Bandit Algorithm (Online Machine Learning)

Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

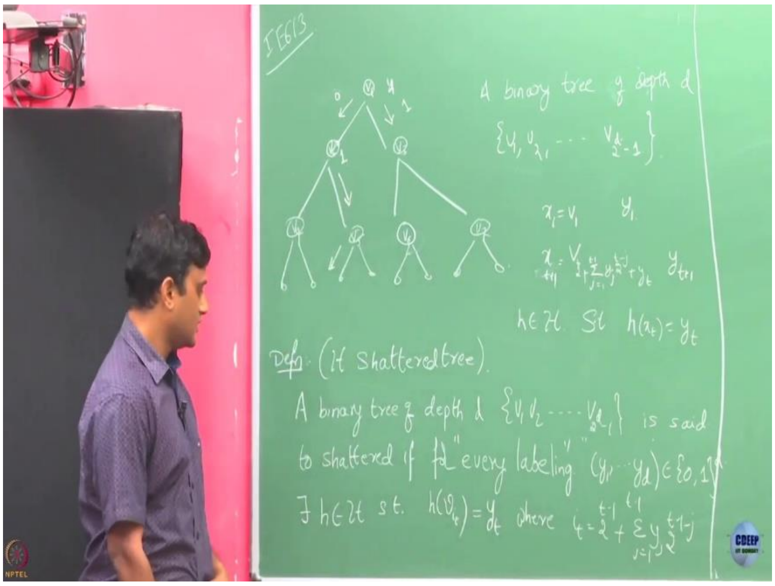
Lecture – 07
Standard Optimal Algorithm

So, let us get started. So, today we are going to bit look into what is the best you can do when you are in an online learning setup. The online learning setup we discussed in the last class that is what is the smallest number of errors that you could incur, and that number of possibility errors is kind of unavoidable for you. What is that number?

So, for that we started introducing some setup to understand what is the best the adversary can do to inflict the maximum number of errors on you. So, for that, we started defining a graph and then we introduce the notion of shattering. So, let us repeat a some aspects are there. So, all of you are able to go through what I asked you to do in the last class that is to understand what is VC dimension.

So, VC basically kind of captured what is the complexity of the hypothesis class you are going to learn right. So, there is a similar notion for the online learning setting also that we are going to discuss today ok.

(Refer Slide Time: 01:43)



A binary tree of depth d

$\{v_1, v_2, \dots v_{2^d-1}\}$

$x_1 = v_1$ y_1

$x_{t+1} = v_{2^t + \sum_{j=1}^{t-1} y_j 2^{t-j} + y_t}$ y_{t+1}

$h \in \mathcal{H}$ St $h(x_t) = y_t$

Defn. ($\mathcal{H}$ Shatteredtree).

A binary tree of depth d $\{v_1 v_2 \dots v_{2^d-1}\}$ is said to shattered if for "every labeling" $(y_1, \dots y_d) \in \{0,1\}^d$ $\exists h \in \mathcal{H}$ s.t. $h(v_{i_t}) = y_t$ where $i_t = 2^{t-1} + \sum_{j=1}^{t-1} y_j 2^{t-1-j}$

So, in the last class we introduced a binary graph right, sorry a binary tree, then which could be like this and we call this like v_1, v_2, v_3, v_4, v_5, v_6, v_7 and like that ok. So, let say this is my initially finally, the leaf node.

So, these are the points which are coming from my sample space. And we said that there are let say associated label y_1. If this y_1 is 1 that label we said we are going to right; and if it is 0, then we said you are going to left.

So, you fix a binary graph, when I say such binary tree, when I say you are fixing a binary tree, your basically coming up with this sequence of points like this which are numbered in this fashion. And then you are going to depending on your labelling sequence, you are going to take one of this paths right.

For example, if you are labelling sequence is y_1 is 0 and let say y_2 is 1, then you would have gone here; and then it is something let say 0 you would have gone here. So, the sequence 0, 1 and let say 0, we will take you to this path.

And if you are going to change this v_1, v_2, v_3, you may end up with an another graph ok, sorry, another tree. So, let say that is why we are going to call this a binary tree with let say of depth d that we are simply going to be denoting by this sequence of points $\{v_1, v_2, \ldots., v_{2^d-1}\}$.

So, this is going to be a binary tree of let say depth d, this will consist of these many points. And depending on what is your binary labelling sequence, you are going to take this path. Now, what we are saying in this we are basically trying to think about a strategy for the adversary ok. Suppose, let say this is at the beginning a sample point v_1 is shown to you, let say x_1 happens to be v_1 ok.

And let say you want to assign a label y_1 for this point let say you assign v_1 whatever, based on you selected 1 or 0, you will end up at some point. And let say at some x_t we have shown this right, what will be the index we have written what will be the index in the t^th^ round right. What is that value is?

Student: That is $x_{t+1} = v_{2^t + \sum_{\{j=1\}}^{\{t\}} y_j 2^{t-j}}\ y_{t+1}$.

Plus y. You are going to we in this and whatever the associated label you are going to say.

And let say you are going to. Now, the question is say you are going through this points, and let say these are the associated label. Now, suppose I have come up with the hypothesis h belongs to my hypothesis plus H such that my h of x_t is equals to y if I can do this ok.

So, what I am basically doing it whatever path we are going to go through this, I have a hypothesis that is going to map this point to that y_t here. If you could do this for any given labelling sequence, then we are going to say that say that this tree is shattered ok. So, is this point clear?

Student: I think y.

No, you give me, you give me like, you, you take a depth d tree and.

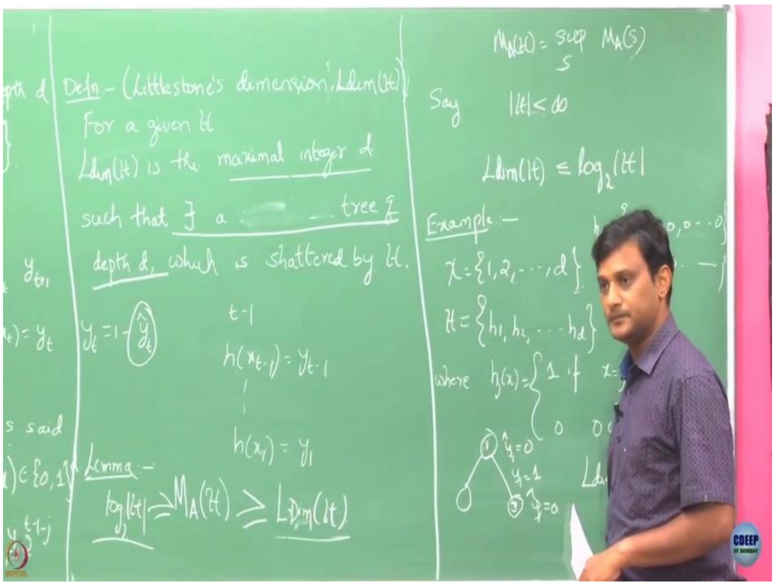
Student: (Refer Time: 07:18).

Now you fix this point. Now, we are going to talk about is this tree. If you are going to change this point it is a different tree. With this points, that is represent by tree will just be shattered. We are going to say that yes this is going to be shattered. If you give me any sequence of this, I will come up with a h such that the path I will be a I have a traverse such that all my x_t's will have this label y_t ok.

Let me just make that more formal where or that we said that ok. So, this is a formal definition of h hat tree. So, notice that what is important here is once I have given this binary tree, I should be able to do this job of assigning mapping these labels to a particular point here on my tree for every label ok.

So, let us think like this. So, let say I have been given this tree. And in addition, I have been given a binary sequence like this d labels. Now, can you come up with a hypothesis which according to a path taken according to this labels will ensure that all the points on this will have the associated labels. If you can do this, we are going to call this, this binary tree shattered ok. Now, let us introduce this definition and then discuss what how this is useful.

(Refer Slide Time: 11:03)


Defn.- (Littlestone's dimension, Ldim(H))
For a given H
Ldim(H) is the maximal integer d
such that ∃ a tree of
depth d, which is shattered by H.
$y_t = 1 - \hat{y}_t$
t-1
$h(x_{t-1}) = y_{t-1}$
$h(x_1) = y_1$
Lemma:-
$\log_2|H| \geq M_A(H) \geq LDim(H)$
$M_A(H) = \sup_S M_A(S)$
Say $|H| < \infty$
$Ldim(H) \leq \log_2|H|$
Example:-
$\mathcal{X} = \{1, 2, \dots, d\}$
$H = \{h_1, h_2, \dots h_d\}$
where $h_j(x) = \begin{cases} 1 & \text{if} \\ 0 & \end{cases}$
$\hat{y}_1 = 0$
$y_1 = 1$
$\hat{y} = 0$

Student: (Refer Time: 11:56).

Capital Ldim(H). So, this is the short hand for little dimension, sorry Littlestone’s dimension. So, is this notion of shattering of a tree is clear? Now we are going to say that you give me a hypothesis class; right now, I am not saying anything about whether this hypothesis class is finite or infinite could be countable infinite or uncountable, I am not telling anything.

Then we are going to say that this notation this is Littlestone’s dimension is the maximum integer d such that there exists a tree of depth d that will be shattered, that means, you keep on taking such trees of different depth ok.

If you can ensure that a tree some tree which has depth d that will be shattered by some hypothesis class in h, but if you go to a tree of depth d + 1 that will not be shuttered ok, then it is clear that, then that integer d is the maximal depth of the shattered tree right. And that value we are going to call as Littlestone’s dimension after the hypothesis class h, so good.

So, what is the usefulness of this then? So, can you think of on this graph what kind of errors that an adversary can force on you, from this can you realize or get some intuition about? It will be at least?

Student: d (Refer Time: 14:41).

d or little dimension d, little dimension d. Why is that?

You first show the learner this point v_1 that is x_1 equals to v_1, learner let say he predict some $\widehat{y_1}$ ok. Then what you can say true label is opposite of $\widehat{y_1}$. And whatever that opposite of $\widehat{y_1}$, you call that y 1, and declared y 1 as you are true label. So, you forced that learner to incur a one error right in the first one.

Now, based on that, whether he made 0 or 1, you are in the opposite direction. So, let say you happen to he happen to say 0 and you happen to declare it as well now you came to this path. Now, you show this to him. Whatever the label he says here now you say opposite of that ok. And because of that, what you are basically generating is the sequence of y. So, you with, so let say user is going to say $\widehat{y}_t$ in round t, you are going to declare the label to be y_t to be 1 - $\widehat{y}_t$.

So, whatever is sequence of predictions you have made, whatever let him whatever let him apply whatever algorithm whatever strategies is applying you will do you are just declaring this to be your true label. But now this labels are not arbitrary, they are still governed by some hypothesis in your class right.

So, the adversary is still sticking to your role. You remember we are trying to enforce realizability assumption here. It is not that adversary we will just look into your prediction and just say opposite of that. Because if we can say that there should be there may not be any hypothesis which will make such labels feasible right.

What to, how to ensure is whatever the adverse is also telling it has to corresponds to some hypothesis for some samples. And this kind of shattering is exactly ensuring that right. Whatever the path I am going to take, I am ensuring that there is some hypothesis which is conforming to that labels that I am going to observe. Is this clear? So because of that I am able to inflict errors on the learner, but now the question is what is the maximum number of errors I can, at least what number of errors?

The learner make can make himself some errors, I do not care about that. What I am caring is what is the minimum that I can enforce on him. By doing this we are ensuring that you will be at least able to enforce this much of error on him. Is this clear?

All we are saying is the adversary have a strategy for him that exist, a strategy which will allow him to enforce this much of error on the learner.

We are saying this is what used by the adversary ok. Remember what is happening in our online algorithm in every time, you are going to show one context, and the learner has to give a prediction for that. Let say learner give a prediction of y_t in round t.

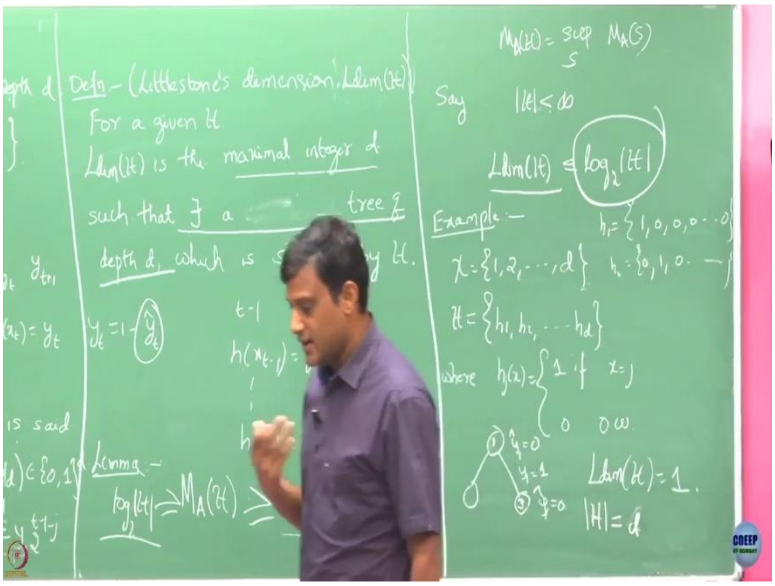
Now, what we are going to do, you are going to actually declare that the true label is compliment of this, and you declare y_t is the true label. By doing this, you have made sure that the learner has made an error ok. But you also have to ensure that this y_t is such that I should have an hypothesis which has been consistent till that point let right, because learn adversary is sticking to an hypothesis and he is generating these labels.

So, let say till t - 1 rounds he has used hypothesis and given y_t - 1 as the label, and till all these points let say right from h_1 to y_1. But in the t^{th} round also you has to ensure that he uses this same ones. But now is it feasible for him to do so? It is feasible for him to do so if he is following such as strategy, if he is using the graph which is getting shattered by this hypothesis right.

Now, we are just saying that what is the maximum depth you can go doing like this and that is exactly the maximal integer we are saying that is what the little dimension ok. Because of this even the learning is perfect, I in the sense like I am myself not making any unforced errors, but adversary is making me at least forced error this much of on this many points ok. You remember we had a notation. What is this notation?

Number of mistakes algorithm makes A while learning on hypothesis class H. Now, can I say anything between this and Ldim(H)? What can I now relate these two things? So, this is the we are saying that the maximum number of errors by our definition.

(Refer Slide Time: 21:37)



So, what is our definition of $M_A(H)$? $M_A(H) = \sup_S M_A(S)$. And what is $M_A(S)$? The number of errors you made in the predictions right using your algorithm A. now this is the kind of maximum number of errors you have you would have made using algorithm A while learning hypothesis class H. And what is this you just said that this is the minimum number of errors that will be forced on you. So, this has to be lower bound on this right. So, is this clear now? Like by using this kind of strategy whatever you are predicting you just make a compliment of that as the true label, the adversary is going to adversary environment is going to make sure that we are going to at least incur this much of mistakes ok.

Student: (Refer Time: 22:42).

Right. So, I have taken it or all possible sequence right. You, this is the worst case scenario the worst case will be at least this much. Before removing the sup, you may end up you may be lucky and maybe making less mistakes on this. But if you are going to take worst case like if you have testing your algorithm or all possible sequences, then either exists one sequences where you will be forced to have this many mistakes ok.

Now, let say the hypothesis classes some finite, strictly finite. Can I say what will be an upper bound on this lower (Refer Time: 23:27) Ldim. So, notice that this Ldim is the maximal integer such that I given tree of this depth get shattered. So, then what can I connect this with size of my hypothesis class?

So, we can always think of suppose whatever points we have, whatever you take whatever points, each of this path should corresponds to different hypothesis right, because it is a, each it is a unique binary sequence ok. So, that should be a different hypothesis class. And how many if I have cardinality of |H| hypothesis class, how many so the total number of leaves here will be equals to the number of hypothesis, but then what is the depth?

If it has a |H| leaves the graph, how may what will be it depth it will be exactly $\log_2 |H|$. So, you have this natural log ok. So, this bound fine. This is one way of looking into this. But we can derive this result from this, and also what we know earlier. We know that if you are going to take this algorithm A to B halving algorithm what is an upper bound on this?

Student: (Refer Time: 25:18).

Log to |H| right. So, we have already this result. So, if you just take this and this, that is what you have derived this. This is just alternate we have saying the same thing. This is specific algorithm I am just saying say this bound is independent of algorithm. Now, if I am irrespective of what algorithm is this bound holds, but for a halving algorithm I know this bound holds. So, because of that, this, this, I can take which is both independent of just, and this since this part is independent of algorithm I can just make this comparison ok.

Now, let us take a simple example. Let say my sample set only d points which have enumerated as $\{1, 2, \ldots, d\}$. And let say my hypothesis class we will be consistent of again only d hypothesis, where $h_j(x) = \{1, if\ x = j\ and\ 0\ else\}$. So, is this example is clear?

Just to realize what is the example I have only d points and I have only d hypothesis. And let us focus on hypothesis 1. This hypothesis 1 is such that, so let say j = 1, it is going to assign label 1 only when x = 1, and everywhere it is 0 ok. So, if you want to represent this h_1 in just in like it is like {1, 0, 0,… 0}; and h 2 will be {0,1,0,…,0} like this. If these are the labels on these points.

Now, can we compute what is the Ldim of this hypothesis class? So how to go about computing the Ldim of this hypothesis class? Well, let us start looking into. So, I know that maximal integer d, I have to look at, but right now I do not know what is the matter.

So, I will start with some small number and keep on looking into till what depth I will it is successfully able to shatter, and after that I will not be able to shatter ok.

So, one simple thing is to do take some initial point x, you show it to the adversary that x is one of this points ok. And after that, so you now let say the learner said a predicted label to be let say he predicted label to be y_1 to be 0, you are going to make him, you are going to contradict him right.

So, you are going to say the label is 1. We are going to say label is 1. The only way you can say label is 1. So, if you have started with let say one then you should have applied h_1 there; if you started with 2, you should have applied h_2 there right.

So, suppose let say you started with showing one, the only way you can say y_1 equals to one had you selected h_1. So, now let say you are here. Now, here you showed let say some other point 2 let say or let say 3. Suppose, the learner says $\widehat{y_1}$ = 0, can you contradict him here? You cannot contradict him here right, because on h_1 you are going to assign label 3; on point 3, 0 only, you cannot contradict him. Because of this, if my learner keeps on just saying let say 0, 0, you can enforce at most one mistake on him, nothing more than that right.

So, now what will be the Ldim or this little Littlestone's dimension of this hypothesis class, it is going to be 1. We have to show that this is the maximum number of errors we can inflict, or this is the maximum depth of the tree that can be shattered, beyond that you cannot shattered it, that is the whole point here right. We showed that he could only shattered up to depth 1.

So, that is what I am saying right that for depth 2 also you can go, but for depth 2 also you should show that for any sequence of this labels he should be able to shatter it. But I am now showing you your sequence your this sequence is let say 0, 0, now you are not able to shatter this.

You could come up with all unit to show come up with there is a one tree where you are not able to shattered 0, 0 sequence. So, all unit to do it see this is for a given graph, this is the definition of shattering. Now, for Ldim to be defined, I want as long as there exists a tree of depth d that I am able to shatter fine. And what I am looking is the maximal depth I can go whichever tree it is, that is fine with me.

So, I am here just showing you one thing which is not able to shattered. Now, I am saying you take anything

Student: There till.

That till then also it is going to not work. You cannot go beyond one step, so that is what this Ldim is going to be 1.

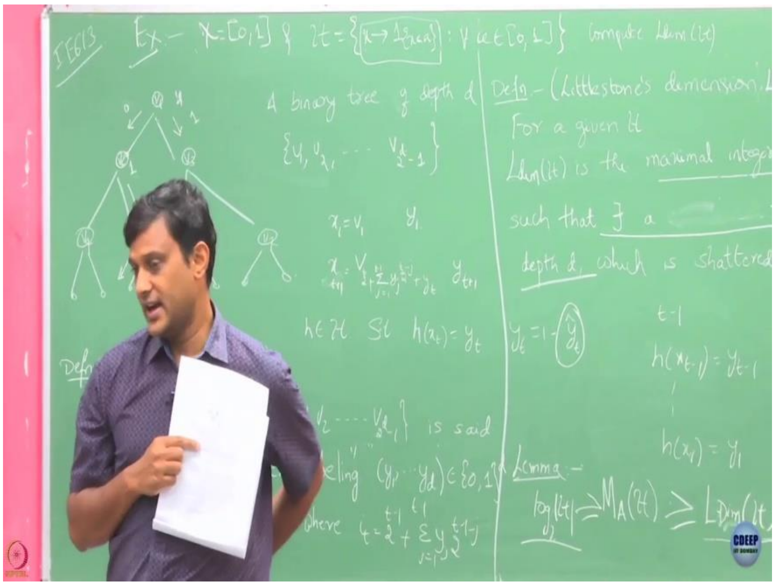
Student: Sir, anything means being any form?

So, this point, if you take any points, in this we will not be able to do beyond one step for this specific example ok. But where what is the |H| in this it is d right because there are exactly d hypothesis in this. So, you see that whatever d here, d could be very large right. Whatever d could be your little dimension is always 1 for this hypothesis class, whereas your cardinality of your size of hypothesis class can be large if we d grows.

So, you because of that you see that for this specific example, this can be very large. If I can keep increasing d, but whereas, this is going to be small ok. So, the only even though you can have a large number of hypothesis, the adversary can only the worst case or that is the best case for you enforce only one error on you ok, even though has large hypothesis.

So, in a way this is again kind of capturing what is the complexity of this hypothesis class right. So, that like analogous to be little dimension in online learning setup Littlestone's dimension is capturing what is the complexity of like in terms of learning it like how many errors you are going to incur before you are going to learn the right one, that is the analogue version here ok.

(Refer Slide Time: 34:22)


IE613
Ex – $\mathcal{X} = [0,1]$ & $\mathcal{H} = \{x \to 1_{\{x<a\}} : \forall a \in [0,1]\}$ Compute Ldim(H)
A binary tree of depth d
Defn – (Littlestone's dimension)
For a given H
Ldim(H) is the maximal integer
such that ∃ a
depth d, which is shattered
$h \in \mathcal{H}$ st $h(x_t) = y_t$
Lemma –
$\log|\mathcal{H}| \geq M_A(\mathcal{H}) \geq \mathrm{Ldim}(\mathcal{H})$

So, I want you to just work out this exercise yourself ok. Just exercise that is there in the book, but you have to carefully analyse it. So, let us take my to be [0, 1] interval ok. And let my hypothesis class to be $H = \{x \to 1_{\{x<a\}}: \forall a \in [0,1]\}$. So, what is my sample space? It is all points in the interval [0, 1] the what is my hypothesis class my hypothesis class is all such, so my each for a given a one hypothesis class is defined ok.

If you fixed a, what it is going to say whatever x were going to give if that happens to be less than you are given a, my label is 1; if my x happens to be greater than or equals that a, you are label is going to be 0 ok. So, depending on what is the threshold you are going to choose, you are going to have many hypothesis right. So, in fact, the cardinality of this h is unbounded ok. Now, compute for this hypothesis class ok. You will end up showing that this is also infinity.