NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

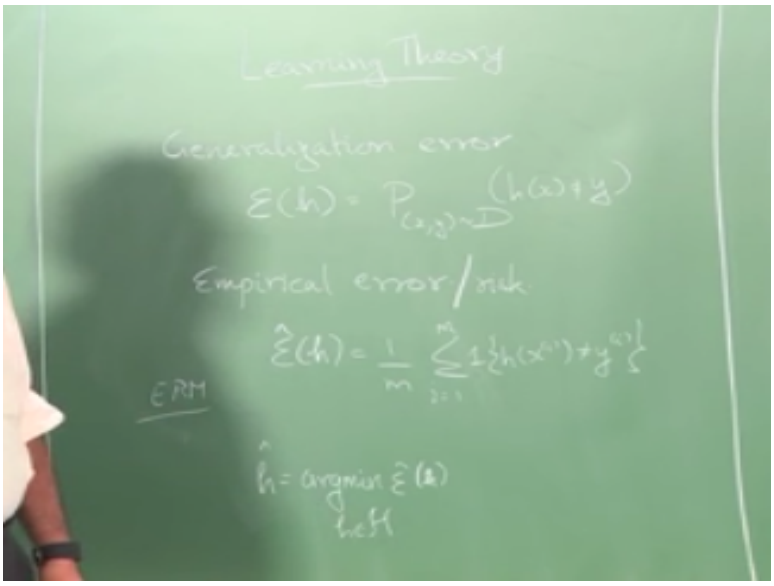
Lecture-80
Learning Theory

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

Scripted science books so all of you are aware that theory and computer science means talking about hardness of problems right, so how are is to solve and looking as space complexity, time complexity right, and approximability right. So how approximate the solution those kinds of things right. So we are going to try and see if we can give such a flavor of theory to machine learning as well right.

So we are going to talk about how hard is the problem to approximate right. So I have a perfect solution right, but then I can get close enough to it right, but how often can I get close enough to the right solution right. So that is the one kind of question that people want to ask. And then there is another question on how hard problems I have to solve in general right. So we are going to talk about hardness of problems right, so different methods of the hardness of problems right.

(Refer Slide Time: 01:16)



So typically we are interested in generalization error right. So what is generalization error? Right, so after the notes right, this will be completed different notation from what I have done earlier in the class right. But just stick with the notes and I will give you these notations. So that is the generalization error of a hypothesis H, so I will denote my ε the error function right, and half h means the error of hypothesis H right.

So the expected number of times the hypothesis will make a mistake or the problem is okay, the error of H right, this is the probability of the hypothesis H making a mistake on a data point X right. When the data points X and Y right, where X is the input and Y is the label right, so where the data X and Y sampled from some underlying distribution D right. If you are achieving that this distributing D is fixed a priori and unknown we remember about all of this right.

So we always talked about this, I talked about the P(x, y) earlier, but here we are talking about the distribution as D. So when the data is sampled according to this distribution D, so what is the probability that I will make a mistake right? So this is also the expected number of mistakes, because several mistake will count once right. So the probability of making mistakes will also be the expected number of, right.

So this is essentially right, the probability of this is what we call the generalization error right. But typically what is the error that we have access to, we have access to something, you know so we often called a, we have access to something called an empirical error right or sometimes

denoted as empirical risk right. So which we will denote by ε^ okay, so ε^(h) is equal to right. So where xi and yi are the i^{th} data point given to you in the training set right.

So the empirical risk for the hypothesis H is the number of times, so what is the 1 is the indicator function. So 1 will be the, this function will be 1 if this condition is true it will be 0 otherwise right. So essentially when it will be 1, whenever I make, whenever h makes an error this will be 1, whenever h is correct this will be 0 right. So I will add it up for all the training data, so here I am assuming that I have m samples in my training set right, I will divide by m.

So this gives me the probability of me making a mistake as estimated from the training data right. So this is sometimes known as empirical error or empirical risk right. So typically I only have this quantity right. So whatever is given to me as a training data right, so I can have many ways in which I can estimate this error, but this is all I have access to. So this is called the empirical risk. What I mainly interested in this is the generalization error right.

So what we want to know is how good is this empirical risk estimate that I make in terms of measuring the generalization error, or how close is it to the generalization error right. This is the question that we want to ask. See this has shades of hypothesis testing right, the shades of hypothesis testing, but, so we are going to do a very different kind of analysis here right, because shades of hypothesis testing, but the kind of analysis we do here will be very different, right.

So what where the question that we are asking is okay, so given that I can estimate empirical risk right, what can I say about the generalization error okay. So before we go on to look at this in more detailed right, I want to introduce a couple of results which would make easier for us to talk about, before that let me talk about one thing. So most learning algorithms that we have do what is known as empirical risk minimization right.

So the answer that you will typically end up giving is, so you will have some hypothesis class H right, suppose you are looking at median classifiers and your input dimension is say some P right. So then you will be essentially looking at classifiers that are defined by θ_0, θ_1, θ_2 and β_0, β_1, β_2 up to β_p and then given by the inner product of that with the data point. And if it greater than 0 you will classify it as one class, if it lesser than 0 you will classify it as another class that is basically what linear classification does right.

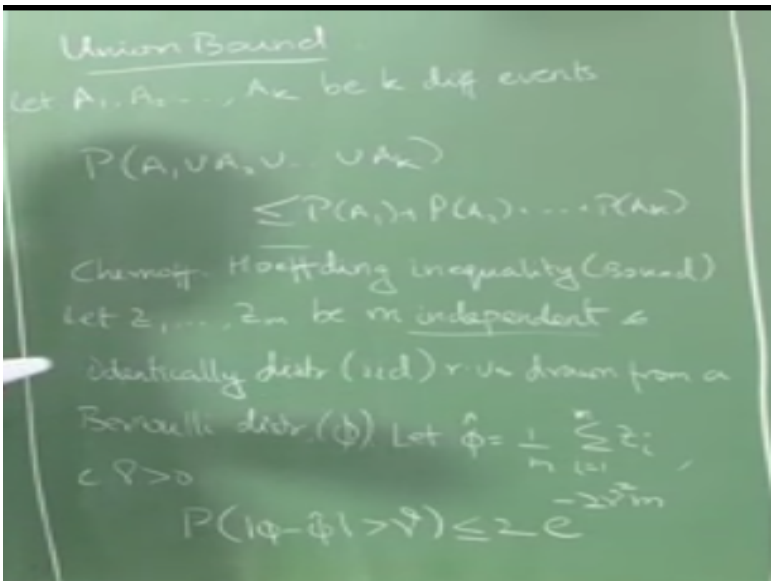
So when we look at the bright of different linear classifiers at the end of it you will have something like that right, you will have some $\beta^T x$ right and then you will have some function of $\beta^T x$ and then whether that is greater than 0 you put it with one class, if it is lesser than 0 you will put it in the other class. So that is essentially what the hypothesis class H would be right. So for example, in case of linear classifiers this is the same in the case of neural approach, it will be all the classifiers that we can implement given the choice of number of layers and number of neurons per layer that we had made right.

When I make some number of neurons, some number of layers choice right, so for different weights that you can set for all those different values you can set for all those weights you will get a different classifier right. That constitutes your hypothesis family scrip age okay. So what I typically would like to report is that H okay, that has the minimum empirical error overall members of that family H okay.

And that H^ is the classifier that we will report by doing empirical risk minimization, I mean so this is essentially called empirical risk minimization. When this is ideal case right, obviously we know that we do not get this, that if you are using neural approach or training using back propagation or neural lens and such like that, we actually do not find the admin right, you essentially have some approximation of it, then you just stay with it right.

So likewise, depending on what classifier you are using you might not actually find the minimum right, what you are trying to do is minimize the empirical risk anyway right, because that is the only thing that we can what, so only thing we can measure right. So is it clear so far the setting is clear right. So we would really like to get a classifier that is good according to this right, but we are not able to do that. So we look at empirical risk right, and then we find empirical risk minimization and we get this right. So before I move on, I want to introduce a couple of results right.

(Refer Slide Time: 10:07)



So the first one is called Union bound, it is a very, very simple, how is it should be familiar with union bound right, yes, no maybe. How many of you said yes? Two, you do union bound into operating tutorial? We did not found in tutorial right? Anyway, right, that A_1 to A_k be K different events, then probability of, okay. In most emergence of probability theory this is taken to be axiomatic okay. So this is called the union bound right, so it is equal right, they are independent right, or they get disjoint, we just thinking of the message they are disjoint okay.

So the next thing is, right so this is something which is variably called a Chernoff Hoeffding bound or the half ring inequality or the Chernoff of inequality when some subset of these two and some subset of that will be sued for describing this result right can you do a Chernoff found also in the probability may be after I write the bound you will know okay so here I am stating this very specific to Bernoulli distribution but the Chernoff bound holds for in general okay.

There are other milder conditions it need to necessarily be Bernoulli but as far as we are concern we are only interested in binary outcomes right so can you guess what is outcome I am interested in correctly classified or not correctly classified so that is the outcome I am interested in so we are only considering Bernoulli case in this case right but this result is much more right in fact there is a version of it where you can also relax the independent assumption but gets more and more complicated okay.

In fact when you relax independent assumption you get you get some kind of you get the chromatic number of the interrelationship graph can enter the picture okay I still have not figured

out how the chromatic number enters the picture in that result okay it get really complicated let so let us hope and pray that all the random variables we deal with out independent or you can think of this like that right but in some case that is not true you will have to worry about it, so in this case essentially what I mean is so the probability that some $Z_I = 1$ is say some ϕ right.

The probability that $Z_I = 0$ is 1- ϕ right so I will just keep it as sum Bernoulli distribution parameterized based some ϕ okay so what do we know about the Bernoulli distribution ϕ is also the mean right we know the ϕ is also the mean right so typically the Chernoff Hoeffding bound is stated on the mean okay but this version of the Chernoff Hoeffding bound is stated on the parameter ϕ yeah but it is when it is not in the role as the probability but the ϕ is here used in the role of mean.

Because that is one thing I want you to remember like I do not want to look at the Chernoff Bounds and then go actually flip through some other place and find that nowhere is the probability of outcome is mentioned right that will never be the case because Chernoff Hoeffding bound is stated on the mean. So what is $\hat{\phi}$ it is the mean estimated from these random variables right I mean this all be familiar to you from the hypothesis testing case right this ϕ is true mean of the distribution and $\hat{\phi}$ is the mean that is estimated from these random variables the samples that I have drawn okay.

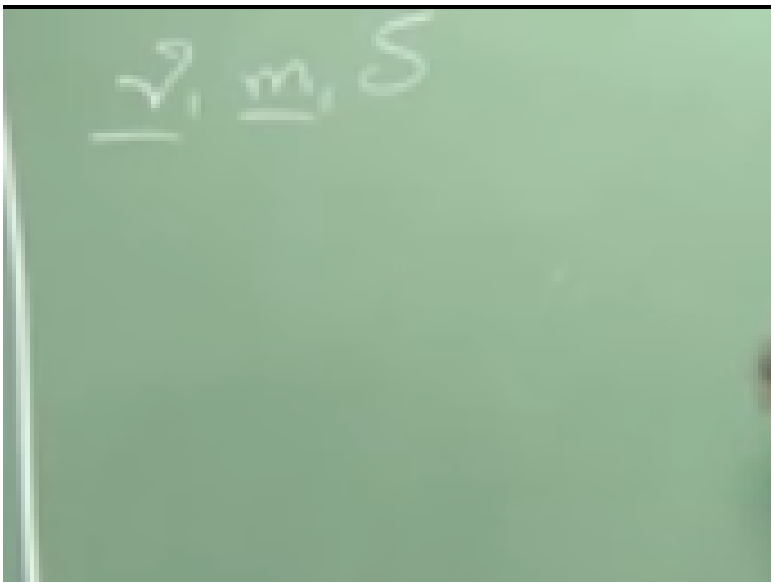
Right and at γ be ≥ 0 there some fixed constant $\gamma \geq 0$ the so the probability that $\hat{\phi}$ is always from $\phi \geq \gamma$ okay so what is this mean $\hat{\phi}$ is always from ϕ by $\geq \gamma$ $2e^{-\gamma 2m}$ right the 2 comes because this is a 2 sided inequality if you think about it right it is $\leq \gamma \geq \gamma$ right both work is it two sided inequality therefore the 2 comes so if you want to look only at 1 side the you can drop the 2 and the use only $e^{-\gamma 2m}$ right.

So essentially what does it say if I have lot of samples right if m is very large right I have $e^{-\gamma 2m}$ so it becomes smaller and smaller right as m becomes large the probability that my $\hat{\phi}$ will be far always from ϕ becomes smaller and smaller right so this gives you a way of quantifying how many samples I need right before my estimate that I am making right before the mean that I have estimated is close enough to sorry the mean that I have estimated is close enough to the true mean.

Right with the high probability right so γ is something I fix a priory okay γ is something okay I need you to at least this accurate forming okay now go and tell me how many samples I need alternative like I can ask question like okay I have so many samples okay how accurate I am likely to be right is it fine is it enough to fix the number of samples, it will be little bit you have to be little bit more work but okay.

So there are more multiple things here one is okay what is the probability that the error I am making is ≥ γ so that is this quantity right so I need supply γ right and an error for me to find m right I need to supply m and an error for me to find γ right is this make sense there are 2 quantities here right actually 3 quantities here.

(Refer Slide Time: 19:53)



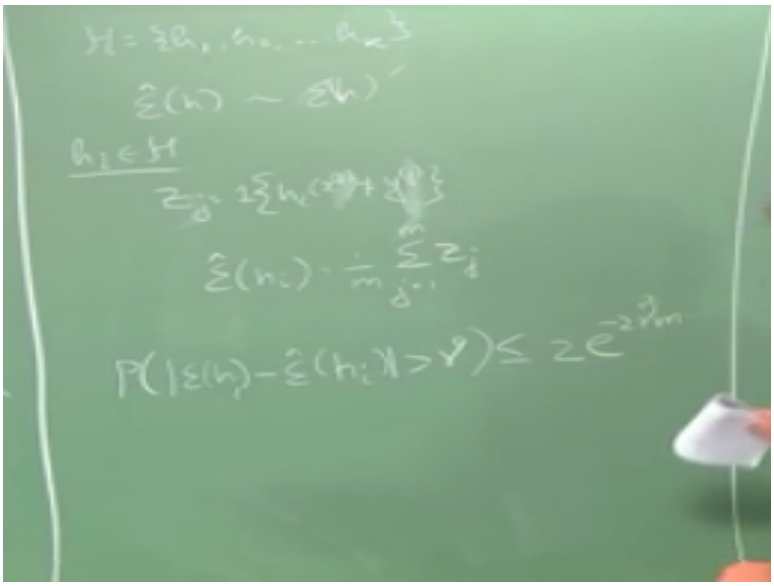
So there is γ okay there is m and there is also the probability of the error right so γ is the magnitude of the error right the people get that and what is the probability that I am ≥ γ, so that is a 3 quantity right, so I have number of samples I have the magnitude of the error and then I have

the probability of making an error of that magnitude okay that is the left hand side right so I have 3 things here okay so this equation as 3 things here so if I want to solve for γ so I need to specify the left hand side.

Okay and I need to specify m the I can solve for γ let I say okay I do not want to make a mistake more than 10% of the times okay that means my probability should be 0.1 okay and I am only giving you 100 samples okay then I will come back and tell you if you want to be correct 90% of the times then you will have to say that even if I am so far away from the right hand side I am correct.

Okay only then I can give you the 90% grantee so that is essentially what I means by saying there are 3 things here you have to specify any two then you can think about it in terms of the 3 one okay great so now we have these 2 results for us okay so usually that probability is denoted by sum ∂ okay so we have this 3 things.

(Refer Slide Time: 21:34)



So we will start of a case where I have k only k specific hypothesis in my hypotheses class right I am only searching thorough a space of k, k can be very larger right I am not tell how small or large k is but I am only I have only k hypothesis in my class and I am going to search through this makes slightly easier for us to developed some intrusion right and the we can go back and go on and talk about the ∞ h okay.

So we want to look at how ε ^h corresponds to εh okay for some hypothesis h how does ε^h right and I am going to some h ∈ h_i ∈ H for the time being right and the Bernoulli random variables that wanted here as we mentioned earlier are going to be defined by so random variable z is if hi(x) not equal to y the it will be 1 it will be 0 otherwise so whenever h_i makes a error then this random variable one whenever h_i does not make an error the random variable will be 0.

So we can go head and write, write zj for each xj as xi xj0 = yj right so if you remember we always make the assumption that the training data was given to us in an IID fashion, independent directly distributed fashion right from very beginning we have been saying that training data is IID each sample was taken independently of one another and we used this fact even in the hypothesis testing case again we will use the fact here and that allows us to apply the often bounds right there allows us to apply the often in equality.

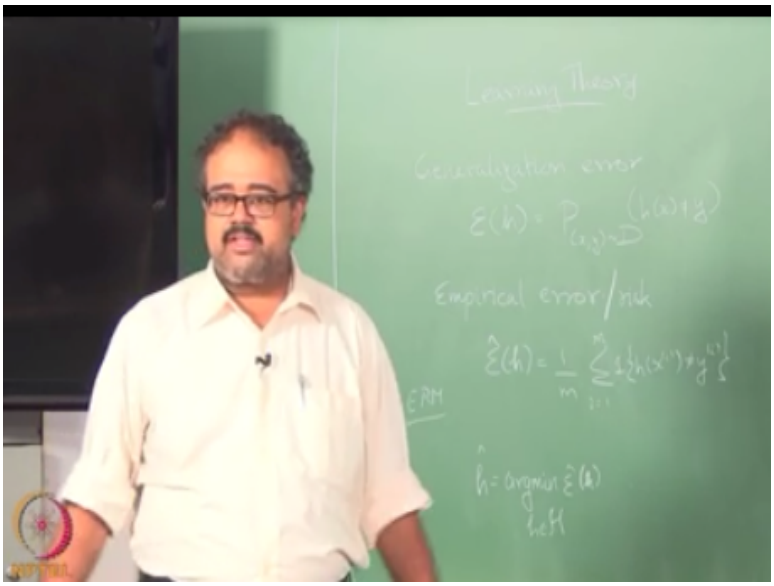
So what is that we have so I have if I have m training points we said we assume there are m training points, therefore I have this m independent identically distributed random variables okay where us the distribution that I a drawing this from, from the true distribution of the data right so I am assuming that all the points are coming from the true distribution of the data, right all the training data points have come from the true distribution of the data right so that is something which is very important.

These all the two main assumptions that we are making here what is the first assumption, IID the second assumption the training data comes from the true distribution right so that is the distribution according to which I want to evaluate the generalization method right so these are the two assumptions you make right, of course we can always relax this assumptions in fact quite of we need to relax this assumptions because in real life we are not be able to satisfy either of the assumptions right.

But this is good enough to give a some kind of a intuition has to how things will work right and then we can worry about relax in this assumptions later on. Right so that is why ε^ ha so we know that right so epsilon is already define there, right so all I have done is take whatever expression there was there in the some there and define that as zj and I get this so already we know that so what is εh that I say true mean right what is ε^ so it is estimated by taking m samples so you can directly.

Apply may some of bounds so next so the probability that what is my ϕ here what is ϕ here εh okay we considering a specific thing okay yeah.

(Refer Slide Time: 27:38)

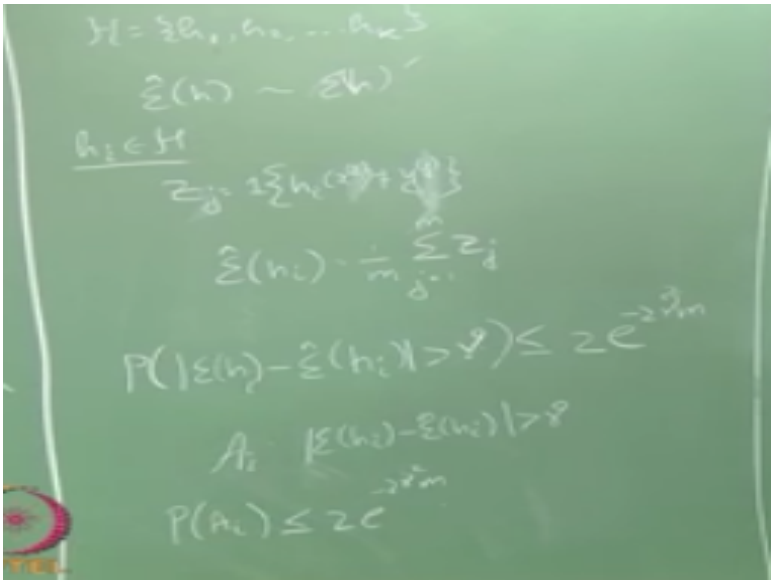


So if you have a very large class right then you have to be very careful about minimizing the empirical risk, because you have you run the risk about fitting it right, so yeah so essentially what you are trying to do in many of those cases is try to get a better estimate of the error on the better estimate of the you know generalization error actually, so when you are trying to do the validation right.

So essentially trying to get a better estimate of the generalization error directly, so with all of these things or essentially to give you some notion of what is it, some notion of the complexity of the problem that you are trying to solve okay. This just gives you some notes of the complexity of the problem that you are trying to solve and we will see that in the minute you will see that right.

So in fact one way to avoid over fitting in neural networks is to have a very, very large training set right if you have lots of ways and we need to have a very, very large training set that will kind of all out of this just give me a second to explain this.

(Refer Slide Time: 28:48)



So what we are now shown here is that for a single hi from the hypothesis class if I have a large m right then the error will be the probability of having a large error will be small, right so or even say I will be pretty close even for small γ right suppose I want to know what is a probability that this greater than ε greater than ε^ is the different between ε, ε^ > than let say 0.01 right so that is what I want to know so my ε^ should be within 0.01 of ε okay then what I will plug in here is 0.01.

Right so $e^{-0.012}$ so essentially getting to 0, 0 or 1 right essentially getting to this since what I am saying is the probability of the error > than 0.001 > 1 sorry < 1 then really tell me anything right is just < than 1 ready at a I mean I know that, right but then if my m is very, very large right suppose I want this to be 0.01 this is 0.012 then I will let say m is a million right so then what will happen is I will have e^{-20} or something, right so that is a small number right so that so if I say have a lot.

Of samples then the probability of ε^ being 0.01 closed ε right will be high right what this states is the probability that it will be 0.01 away from ε is slow right, so the converse is if it is the probability that it will be 0.01 closed ε will be high so that is essentially the result that we have what we have shown here using the hoeffding bound so I will show the proof of the hoeffding

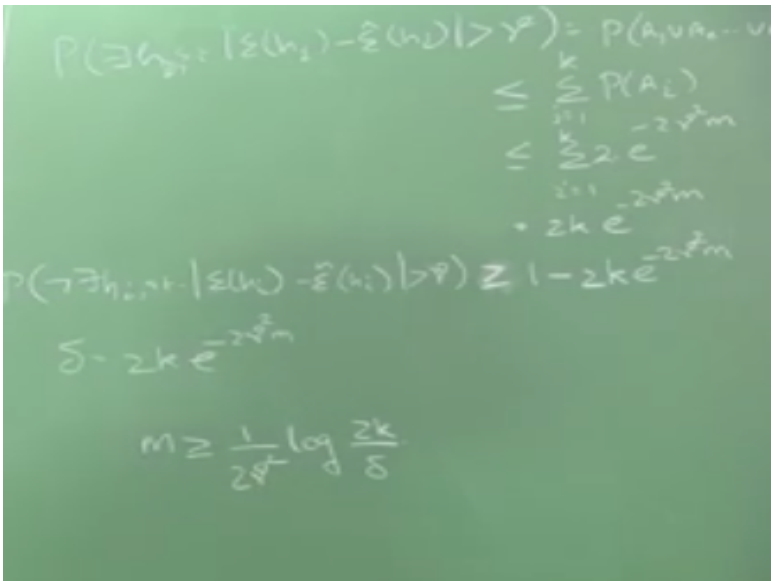
bounds it is not real right so interested you can look it up right this is not very hard okay it is just have to work it out right that is all.

So but once you accept that on faith right though you have the result but unfortunately this holds only for one particular ha right that is not very interesting I am essentially what I have shown is you can give me like a 10, 000 different hypothesis and I can show you in for one of those hypothesis.

If I have a lot of samples then I will be close, what I really want to show is okay if you give me 10, 000 samples are give 1m samples for every hypothesis in this hypothesis class, I will be close right so every hypothesis is in hypothesis class I will reclose so how where you going to do that use the union bound right so what I will say is I will define my event ai is I need those a1 to Ak right, I will define the event Ai as so define this event As ε and ε^ being more than γ away right so now prove this becomes probability of Ai.

Right now this is essentially because of Ai is $< = 2\ e^{-2\gamma 2m}$ right so now I do union bound so what is the union of A1 to Ak right what does it mean at least one of them giving a higher error right, so essentially this reduces to.

(Refer Slide Time: 33:43)



Where exist h right so it is that this is some ha okay at least 1h right so probability of at least 1h so that is what there exist stands for right there exist me there at least 1h for which the error is large right so this is essentially equal to the probability of right so the probability that there will be at least one error there may at least 1hi which has large error is bounded by 2k $e^{-2\gamma 2m}$ g for one of them it is to $e^{-\gamma 2m}$ right into 2γ2m now for k of this it us just k x k this is essentially what w get from union bound okay.

So now what I need I need the probability that, that does not exist any classifier that makes a large error if I give it m samples okay that does not exist any classified that gives a large error so what will that be 1 – this right similarly 1 – that right so the probability that does not exist hi, okay. Yeah is it clear what we have done here right this is simple algebra here so we got this and so this kind of a result which holds for all H in the complexity class where mean in the hypothesis class we are taken right if because this now this results holds for all H and capital H right these are called uniform convergence results, right. So this is the uniform result.

Because it holds for everything in the this is more like a single result right this is for a specific hypothesis well this result that we are giving that $(1 - 2k)e^{-2}\gamma^2 m$ that bound is for all hypothesis in this hypothesis class so this is called a uniform convergence result okay, when somebody says uniformly convergent that means set it works for all classifiers right it works for all classifier, okay. So far note that I am not actually talked about finding the classifier, right.

So what is ε, ε given a classifier what is the error in the classifier will be making in the overall population right and ε^ is the error that it makes on the training data so I am just comparing the two right I am not actually talked about finding the classifier right so what we should not be looking at is, we should be looking at comparing okay I will come to that in a minute. Right if you remember I said there are like three quantities in the beginning right.

So now I am interested in solving for m I want to know how much how many samples I have I want I can draw before I can give a certain guarantee on okay I can give a certain guarantee on the performance right what I mean by performance here, whether my empirical error i8s closed to the generalization error I am not talking about the best generalization error when I am talking about performance here I am talking about whether my estimated error is closed to the true error, okay.

So how many samples should I draw before I can give some guarantees on the performance of the estimator okay. So I want to solve for AM right and do I try to fix, I need to fix γ as well as the error probability as well as this guide right so, k is fixed for me right I give you the hypotheses class right as soon as you give me the hypothesis class K is fixed for that right so I am going to says that the probability should be utmost this quantity should be utmost δ like some δ I will give to the δ so this quantity I will fix so the probability should be $1 - \delta$ right and I will give you the γ also right solve for M right δ will be right, now I can give you γ also so I have given some value I will give you some number for δ.

I will say δ should be 10% right so I will δ should be 0.1 right I will give you the δ I will say δ should be 0.1 I will give you the γ also okay, now solve for it. So what does $\delta = 0.1$ mean that 90% of the time this event should be true okay so I will give you a γ I will give you a δ you find M for me. Yeah, so ensure uniform convergence yes, no but then what this tells you is you chose a small hypothesis set to ensure uniform convergence.

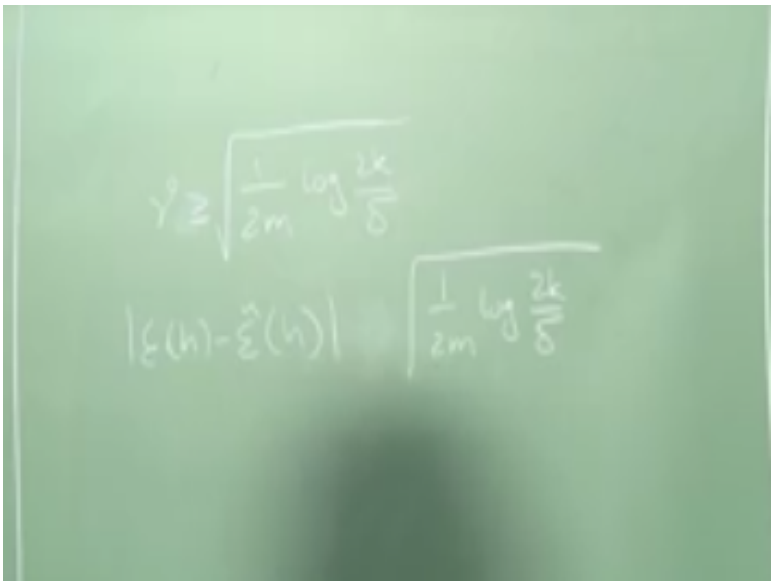
So that it will tell you what the true error is it does not tell you anything about how good the true error is okay right so that is so that is what I kept re-iterating to say what performance means in this case, performance is not minimizing the error performance is minimizing the error in predicting the error okay this is kind of a circular but this is trying to minimize error in predicting the error, okay that is all.

You are in measuring the error okay so you solve the remaining that is easy so this is called the sample complexity right this tells you how many samples you have to draw so that what, so you are all your classifiers are within γ of the true classifier right, the probability of that happening is at least $1 - \delta$, right. So this kind of formulation are called pack formulations PAC so you know what is PAC is?

You know what PAC is now probably approximately correct right, so no that probably path comes from that, right. The approximately part comes from that so I am not telling you it is correct okay it is approximately part comes from that so I am not telling you it is correct okay it is approximately it is within γ of the right answer but is it always within γ of the right answer no, no it is with high probability it is within γ of the right answer, right.

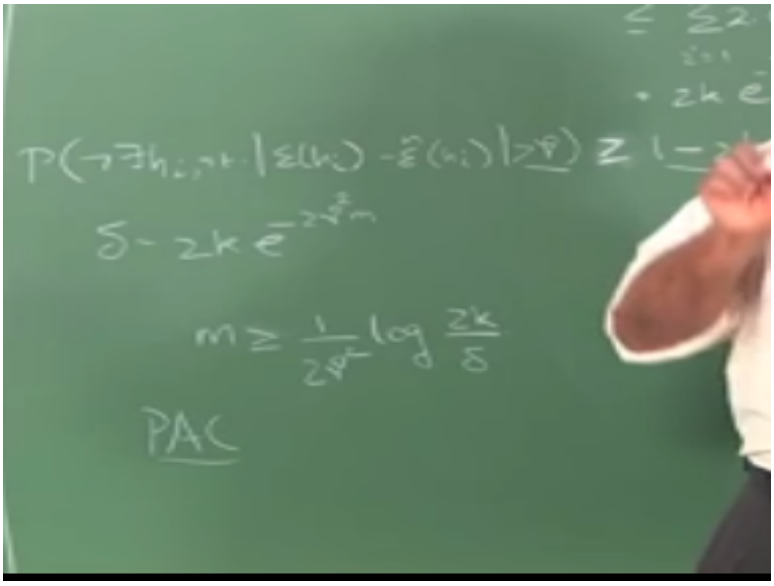
So it is probably approximately correct right but in many cases this is the best I can tell you because there is so much variation in the samples that you are drawing right and the problem itself has inherently has noise in it say there is only so much I can do in terms of predicting it correctly. So in this we have saw interesting so the another thing which I want to look at is I give you m and δ okay. Can you solve for γ? Right I fix m and δ solve for γ what you get fairly straight forward right.

(Refer Slide Time: 44:50)



Okay likewise if you want to solve for δ you can try these things right but this is fairly easy system right so what is γ really it is the error in the prediction of the error right so it is ε – ε ^ as γ so essentially I am saying that ε – ε ^ will be wait, a γ will be yeah.

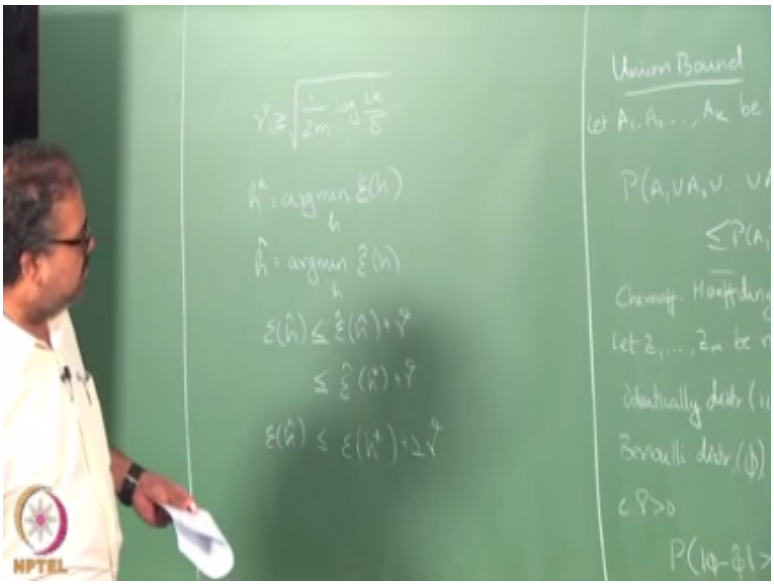
(Refer Slide Time: 46:50)



So δ is this quantity 1- δ is the probability of not making a mistake so δ is a probability of making, this one. So what happen I just solve this expression here and so what do I get? And δ we are ready to take E, yeah so see M should be equal to this to give you the exact δ probability right we give you exact $1 - \delta$ probability but I wanted to be at least $1 - \delta$ right, so that is why I say it is $\geq$ right so I will get at least $1 - \delta$ if M > than, if M equal to I will get δ exactly $1 - \delta$ for M $\geq$ I will get this something greater than $1 - \delta$ right. So that is essentially what we are looking at here and so using the same argument what should we be looking at, right?

So my γ yeah so should be greater than or equal to here right, γ should be greater than or equal to so my γ should be at least yeah, γ can be at most this small right, γ can be at most this small but it can be greater so then I can give you the guarantee so that is essentially what we are looking at here right, so this is no needed.

(Refer Slide Time: 48:50)



So we will define h* to be, h* is the true, the truly the best classifier that it have right, in hypothesis class h right, we will define ĥ to be that classifier you will pick by doing empirical risk minimization right, ε had your classifier you are picked by doing empirical risk minimization okay, so knowing whatever we know can be write things right, so $\hat{\varepsilon}$, ε(ĥ) is less than or equal to $\hat{\varepsilon}$(ĥ)+γ why is this true because of whatever we have shown all the value right, so I am saying that right, I have taken enough samples EM so I can say with some probability this event will hold, right.

Because of my uniform convergence with some probability that some probability 1-Δ this event will hold right, because this will be γ close to the two error right, the $\hat{\varepsilon}$ will be γ close to ε, right. So why can I say this so ĥ is produced by doing r mill over $\hat{\varepsilon}$ right, so that means that h* should either have a higher error than ĥ or at least equal rate cannot be a better error than ĥ because we have better error than ĥ then ĥ would not have been chosen okay, does it make sense right, this I using the visual of the fact that I did minimization here right, and therefore $\hat{\varepsilon}$ (h*) should be what is then $\hat{\varepsilon}$(ĥ).

Using uniform convergence again right I can peel out one more γ from there right, did it may sense what I have done here is it is exactly I went from here to here right, from here to here how did I write this γ because of uniform convergence right, similarly what I did was I took $\hat{\varepsilon}$(h*) and I said that it will be within γ of ε(h*) so then I add up the γ here so I get 2γ right, so all of this

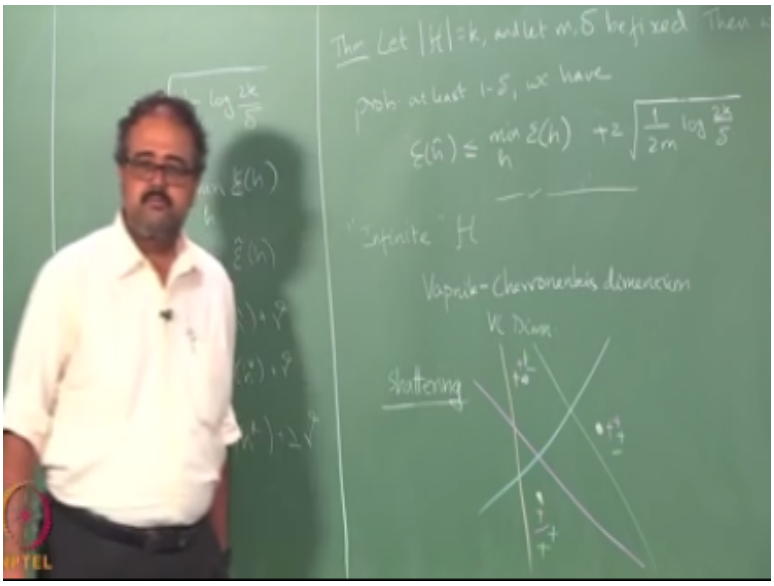
will hold with some probability but I am assuming that I have taken enough samples and I have converged to whatever degree of accuracy I need, right.

I have converged to some probability 1-Δ I have converged here, so once I have converged so this will hold so essentially what I have is that so what does is mean so the true error of the classifier I produce d by doing empirical risk minimization is within 2γ of that true error of the optimal classifier okay, does it make sense the true error of the classifier I produced by doing empirical risk minimization of which is ĥ is within 2γ of the true error of the classifier that is produced by minimizing the true error is essentially the optimal classifier, right.

So essentially this gives you the guarantee that if I do empirical risk minimization right, taking that many samples right, then I will be within 2γ of the true classifier anything else I need that, with probability 1-Δ right, so if I take at least this many sample EM right, so the classifier that I find by doing empirical risk minimization will it be within 2γ so our γ is say number you plug in here right, within 2γ of the optimal classifier at probability 1-Δ, so that is essentially the result that we can have, right.

So let us put this together and write small theorem and I can erase this side I would not be using this again so.

(Refer Slide Time: 54:32)



Right, so essentially I have written our γ from wherever we solve for γ yeah here, right I wrote the γ expression down there right, and we get this okay, right. But what is ε(h*) right, so ε(h*) is actually minimum over h of ε(h) so this is essentially the best possible classifier that you can build okay. So now we have this result we can go back and think about the question you asked sometime back right, if my hypothesis class is small right, essentially what I will be doing is I will be searching over small number of things to find the lowest error so it is likely that my solution the best solution I can find in this class itself will be bad, if my class is small, right.

But if my class is small the second term is small right, if the number of think if the case small the second term is small, but my case large right, then the likelihood of me finding a small error solution is high but the number of samples enable will also become larger, right. So this is one form of bias variance trade off. So if the hypothesis class is small that means that regardless of how much data you give me I will always be making some error because I have only a few hypothesis and I search through.

So that is like a bias you know it is like doing linear regression kind of thing right, and this is variance because hypothesis class is very large then I will leave the lot of samples for me to estimate the error properly, right so that is the variance path so this is like one version of the bias variance trade off that comes in so that is the reason you need large hypothesis class so that you can be sure that you contain you have the right solution in there.

But if you know exactly what is the solution we are looking for then you are better of using a much smaller hypothesis class okay, so that is essentially the to take away message here, right. So we already know what is the sample complexity you need for this result hold for a specific γ right, so here I mean the γ is given by this expression but if I give you a specific γ then say something like okay, with some probability 0.1 I want to be at least 0.1 close to the true answer right, the probability 0.1 I want to be at least 0.1 close to the true answer.

So what is the 0.1 close here it is 2γ right, so it is 0.05 right, γ should be 0.05 right, so I will plug in 0.05 and 0.1 here right, and I know what k is because of the hypothesis class I have chosen I can always find out what the sample complexity m is, so given a Δ and a γ I can find the m, okay. So this kind of an analysis, this kind of sample complexity is sometimes called is usually called εΔ where because people usually use ε has a symbol for γ.

But in this case it will be γΔ sample complexity or γΔ pack analysis because I fix the γ I fix the Δ and I ask you for the sample complexity right, so this sometimes called γΔ pack, okay. So this is assuming you have a finite hypothesis class what do you do in the case of a infinite H, any thoughts about how it extent this analysis an infinite H?

So in a practical setting right I will be typically implementing all my machine learning algorithm in a digital computer right this is kind of a cheating argument but it is fine so I will be implementing these in a digital computer and digital even though they are implementing and infinite class of things they are limited by their by the numerical procession right.

So let us say I use 64 bits to the present 14 point numbers then only I finitely number of finitely many classifiers I can represent which 64 bits right. So the problem is it is a large finite number right but I can still go back and if you look at that number I have there right if you look at the m that I need right you can think of m is being actually ignore a whole bunch of things here but m is order of right.

So you can always say that okay regardless of how large hypothesis class becomes right this is going to be log of that order of log of that so that will be a significant reduction. Unfortunately the hypothesis class becomes exponential suppose I have d I have d numbers I need to specify one classifier right and have 64 bits right, so how many hypothesis I have? So how many bits to I

need for representing one classifier? 64 times d so how many do you have and then how many classifiers I could have 264 times d that is K.

I plug in K here then it becomes D right it becomes order of D. So if I have a 1000 parameters then I need about order of 1000 samples not bad 1/λ log δ all I mean it depends on what you except out of it right you chose a large λ large δ I am just joking right, yeah of course you always have those things the λ δ and other things actually play role there right but sop think about it right I mean I have d parameters I mean to need to get at least order of D samples to solve the problem okay, how do you think it can do it less than order of d if I am do it less than order of D? Then many of my parameters are all tie together they are redundant.

So this the power of the big o notation you can hide a lot of things under the big o umbrella because I hiding all your $1/\lambda^2$ and log of 1/ δ you are hiding under the big o umbrella but it is not a bad thing okay so this is 1/f thinking about it right it is not the greatest way of thinking about it but this is one way of thinking about it in fact people use a rule of them right suppose you are training a neural network which has a say 10000 weights they use the typical rule of them they uses you need at least 10x the number of x, so if you have 10,000 weights you need 1,00,000 data points.

At least right in fact this is a very useful rule of them if you are only using feet formal neural networks so remember this if they count the number weight you have and ten times that how many data points you need at least for you to give anything useful. But then there is a better way of doing it right call the Vapnik - Cherropnlics dimension otherwise known as the VC dimension right.

So given a hypothesis class we can defined the VC dimension of then hypothesis class I will finish in a few more minutes okay. So before I define a VC dimension I need to introduce the notion of shattering right, so give some set of points okay let us say some set s I give some point x1 to xk let us say a hypothesis class H is set to shatter the set s if for every labeling you can give for that set s, there is some element in the hypothesis class which actually separates the classes a binary classes okay.

So every possible binary labeling that give on the set okay I have a hypothesis in my hypothesis class that separates it from one class from the other okay is it clear, so let me draw a picture that

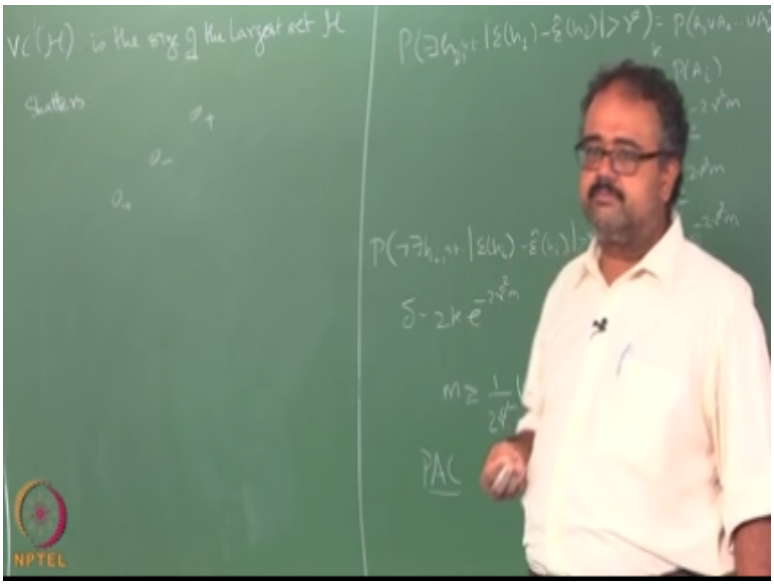
will make it clear let us say that, so that is my set s okay and my hypothesis classes all straight lines okay. So think of all possible labeling I can do for this right. I can basically set on say everything to one side of the line is + everything to other side o0f a line is minus great.

So now what we do everything to one side is plus everything to other side is minus right so likewise I can keep going I keep going as long as there are different number of colors here and then guys can kind of intrude right. So give the set of three points you can just see that right so if I flip the plus and minus it is exactly the same right so you have to only worry about the unique things right so if the flip the plus and minus it is just the same.

So if I make this plus and these two minus the same thing will work right. So is there anything else that we need to consider? So ++- +++ ++- that should be a – right, that will work anything else so I have to consider 2plus and one minus two minus and one plus is just flip of it okay and three pluses I have considered and three minus is just the flip of three plus okay, so anything that we can to consider? That is it right; I will leave anything but anyway even if I left out something you can make it up right, so you can easily see that.

So straight lines so hypothesis class of straight lines shatters three points in space right, what about four points? Yeah we talk about single straight lines, so what about four points? There any configuration of four points which can be shattered by single straight lines, no configuration of four points that can be shattered by single straight lines that then automatically applies a five six seven eight nothing right. So the VC dimension of a hypothesis class H okay is the size of the largest set that the hypothesis class shatters.

(Refer Slide Time: 01:10:50)



Note that this is the size of the largest set high H shatters that does not means that H ahs to shatter all point of that size right even in the three case collinear points I cannot shatter right if may three points like this and I label this +-+ I know straight line that can separate this okay but there is some configuration of three points in fact lot of configuration of three points which I can shatter and therefore the VC dimension of straight lines is three right.

What about VC dimension of tars of straight lines parallel lines not arbitrary pairs. Let us say I will give you parallel lines what is Vc dimensions of parallel lines, 2 parallel lines, they have to be parallel. However you want you can rotate it but it has to be 2 parallel lines, you can 4 sure. Can you do 4? You can always keep one of them as radiant straight line you can do as much as you can, and then you get Xr you get straight lines. As everything else can do by 1 line, only the x r case ++, -- you cannot scatter with the single line right.

Remember perceptron, so it is exactly that, so for that you need parallel lines. Everything else you have 12 lines there and others you keep it as infinity or fine. What about 5 points? Now you can see why we can ask all kind of interesting questions right. So I can give you this kind, next I can give say okay, instead of looking at parallel lines, just look at quadrilateral right. So if it inside the quadrilateral it is class 1, class positive is outside the quadrilateral- okay. Now what is the Vc dimension, in fact there are uniform conversions Vc dimension as well.

So I will have put up this or write up on online and there off course other material you can find out online. I am not going to do the result because it is pretty complicated derivation right, but

then the nice thing about this is, at the end of the day just like we have staple complexity in the number of parameters right. You can show even with the Vc dimension that it is polynomial in the Vc dimension right. So if the data as the hypothesis finite Vc dimension right the uniform conversion, require you to have order of the Vc dimension of samples.

And typically it turns out that for most kind of classifiers, most kind of reasonable classifiers are out there, Vc dimension will be of the order of the number of parameters in the classifier. So we think about it straight lines. How many parameters are there in this case? 2 or 3? 2 slope and inter slope that is all you have. 2 parameters and the r in the Vc dimension is 3 close. This will be close to the parameters you have right, so and that is why I said sometimes the number of weights in the neural network rights all of these things and you can get that kind of rough intuition.

So I will stop here so if you have any questions feel free to fire away. So you could do pack line regression right, but Dc dimension is different for classification. Yeah you can corrosion the regression problem little bit then try to do dimensional classification and pack you can do for regression, essentially you are trying to look at. We defined a very specific variable random and did it, so you could define any random variable. Whatever the distribution is there, it is the amazing thing about the Sharon often bonds right.

So the result holds on the expectation, the parameter μ whatever is the expectation of the distribution, so the empirically evaluated expectation will be the close to true expectation that is the result we have. With that you can change your random variable definition and you can get something appropriate for the regression as well. What does it mean? When you change the parameters it is the different classifiers that are what I am concern, Do you mean same family of classifier? It depends on how you define hypothesis class right.

So in at the end of the day what I am really interested in this is? What is the decision rule the hypothesis class is entailing for me. So I can say that I am going to define hypothesis class is the mix of decision tress and something it is up to you but typically you define hypothesis class as a single family which is differentiated with the parameterization. But we do not actually take a call on that, all I am telling you is that okay given k hypothesis right, how are you are going to find out the.

So what I am finally, at the end of I am interested in the decision rule, okay given a data point what does it assign it to you? You derive the decision rule by the means of using a logistic regression or whether you derive the rule by means of using I don't know LDA it does not matter. For this kind of complexion, how you got to it that does not matter for me. Only in this context right otherwise it matters, with the context of the complex analysis it does not matter.

Why is our finite procreation arithmetic is little disappointing? Because there I chose D parameters right, if you have got it from some other I could re paradise thing and I can increase the number of parameters for representing the same decision surface. I want to define the straight line but I can actually increase the number of parameters that I am going to use to define the straight line, I will be saying that where ever I have, I will have –ax right. so now I will have two more additional parameters for define one should a and other – a.

And it will take out the difference of x^2 so it will still be a straight line but I have increased the number of parameters and going back to our finite procedure, arithmetic argument I also have blown up my complexity, which is casually not correct, all I am interested is the decision rule, that is the finite thing is little unsatisfied. But here we did not tell you how you could represent the line. So Vc dimension definition does not require you to know to represent the line. At the end of it you know that it is the hypothesis that I want to represent.

But the most compact way of representing the hypothesis parameters I would need. You do not have to worry about how we get to it. Anything else, I can chose any classifiers that I want from my set right, foe shattering the data points, so any effect it will 1 corresponding to the, I mean what is the most powerful, so other one will shatter that would be Vc dimension. So Vc dimension if it can shatter of any arbitrary VC dimension will be infinite. So Vc dimension infinite none of the analysis will work, all the Vc dimension analysis works assuming that Vc dimension of the classifier is finite.

I should point out that most of the classifiers we looked at do empirical minimization except? Anyone know? SVM, is called structural risk minimization because they have an additional constrained that is there apart from the empirical they also try to minimize the solution size. They try to minimize the nom o the weight factor right so that actually gives rest to a different kind of minimization. So it does not do empirical, they called structural risk minimization. So you know who came up with SVM? So vapnik from these Vc dimensions stuff.

So he said came up with structural risk minimization this is the best you can do, so we need to have a different way of doing the minimization, then he motivated and then he said and came up with empirical structural risk minimization you can do. We will go and do something else and he can do something else structural risk minimization and then he divided SVM with them, if you read original presentation of SVM right it will not look anything like how we presents SVM now days. If you remember we started off with perceptron and then we went from there.

He starts from the theoretical things, he said okay if you do empirical structural risk minimization it is the best you can do what else I can do and then I can improve on it and then he came up with structural risk minimization okay and then he derived all those systems okay.

IIT Madras Production

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

www.nptel.ac.in