NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

Lecture 53

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

Introduction to Hypothesis Testing

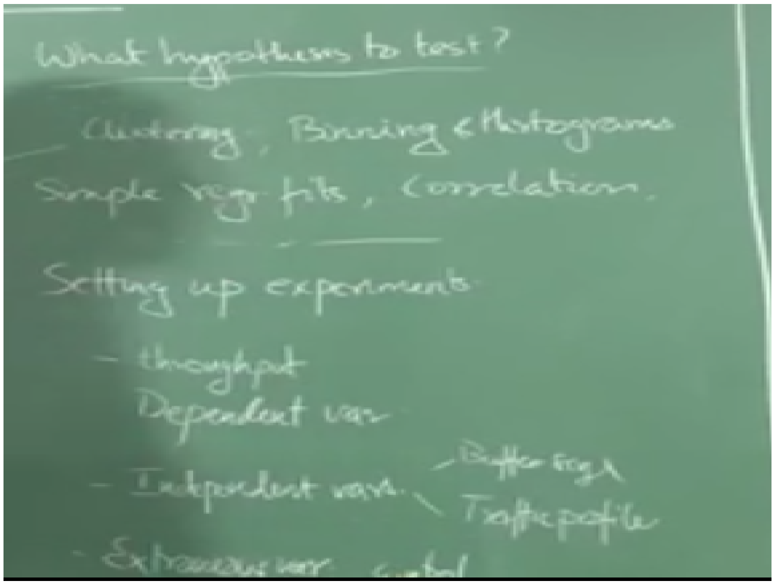
Okay let me stick with whatever I returned down there, so let us say that I am trying to build an intrusion detection system right, I am trying to protect my computer network, so I am going to be looking at all the packets that are coming into my network I am going to say okay all these packets are fine these packets are not through the more block them right. I want to build a system like this right and then so I deploy your system in the IIT network for the first15 days of a month okay, I know it will crash everything will be everything will be malicious or whatever.

But let us say this let us in an ideal world and then it catches like 84% of the malicious traffic okay and then from 16th to 38 I deploy his system it catches 87% of the malicious traffic is a system better than yours? Yeah, when can you say that? can you say something more than it depends, that is what we are going to do now looking that so that we are just going to look at a formal way of trying to say how sure you can be his system is better than your system. That is essentially what hypothesis testing let us you do okay.

It looks at the underlying data distribution that you are operating with and it should be able to tell you that okay with some confidence his system is better than your system okay. Typically what we do in hypothesis testing, we set the confidence level prior okay, so unless with 95% confidence you can tell me that his system is better than his system I am not willing to buy it and I am just going to consider they are all the same system. I need at least 95% confidence for his system to be better only then I will accept it otherwise I am not going to accept.
Because there is so much variability in the in the whole process that 95% is something which I can be comfortable with people usually ask for 99 right people, usually ask for 99% confidence because of the, because inherent uncertainty in the whole thing. So that is essentially what we are

going to look at, so how can we set up experiments okay so that we can answer such questions but before that we really need to know what experiments we need to set up right.

(Refer Slide Time: 02:30)



What hypothesis to test?

Clustering, Binning & Histograms

Simple reg. fits, Correlation.

Setting up experiments

- throughput

Dependent var

- Independent vars. — Buffer size
  Traffic profile

- Extraneous var. control

So like I already gave you two examples right I said that your system is better than his system that is first question right it is your system better than his system then the second question is your system better than his system under high load that is intuition direction right so traffic lot of traffic is coming right, so instead of so maybe your system is not very different from this system and the traffic is 10 Mbps right.

The traffic is 1 gbps maybe you start becoming better than you may be yours is a lighter system therefore you are able to respond faster and then his system starts dropping packets because of the heavy load so that could be a question right that that could very well be the thing so but then you have to think about it you have to figure out hey what is happening and then you can basically what you do in such cases is you observe the observe the system like you have to make some kind of exploratory behavior.

And then you can say okay the mean number of packets I let through when it is at 10mbps is the same for both case but then mean but whatever some rough estimate I have seems to be slightly different when it is 1 gbps maybe I should run the more careful test will figure out which is which one is different whether it is with the high confidence whether it is different or not right so that is a things like and there are other things which I could do right I can say that your algorithm is better when you run it with this parameter setting as opposed to that parameter settings.

When I say the parameters θ 1 versus when I say the parameters θ2 at your algorithm is better when it is θ 1 versus when it is θ 2 so there is another question to ask or your algorithm is better than his algorithm when you use θ 1 so when do you get to these questions so that is where our exploratory analysis comes in right so you have to do some amount of X exploration with the data you have to talk to an expert right who understand this you have to ask you hey by the way will θ being θ one versus θ 2 will it actually make a change to the performance.

And then that gave myself okay yeah maybe so maybe you should not throw out all the packets which are having parameter θ 2 maybe you should include them right so maybe that does mean that could be something really we can do all kinds of things so some of the simple things you do are well could do clustering right.

So what would clustering help you to find so helps you find how the data is clumped up right when you do clustering you can figure this out right you can figure out whether the data is coming from a single distribution or whether coming from a mixture distribution because you will find different clumps of data corresponding to the same class right so now this is all of you to tailor your classification choices accordingly okay so this is one thing that you could hope to get right in some cases in fact people use clustering to even generate the labels.

So I will give you a lot of data right I do not know I have not labeled the data into anything right but I can do clustering and figure out which are the major clusters and then okay there are three kinds of people in my customer base now I can build a classifier that will predict which of these when a new customer comes in I can make I can have it predict which one which category he belongs to so those kinds of things right.

I want to get some rough idea of the frequency of occurrences of features in my data right I can do some kind of simple bending on the features and I can build histograms that allow me to understand how often something's occur so if the data is concentrated suppose I do this thing and then I find said only some bins in the histogram I have very large numbers that essentially mean seven though my the feature can span a very large range it is only some very small values are actually present in the data.

So these kinds of observations I can make right so this will essentially help me do those kinds of things right and then I can do simple regression fits I can do simple regression fits and figure out if there is any turn to the data already that lets me to figure out whether I should be using a linear classifier or whether I should be doing something else where the data is more complex right and we already talked about correlation analysis you should do correlation analysis for what for throwing away features right we already talked about if the two features are highly correlated you should throw them away because otherwise it will lead to numerical instability in many of your algorithms right so apart from that you can actually use this correlation analysis to figure out what are the kinds of questions to ask as well I think about it right.

So once I know what is the hypothesis to test right once I know what is the hypothesis to test then I have to set up a proper experiment right so I have to set up a proper experiment so here I have to be very careful about right what is the question I am asking and which of the variables in the system okay are important for the question that I am asking which of the variables in the system are important for the question I am asking.

So for example that is stick with our intrusion detection system so I want a good intrusion detection system to have a high throughput right so acid when the things come in I should be should be able to put it out right it should have a high throughput let us say I want to test the throughput alone I am NOT interested in the accuracy or anything I just want to make sure that the traffic is not being delayed by the inserting the system.

So I can take this throughput I can make throughput as the variable of interest right or if you are looking at classification accuracy I can take classification accuracy as my variable of interest okay this is essentially known as the dependent variable there could be more than one dependent variable that you are interested in okay then I would have many independent variables right it could be the parameter θ1 θ2 θ3 it could be something else right we are talking about throughput it could be something like a buffer size right.

Or if I am talking about classification accuracy it could be a variety of different parameters right so these are all independent variables of interest so independent variables could be something like buffer size traffic profile and so on so forth right and then there might be other variables okay call extraneous variables right, so for example time of day right so time of day can actually affect the network traffic significantly.

But there is nothing I can do about it right and more people are awake in the morning and they will be doing something in the morning and well more people are awake in the night and they will be doing something in the light and more people are less people are awake in the morning or they are in classes okay so that will affect the traffic right maybe that is not something I can control right so I am not going to I am not going to worry about it.

But whatever I do is whenever I do comparison between algorithm A algorithm B I only do it during daytime or nighttime right so whatever these extraneous variables are I will control for them in the sense that I will make sure that they are the same that even though I cannot independently set it to whatever value I want I will make sure that they are the same so that they do not affect the outcome of my experiment okay.

So there are extraneous variables for which we should control for does that make sense right so these are the things that we should look at right so there are dependent variable independent variables and then extraneous variable which we should make sure you are controlling against okay great there could be other variables in the system right like temperature pressure humidity and all that which does not really affect your network thing this is maybe this if it is very hard people do not less likely to sleep in the night right.

Maybe you should control for that as well do this only on hot days or cold days okay matter has the second does not exist but yeah so that is essentially have to set up the proper experiment making sure you know what are the variables you are paying attention to so I mean all of this is very basic fundamental stuff which should all of you should learn in a proper design of experiments course right.

Once you are set up this experiment right you have to make sure you are avoiding any kind of spurious effects what do mean by few spurious effects the people know the floor effect and the ceiling effect floor effect as in they are close enough so suppose I am setting up an experiment to measure whether his algorithm is better than this algorithm right so and then let us say throughput again let us say take throughput.

And the traffic is flowing in at 10mbps right and your algorithm let us the traffic through a 10 Mbps how can you hope to be to match you can but you cannot beat so I do not know if it is better or not right so both of you can at best achieve 10mbps this is called the ceiling effect so you might be capable of achieving 30mbps but I do not know that because 10 mbps is all that is there in the system so this is called the ceiling effect.

So likewise the floor effect is at the other end of it right so one of the main so I learned all of this in actually a empirical methods course when I was doing my PhD long time back right but the person who taught it was a very strong believer in avoiding ceiling effects so he used to set question papers which could never be completed in the time allotted for them, so there are no ceiling effects so there is always if you are good and you finish the question paper early mean if you finish 10 question early you are always another question for you to attempt right.

So people typically end up I mean you are the best person ends up finishing about sound of a percent of his paper so you can see is an incurable optimist right I mean but he just wanted to make sure that there is no ceiling effect yeah so and likewise there are order effects you know the order in which you actually test things could matter so one example is not exactly an experimentation but very interesting effect that I thought I will mention right.

So when you when you are bargaining we are trying to bargain with somebody so the first thing that you put on the table right actually determines a path in which it is going to go right suppose you want to somebody is trying to sell you something the first thing you should go if you go and tell him okay I will pay a 10 rupees for it now he is going to feel a little bit bad about asking you for 50,000 rupees hey know is that this actually happens in you bargained in Bombay okay.

If you let that guy first give you the money he will say 50,000 rupees now you are going to feel bad about asking him for 10 rupees right so first I went with a friend of mine so he took us to some shop and there was his thing he said like I said what is this thing he said he has 3000 rupees I know I will give you 15 rupees for it okay no they actually bargain with this you know so I ended up buying it for something like fifty five rupees.

So all order effects matter right so this is not really that but depending on which order you make measurements in all right yeah I mean there are other examples I can give but I thought this will be more funny anyway so those are things which you should avoid and there is a third thing which you should very be careful to avoid a sampling bias right suppose I want to know whether algorithm a is better or algorithm B is better in playing a particular game okay.

Then I look at the average moves that were taken across games that were one right and then I find that there is no statistical difference between algorithm A and algorithm B both algorithm a algorithm B win in similar number of moves right but I did a very big cardinal sin I made a sampling by said I only pits games which both of them one right, so a1 b1 so essentially this probably are simpler games right.

Both of them won and I am comparing them and so they all won in similar number of both I should actually be looking at all the games at they played right how many they won how many they lost so all of those things I should be comparing so I have to be making sure that the sample on which I am running these experiments or not biased in any particular way it is very important in fact quite often when people do all this phone in surveys and things like that right that is always this criticism of what the doing phone in service.

Like when somebody calls you and say he do ok are you going to vote for Modi or Rahul Gandhi in the next election so you give some answer right so why is this a bad survey to run you asked for me those people who have phone so I am most likely not going to vote so that is a different issue but you are asking people who have phone threat you are essentially skewing your sampling so you can say anything you want it I managed to control for income level so I only ask people who make so much money.

And so on so forth but then that still means you are leaving out a whole set of people with the same income level who do not have phones I mean right so when you could have very low income level and still have phones nowadays right so that does no correlation to having fun maybe it has correlation to how much you waste on the phone but the mere possession of a phone no longer has a correlation with many of the demographic factors.

But still there is something very selective about curly calling people have phones later than India any is any meaningful survey should be done door to door or straight to street and so on so forth there are complaints but you see many of the surveys that people put in all your magazines and things like that are mostly phone in surveys even women in India so in the U.S. it does not make a sense if it does not make a difference.

Because every household needs to have a phone right there the number of people who do not live in households is small in India that is not the case right so these are these are sampling by so this things enter very often we do not even think about it we do not even think a second time about all the sampling bias that we introduced okay, so I will stop here.

IIT Madras Production

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

www.nptel.ac.in