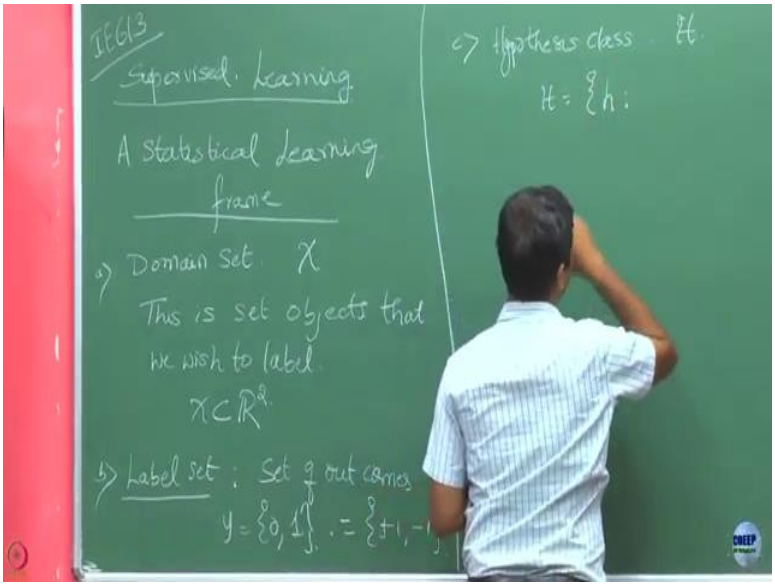
Bandit Algorithm (Online Machine Learning)
Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

Lecture - 03
Basics of Statistical Learning

So today I am going to start with some Basics of Supervised Learning this is actually not online learning, but supervised learning that just as a precursor to our online learning ok. So, we will try to map what we are going to do in supervised learning in the offline setting, how the kind of matrix, the kind of algorithm the kind of strategy you are going to see, whether we will have a natural mapping when we move to the online learning settings.

(Refer Slide Time: 00:55)



So, let us do the revisit our basic things that we would have done in the supervised learning methods. As I said in the beginning this course is about learning setups learning different learning setups, looking for the best algorithms and as a best if I am saying something is best I need to show it like this is the best. So, how one is going to show something is best that you have to say that you can not do better than this ok.

So, our focus of this course will be like coming out with algorithm and then showing they are best through giving a detailed analysis because of that as we grow on this as we go into this course we will get into lot of analysis. So, you have to be like bit focused sometimes the proofs and all get dry. So, if you loose track then you will be totally lost for the rest of

the class. So, interrupt me at any point if you are not following anything so, its not necessary that you have to cover all the proofs, but till whatever steps we are going to cover in the (Refer Time: 02:29) its important that we are all in sync with that.

So what is our supervised learning method? What we have? So, basically supervised thing is you have let us say let us take the simplest case of classification further simplest case that is a binary classification, you have given an instance you have to say whether it has a label 0 or 1. So, let us try to setup this simple thing in a bit more formal way.

So, for this what all the things we know? First thing we are going to start with we are going to assign a label to something right, what is that we are going to assign a label? We are going to assign a label to let us say some instance. To be concrete let us take some example, let us say you are a stock enthusiast and you want to predict, your stock is going to go up or down your share value is going to up or down.

Then if I give you a stock that stock is what you are basically predicting the label right you are basically. So, that we are going to call the point the set from which these stocks are coming we are going to let us say we are going to call it as some domain set. And we are going to denote it by X, and this is set of objects that we wish to label. So, in our stock example this X could be a collection of stocks to which I want to say whether this we want to label, whether their share points are going to go up or down.

Another example could be let us say you have some information about the weather. Weather let us say you know let us say weather information has only two things like how humidity is and how what is the temperature. If I tell you this information you have to predict, whether its going to rain or not your job is to say whether its going to rain or not.

In that case this X case what in this case the X could be collection of all these pairs. The pairs is like humidity value and the temperature value ok. So, in the in this case this $X \subset \mathbb{R}^2$. So, you give me any point in this; that means, somebody told ok, today is this is the humidity level and this is the temperature and you would like to tell whether its going to rain or not.

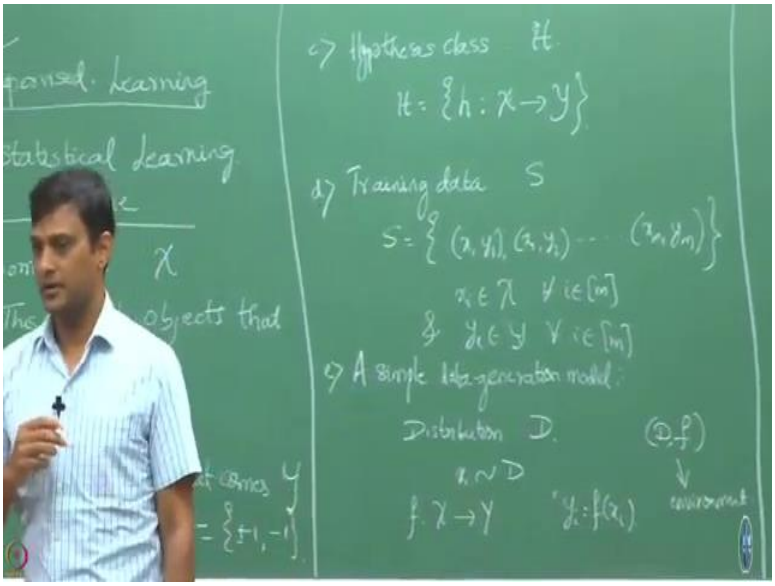
And second thing is label set, but the label set we have already fixing we have kind of already restricting our self to binary labels, but labels could be any number of labels. So,

this could be like let us say set of outcomes. And I am going to denote it by Y and in the binary case $Y = \{0,1\} = \{+1, -1\}$.

And now the one who is going to give it labels based on the input you give from this domain set let us call him a learner. So, learner is basically trying to do what he is coming up with a rule that will tell if this is the input this should be the output. And he may have access to multiple rules right it is there is if you give me weather that let us say humidity and the temperature value, I may combine them in some various ways to come up with a output that is saying ok, whether it is going to rain or not. So, I have multiple possible rules right.

And the learner may have access to finite set of rules and he may want to pick one from that to give you the outcome. So, we are going to that set of rules that that is accessible to the learner, we are going to call it as hypothesis class.

(Refer Slide Time: 08:59)



And we are going to denote it as H and this H is nothing but a collection of rules which are maps from X to Y. So, these are all functions which maps your input to some label in set Y ok. There could be multiple such hypothesis and the all so, that is why I am doing it to denote it as a set this set is collection of these rules.

Now, the learner here he has to find which is the rule is the best or like he have to basically let us say for time being he has to pick one rule from this. So, how he is going to decide?

He is going to possibly decide based on the past observations or the past history. For example, he might have from historical data from weather department he would have seen that with this temperature with this humidity this happened rain happened or not.

So, he may have all the data observed from past and he is going to use that to see which is the which rule or which hypothesis he likes to pick ok. So, he the learner has access to training data, we are going to call it as training data. And that this we are going to call it that is S and $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ $such\ that\ x_i \in X\ and\ y_i \in Y\ \forall i \in [m]$ (Refer Time: 10:58).

So, there are m points we are going to call this sum training data, there are m points in the training data each point here is a pair the first part denotes the object here, that is coming from my set X that is my domain set and the second component is the label that is coming from my label set Y.

So, we have these components in this the learner have access to this training data, he has access to this hypothesis class. And now he has to figure out which is the best hypothesis for him that does a good job of prediction on a given data point. So, when I say data point it will be it can means two things, data point could be it will have this object (x_i) as well as associated label (y_i). Sometimes data point I mean to say only the first part (x_i), it may not have the associated label (y_i) and in that case you have to come up with the label.

So, you will either say simply data point or a sample when I say sample, if is we do not have label for it our job is to predict it. Then the once we have already have it that is part of our training data.

Now, when I said the learners objective is to pick an hypothesis from this, that in some way does a good job right on the prediction task. So, let us try to quantify that like what is that the learner's objective would be. So, we are going to say here learning for the learner is pick an hypothesis from class H which is good in some sense, that we will make it clear in what sense ok.

So, now learning boils down to learning this hypothesis. I want to identify which is a good hypothesis from class H. Now, let us say, now I am learning on a data set, but then the question is how this data set looks like or how it gets generated? Is this data set arbitrary?

Like if I ask you somebody to give me data, can you just generate data and give it to me anything we will do.

We are going to assume that there is an underlying process that is going to generate this data. And actually that process is known to me and my job is to kind of figure out that underlying process that is generating data.

So, for just for a beginning we are going to assume a simple data generation model. We are going to assume that data is generated according to the some distribution D, We said that the data points are coming from the domain set X, and how they are picked? They are picked according to distribution D. Let us say I have said that I have a training data. This training data when I said how are these points generated in X?

Each of these points I have told you they belong to X, but when you generate this point it is going to be sample from this set according to this distribution D. So, we are in this case we are going to say that $x_i \sim D$. So, this is about the first part of S. How this samples are generated.

Also part first part of this data point in S (x_i), they are also called features ok. And so, henceforth we start calling them as features and we are going to call this as domain set as also feature set. So, for example, in the rain prediction problem we said the humidity level and the temperature level right, these could just act as a feature for us.

So, given a feature I want to identify whether it is going to rain or not that is I want to assign a label. And these features are generated according to this distribution D. We are going to say that this distribution D is fixed throughout its not going to change, but I do not know it ok. What I get to see is the sample is generated from this distribution D. This is one part of this data set right, I told you how to generate the features.

And we are going to say that there is a label function $f: X \rightarrow Y$, which will make it $y_i = f(x_i)$. Again I am assuming that this once you tell me a feature the associated label is governed by this function f. And we will assume again for simplicity for just in the beginning that these f are again fixed, but I do not know that . So how these features are generated and how the labels are generated both I do not know, but they are fixed.

Now, and we are going to usually say that this is the environment. The D together with f this constitute for us environment, both of this I do not know, but the data generation is governed by this environment .

(Refer Slide Time: 18:47)

So, now we going to draw measure. So, we said our goal is to pick a good hypothesis from this hypothesis class right. Then, we need to quantify what is that good we are talking about. So, that is the measure of success.

So, the measure of success we are going to use is called if you let us say give me hypothesis h from this hypothesis class H. Then, $L_{d,f}(h) = P\{h(x) \neq f(x)\}_{x\sim D}$. So, what does this tell? The probability that your hypothesis does not give the same label as the underlying label generating function f.

And notice that I am taking this probability with respect to this distribution D. So now this x is generated according to this distribution, but this is kind of any arbitrary0 x, it is it could be any x here right. So, whereas, when I did this training that these the some specific data points which has been sampled from distribution D and that is made available to me.

When you are trying to measure the success of the hypothesis (Refer Time: 21:10) I am going to look at all x's entire x, but I will be looking at by sampling them according to the distribution and look over this.

So, what we are basically do is whatever the training set that is given to me, I am going to use this let us says and I will come up with some hypothesis. And I want to measure the success of that hypothesis, I am going to see how good it performed on a random x that I picked from some distribution D. So, when I pick a x randomly from this distriution that could act as my test point. And points in S are all my training points let us make this bit more formal.

So, if I have defined my measure of success like this for a hypothesis h, what kind of h you would like to use? So, let us say you are a learner and I am like a so, you have access to the hypothesis class and let us say I am a measurer like I am going to evaluate your performance. And my evaluation performance criteria is this. What kind of h you would like to give me? The one that minimizes $L_{d,f}$ this is called prediction error or could also be called test error. As a learner you would like to find an hypothesis which has the smallest test error.

So, my goal could $Min_{h \in H} \; L_{d,f}(h)$. Now, let us we have not made any assumption that whether this hypothesis H class is finite or infinite or countable infinite or uncountable anything, but let us lets be bit wake here and still write minimum of this quantity here.

So, is this entire setup is clear now, what we wanted what we usually try to do in the supervised learning is like under the simplified case like binary classification and everything . So, let it X is the random variable. This X is a random variable here. And that X has distribution D, that is what I mean.

So, let us say nature has its own rule mother nature it kind of it relates, whether the rain happens or not this humidity and temperature in some fashion. It follows its own laws we do not know about those laws, but we have observed this like maybe like for the past 50 years or something we have been observing what is happening, then the humidity level or this temperature . So, let us say somebody has recorded this information.

So, whatever the natures law it has been using, we have seen the samples of this over the last 50 years ok. And that law is (D, f), that is the environment, it is generating this features and the associated labels according to this distribution which we do not know. But what we are interested is suppose if I just get to observe humidity and the temperature level, will I will be able to predict my rain.

The way the nature would have done it or the way in it would have happened in reality, that is the goal here that is what I am trying so, this will like it always happens that you will see the feature vector and at the end, you can think of label as the result. So, in this case whether rain happened or not is the label right that you are going to see the result.

So, let us say you saw both temperature and the humidity growing in some fashion and at the end of the this as result of these two the rain is going to happen or not. So, that is like a outcome right like rain happened or not and that depends on these two value. You have observed let us say humidity and temperature and now you want kind of predict, whether it is going to happen or not. And also like in the stock market case right like, you want to see tomorrow my stock price is going to share price is going to go up or not.

From your so, market has its own dynamics if market loss are governed will be very complicated you do not know it. So, based on that you have past historical data from the stock exchange you have observed that based on theirs like, if this has happened like let us say whatever the feature vector. If this has happened, the next day the share price went up or not.

So, you have that past observation and now let us say today something happened like you have observed the current market scenario. And, now you want to predict whether tomorrow your share price are going to up or not. So, that is what we are trying to do.

So, this is based on the samples generated by the underlying environment, which I do not know and now I want to see that. If any features are generated on this how good the predictions are $h(x)$ is what is basically prediction or the label assigning right, how good I am assigning a label that is as good as the nature would have assigned it.

If I can do this so, suppose I find a h such that $L_{d,f}$ is very small. If this probability is very small that whatever the feature vector I am able to make the correct prediction of its label or its outcome right; that means, in a sense I have figured out what nature is doing; that means, I have learnt these things from my observations.

Now, the next question is fine you have setup the problem you have setup your learning environment, you have setup you measure of success, how you go about achieving, how you are going to get a hypothesis from your observed data? Ok. So, all you have this somebody generated a data and gave it to you, there have been also you had let us say

learner has said ok, I am going to do this much hypothesis. Now, this is your objective how you are going to go about it.

So, basically I want to find a hypothesis that kind of minimizes $L_{d,f}$. Suppose for time being written that you happened to be the god, or you happened to be the nature or you happened to the oracle or god revealed you what is the distribution he is going to use to generate this feature vectors. And, he also told you what is the labels function that he is using to assign labels to the feature.

So, I am saying god told you these two things (D, f), then what would you have done? In that case for any h can you compute $L_{d,f}$? If you know the distribution, if you know your function for any h you can compute this right, then it simply (Refer Time: 30:36) to minimizing this, but god why should god tell me right I do not know this.

What I have been like, but what I have been doing is I am observing what like oracle or god is doing and now I am trying to imitate him right. So, how best you can imitate him? So, you have access to only this from this, you have to come up with something. So, how you how you will do this?