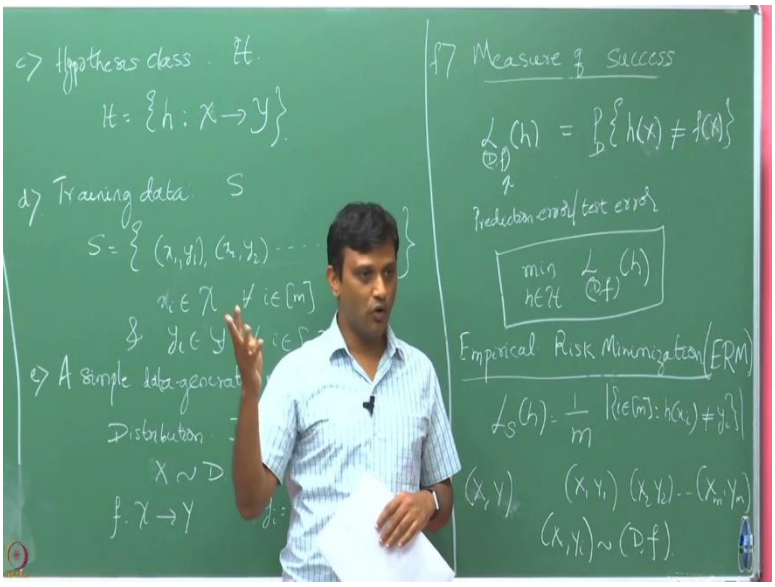
Bandit Algorithm (Online Machine Learning)
Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

Lecture – 04
Empirical risk minimization

So, this is the standard trick we always use in probability. If we do not know the underlying distribution, we go with the samples of this distribution and use the two have some empirical estimates of the quantity that we are trying to optimize ok. So, empirical estimate of the quantity we are trying to estimate.

(Refer Slide Time: 00:48)



Here $L_{D,f}(h)$ is the quantity if I know I could go into this optimization right, but I do not know this. So, I want to get the empirical estimate of this using my data samples. So, the standard thing here is to do empirical risk management.

So, now, what will be a good empirical estimate of this quantity from the data samples I have? One thing is or what I can do is I can say my $L_s(h)$. So, S is the training data that I have been given, I can use this and I am going to write an empirical estimate for this as

$\frac{1}{m}|\{i \in [m] : h(x_i) \neq y_i\}|$. Let us understand what is this doing.

I have S is given to me S is given to me what I am doing is and I am trying to compute this quantity for my hypothesis h. So, I will just see on how many points I agree with the value y_i; y_i is the label right. So, this quantity is just give me this. So, this is the set. So, the value within these braces is going to give me a set of all the points, the data points on which my hypothesis h does not agree right. So, y_i is the corresponding label of x_i I am just counting how many points my h does not make a correct prediction and now I am dividing it by m.

So, what is the size? So, I am taking the cardinality here; that means, the numerator will tell me the number of points where my hypothesis h has made incorrect predictions and I am dividing it by m. So, what is this quantity is telling? The fraction of the points on which my hypothesis has made a wrong prediction.

So, next; so, henceforth what I will do is this x_i; so, (x_1, y_1) these are actually one sample right, but instead of right I am just going to write them as (X,Y); that means, this X is a random variable, Y is also random variable, but this Y depends on X. So, with the function f; so, this X is a random variable which is drawn from my distribution D. Once I know this, this Y is simply the function f of that point X. So, these are my points like I have this is my random variable and these are just like realizations of this ok.

Now, suppose this is a particular realization, but let us say let me say that my I could have $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ This could be all of them drawn from the same distribution right. So, this (x_i, y_i) coming from the same environment. So, these are just m points drawn from the same environment. Now, if I am say that these points are drawn independently and anyway they are identically distributed as m goes to infinity, what you what this quantity is equal to?

Student: (Refer Time: 05:55).

Yeah?

Student: (Refer Time: 06:58).

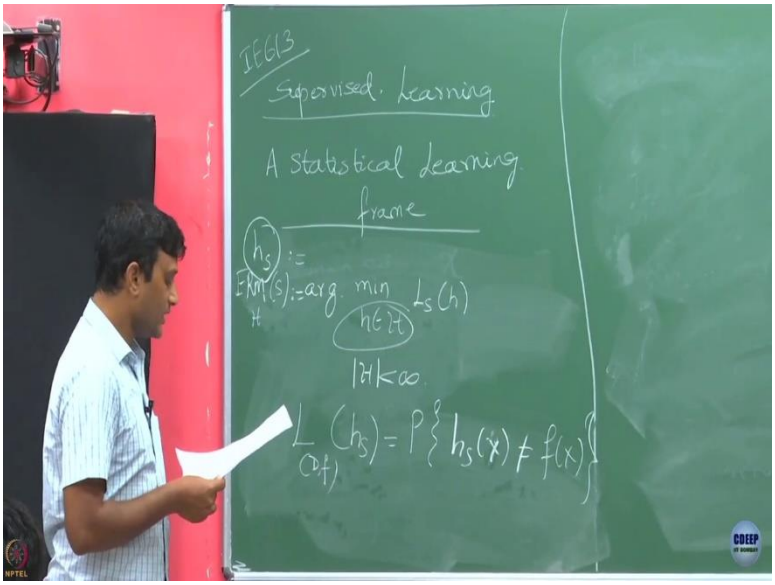
So, is this quantity is going to be equal to this the probability of the disagreement? Why is that? Which what did you what is what did you apply to conclude this?

Student: (Refer Time: 06:09).

So, if this x_i pairs are IIDs, independent identically distributed I can apply my large law of numbers after number says that this has to be equals to this right, but it is just that m is not infinity here, m is some finite number that is why we are going to call it empirical risk. And, now this quantity $L_s(h)$ I am going to use it as a proxy for the quantity $L_{D,f}(h)$ and now, I am going to minimize this over h, that is why we are going to call this method an empirical risk minimization.

Now, so, what I will do is I will do this empirical. So, this has in general popularly just we are going to call it is as ERM.

(Refer Slide Time: 06:59)



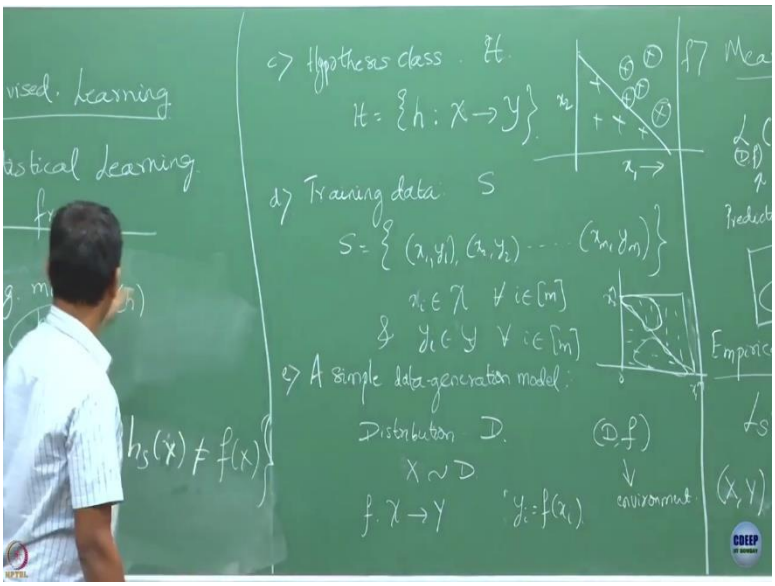
ERM technique what it is going to do is $\underset{h \in H}{Min}\, L_s(h)$. Now the goal I am trying to I am my goal was to minimize $L_{D,f}(h)$, but it is just that I do not know this, but what I did is I found a proxy ($L_s(h)$) for it based on the quantities I know and then trying to minimize $L_s(h)$.

Next, the question is, fine, if you do like this whatever this quantity is the minimum value of $\underset{h \in H}{Min}\, L_{(D,f)}(h)$, if you do $\underset{h \in H}{Min}\, L_s(h)$ how far these two quantities are away from? Are they very close to each other? Or they will be very far from each other? Your goal was to

minimize $L_{D,f}(h)$ right, but you have ended up minimizing $L_s(h)$. Then, you would like to ask the question whatever I got get from doing this how far this value is going to be from this value ok.

And, whatever the hypothesis I get from this let me call this as h_s, the hypothesis I am going to get I am going to call it as $ERM(s)$ and this is hypothesis. So, notice that it depends on the training set you got and over the hypothesis class you are trying to find your all your hypothesis ok.

(Refer Slide Time: 08:49)



So, this is one training set (S) that I have given to you from which you got your empirical risk minimization hypothesis. If I change this S can your empirical risk minimization hypothesis can change? It can potentially change, right? Because, these quantities are being explicitly used; all these data points have been used coming from set S if you are going to say it S it is going to change ok.

Or may be like I will instead of this I will also; I will also going to use it as $h_s = ERM(s)$, this empirical risk minimization. So, subscript s denotes makes it explicit that this hypothesis depends on your training data ok.

So, now your training data whatever that is given to you it is generated from this particular distribution D, right? Every time you are going to generate this data, you may end up with different set of data points because of this, this set S itself is a random quantity and because

of this h_s can be also random right it depends on the data set you have used to generate to learn this hypothesis. So, S, this h_s is a random quantity depends on the data set, but as I said what kind of guarantee we can give about the hypothesis we got by doing this empirical risk minimization with respect to whatever the best one ok.

So, there is when we do all these things the hypothesis class matters. If we are going to change hypothesis class things can also change. What we are going to now just to understand what we wanted to see like the difference between this minimum and this minimum, let us focus on the case where we have only hypothesis class which has only finitely many elements in this ok. Before that, how many of you know what is over fitting?

Student: (Refer Time: 11:23).

So, what happens in over fitting?

Student: (Refer Time: 11:28).

No, just tell me what happens on the training data?

Student: (Refer Time: 11:33).

So, what will be this quantity ($L_s(h)$)? Let us say if you have a.

Student: Very low.

Very low?

Student: Yeah.

If you do if you are over fit what is this quantity for some hypothesis you can always find a hypothesis which will ensure that on your data points it exactly does the current job. If it does the correctly job this quantity is going to be 0, right?

Student: Yes.

It may do a very good on the training one, but when I look at the actually the test error here where the points are not just these points right the data points they could be also coming from other parts it may not do well. So, there are some things like the how big is

your hypothesis class and all they are going to affect your over fitting issues. So, for that like that is not of our interest. So, we will not go more into that, but what we are going to assume is this hypothesis class is finite size, that is the cardinality of the set (H) is finite ok.

So, if this hypothesis has like infinitely many components in this right, you will always come up with a hypothesis class which will over fit and make this loss 0 right and in that case you will your empirical risk minimization always end up showing you a hypothesis class which is like a over fit on the data. So, when you have only finitely some you restrict your hypothesis class that kind of over fitting you may avoid. So, let us for time being assume that this hypothesis class is has a finite to many elements in this ok.

Now, let us say whatever hypothesis you got from your training data, I want to now measure its performance. So, what I when I will then how I am going to do this? So, here the h data has been you have found out is your h_s based on your training data, right? Now, you want to see what is its success. So, what you will do is then whatever h_s you have got, now I want to I am interested in this quantity ($L_{d,f}(h_s) = P\{h_s(x) \neq f(x)\}$). So, you have done this empirical risk minimization using data set and you have come up with this hypothesis h_s. Now, I want to measure its performance.

Now, I would say your empirical risk minimization has given you a good hypothesis, if I can show that this quantity small right or like if this quantity is 0 that is the best may be like, but I want to can I claim that this quantity is going to be small. So, what all the things that governs how good your hypothesis h_s? So, what do you think what you think you will come up with a better hypothesis h_s that will make this quantity smaller?

When you think what are the things in your setup that are going to affect how good my hypothesis h_s? So, I am assuming that these data points are generated in an IID fashion, right. So, instead of giving let us say let us take m equals to 10 let us say I just give you 10 points and then using those 10 points you come with your empirical risk minimize. Let us call that as a h_{s1} there.

And later I will give you hundred points I added 10 more 90 more points to your 10 points. You have 100 points and using these 100 points you come up with this another h_{s2}. What do you think h_{s1} will be better or h_{s2} will be better?

Student: (Refer Time: 16:14).

Why? So, when you have larger yeah so, you will be approximating this function more up more accurately and may be because of that you will end up with this. So, naturally m has a role to play here right, how many data points you have on what how good h_s is going to be. So, notice that like there is also I said this could have m points, but each of the points here itself could be generated independently right.

So, let us for time being let us say that let us take a 2-dimensional case. Let us say my x-axis denotes humidity levels and my y-axis denotes. So, instead of this let me call it x1 component and x2 component denotes my humidity levels, sorry temperature levels and let us say these are the points. So, these are some points I have observed. Let us say when it rains? When high humidity and high temperature when it is going to rain?

Student: (Refer Time: 17:48).

High humidity, low temperature ok. Let us assume we are like some weather say it weather experts I do not know what are. Let us say in this region let us say so, the in this region it going to rain. All these points has the label that they are going to rain and below these point these are like not going to rain. Yeah, I am just like this is a hypothetical case it is not necessary there should be an I separation like that it could be much more complicated.

And, like so, these are data points, but let us say. So, this is let us say this is my range of humidity. Let us say humidity goes from some value 0 to maximum some value let us call this x. So, this is like $x_1 \max$ and my temperature also 0 to let us say some $x_2 \max$ something like this ok. Let us say the true the true like let us say weather the true in the nature the actual way it happens is they never let us say in some fashion. So, in this let us say it is all going to rain and in this it is not going to rain this is how let us say the actual environment behaves.

Now, what we are doing is you are basically getting samples from this right over the past history you are basically getting. So, this is rain and this is no rain. Let us say you are now

collected data from the last 10 years and the last 10 years happens to be like a drought years may be of. So, that means, there are not enough rain in those case itself. So, because of that most of the data you collected happens to be from this portion from the last 10 years, where data set will have data mostly coming from this portion ok.

If you are going to train on this portion of the data, do you think you are the hypothesis you are going to end up will have a will do a good performance? No, right? Because it did not get the full things here like it did not get the full picture. So, it may not be able to do. So, and also like if it. So, if you just see this what it will feel that it looks like this in this part of the world it is always going to be a drought. Every year it is going to declare that it is going to be a drought, drought, drought and everybody will panic.

But, let us say if all the data point like last 10 years happens to be again like flooding or something and you got lot of points from this, this is also going to be bad. You are whatever you are going to get it is going to predict it is going to rain, rain, rain and it is not going to be a good thing. So, which from which portion of your distribution the data came also matters right because of that the S itself is going to matter how good or bad your hypothesis that is why you have also subscripted.

Notice that we have saying that this points are IID generated right when and I am saying this points are getting IID. So, when I am saying this IID generated, it is unlikely that all the points are sampled they concentrate on one area. You get this? So, maybe when I generating this IID maybe possibly they will be spread more in this and maybe they will give me a true representation of what is happening.

But, still when you are generating IID with some small probability you may still get samples only from one region of this, right? Right it is not like even if you generate IID, you will get with high probability samples from all over the region high probability run. It you may end up with the some small probability you it may happen all the data points you generated could be coming from this small region which could be bad representation of this.

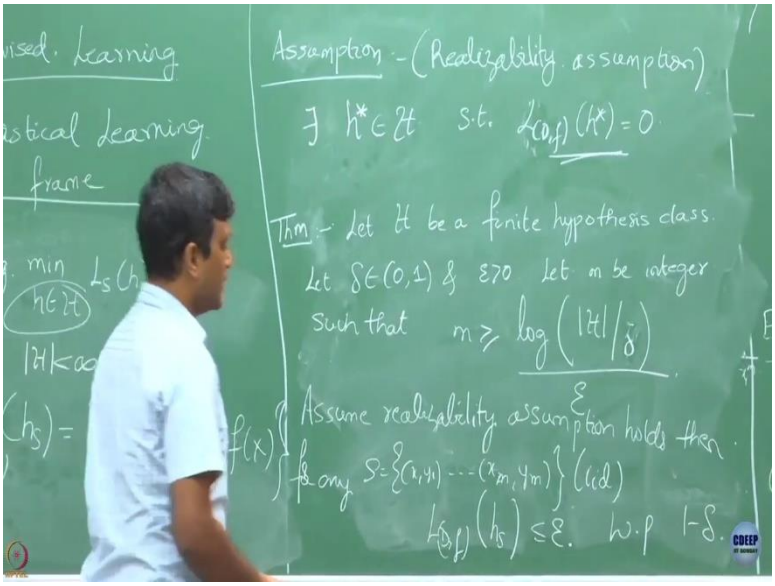
Because of this whatever I have here this quantity here itself will be determined by the data samples I am going to see and not only that what is the size of the data set. You may have large samples, you could increase m to large number of points, but with some small probability you all of them might have come from a small region, small area which could

make your empirical risk minima hypothesis is really bad because of that this value could be also bad.

So, in this case we want to see that whatever this hypothesis we learned from this given data set S what kind of guarantee I can give on this. So, is it true that if I have this large number of points in S that is enough to make it arbitrarily small? Need not right because as I said even you may have large, large, large number of points, but those sample points might be coming from some small region which could make this a bad representation of your space ok.

So, the kind of guarantee we are going to give henceforth that is why depends a probabilistic guarantee we are going to give and we will only going to say that how small this is going to be clear ok. So, before we give any such things we to make a such reasons concrete we may require some assumption.

(Refer Slide Time: 24:12)



I will just tell you what I mean. We are going to assume that there exists a hypothesis in a hypothesis class such that it is going to give me 0 loss on all possible realizations of my data set. So, let me make it formal. So, with this assumption you are going to call it as realizability.

So, we will assume that there exists a hypothesis class h^* in my hypothesis class h such that whatever this quantity I have whatever the prediction error I am going to get whatever written as such that. So, what does this imply? We were basically saying that.

Student: (Refer Time: 25:58).

h^* equals to f with probability 1 ok. So, it may happen that on some points it may this h star may not assign a same level of f, but those number of points have 0 mass. So, let us say this assumption holds ok. If let us say if this assumption holds, is it true that if I compute this quantity for any S on h^* what will this value be? That will be also 0, right? So, this basically means that you take any data samples I will have an h that will have a 0 loss ok.

Now, once I assume that there exists such a h^* which is belongings to my hypothesis class. If I go and do this empirical minimization whatever that minimizes this value. So, we know that at least there is one h^* here that is going to make this quantity 0, right? Whatever h_s we are going to have that is going to have this quantity to be 0 and there could be more than one, but what we are interested is the one which minimizes this ok.

So, let us assume this. This is can be all easily realized, but for time being let us say that my hypothesis class is expressive enough such that it capture the true labeling phenomena; that means, the true label it has it can do as good job as a true labeler that is it can be as good as a the true labeler labeling function f that is why we are saying. So, if this happens then we have the following result.

Theorem : Let H be a finite hypothesis class. Let $\delta \in (0,1)$ and ò> 0. Let m be such that

$m > \dfrac{\log(|H|/\delta)}{\text{ò}}$; realizability assumption holds the for any $S = \{(x_1, x_2), ..., (x_m, y_m)\}$ (iid)

Then , $L_{d,f}(h_s) \leq$ ò with probability $1-\delta$.

So, you have to get used to such the statements now onwards. So, let us try to understand what is this result saying ok. It says that let us say H be a hypothesis class, the H as finite and a δ is given to you and òalso given to you. Then, if you are going to choose your m; m here is going to denote the number of data points to be larger than this.

And, if you are going to give me a data points which are generated IID and it has at least this m number of points in it, then whatever this error whatever the hypothesis class I am going to learn by empirical risk minimization. It is (Refer Time: 31:24) to have a loss which is smaller than ò and that is going to happen with probability $1-\delta$. As I said I cannot give a deterministic guarantee on this because even though I can give a lot of samples the samples may come from some bad region. So, I have to I can make this guarantee I can only give probabilistic guarantee.

Now, let us try to understand how to interpret this result. I have let us say I you are in a company you have been told to find give me hypothesis whose value is going to whose true value this quantity is going to be less than ò with probability $1-\delta$. So, let us say your boss has given this parameter to you we cannot guarantee anything in world with probability 1, right.

In being realistically we can guarantee something to happen with some probability, may be that probability is very high. Like, me being alive tomorrow is very high, but it is still probabilistic right there is always some maybe there is also some case that probability that sun may not raise tomorrow is not with probability 0, it may not rise some with some positive probability.

So, we have to give similar like. So, I am let us say I am also asking the same thing keeping that in mind I will ask give me this guarantee with this probability and my guarantee has been specified by two things. I want loss to be not more than ò, it should be less than ò and this should happen with this probability $1-\delta$. Yeah, this delta.

Student: (Refer Time: 33:16).

So, these are the parameters, that has been given to you. If anybody ask you to guarantee anything you also can not give him that guarantee right, you will be also demanding. You are going to tell see I will guarantee you this, but you give me this many data points m number of data points. And, how with this m determined?

Like this. If you are asking for òguarantee and with δ accuracy, I want you to give me this many samples, then I will guarantee you that whatever the hypothesis I am going to give that will have a loss which is less than òok. So, if somebody ask you get me this job done

in 1 hour or something then you will ask give me 10 GPUs, 10 servers I will get it done in 1 hour.

So, like this somebody is asking you to give you with this guarantee then you are going to say if you give me these many data points then you will be guaranteed your guarantee ok. So, you see this. So, this is a kind of wanted to go through quickly this supervise setting like all of you who have done this course in machine learning when you touch up on this learning theory this is what you are going to do there right.

So, this is kind of this is touching upon the sample complexity. If you want to learn something with some guarantee you need certain number of data samples ok. So, this kind of results lead to sample complexity result. Now, but this is not our interest right, this is just like a precursor.

What we are interested is I do not have a priori access to this m data points that are given. Like in this case you have already been given m data points, you train on this and try to apply it on a new point and you are trying to guarantee that how well you are going to perform on the new data point right and you have been guaranteeing such things.

Now, I am going to say I do not have luxury of collecting all these data right. Suppose, let us say in the weather case right like somebody was collecting data for the last 50 years and then give it to you and he is asking you to predict, even that we do not do well right like we have so much of historical data, but our weather forecast are always wrong. But, what so, that is one case fine.

But, what it may if happened like what happens if you do not have any such information at all, but you want to make a good prediction right from day 1, how you all going to go about it? Fine, on day 1 you have nothing. You cannot do anything, but after doing something on day 1, you have some information for day 2 right. Can you think all doing something better than what you did on day 1 and you do it is something on day 2?

Now, for the day 3 you have information about day 1 and day 2 can you use this to do something better than in day 3 than what you did in day 1 and day 2? And, as information gets to you can you quickly try to do what is the best you could have done. So, that is where we are going to get into this online machine learning here. So, this is in this case we have already bunched our data we want to train and guarantee what we are going to do

from the whatever if we learnt from the past, but in my online setting I do not have this luxury.

Student: (Refer Time: 37:30).

But, we are going to be ambitious and we are going to set our goal as I want to do quickly as soon as possible. And, you will also see that we are going to set our goal as what we are doing is it how good it is compared to the case when I knew everything. Like I am the oracle if I am the oracle I like nature myself; so, when I say nature it is like about the data generating process. If I knew everything I know what is the best how could I have done right.

Now, I am going to always compare myself to that like I do not know anything I am started getting this incremental observations using that how quickly I am going to do as if I knew everything. If you could do as if you knew everything; that means, basically you have learnt. And, now the question is how to setup this and how to characterize all these things. So, that is what we are going to start looking it from the next class and there we are going to see various settings ok.

So, let us stop here.