NPTEL

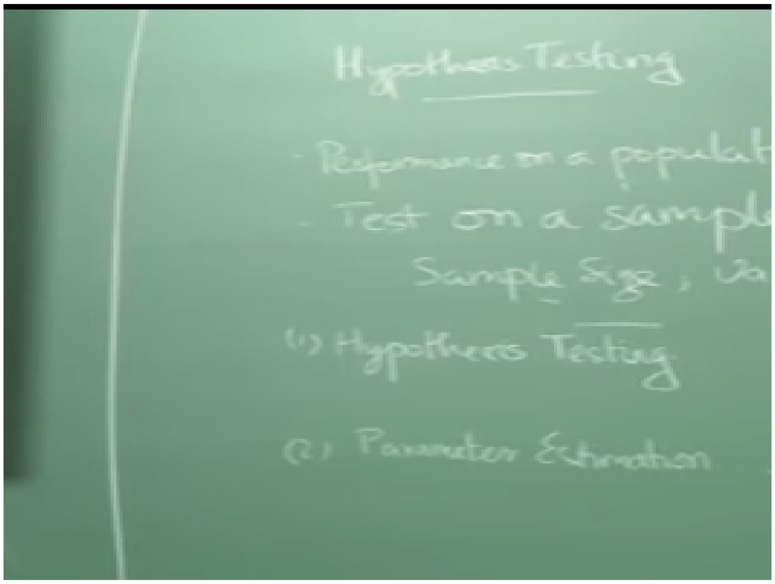
NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

Lecture-54
Hypothesis Testing – I

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

(Refer Slide Time: 00:16)



Right, so last class we were looking at performance measures right and then I started talking about setting up experiments in order to say something about, in order to measure performance of algorithms and why you will want to setup experiments right, and being an empirical subject how experiments are very important right. And so the whole idea behind all of these experiments is that we really want to measure.

So we want to measure performance on a population right, I want to measure performance on a population, but all I get to do is test on a sample right. So whatever is the situation right, whether we do cross validation, or whether you do bootstrap, or whether you just set aside a validation set, whatever it is, it is a sample that you are testing on. And what I am really interested in knowing is, how will my algorithm perform on the entire population as a whole right.

So I give you this P(x,y) right, so I want to know, well I do not give you the P(x,y) that is the whole problem right. So there is this P(x,y) and I want to know how the performance will be with respect to that underlying sampling distribution. And I do not have axis to that P(x,y), therefore I will always be testing on a sample right. But I am really interested in performance on a population right. So what we are doing with hypothesis testing here, this is essentially trying to say that how much can you infer about the performance on the population from the test results on a sample right.

So how confident can you be that whatever you are getting as the test result on a sample is the performance on the population right. So that is essentially what we are trying to here right. So in statistics terminology the test on a sample right, gives you what is called a statistic right. And the performance on a population is in some sense that is a kind of a parameter that what is the average prediction error right, on the entire population.

So that is the parameter that you want to estimate and what you have is, what is the prediction error on a sample okay, that is the statistic okay. So the more common restriction that people can make is average versus mean right. So average is essentially a statistic right, so the mean is a performance thing right, it is actually over the entire distribution right. And you take samples and you take the average of the samples, you use that as the mean of the distribution right.

So we just use it as it is, but that is not correct right. So because when I take a sample average okay, there is some probability that it will be close to the true mean of the distribution right. So the statistic will be the average and the parameter that you are interested in would be the mean okay. So what are the factors that will influence this, how confident you can be about the parameters from the statistics?

Sample size is 1, anything else? How? Yeah, no but I am going to take a lot of samples, somebody else said something else. No, no variance, who said variance? Yeah, so the variance of the underlying distribution right, so how variable is underlying distribution. So for that I probably have to compensate for that and I need to take a larger sample and things like that. So the variability also is assured right.

So this is something under my control, this is something that is not okay. So these are the things you should remember right. So we talked about two things that you wanted to do okay. So in the hypothesis testing, so what we are really interested in doing is actually answering some kind of yes or no questions right, I have an hypothesis okay. So my learning algorithm is better than the other learning algorithms. So algorithm 1 is better than algorithm 2, yes or no okay, right.

And I give you an answer, I say yes okay. I also would like to know what is the probability that the answer was wrong okay. So that is essentially what I am trying to do in hypothesis testing. So I will ask you an yes or no question right. So this question usually is of the following form, people have already done some amount of hypothesis testing and have it done something some null hypothesis alternate hypothesis reject one in favor of the other no yes okay.

Yeah people have done the course in terms would know this but apart from that nothing in satirical signal processing no okay right so the basically is yes or no question will be of this following form right so I will have 1 basic assumption right which is both the algorithms or the same right and then have an alternate assumption which will say that algorithm 1 is better algorithm 2.

So the question I ask is should I aspect te3h basic assumption or should I reject it but not blindly reject it should I reject it in favor of the alternate assumption that I have right are they equal are is 1 better than 2, right I could also post my alternative question in different way I can say or they equal or they not equal okay so that continence with which I can answers these two questions will be different for the same data right so 1 case the question was or they equal or is 1 better than 2.

So in other case the question was or they equal or they not equal so in the in both these cases the confidence with which I can answer this will be different for the same data that I have, right so we will see why that is the case as we go along but the questions will be of this form right so yes do you aspect this or do you reject this okay and if I choose to aspect this what is the probability that I was wrong okay.

So I do not want to aspect the something if the probability is to blow I will just basically say I am sorry I cannot say anything that is statistically sound about these two algorithms given the experiments that we have run okay you will give me some data it will say I cannot say something strategically sound about this given whatever you have told me because the probability of me making errors is fairly large so how large is fairly large.

Yeah but typically I do not wanted to be even large than 5% okay usually I want to be even smaller 1% right why is that the case because as you will see which we go along we will be making a lot of approximate assumptions so that we an get things in tractable form so given that we are making so many assumptions we would at least the probability of error to be very small so that we can be assume of something good okay something reasonable.

Right so is that fine so we ask yes or no question and you look at the probability that is hypothesis testing the second one is parameter estimation right so here it is not enough for me to answer weather program 1 is better than program 2 right I want to know what si the average performance I program one right let us say it I just looking at running times okay so I have some program that is suppose to crutch a lot of numbers and give some output and I want to look at the running time of this program,

Right so I want to find out what is the average mean running time or the expected running time of the program on any sample given from a population, right but then I only have some 20 samples on which I run this program okay I can take the average of the running tine on this 20 samples but I want to know what will be the running time on any sample I give you from the population right.

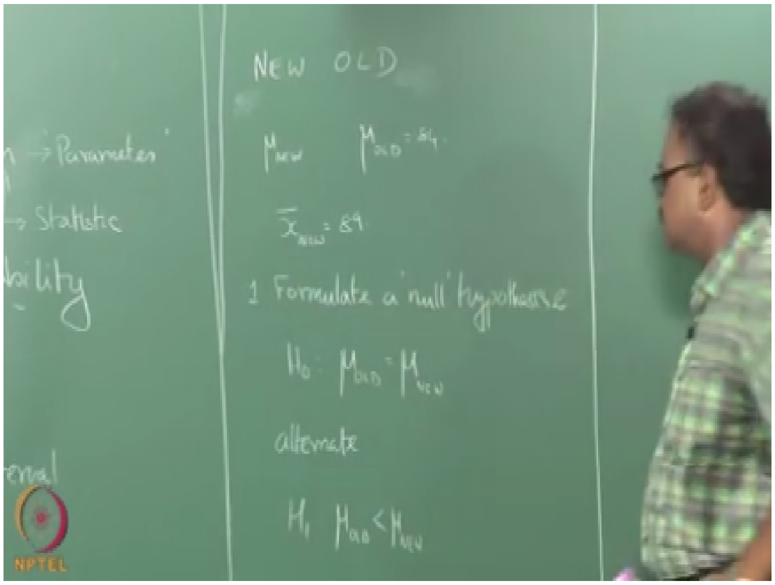
So how like right how far away is this estimate on 20 samples from the true mean ruining time of this program so this is what we mean by parameter estimation right this is why I code sub here okay there slightly different usage here really not at very fundamental level they are not but at least they are very different from the way we have been using it so far right so we have been talking about when you say parameters we have been taking about like wait say in network or the alphas in support vector machine and so forth.

But here when you talk about parameter I actually mean the performance parameters that we are interested in right so the second thing which we want is essentially parameter estimation where I am looking at some kind of a interval right around my statistic right and I should tell you that okay with some amount of confidence right the true ;parameter lies in this interval around the statistics so it is like saying that okay so I run this all my tests on this sample data and I get the performance as say 3 .3 seconds okay.

Then I will say it is 3.3 + - 0.5 seconds okay so then the true mean will lie somewhere in that interval okay with the high probability right you can see that the 2 question are related so the first question again says how can I reject can I say 1 is better than 2 right second 1 I am saying no I want to know what exactly is a performance of 1 and in both cases I am looking at some kind of a confidence core of comparing these two does it make sense great I can reaped confidence core.

But I will tell you more about later okay that is the rest of the lecture is going to be telling you about how to get this confidence mission right. So I will repeat something which I gave as an example in the last class right, so let us say I have two algorithms right, I am going to call them new and old okay, so I have two algorithms.

(Refer Slide Time: 12:22)



Okay, this two algorithms new and old okay, so the old one is running for a while okay, the old one I have used the old for a while and I know I am running for a long time right, and I have some measure of how good the old ones performance is going to be right, so I know them mean performance of the old algorithm because I have been running it for a long time right, and I also know the kind of this standard deviation of the performance because I have to chained a lot and lot of sample let us assume that right.

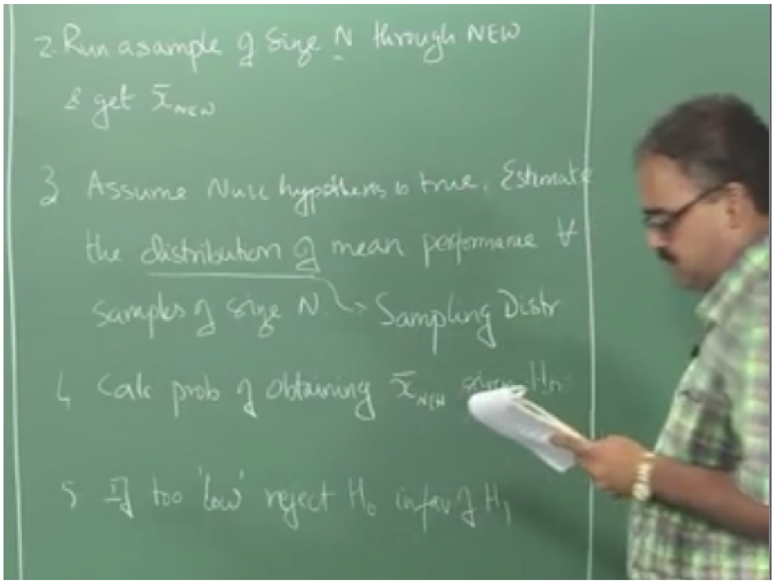
And the example I think I gave in the class so it is on intrusion detection right, I said there is some algorithm that has been running for a while right, and then it gives you some performance it catches and say 84% of all the intrusions right, and then I propose a new algorithm it is runs for like 10 days and it catches 84% of all the intrusions, so it is new one is better than the old one right, is it clear. So the question so the old one has a okay, I have numbers here.

So I do not have fully worked out things the old one has some 84% performance the new one has 89, so it is a new one better than the old right, does not matter this number do not really matter okay, do not get hang up on these just for illustration purposes. So I am going pose it a little more formally now right, so by so this is as we had 84 and 89 I am going to call these as μ_{new} and μ_{old} right, so I am sorry, I should be very careful okay.

And I do not know μ_{new} okay, I only μ_{old} I have some more estimated it to be 84 because I have a lot of experience with the old thing and μ_{new} I do not know, what I do know is a statistic right, so I already μ_{new} is do you know some $\bar{x}_{new}$ which is 89, I have a statistic where it run it on some 10 samples and I know that the performance is 89 right. So now what we do is I formulate a hypothesis, right.

So what is the base hypothesis I am going to formulate, seriously I mean all of you have done some probability in statistic course right, okay did you guys did not do all of this, not in the PRP is it, okay I guess it is not a statistic course okay, it is probability and random process there is no statistic in it, okay fine. Because I did it in my very first maths course in under grade and I am not CS student so, right so you formulate a null hypothesis I am going to say $\mu_{old}=\mu_{new}$ okay. Then I am going to formulate an alternate hypothesis okay.

(Refer Slide Time: 16:35)



2. Run a sample of size N through NEW & get $\bar{x}_{NEW}$

3. Assume Null hypothesis to true. Estimate the distribution of mean performance ∀ samples of size N. ↳ Sampling Distr

4. Calc prob of obtaining $\bar{x}_{NEW}$ given H_0

5. If too 'low' reject H_0 in fav of H_1

So I get a statistics okay, I get one measurement of new right, of this right so x bar new so here is the question is sample of size N so that is the important thing that we have note here. So next thing I want to really figure out is suppose my null hypotheses is true, okay what is the probability that I would have got a performance of x bar new on a sample of size N sorry, this new is hold is less than μ new I am sorry.

M new is exactly equal to μ new that essentially there is no difference in the tow algorithms, greater than μ new when you are do it okay proposing a new algorithm right at least you are assuming is not better that is a very settle point here right so the question I really want to ask is μ new better than new old right. So if new was lesser than μ old right what is the question I am asking can I accept the null hypothesis right or can I reject it in favor of the alternate hypothesis.

So that is the question I ask right so μ new is actually less than μ old as we will see when we go along then we will say that no I cannot reject it right I favor of the alternate hypothesis basically then we have to go back your test basically falls up here your basic assumption was wrong then you have to go back and redo in the test right. So a safer question to ask is new old not equal to μ new but that not of interest to you right you really want to establish whether new is better than old or new is worse than old you do not want to know new is different from old.

That is not interesting question for you right so you remember yesterday or the last class I was telling you about it need to be very clear of what is the question you are asking in the experiment right. So running in experiment you need to be very clear out what is that we are looking for in the experiment right, so for example so I am going to point you next to a really fantastic book on empirical methods in AI right.

Explaining all of these things to you which will usually the statistic book is in a very dry statistical sense right and it very mathematical sense they actually trick real experiments at they run on different kind of machine learning and AI settings right and then talk about introduce this topics very slowly to you right in fact I think I have already did one chapter from this book last class and today we are going to do another chapter from the book.

So I will just want it read it I am not going to the full book do not worry, but I did a course during my PHD I did a course entire course based on that book, so if this is person how quite told you does not when believes very strongly in feeling effects and the course itself was on empirical methods right and so he ask this dialog nice dialog that he ask in the book right so there are two people talking and then one guy say hey what are you trying to do? And he say then the researcher to replies I am trying to run this experiment I want to figure out if algorithm one faster than algorithm two okay.

Then he says how will do how die algorithm one is better than algorithm two, then he say how wily u do this? Then he says this not describing I am going to set up this experiment so that on this data set I will run this algorithm ten times and this data set I am run this algorithm and then I will make this measurements. So then he ask why are refined or why are you doing this experiment?

Again that gut is replied oh! I am trying to do this to figure out if algorithm one is faster than or algorithm one better than algorithm two okay. But then there is a other conversation between two people and they asking hey what are you doing this? Oh! I have this new method I heard this new method for estimating some significant of some biological markers I am trying to figure out whether this is better than that.

Then he says how are you going to do this? Then he goes about describing the experiment setup then he says, why are you doing this? And then he says, oh! I heard that this particular method uses technique X for doing this and therefore that is supposed to be better, so I am trying to figure out whether that assumption there on which this algorithm is based on right is that valid or not?

So there is very certain difference between the two conversation right the first one essentially that guy wants to know it is faster or not right, does not really have any deeper scientific that he is asking right. So in the second case this person actually has some other valid scientific question that they are asking and reducing experimentation as a way of answering this scientific question right and that is the really reason you should do this experiments not just for making measurements for measurement sake okay.

So it Is not directly related to your question but I have just using that as a excuse to talk about the story so you should be very careful about why you are setting up this experiments and what you are alternate hypothesis depending on how we studied of then you have to interpret the results you are getting okay.

Let us move on to point 3 okay assuming that you are null hypothesis is true right how lightly is it that you would have seen this performance this statistics x bar new right so assume null hypothesis is true and then you try to figure out how well the mean performance be distributed so if I run the algorithm so when you are old that assuming as null hypothesis is true so new or old should give me the same performance if I run this on sample of size N okay so I take sample one size of N run it I take a sample two of size n and I run the algorithm and I take sample three of size N and I run algorithm and so on so forth.

And for each of this I am going to get some average performance right so how well those averages be distributed right for every sample I draw I am going to get a different performance and how well that performance be distributed right so that is the question that I want to ask so I assume so I have to set up this distribution and that is called the sampling distribution okay so what is the sampling distribution again.

I heard some voice from somewhere I cannot locate who what is the sampling distribution it is the distribution of the mean performance on samples of data okay of the whatever algorithm we are talking about so it is the distribution of the mean of the performance on samples of data of a particular size if I change N the sampling distribution could also change right I am looking at a distribution right note that the means by themselves do not mean much okay.

So then the use the sampling distribution to calculate the probability of obtaining X new so once I have a distribution I can figure out what is the probability of seeing bar new under this distribution right okay so couple of things that we have to decide on here so the first thing is the most tricky part of all the hypothesis testing is how to come up with the sampling distribution right how do you come up with the sampling distribution that is the tricky part.

IIT Madras Production

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

www.nptel.ac.in