})}$$

A message has the words 'Viagra' and 'Unsubscribe', but not 'Money' and 'Groceries'

Find the probability of being spam:

$$= \frac{4/20 \times 12/20 \times 10/20 \times 20/20 \times 20/100}{5/100 \times 35/100 \times 76/100 \times 91/100}$$


Now, imagine we do not a create spam detection technique so we do not create, I would say classifications of spams or hams. Based on only one word, we generally do with multiple words. So, let us say, whatever I did for that hundred things, I got 20 spams, 20 hams and instead of one word, I did it for 4 words. So, Viagra, then Money, then Groceries, then Unsubscribe.

These are the four words, very commonly found when people saying mass mails or mass SMSs, in the mail in US probably. These words are very common and correspondingly the probability distribution is given there, yes-no blah-blah-blah. Now the question is that, a message has the words, Viagra and Unsubscribe and but does not have the word Money and Groceries. What is the probability of being spam?

So, let us write it in mathematical way. So, I am writing, what is a probability of spam given that v and u, these two words are there and \bar{M} and \bar{G} that means M and G these two words are not there? So, if I just follow the normal Bayes formula, $\frac{P(v, u, \bar{M}, \bar{G} | s)P(s)}{P(v, u, \bar{M}, \bar{G})}$, so P (A,B) means A and B are both occurring which is basically P(A ∩ B). Okay, so this is something. Now, I do

not have V, U, \bar{M} , \bar{G} , that value I do not have. So here the assumption comes in, and that is a failure, relies on often faulty assumption that equally important and independent features. So, we are, we are bringing that these features which is the words or independent to each other. Might not be the case, probably sometimes Money, Grocery or Unsubscribe might be related.

In a same the probability of having a grocery related email and having unsubscribe word written there, might be very high. So they might be related but here for our simplicity purpose we are assuming that these words are unrelated. Now, if $P(A \cap B) = P(A)P(B)$, if you remember that formula.

So, using that what I am writing here is $\frac{P(V|S)P(U|A)P(\bar{M}|S)P(\bar{G}|S)P(S)}{P(V)P(U)P(\bar{M})P(\bar{G})}$, what is the assumption, you should write this thing, assumption is the words are independently occurring. Now, tell me the value of v given s, what is the value of v given s given that it is spam what is the value of, what is the probability that will be Viagra? It will be 1/80, right?

Check carefully, it is 1/80 No? No, it will not be 1/80, it will be 4/20, right or wrong? Yeah. So then what is u given s? U given s is unsubscribe, you see this, this value, this value. So, basically 12/20, then what is \bar{M} ? M is not occurring so this 10/20. Then, then the last one is \bar{G} , G is not occurring so 20/20 into what is P of s, that is 20/100. Divided by P of v which is 5/100 into P of u which is 35/100 into 76/100 into 91/100.

$$\frac{4/20 \times 12/20 \times 10/20 \times 20/20 \times 20/100}{5/100 \times 35/100 \times 76/100 \times 91/100}$$

Whatever value comes, that value, if it is higher than whatever cut of you have decided, then it is high otherwise low, otherwise it will not be a spam. So, this is how we calculate. What is one of the limitations? One limitation is that, that I am assuming that the words are independently occurring. Now another small problem that I will discuss.

(Refer Slide Time: 22:56)



$P(A, B) = P(A \cap B) = P(A)P(B)$ A and B are independent

Naïve Bayes Algorithm

Likelihood	Viagra (W_1)		Money (W_2)		Groceries (W_3)		Unsubscribe (W_4)		Total
	Yes	No	Yes	No	Yes	No	Yes	No	
spam	4 / 20	16 / 20	10 / 20	10 / 20	0 / 20	20 / 20	12 / 20	8 / 20	20
ham	1 / 80	79 / 80	14 / 80	66 / 80	8 / 80	71 / 80	23 / 80	57 / 80	80
Total	5 / 100	95 / 100	24 / 100	76 / 100	8 / 100	91 / 100	35 / 100	65 / 100	100

Assumption is the words are independently occurring.

$$= \frac{P(s|v, u, m, g) P(s)}{P(v, u, m, g)}$$

$$= \frac{P(v|s) P(u|s) P(m|s) P(g|s) P(s)}{P(v) P(u) P(m) P(g)}$$

A message has the words 'Viagra' and 'Unsubscribe', but not 'Money' and 'Groceries'

Find the probability of being spam:

$$= \frac{1}{\frac{5}{100} \times \frac{35}{100} \times \frac{24}{100} \times \frac{8}{100}} \times \frac{4}{20} \times \frac{16}{20} \times \frac{10}{20} \times \frac{0}{20} \times \frac{20}{100}$$

= 0

Let us say somebody ask me this question, that a message has all words, Viagra, Unsubscribe, Money, Groceries, all these 4 words are there, what is the probability of being spam? Okay, so if all words are there, what will change in this whole story? In this whole story, this will be M, these dash signs will go away.

So, here also this dash signs will go away, fair enough, if this dash signs go away, then how much will be the value at the bottom? This value and this value, this value and this value becomes let us say 8 / 100 and 24 / 100. Are you with me? What happens in that numerator? The numerator, this remains 10 / 20, no issues, but this guy all of sudden becomes 0 / 20.

The whole value thing becomes 0. So, if the whole value becomes 0 even if these are very high, let us say there is one, instead of 10 / 20, these are 1, still it will become 0. So, all other


words are saying that heavily saying that okay this is a spam, this is a spam, this is a spam. Still if one observation is by chance coming 0 and that is one of the thing that you, that is occurring, the whole thing becomes 0, that is a limitation. It has a zero limitation.

(Refer Slide Time: 24:34)

A message has all the words 'Viagra', 'Unsubscribe', 'Money' and 'Groceries'

Find the probability of being spam:

Laplace Estimator



$P(A, B) = P(A \cap B)$ A and B are independent

Naïve Bayes Algorithm

Likelihood	Viagra (W_1)		Money (W_2)		Groceries (W_3)		Unsubscribe (W_4)		Total
	Yes	No	Yes	No	Yes	No	Yes	No	
spam	4/20	16/20	10/20	10/20	0/20	0/20	12/20	8/20	20
ham	1/80	79/80	14/80	66/80	8/80	71/80	23/80	57/80	80
Total	5/100	95/100	24/100	76/100	8/100	91/100	35/100	65/100	100

Assumption is the words are independent occurring.

$$P(s|v, u, m, g) = \frac{P(v|s) P(u|s) P(m|s) P(g|s) P(s)}{P(v) P(u) P(m) P(g)}$$


A message has the words 'Viagra' and 'Unsubscribe', but not 'Money' and 'Groceries'

Find the probability of being spam:

$$= \frac{1}{100} \times \frac{4}{20} \times \frac{12}{20} \times \frac{10}{20} \times \frac{0}{20} \times \frac{20}{100}$$

$$= \frac{5}{100} \times \frac{35}{100} \times \frac{24}{100} \times \frac{8}{100}$$

Result = 0

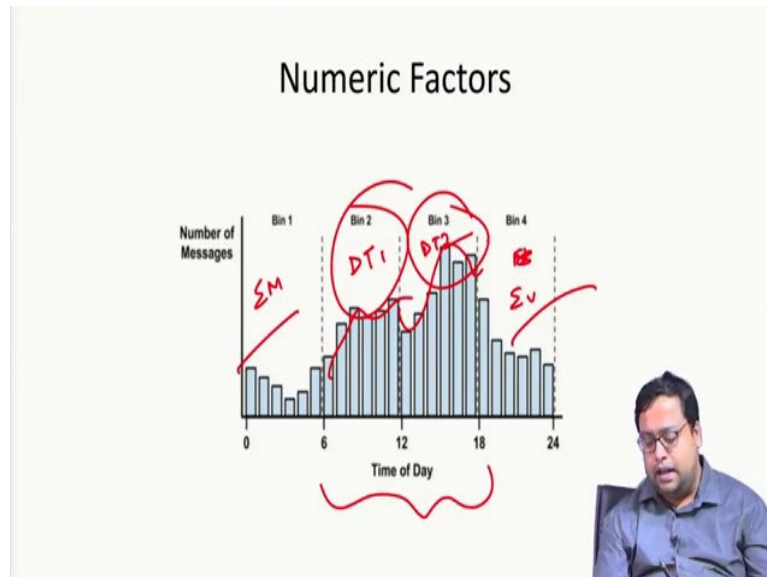


So, to solve that what we do is we use a Laplace Estimator. What is Laplace Estimator? We say, that this is not zero, this particular value, we say this is not zero. This is small value, let us say 1. If this 1 and this is 19, let us say if the instead of 100 data points, if there were 5000 data points or 50000 or 1 lakh data point, then in that 1 lakh data point, if I change, if it is 0/20 and this is 1 lakh/20 instead of that.

Sorry, 1 lakh/1 lakh and that was 0 / 1 lakh instead of that if I just write 1 / 1 lakh and here 99999 / 1 lakh. Will that change the whole picture a lot? It will not change. But we will solve

save me from this 0 thing. So, that is what we try to do, we call it Laplace Estimator and we put a small value for this thing. Now, somebody asked me, or we discussed in the limitation that we cannot handle this particular thing, cannot handle a categorical, continuous data.

(Refer Slide Time: 25:50)



So, for continuous data what we generally do is something called binning. What is binning? So let us say at our childhood days, I am not sure that was the case in your case also, whether you have faced this kind of situation. For when I started using email for the first time, it was long back and at that time, all the servers of email were in US countries or Russian countries and etcetera.

So, when they, it was working hour at that time, it was a sleeping hour in our time so, what I used to see that, in the morning when I check the email, there would be lots of I would say spam emails because all these spam emails are post in the night, when it is the working hour for US and Russia.

Now, spam email comes at every point of time, it does not matter whether it is day or night or what kind of a time, but at that point of time, there was a time variable. The time of the day was also important to measure that whether a particular mail is spam or ham. So, what we do given that is the time is 24 hours, so I can break that 24 hours in 24 bins or at least in 6, 4 bins.

So, see this 6 to 18 is when the frequency is high, so I say this is, this is night or let us say evening. This is early morning, this is daytime 1, daytime 2. So, whenever daytime 1, daytime 2 comes there will be higher chance of being a spam and here it comes. So, we

basically change numeric factors to beans. But the moment we do beaning, we lose a little bit of variations.

So, here that is some variation, here there is some variation which we are losing but we have to take care because otherwise it becomes very difficult to handle numerical data in this particular algorithm. So that is all on Naive Bayes, in the next video we will actually discuss about how to use Naive Bayes algorithm to spam detection using R as the code. Thank you very much I will see you in the next video.