

NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

Lecture-57

Hypothesis Testing IV – The Two Sample and Paired Sample t – tests

Prof. Balaraman Ravindran

Computer Science and Engineering

Indian Institute of Technology Madras

(Refer Slide Time: 00:17)

Two Sample t-test

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$
$$\bar{x}_1 - \bar{x}_2$$
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$
$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}^2 = \hat{\sigma}_{\bar{x}_1}^2 + \hat{\sigma}_{\bar{x}_2}^2$$

Okay so in the two-sample t-test the question that we are going to ask is I am going to take two different samples right I am going to take two different samples and I want to know if both of these samples came from the same distribution or not and they have some underlying distribution from which I have done two samples I want to know if both of these samples came from the same distribution or not right.

So it could be the distribution of errors right so I could actually draw two sample I can say I am going to run algorithm one okay I am going to get this many errors like ten I am going to run algorithm one not at ten different times right so does it remind you of something would think of

something like 10-fold cross-validation I run algorithm one on using 10-fold cross-validation I get 10 different numbers right.

I run algorithm two using 10-fold cross-validation I get 10 different numbers right now what does what do I mean by the question do they come from the same distribution that means that if I run this algorithm one again and again and again and again and again I am going to see some distribution over the errors right if I run algorithm two again and again and again I am going to see some distribution over the errors right or these two distributions the same right.

So the question I am asking is I have algorithm one I have algorithm two or the errors similarly distributed that means there is no statistical difference between algorithm one algorithm two okay so that is what we are asking is a kind of questions that we would like to ask right, so two sample tests t-test allows us to do that compare means of two samples to see if they are drawn from the same population or different and again remember when you are talking about same population or different we are only asking the question or their means same or different assumption we are making is the standard deviation at all same right.

So another hypothesis is yes they are drawn from the same distribution so $\mu_1 = \mu_2$ right and the alternate hypothesis is δ for a change okay let us do a two-tailed test so this is called two-tailed because I am going to look at both ends of the distribution right so the greater than or less than were called one-tailed or single tail because we are looking only at one end of the distribution. So what I am really want now is to look at the look at that right.

I want to look at $\bar{x}_1 - \bar{x}_2$ and what it should be zero if the null hypothesis is true right so I am going to have I will compare it with a zero-mean Gaussian or a zero mean yeah so zero mean T distribution right so with some number of degrees of freedom but I need to really compute the T statistics right so the t statistic look something like this in this case right, so I am going to look at so this is zero mean right.

So $x_1 - x_2 - 0$ right divided by the variance so how will I compute the variance right, so the variance of the difference is actually the sum of the individual variances intuitively that makes sense right so here is we will do something these are these are details okay there are nothing to get hung up about right so I basically had to estimate this variance but how will I do this variance

I can do one of two things I can take the samples have drawn right under algorithm one I can estimate $\sigma^2 \bar{X}_1$.

I can take a look at the samples I drew hundred algorithm two and I can estimate $\sigma^2 \bar{X}_2$ so I can do that independently and I can get this variance and then what I do, I can plug this in here and I can get away with it right but the problem is not really problem that is a small advantage that I can take care what is advantage or what is it what is they can do I am assuming that the variances are equal right so what did we do earlier when we had a situation where we had this thing and you assume the variances are equal people remember that we did something called a pooled estimate right, so the pooled estimate what you do is you essentially look at the variance across the entire population right and we compute the variance so you can actually do a pooled estimate right. It is my σ^2 , right.

So how many degrees of freedom is going to have so this essentially I will plug this in here wait I will plug this in here and they will compute my T statistics right once I computer the T statistics like I said these are all details if you understood everything so here so far to hear everything is fine here we are just computing the variance this just looks little complex where is nothing but this computing the sample variance by using a pool of estimates right.

So now how many degrees of freedom I am going to have here we talked about the last time also $n_1 + n_2 - 2$ right that is $N_1 - 1$, $N_2 - 1$ so it is $N_1 + 1 - 2$ so the number of degrees of freedom is n_1 plus so you take this T statistics look up that table and figure out for whatever p level you want right so that is basically it so this is called the two-sample t-test and it is very useful when you want to compare performance of two different algorithms on some sample that has been drawn right you remember the example I told you right.

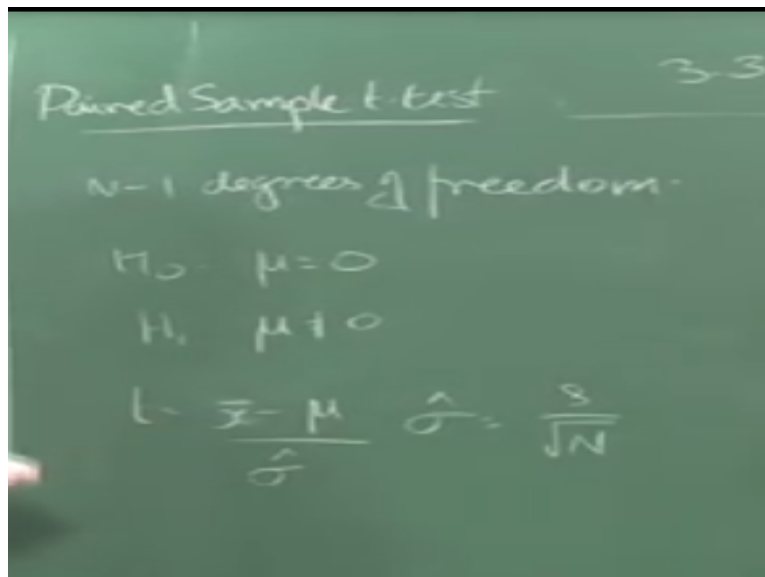
In fact the nice thing about the two-sample t-test is I do not really need to do 10-fold cross-validation on both algorithms let us say one algorithm is significantly more expensive to run than the other so I can do a five-fold cross-validation on one on a 10-fold cross-validation on the other because I am not expecting the n_1 and n_2 to be the same here right but the variance is going to be higher right.

If you think about it so the variance will be higher if the samples are very different right because the n_1 samples I run on the algorithm one right on the n_2 samples on which I run algorithm two

if they are different sets of samples then if I look at the pooled estimate of the variance the variance will be higher right because there will be some underlying variance because of the change in the samples itself and if I run the same algorithm again and again on the same on different samples I am going to get variants I am running different algorithms on different samples.

So the variance will be larger so what will that mean so naturally my t statistics will become smaller right so the larger the T the more by p value can be right so the T statistics might become smaller if the variance is larger so in some way I can get rid of at least some of this variance right so we do something called the.

(Refer Slide Time: 10:29)



Paired Sample t-test 33

$N-1$ degrees of freedom.

$H_0: \mu = 0$

$H_1: \mu \neq 0$

$t = \frac{\bar{x} - \mu}{\hat{\sigma}} \quad \hat{\sigma} = \frac{s}{\sqrt{N}}$

So Paired the sample t-test so what does this mean so I am going to run algorithm one and algorithm two on the same sample right so if I am going to take ten different samples I will run both algorithm one and algorithm two on the same set of ten samples right instead of running them on different samples right, ideally if you have the control over how the sampling is done and how the experiments are done then you should run Paired sample tests okay.

The two sample tests is appropriate only when somebody gives you the performance on different samples a priori right they do not allow you to sample and run the algorithm somebody says okay I have might I have the have access to some 15 samples I have run my algorithm on it here are the 15 performances and you can do whatever you want on the on samples that you draw will not tell you what the samples I drew is okay.

Then you can run your algorithm on 10 different samples that you draw from the same data and then you can compare the two then you do two sample T – test but if you have complete control over what you are doing then you do paired t-tests right paired sample T test and so essentially what does this mean it means the following right suppose I am doing 10-fold cross-validation so what do I create these 10 folds right.

People know what the folds are right so I am dividing them I will do some stratified sampling or whatever it is I create these ten folds I keep them I write them on to disk so whenever I run a algorithm and there is going to read the folds from the disc and done it I am not going to regenerate the foals every time I run the algorithm so that would mean that for every fold I will have results from both algorithms, so in fact this is catching on so much in the machine learning community now that for many of the newer data sets that are being published okay people are actually publishing the folds on which they run the experiments, so that you can also run them on the same Folds.

So you do not you do not generate your new folds and start running because then the comparison becomes little iffy right you can use the same false at a Rand experiments on therefore you do not have to repeat their numbers I can directly compare it with their numbers and I can report right so that is that is why people are actually publishing the folds also right, so when you do pair sample t-tests what you are doing is here what are you doing here you are taking the mean of X_1 and the mean of X_2 and comparing it against zero right.

In this case what you can do I can take the difference because I am running it on the same sample right so I can actually the difference of the performance now makes sense on one sample here instead of averaging the samples and taking the difference I can first take the difference okay and then compare it to a zero-mean distribution right, so instead of instead of having a lot of excess and then getting \bar{x}_1 I am going to have lot of excess lot of X_2 and then I will get have a lot of $x_1 - x_2$.

And then I will take a $x_1 - x_2$ the whole bar okay and then compare it to a zero-mean distribution so that is what I do in pat sample tests and so this is going to have $n - 1$ it is going to have $n-1$ degrees of freedom right my H_0 is right so what is this μ so this is the difference of the means when I saw the difference of the performance right so that is 0 right the mean of the difference of the performance is zero across many samples that means they are the same as my null hypothesis right.

And or I can do $\mu > 0$ which case which case depends on which one I am subtracting which $x_1 - x_2$ if we say μ greater than 0 that means X_1 is better than x_2 but there should be some something from the data that supports your alternate hypothesis, so this basically the standard stuff right so you do this and well if $\mu = 0$ then this is 0 this is just $\bar{x} / \sigma \wedge$, $\sigma \wedge$ is the samples standard deviation by \sqrt{n} and we haven m minus 0.

So this is actually a lot lower variance because you do not have any problem generated variance you only have the variance due to the performance of the algorithm right the samples themselves are exactly the same so that gives you a much higher t estimate than you would get if you run the two-sample t-test okay, so we explained all about all of these things too but almost all packages that you can use right have all of this built in so you can do t-test z-test whatever it is you want you can run.

You do not really have to worry about the internals of it right you just need to specify the sorry the p level you just need to specify the p level you what is it a apart from center before a given test you have to specify what is at p level right what is the P level that you are looking for so if you say I want a P level of 0.001 atleast right then some of these could actually reject it saying that no I cannot reject the null hypothesis at a level of 0.001 or something it could come back and tell you, okay.

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

www.nptel.ac.in

Copyrights Reserved