NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

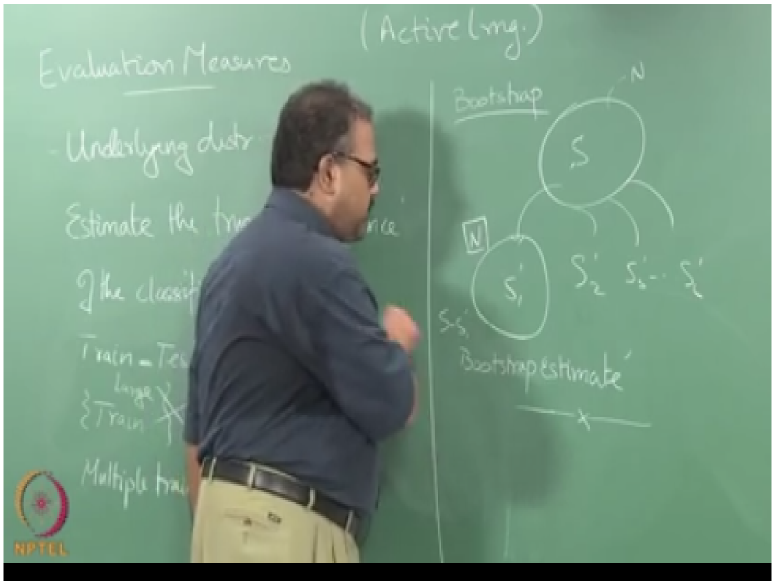
Lecture 49

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

Evaluation and Evaluation Measures II: Bootstrapping and Cross Validation

Right anyway so I will just talk about11 news of bootstrap here so the idea behind bootstrap is very simple right so I have a large sample right I have a large sample of data so what I am going to do is I am going to sample from this data with replacement.

(Refer Slide Time: 00:43)



Let us assume that the set S has N elements in it so what I will do is I will sample from it another set S prime that also has N elements then sound like a big thing right. I mean if I wrote sample with the if a sample without replacement I will essentially be duplicating it but I am sampling with replacement so what is the idea behind doing this no so if you remember I said that the assumption that we are making is that exactly the assumption that we are making is that the data is truly representative of the underlying distribution right.

In which case given the data right the best approximation I can construct to the underlying distribution is the discrete distribution right defined on this data in one sense if I do not make any other assumptions all I can do is I can construct a discrete distribution on the data. It is a probability of sampling this point x_1 is equal to the number of times x_1 appeared in my state set S divided by the size of s right.

So how do I simulate this distribution sample form S with replacement is it make sense to people right I am going to assume that yes in some sense the set s is representative of the underlying data distribution and I am going to simulate the underlying gator distribution by using the discrete distribution form by S. So what do I mean with the discrete, discrete distribution I do not know may be the underlying p is actually Gaussian or whatever.

But may s has only N elements so only these end points will have some nonzero probability of occurring so that is what I mean by the discrete distribution so I can just construct the discrete distribution from is and I will sample from that that will give me a spring right so I will call its 1 prime like that I can do that multiple times to get right I can go up to SL prime right I can do that you can create many many many such samples okay.

So now what I do to get a bootstrap estimate of the classification error so this kind of a sampling to produce this L sub subsets I have done is called bootstrap sample oaky. So wonderful once I would read such samples I can find out the bootstrap estimate of the quantity that I want right so in this case error so what will I do I will try on s1 prime right and what will I test on I will try non S1 prime and I will test on s - s1 prime right.

Because I am sampling with replacement so some of the data points will get left out right so whatever gets left out I will sample I will test on that so likewise I will try another classifier on s2 prime right using the same method if I am using back drop for training I will use back prop and train on s1 prime okay and test on s-s1 prime likewise I will try non s2 prime and test on s-s2 prime.

Like way so I will get how many estimates for the error L estimates I will take an average of that that gives me the bootstrap estimate for their right and you can show that the bootstrap estimate will have a lower variance then just the error estimate on just using s and just randomly splitting into test and train it will have a lower variance right so what I go again--again I want to be clear what do I mean by lower variance here.

If I give you another training data point of size N right and then you do the same thing you do you do to estimates one just train on the original set that is given to you and test on the test set once or blue this bootstrap estimate likewise I give you another data point net the data setoff size n another data set of size and so on so forth so now you have to to estimates for each of these okay the second estimate will be more consistent than the first estimate.

That is what I mean by lower variance okay so that is a bootstrap estimate okay so this is sometimes I forget the exact number so its estimate of the error then but it is some parameter that i am estimating about the whole process that's why I said right so this is one way of estimating the performance right or thing right I mean you can estimate anything you want on s1 prime s 2 primes 3 prime s 4 prime you can do whatever right.

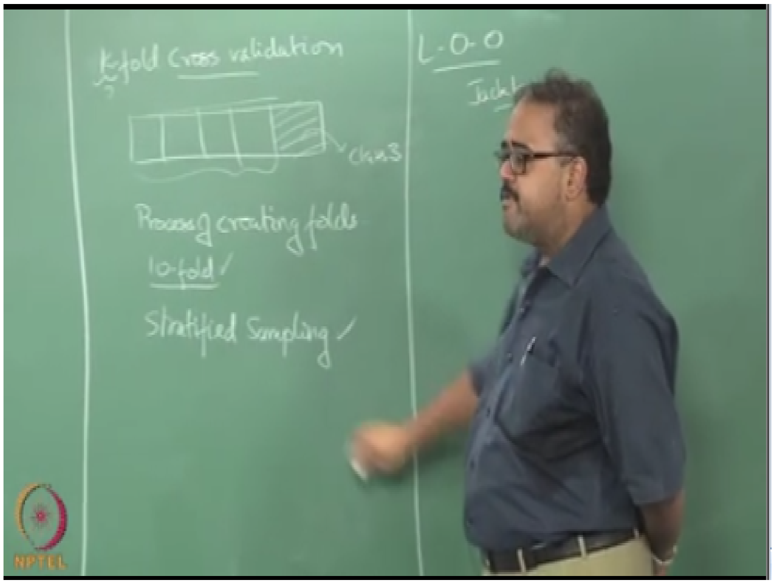
You can estimate the variance of the data on s1 prime just the variance office one prime not all here for anything that you can do this for each of those yell things and then you can have a bootstrap estimate of the variance right so you can do whatever--whatever statistics you want you can measure on these L sets right and then you can call that the bootstrap estimate okay so good stuff is a very general technique it is not just for error measurement.

Yeah so if you have been in--in a proper statistics course by now should have gotten into the variance reduction properties of bootstrap and we could have shown some interesting results right but I'm just going to tell you that I variance Goes Down and you can see intuitively it goes down and I will leave it at that okay so roughly about sixty--sixty nine percent of the data okay on an average rate sixty-nine percent of the data will be in s1 Prime.

And the remainder or sixty-three percent of the later sorry 63 will be in s 1Prime and remainder will be in the test data right. So this is also sometimes called that ever get 0 point 632 bootstrap or something like that because sampling with replacement leaves a certain fraction of the data in the in the sample and leaves another fraction of the data in the test set okay. So it is sometimes that fraction also denotes what bootstrap estimated this right.

Remember okay so this is one way of doing it and this works fine provided I had a large enough sample to begin with right so it had large enough sample to begin with so suppose your sample is smaller suppose your sample was smaller you do something called cross-validation K fold cross-validation.

(Refer Slide Time: 08:25)



So what you do is you take your sample okay you divided into multiple bins and it divided into cables know what I do I train on some K-1bins and I test on the last wave right the clips I use the first k-1 bins as my training data and one bin has the test data next what I do and servicing the last bin I use the second last bin as the test data and everything else has the training data right suppose I break this into k bins I will have k different estimates right.

I will take an average of those so which will give you a better estimate bootstrap or cross validation depends yeah okay that is not that is not the end of it we are done we want it depends on the size of the data I mean if you have a sufficiently large sample that your bootstrap assumption is true right so you might get a better estimate with bootstrap right one of the nice things about cross validation is that every data point is in the test set at least once.

Is it correct exactly once every data point is in the test set exactly once so in some sense whatever number that you are reporting is essentially the average over here performance on the entire sample that has been given to you right entire sample that has been given to you at some point you are using this so there are a few caveats that you have to worry about the first one is what is K right and the second one is the actual process of creating folds so what should be k 5 to 10 yeah okay.

Yeah so those are the numbers typically used five or 10 okay those are the numbers typically used and so depending on the number of foals you have okay you have stronger variance reduction properties the more the number of folds the more the variance reduction property provided the folds are large enough that if you have K data points there is no point in creating k folds but people do okay.

That is a very special kind of cross-validation I will come to that later and I'm just going to leave you with the short form of it okay so we will come to that so which is exactly creating k folds if you have k data points it is called leave-one-out leave -one-out cross-validation okay leave one out that is what L over--over stands for right that essentially means that you will train on N- or K-1 data points in tests on one data point so in some cases.

No--no there are actually in there are in cases where this still gives you good estimates okay and earlier version of this was the one of the first do not ask me why it is called jackknife but one of the earliest is this kind of variance reduction technique used for a parameter estimation it was called jackknife a jackknife is very similar to leave one out okay. So going back here so I would recommend that do not split it into so many faults.

That you have very little data point left in each fold right and so the typical number do not do not go more than 10 folds right if you manage to get 10 folds out of your data right then you should just be happy right that gives you good enough variance reduction so typically people in a report empirical results do not expect you to do more than 10 folds right. But then if your data size is small right people end up doing only five folds right.

There are extraordinary cases even when your data size is not small when you have to do fewer folds right. So let us think for a minute suppose I have am solving classification problem right so I have created these folds right and this is entirely of class3 okay and there is no class3 in this data right. Since you think this is odd this can cope happen quite frequently if you are dumb about how we split to your data so the data has come to you sorted by class level.

I will give you the class well sorted by class level and you do the fivefold splitting by serial number right so what will happen is you will have all your class go into one fold right the other four folds will not have any data point of class3 now if you try to test on this what will happen yeah right because you do not even know that exists class3 okay a training you didn't even know there was class3 so we are going to get hundred percent error right.

How do I avoid that shuffle the data in fact I do better than that I I do what I call what I call stratified sampling so I am not going to make much progress today so we will have to stop with the cross-correlation I have so much more that you need to talk about will do the next class so stratified sampling so stratified sampling essentially says that when you create the folds try to make sure that the class distribution that you had in the original data is maintained right.

Suppose I have five five data points of class 1 and 10 data points of class to right and I am splitting it into five folds right I should make sure that there are two data points of class two and one data point of class1 in every fold right so the class distribution the 2: 1 ratio is maintained in every fold. So this is called stratified sampling so this is something that you have to do so the recommendation is due 10 fold then you do stratified sampling.

And now can you answer my question why even though you have a lot of data you might be forced to have smaller number of folds class imbalance right so i might have very few data points of one class what if i have only 10 data points of one class right if I do tenfold sampling then forced at the cross-validation right I will essentially put one data point of that class into each fold that might not be sufficient for me to get a good good enough estimator.

And I might want to do a smaller number of force five may be 3 of course the other things which I could do but this is some case where you might want to you work with a fewer number of folds then what your data would suppose right so if you want to have a more formal description of cross-validation bootstrap and leave one out and all of that you are encouraged to read has t write the elements of statistical learning book has very very nice discussion on all of these things right so right now I will stop here right.

IIT Madras Production

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

www.nptel.ac.in