NPTEL

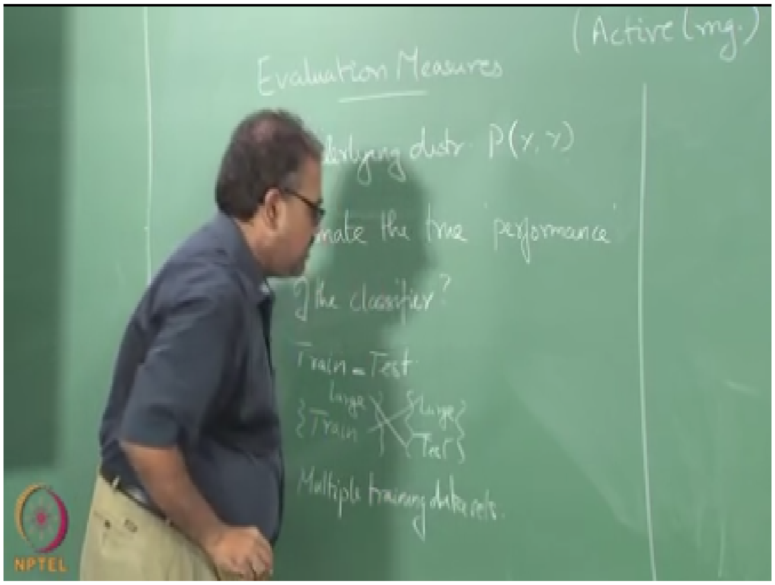
NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning
Lecture 48
Prof. Balaraman Ravindran
Computer Science and Engineering
Indian Institute of Technology Madras

Evaluation and Evaluation Measures I

Okay, so that leads us into some things to talk about.

(Refer Slide Time: 00:26)



I am going to talk about evaluation measures for a bit today so there is maybe like a little bit of hodgepodge class okay, I am going to talk about a few things which do not really fit into a bigger theme right, so for as far as evaluation is concerned we have spoken about some very standard think so far right, so in classification we talk about come on give me something yeah ,what would be the evaluation measure in classification, miss classification error then cross entropy then Gini index okay.

Gini index is not a evaluation measure right, Gini index is more of a parameter selection mechanism right, I can after and finish the classification I can actually compute some kind of a deviance right, and I can say okay this method giving me better deviance then that method and things like that or the same thing for 0 1 error I can use squared error right squared error right, squared error of your thing to a target variable and anything else can I use, penalties in what sense, okay.

So there are a couple of things that we had to be careful about here, so there is something which I optimize right, to get to what I want right and there is something which I use for evaluating what I finally get if you limit ourselves to the classify supervised learning scenario right we are not even started unsupervised learning limit yourself to supervised learning scenario more often than not what we really want to evaluate ourselves with is the performance on the entire data distribution right, so I have some distribution of data right, I do not know the distribution apriori it derive go back to the very first class we started talking about something serious I mean not the one with the pictures right, the one with the Greek in it, right.

So in that class we talked about that being an underlying data distribution right, and that we did not know about this data distribution the only way we know anything about this data distribution is through training data points, right is through the samples that are given to us right, so there is an underlying so we had this distribution so what I am really interested in is finding out when you give me a classifier how well it is performing with respect to this distribution, right.

So in that sense how well it is performing I am not really interested in figuring out the square the ridge regression loss or anything like that right so I am using that to come up with a single classifier, but at the end of the day when I am looking at how well this classifier is performing with respect to the underlying distribution I have certain measures, so one of them is the 01 loss, so I do not care how we are red with the classifier right I just want to look at the 01 loss and then I can do that right, that 01 loss gives me the miss classification error, right.

So ideally that is the evaluation measure that you should be using okay, so sometimes what people do because they are optimizing a different objective function they choose slightly different evaluation measures that may can make their method look better right, so squared error could be one evaluation measure, if you are doing classification 01 losses the measure that you should be looking at if you are doing regression while little tricky but square error is the most widely accepted measure for looking at regression.

But then you can look at other things also like deviance and other things you can use it for classification right, but having said that how do I estimate it is easier for me to write classifier or I will pick one I am going to pick classifier but some of what I talk about now works for regression as well right, and how do I estimate the true performance of the classifier, so what do I mean by that, I give you a sample data right I give you some sample drawn from P(x,y) right.

So based on the training that I that is all the information I have right, and I can use some of the data for training. I see find the parameters so how do I find out how good these parameters are or how good is my there are two questions to ask right, so the first question is how good are my parameters that I have found okay, the second question to ask is how good is the method that I use for finding the parameters is if you give me a slightly different data will I perform better or worse right, how will I perform right.

So I need to know something about the technique right, suppose I am proposing a new technique and I want you to go use it on your data later on right, but I should convince you that you can use it on whatever data you have right, so that means I have to convince you that my technique is good for finding the parameters this is the two things here for a given set of parameters I have to figure out how good they are right, and I also need to tell you how good my overall mechanism for finding these parameters are, right.

So for a given set of parameters how do you find out how good they are, on the training data I said good enough on the testing data, cross-validated okay, so if you okay, so we will get to this thing right, so one thing if we spoke about earlier was to split the data into train and test right, so if I estimate parameters on the training data again write it on the test data that will give you some performance, okay is that a good estimate of the true performance of the classifier.

Why not, fresh data might not be independent of the training data okay, the training data may be biased I met a wharf the model then you are doomed right, no, that is actually a very, very valid point in fact that is something which you will face in real life right, but the assumption mostly we make in theory is that the training data that is given to you is a sufficiently representative data of the true distribution okay, that is not the case then you are doomed anyway, right.

So you assume that it is the, so in real life that happens what do you do, in real life that happens okay mean you cannot avoid it, you cannot just say okay, I am assuming it is a properly representative sample of the underlying distribution then what do you do in that case, come on he is telling that you are not sample the entire range of P(x,y) so what is exactly so figure out where you are deficient right, so sometimes the most obvious thing is what you have to do, we have to sample more right.

But then do not blindly sample I mean of course you blindly sample you may actually return the same samples from whatever region you already have right, so what you should do is you should be more careful in how we do the sampling so you can use this is where you try to understand what the data is all about right, so you try to understand how the data is distributed the data that is given to you is distributed and figure out if there are parts of the input space which you believe we are important but are not covered in the data right, go and try to sample from that region, right.

So there are the different names for this okay, so one popular thing that people call this why this call, is called active learning, because I actively I am asking you for samples so I am not passively learning from the samples that were given to me okay, the learning algorithm comes back and says hey I want to know more about this part of the state space give me some samples from there right, I want to know more about this part of the input space give me so this is called active learning methods.

So your question this valid point of this discussion is yeah, one train one test is usually not a good idea right, so what do you do there are two things which you can do we can try to get multiple training sets from their data right, you can try to get multiple training sets from the data and try to so nobody is asking the obvious why are you getting multiple training sets from the data why not one large training set, okay why not pull everything together and then create one large training set.

Yeah, close yeah, so I will have to spend a whole week on weak classifiers right so we will come to that right, so it is a see that is one very, very amazing property that will look at later which probably the next class and later means pretty soon, on how you can take a lot of not so good classifiers right, which are just better than random of course it has to be better than random right, it cannot be worse than random so classifies that are just better than random and give you an accuracy of 51% okay.

So in the two class problem that is just better than random okay, I can take classifies at giving accuracy 51% and they can produce arbitrarily powerful classifiers okay, it is an amazing, amazing insight that came about a couple of may decade and a half ago now maybe do not I am old more than two decades ago right, and it is one the girdle price and things like that is it an amazingly wonderful inside and we will talk about that right.

So that completely revolutionized machine learning once right, people then started saying so I really did not, do not have to build this super optimized classifier I can build a lot of this almost moronic classifiers but the operational word is almost right, I have a lot of them right, I have a lot of them and I will be able to do really well.

In fact in many, many applications that we have worked on right in real life where I have worked on with real data, I find it very hard to beat these kinds of classifiers you can think of whatever optimal classifier you want to come up with right, but beating these kind of groups of weak classifiers is actually very hard in practice alright so we will talk about that but no, that is not what I meant here I still have a point to make, right.

So yeah, so even if you do one large training set, to do one large one train one large test set right, you can get away with it provided largest large enough right, provided large is something that is dense in your input space right, if the large is so large that you essentially plaster your entire input space right, so any point in the input space if a pick there will be one point very close by in the training set that is what we mean by dense I mean there is an actual more mathematical characterization of dense.

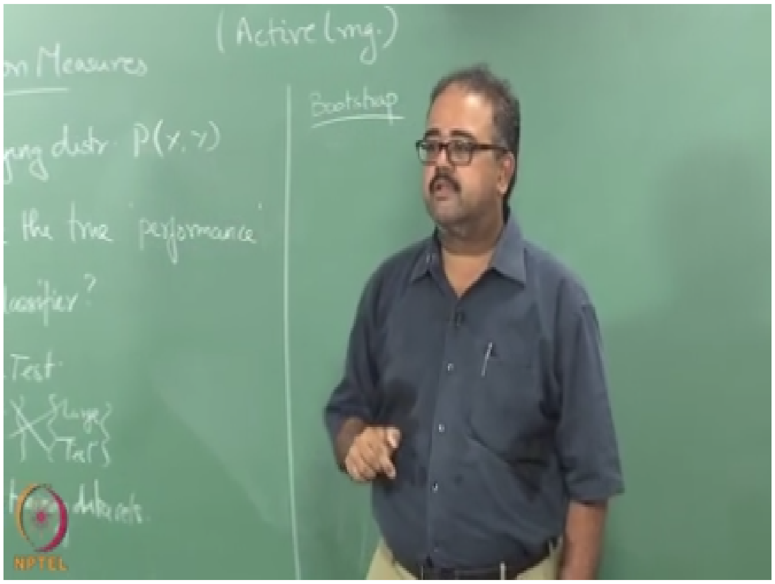
But if my training data is really dense in the input space then it is fine right, then I can get away with just doing one sample like one very large sample, but usually what is going to happen is you are not going to get such a large sample, so you are going to get a much smaller sample than that and therefore if you just use one sample and try to make an estimate okay, then the variance people remember what is the variance of the estimate, we talked about this again no, no that is unstable I am talking about variance yeah sorry, on data of similar the size we train a lot of models on data of similar size the parameter estimation I am going to make will be varying a lot right.

So it turns out that instead of doing one sample and then trying to train this if I take many, many samples right and then find the parameters on these samples individually right, and then take an average of those okay, turns out I can show that the variance will be lower in that case then what is the variance you are talking about the variance in the parameter that we are estimating, okay and what is the parameter we are talking about estimating here so here is the point where I am going to confuse all of you.

But the parameter I am talking about estimating here is the error is the miss classification error right, so I have the classifier right, what I am trying to measure is a miss classification error and that is what we have the whole discussion was all about I am trying to estimate the miss classification error right, so what I do is I start off with many samples of data right and on each sample I train a classifier separately again then I look at how the performance is on the test data and then take an average of all these performances and then I can tell you okay, if you give me a new data new set of data I expect to make this much error on the test data.

I am trying to figure out what the performance of the algorithm would be on a unseen train later I remember I was telling you I want to know how good my algorithm is right, so this way I can estimate the performance of the algorithm on the unseen data right. so there are many ways in which you can generate this, the many ways in as you can melt this multiple training data sets right, so many of these or I have strong roots in statistics and were typically designed in errors where the amount of data available to you was small, right. The amount of data available was small and we are trying to see how you can fake multiple data sets with a small amount of data okay, so the first technique is known as bootstrap okay.

(Refer Slide Time: 17:33)



So bootstrap is actually a very powerful statistical technique it is used in a variety of different places we will come back to another use of bootstrap a little later.

IIT Madras Production

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

www.nptel.ac.in