NPTEL

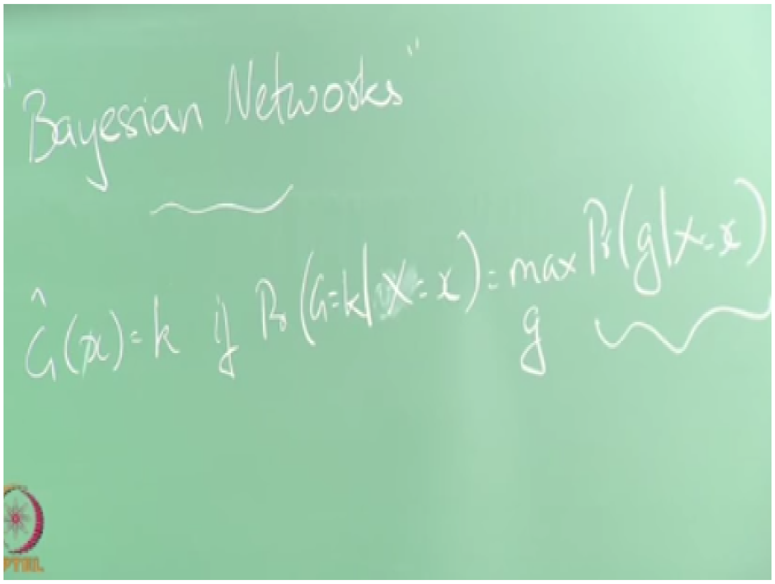
NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

Lecture-63
Naive Bayes

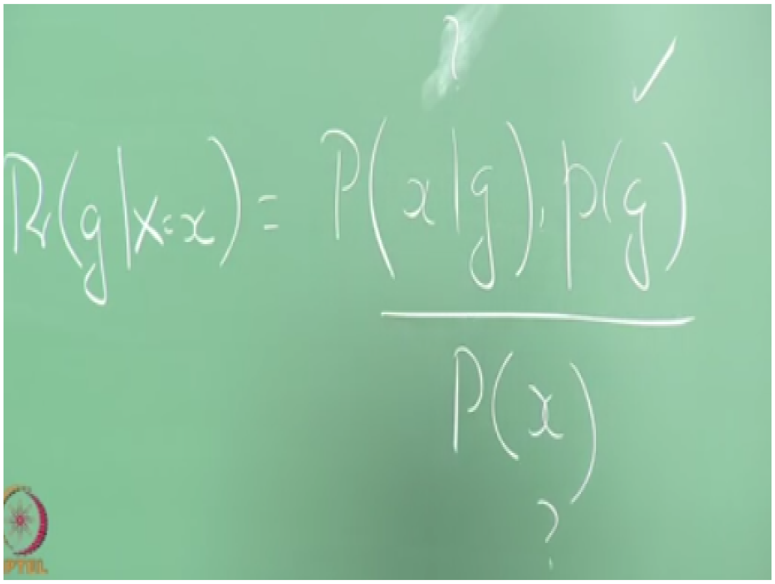
Prof. Balaraman Ravindran
Computer Science and Engineering
Indian Institute of Technology Madras

(Refer Slide Time: 00:17)



Okay So people remember what we mentioned by the bayesan classifier or the Bayes optimal classifier this is the very first class and we talked about nearest neighbor methods and so I told you right something like this that if people remember right and then we said hey we cannot find this normally because this is not supported in our training data right we will probably get 1x it how do you know what is a class of problem the probability of the class is given a data gram point because you are going to get only one data point in right one sample of the data point in your training and therefore we said we will take an average over a region and so on so forth. I so we will use that estimate over a region for finding the probability so this is this is how we motivated what knn right the k nearest neighbor classifier right.

(Refer Slide Time: 02:15)

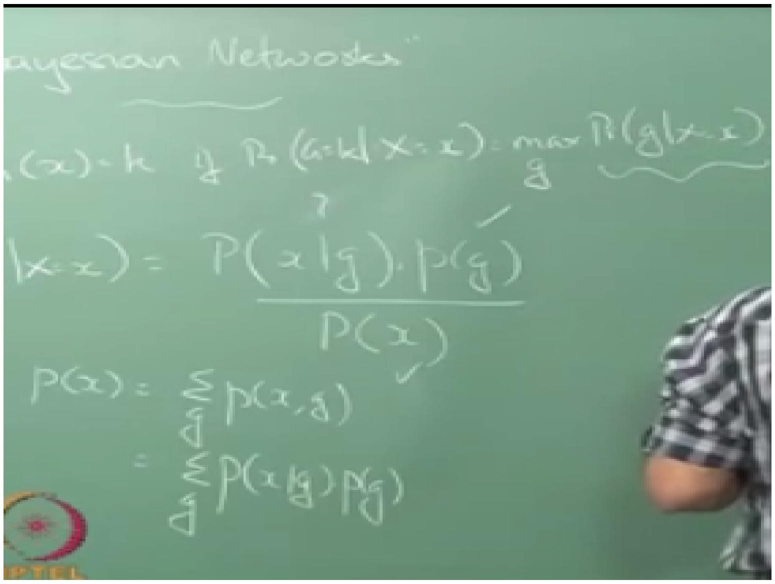


$$Pr(g|X=x) = \frac{P(x|g).\overset{\checkmark}{p(g)}}{\underset{?}{P(x)}}$$

So now I am going to take a slightly different tack right, so what we want we want probability of right g given x nice we want probability of g given x right so we have our friend the Reverend how many of you know that thermos space was actually a ordained priest okay, so we have Reverend base to help us out right, so is essentially right so then we have a lot of quantities that we can estimate here you can estimate this from data right how will they estimate that from data I will estimate that from data.

Okay we will come to that can estimate this from data obviously right whatever fraction of a particular classes you can do that or you can make assumptions about the class densities sand you can estimate the parameters of those from the data and so on so for that is fairly straightforward to estimate right, so what about this guy but in general I mean if you want to do the if you wanted to the max, yes if I do not want to do the max then this becomes a question is how do I go about estimating this right.

(Refer Slide Time: 04:19)


ayesian Networks"
(x) = k if Pr(G=k|X=x) = max_g Pr(g|X=x)
|X=x) = P(x|g).p(g) / P(x)
P(x) = Σ_g p(x,g)
= Σ_g P(x|g) P(g)
PTEL

So that is one way of doing this how do you do that, so you can assume that you have a is what is probability of x that essentially some of the numerator for all possible g so I get my denominator so this I can do so all we need to know is how to estimate that right, so you can estimate that it is enough find the distribution of data points given the class see the main problem that will happen in this case is that you will run into sparsity issues right, so what is going to happen you say and what is going to happen is I am going to run into sparsity issues because you will not see enough of the data space right.

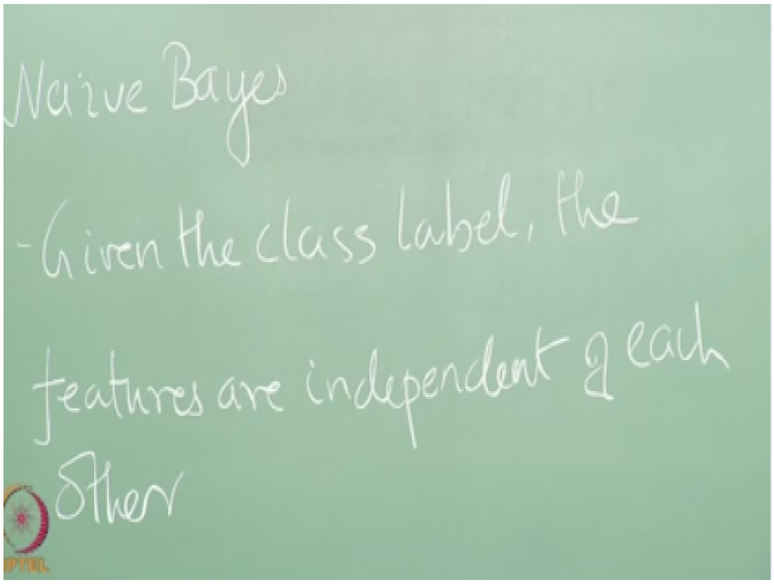
Often enough it for me to know I might get one sample here one sample there and so on so forth but I might not cover the entire data distribution for me to get a good estimate of the probability distribution this is especially true when you are talking about really high dimensional spaces right suppose I have now be we are assuming X comes from some $r^{\wedge}$ p right so if p is a hundred dimensional thing XP is 100 , so if X is a data point in our power 100 right, so what is going to happen data points are really sparse in this vast space right.

So for me to estimate that becomes hard, so what I have to do is I have to start making some kind of assumptions about this distribution so I cannot just say that I will be able to estimate it off hand right I have to make some kind of assumption suppose this distribution I have to say hey so what do we call these distributions they are called somebody louder put your hand up okay that is not bad idea just say louder class condition distribution so we typically call them class condition distribution because their probability of x conditioned on the class right.

So sometimes you also call them what likelihood yeah I thought somebody would say likelihood before anything else but people are just keeping quiet, so this also called the likelihood but is the class conditional distribution right so what is the difference between likelihood in class condition distribution so likelihood is a function of g right I remember I kept repeating that multiple times when we looked at likelihood and I am conditioning it on some parameters θ okay but it is actually a function of θ say x is the same.

But when I am talking about class conditional distribution I am talking about probability of X okay it is a actually a function of x conditioned on g okay all good I will just go back so the most the simplest assumption that we make.

(Refer Slide Time: 08:13)

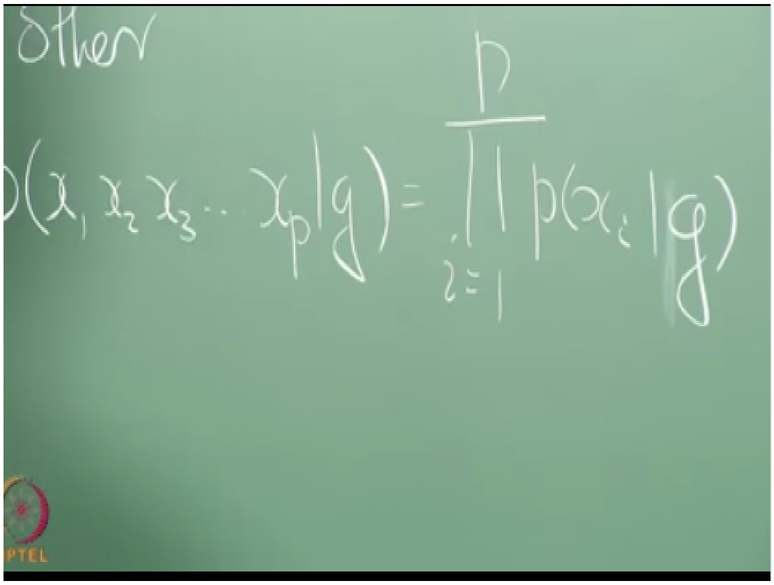


Naive Bayes

- Given the class label, the features are independent of each other

To get this to be tractable is called the Naive Bayes assumption right, so what does a Navy base assumption tell you it says that given the class label given the class label given the class level the features are independent of each other right.

(Refer Slide Time: 09:17)


other
$p(x_1, x_2 x_3 \cdots x_p | g) = \prod_{i=1}^{p} p(x_i | g)$
PTEL

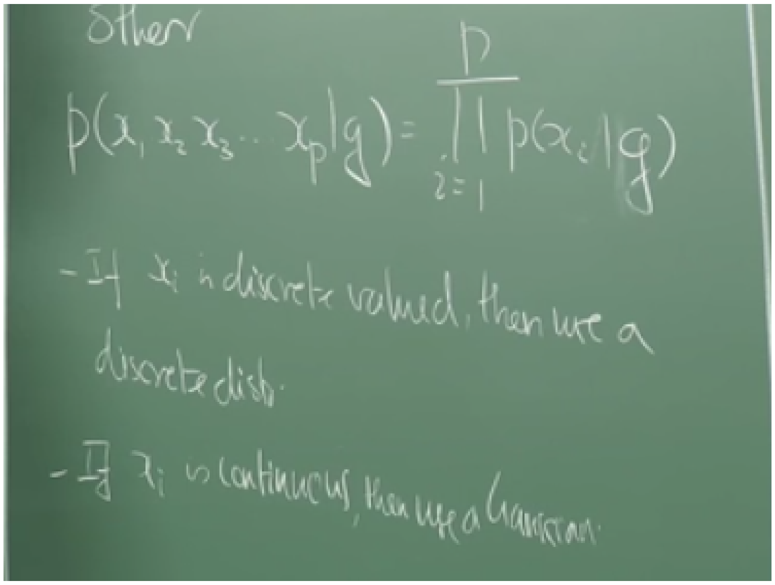
So what this is tell us it tastes that if I have probability of that i can write this as this is called as Naïve base assumption right. So once I do this not becomes very easy to for me to estimate parameters right, so what I will do now okay in how many data points has xi taken a particular value of that particular class right, so I first segregate all the data points by class and then okay in class one in how many data points did xi take the value of zero in class one how many data points did xi take the value of one and class two how many in the how many data points xi take the value of zero and xi I take the value of 1 so on so forth.

Example I want me to say XA takes the value of 0 XA takes a value of 1, XA take value 2, XA takes the value 3 I know that just takes a lot of time so just making it binary so that it is easier for me to speak right, so it could be anything right, so xi could be real value this for example our always our setting is our r $^\wedge$p right in that case what do you do some billing you could look good in fact that is one very valid the option infact even though many text books do not recommend that but not that they do not there is not that actively not recommended they do not even talk about it okay.

But that is actually a valid option and I will tell you why in a minute but the usually recommended option is to have some kind of parametric form for these marginal distributions these are conditional marginal's and if you think about it these are basically marginal's if so this was the Joint Distribution right then this is just the marginal that is the conditional Joint

Distribution, so this is a conditional marginal so for the conditional marginal they ask you to assume some kind of parametric form and the usual one that they suggest is a Gaussian right.

(Refer Slide Time: 12:00)



Other

$$p(x_1 x_2 x_3 \ldots x_p | g) = \prod_{i=1}^{p} p(x_i | g)$$

- If x_i is discrete valued, then use a discrete distb.
- If x_i is continuous, then use a Gaussian.

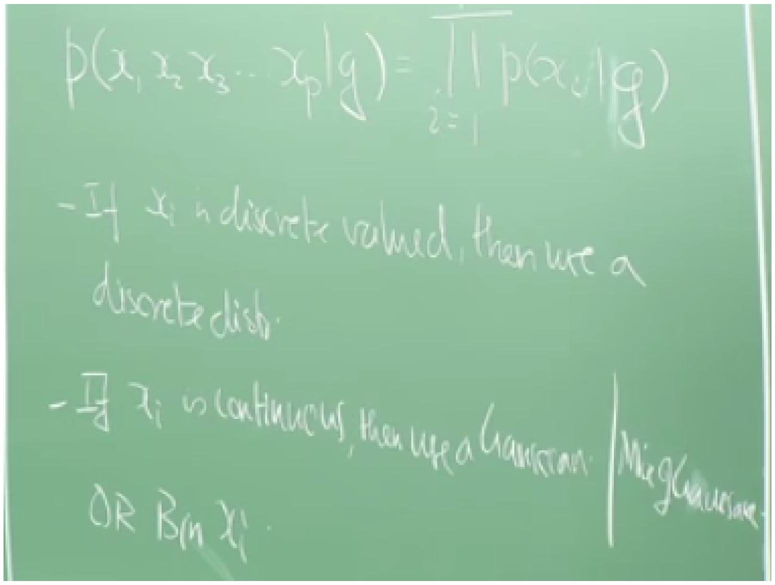
Usually this is some kind of a Gaussian form for this conditional marginal’s right can you read it at the back can okay we can read it okay fine now because the thumbprint stay up long enough so I was not able to make sure which direction it was you know is it something like that and right but then what is the problem in using a Gaussian all of you know the problem just think for a minute not lesser okay not a big issue as much bigger is not it too much affected by us but the wrong kind of inductive bias why is it wrong what do you know about the Gaussian.

Sorry close so what does it mean in terms unimodal let the Gaussian is unimodal right suppose there are two different values of xi which are separated right which are very power probable right for this particular class if use the Gaussian what will you estimate the most probable pointers there are mean okay, so if say 3 is very popular very likely to occur and 5 is very likely to occur your Gaussian will say 4 is the most probable value for x given the class right , so you do not want that to happen right.

So that is the reason I said winning makes sense but then the problem is you have to find the right kinds of bins because when you are using the discrete distribution there is no I mean it can be multimodal right I mean it is no notion of like you unimorality there right so I could have one output having very high probability another output having very happy I could have like 10 different outputs having high probability is everything else having low probability it could be anything right.

So there that I do not have to worry about these but then the problem here is hard to find the right binning so that is a whole set of lectures by themselves, right so how do you how do you bin your input variables there are many ways in which you can bin input variables in there you have to keep coming up with clever tricks depending on the application you have and so on so for this is actually not trivial right and you could do that.

(Refer Slide Time: 15:32)



$p(x_1 x_2 x_3 \ldots x_p | g) = \prod_{i=1} p(x_i | g)$

- If x_i in discrete valued, then use a discrete distb.
- If x_i is continuous, then use a Gaussian. | Mix of Gaussians.

OR Bin x_i.

But of course the Gaussian is just a simple example if you know that the data is going to be multimodal right what should you be using mixture of Gaussian is not a single Gaussian is a mixture of Gaussian and we will get you will get there right, so how well you think Naive base would work seems to be a very simplistic assumption is even called Naive base even though I have actually never heard used in any papers or any literature so has to Tiffany Friedman says that it is also known as the idiot based algorithm.

So Naive Bases sounds a little better more sophisticated right yeah do you think Naïve based will work well in did work when really well right when India missed you did know how simplistic it was at the time compare it against SVM's did not we compared against the SVM’s should so it turns out that the let us say for examples like text classification where you are the data dimensions are inherently very high right it is incredibly hard to beat naïve based you know it looks like it is something so simplistic right.

I mean look at the assumption you are making I have a lot of texts I want to classify them as politics or sports okay that is one very simple problem that is out there is a standard problem that people use for text classification is a standard data assets I want to classify a news article being sports on for politics and what I am I saying here so given that it is ports the probability that i will see cricket is independent of the probability that I will see football.

Not only that the probability I will say cricket is independent of the probability i will see say Dhoin in the document sounds like nonsense right well if you are talking about Indian media is right but in general right so it seems very surprisingly, so that is because we are trying to assign all kinds of semantics to what is happening and but the algorithms that we are trying to use whether it be SVM or anything else or really not into the semantics of these things right, so we are only worried about it because we have all these others the knowledge base super structures that we have built.

And that we are trying to look at the data through that right at the end of the day if you look at it is more a question of things occurring together co-occurring and in a very large document space right so the probability of any words co-occurring right this is kind of diminishing right, so if i do not know it is a document about cricket right and if i see the word Dhoni then I will say okay now maybe the probability of me seeing something with actually a sports related term goes up because I did not know whether it was a sports document or a politics document before I looked at it right.

But given that I know it is a sports document and the probability of me seeing this anyway be higher it is not going to change appreciably because the word dhoni appeared you see the reasoning if I had not known anything about the document because the words Dhoni appeared in it right the likelihood that it is a sports document goes up right well given the nature of Indian cricket the probability that is politics has not gone to zero yet right so but if I had known that it is already at the sports document knowing that okay it is Dhoni the words Dhoni appeared in it really does not appreciably change the other probabilities okay.

So that is essentially the idea here because the number of words are very large if there are only like ten possible words or ten possible values of these features can take then there will be appreciable change even if I know that value for one features but each of those words can take something like 10,000 different outcomes so it ends up becoming actually a pretty good approximation right when you are doing it in practice if you remember I told you something like Knn would be bad if you are in 100 dimensional space or a thousand dimensional space.

And imagine text is a very large dimensional problem right, so what will be the typical dimension for text classification yeah so naively modeling it you can get 24,000 right you can do some kind of feature reduction and so on so forth try to reduce it to smaller state space but still it I pretty large right 10’s of 1000 dimensions so if you try to do k in that things are not going to work that well but people still do that right so all this cosine similarity based retrieval systems and other things that people used to do in the past but all nearest neighbor kind of techniques.

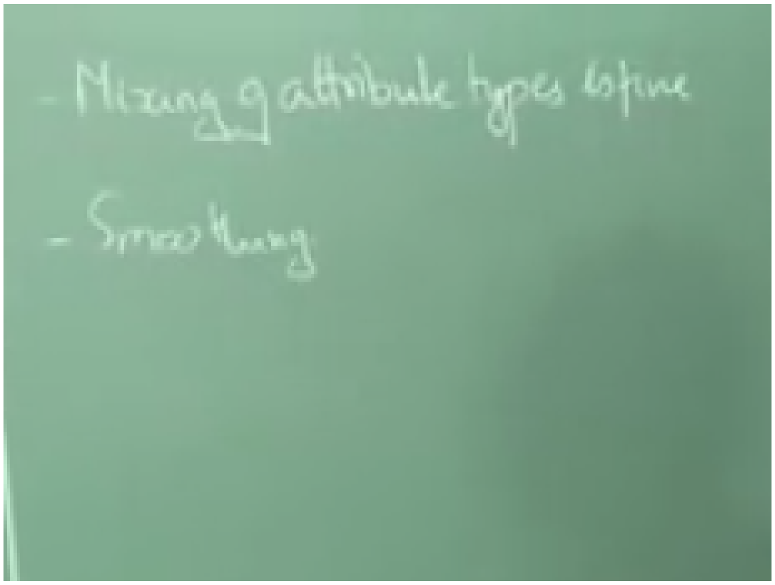
And not that they do not work but something like Naive base when you want to do classification works tremendously well in this you can see this name challenging you go try it out against SVM right you will find that maybe there is like a couple of percentage difference right and the math is so much simpler there are a few things which I should point out about naïve based one of the biggest advantage of Naive is that I can have mixed mode data and I can have some of the attributes being discreet valued some of the attributes being continuous valued and everything.

And it is very it is fine for Naive base so what other classifier can you say that trees right anything that is based on trees are also pretty robust when it has kind of mixed and you do not even have to do any kind of normalization of those attributes also and I can just keep each attribute one thing can run from one to a million other one can run from 0 to 0.1 is all fine right when I am looking at other more numeric classifiers in the sense of the kind of computations that they do they do distance-based computations and things like that.

There I have to normalize things a neural networks if something goes from one to a million and other one is going from 0 to 0.1 the feature that runs from one to a million will overpower the feature the runs from 0 to 0.1 right I should Ideally be normalizing everything to 0 to 1 right so those kinds of things I do not have to do in naïve base it is all clean and nothing the second thing I do not have to do is any kind of feature encoding normally i do not have to do any feature encoding I do not have to convert this into some kind of code.

So I do not have to take red and convert it into some code words so that I can feed it into my neural network wait how will I feed right into my neural network you have to encode it some more right whether you use the RGB value for the color or whether you use some other encoding for the color we will have to do something about it right I do not have to worry about all of that thing this the same thing in addition trees as well right.

(Refer Slide Time: 23:45)



- Mixing of attribute types is fine
- Smoothing

So that is one thing the second thing I want to talk about second thing I want to talk about is how do you handle missing values right if X is continuous if Xi is continuous we are all fine right where anyway using a Gaussian the Gaussian has a infinite supports or something else which you have never seen in the training data comes will always assign some probability to it but what about discrete value things right, so I have trained you I have trained my classifier by looking at data right that has red, blue and green.

In that thing and in the test time somebody comes along with yellow what do I do I can assign it a probability of 0 and it says different so what if I do not have all the exercise one more exam I mean things like neural networks would cover the tried I mean I am something I have not seen before hand as long as have some encoding for it will just take care of it I do not need to have actually seen it in the training phase we do not know xi is going to come or not say we have to compose account for all the unseen excise somehow.

So there are multiple ways in handling this in naïve base one thing which you can do is you can just ignore that attribute do not multiply it in and make it zero just ignore it look at all the attributes you do know the probability is for okay and then multiply that is a problem with it what is the problem sorry you are assuming the probability is 1 right, so your role will be overestimating the probability, so you will have to come up with some mechanism by which you will normalize that right if you are using lesser features that they come up with some mechanism by which you are normalizing them right.

So that is something which will have to think about the other one is called smoothing right smoothing is essentially similar to what issue is saying earlier is that you assume that everything is everything that you could possibly see right has occurred at least once in the training that means that you will give it some probability at least you will not make it zero and will also take care of this overestimating problem you will not make it one it will make it like 1/10,000 or something that will be very small probability that you are signed to all the unseen values in the training data.

So when it test time at least you will not assign zero value to that data point right if everything else is very probable except that the color is yellow right so you will not make the probability 0 the probability will get depressed significantly but it at least not go to 0 okay, so this is one thing the problem is smoothing is the following if there are lot of values that you do not see I suppose that is color I see only four colors in training but my actual color spaces 256 right so I mean obviously rays are my actual colors fade 64,000 okay and I see only four colors in training what will happen dreaming.

I am just talking about practical issues here if I do smoothly in such a situation what happens obviously it means you are training data is messed up so you have to think about it but if you use smoothing blindly in this situation what will happen exactly you will smooth the head out of your probability distribution right you are taking this probability of one you are dividing it among 64,000 guys, so everybody should get a count of one at least right suppose I have 10,000 training points right the back best account I can hope for is if all the 10,000 of same color it will get 10000 x 64,000 .

No I will get10000 x 74000 each one of the 64,000 i will count at least once and actual 10,000 as I happened so 74,000 le 10,000 by 74,000will be the probability for the color which occurred in all your training data points okay that is a really small value for the color so smoothing you have to be very careful when you apply smoothing right so if there are too many unobserved values and smooth it blindly like so you will essentially lose all the information in the training data you will have to come up with other mechanisms like she was pointing out there is something wrong with your training data first right.

So you have to go back and try to fix that see if we can generate more representative sample so that is those are the things that you should look at right, so anything else right so we have we know how to do p(g) right so we know how to do p(x) once we know how to do this and this we know how to do by doing this so parameter estimation is taken care of and this all the ancillary things.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in