Indian Institute of Technology Madras
Presents

NPTEL
National Programme on Technology Enhanced Learning

Pattern Recognition

Module 06

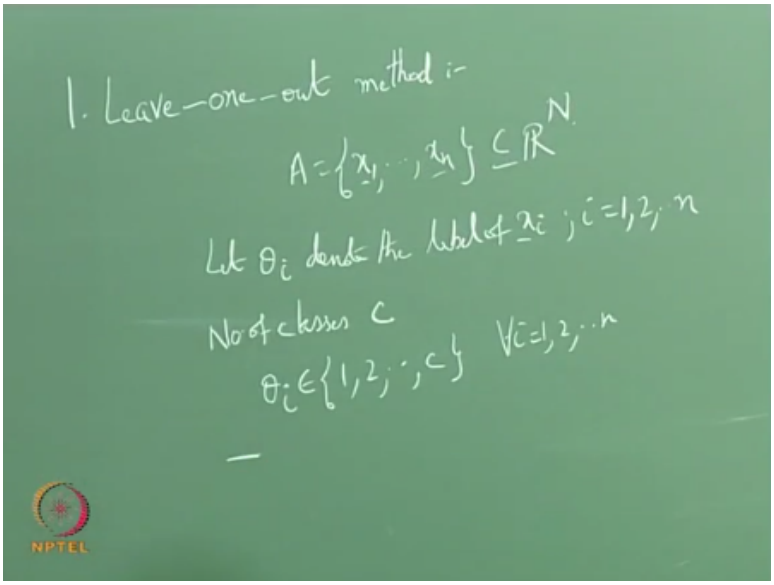
Lecture 01

Comparison Between Performance of Classifiers

Prof. C. A. Murthy
Machine Intelligent Unit,
India Statistical Institute, Kolkata

Till now we have been discussing about classifiers and we have been discussing also about feature selection in this lecture I shall discuss about how to do comparison between the performances of classifiers how to do comparison between the performances of classifiers there are basically a few methods available I shall discuss two of them.

(Refer Slide Time: 00:43)



The first method is what is known as leave one out method what is known as leave one out method suppose we are given a set of points a subset of n dimensional space and there are some

number of classes and let θ I denote θ I denote the label of xi that means θ I denotes the class of xi for all i = 1, 2 up to n let us just say the number of classes is K number of classes is let us call it C number of classes is C so each θi it belongs to the set 1, 2 up to C.

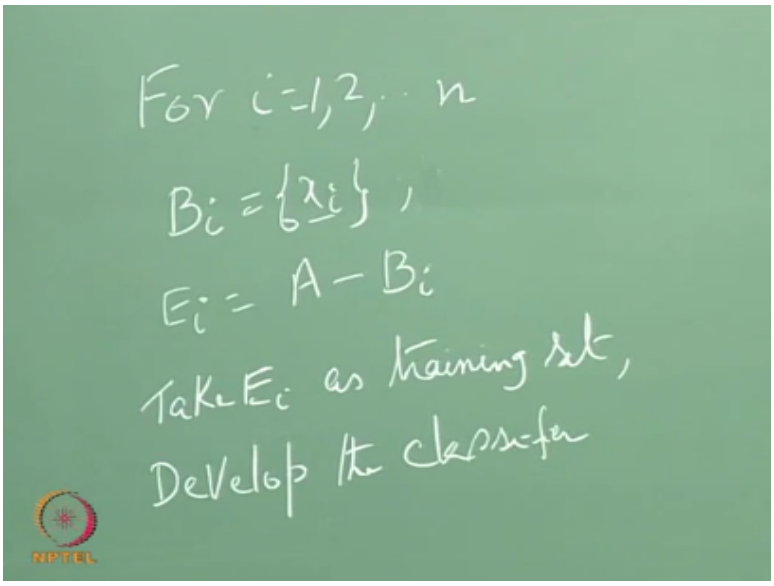
If the number of classes is C then each θi it belongs to the set 1, 2 up to c and we know xi we know θi for each i =1 to n we have the complete information we have the information now in the initial stages I talked about training set and test such where we divide this whole set into one part which we call it as training set in another part which we call it as test set and then using training set we get the classifier and using the test set we get the performance of the classifier that was this is true but you should there is some small point here the point is that the performance of a classifier depends on the training set.

Okay the performance of a classifier it depends on the training set for the same data set if I take a different training set probably the performance may go down okay now for the same training and test set if I use and I use two different classifiers then I can say with respect to this training set and with respect to this test set this classifier is better than that this sort of statements I can make then that the statement has this tree with respect to this training set and with respect to this test set.

Now suppose I do not want to put those two I should say I do not want to put those two phrases with respect to this training set with respect to this test set I would like to say with respect to this data set then how do I say that we are given a dataset then how do I say with respect to this data set this works better than that so basically what one needs to do is that somehow you need to change the training set if you keep training set fixed then with respect to that training set only you are will be in a position to say but the moment you start varying the training set.

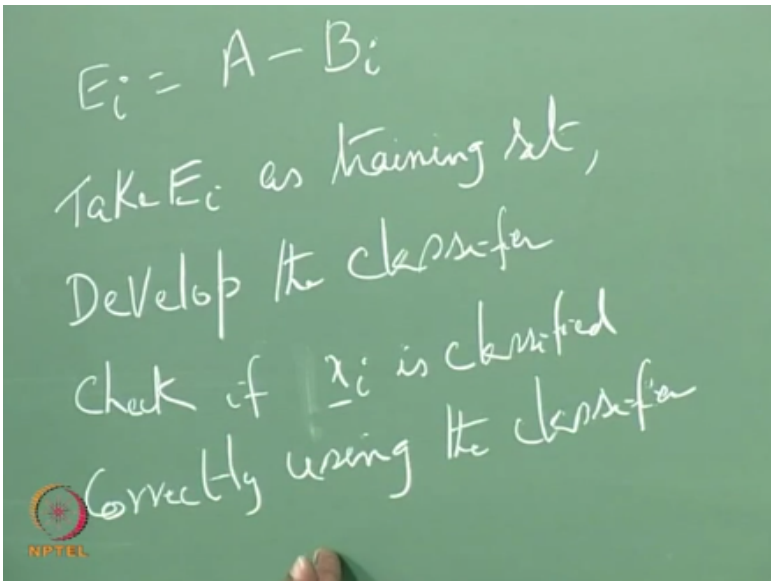
Then you cannot make the statement that with respect to this training set this is the, so somehow you need to vary the training set so then how do you do it in leave-one-out method as the name suggests we are going to leave one observation so how to do it so this is the thing that is given to us.

(Refer Slide Time: 05:31)



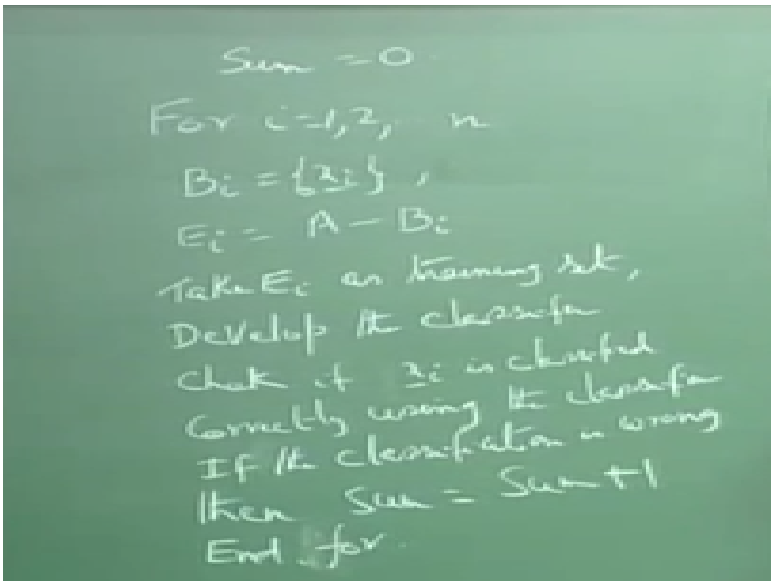
Then I will write a loop for i = 1, 2 up to n okay now let Bi = just the singleton xi Bi a singleton xi okay now let us just say Ei = A - Bi i = 1 to n okay so for i =1 find out B1 is just singleton x1 y1 is A- B that means x2, x3, x4 up reaction take Ei as training set develop the classifier that is if you are using something normal distribution based classifier and if you want to measure your means and covariance matrices then you can do that using this Ei or if you are using something like a k-nearest neighbor decision rule then your training set is EA.

(Refer Slide Time: 07:44)



Then what you will do is that check if Xi is classified correctly or not if Xi is classified correctly using this developed classifier using the classifier so you have got a classifier then check if F X I is classified correctly okay.

(Refer Slide Time: 08:14)



Now before that let me just write something as sum =0 then if it is classified correctly then there is no change if it is classified incorrectly then if the classification is wrong if the classification is wrong then sum is equal to and here end and the loop that means end for this I will come to that I will come to first you ask you tell me whether you have understood the method or not I will come to the remarks part later how much your main question is that how much it is feasible to implement it observation from the training I am understanding about the feasibility part.

Okay feasibility means if the number of observations small n is say of the order of say 1 lakh then you have to do this thing 1 lakh times my question is considered to be a 100 number of samples right and 100 number of training some samples and one training sample 100 that means small n = 100 yes sir and why one observation we are removing at a time right then we are calculating me and variance kind of thing is it really makes sense means mean will be somewhere near to that means original new to original mean only if we are considering mean and variance is a classifier.

That is if you are considering the normal distribution based classifier okay it may be or may not be solid if we are removing some kind of outlier then we can say if there will be in variation there will be some change in the mean but if BR removing some example from the cluster it mean will be somewhere near the original mean only what you are trying to say is that if you are doing this thing you what you are trying to say that most of the points will be classified correctly yes that is what you are trying to say.

I am trying to say sir if we are going to remove one observation from the big set of trin examples is it really change in classifier, that means so does it really change the classifier that means basically what you are trying to say is that most of the points will anyway classified correctly that is basically what you are trying to say that means there is no significant change in the performance of the classifier that is what you your basic contention eats you try to look at this thing from another way ultimately you would want every point at every point in the data set at some time or other to be included in the test set.

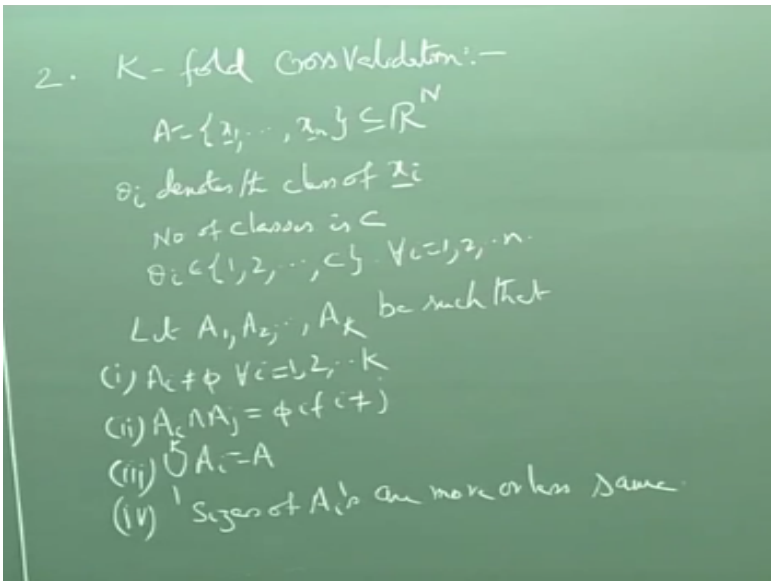
How are you going to do it I think you have understood the problem here if you want to make the performance for classifier to be in some way not independent of the training set then that means at the there has to come a situation where every point should at some time or other should belong to the test set how are you going to do it this is a simplest way of doing one of the reasons we can remove more than one it will be give a better but then at a time suppose you are removing two observations then how many sets you how to make NC2 that will be huge.

Even this will be huge for large values of n even this is huge for large values of N and if you want to remove two at a time then it is going to have NC to such sets you are going to have in NC is a very large number actually the question that I am trying to put to you is that some point of time you should make and every point should belong to the test set sometime another then how are you going to do this training 10 test set different distinction that is the basic question that I am trying to ask you.

This is the simplest way of doing it leave one out method then in that way you are not having any bias on any point okay there would not be any bias on any point but then this has some nice properties about which I am not for the present moment I am not going to the details but then the implementation wise if the value of small n is very large then implementation becomes a real issue as you can see it.

So now then the next question is that how do you do the implementation in a nice way without really changing the soul of this one how do you do the implementation then people have come up with the second one.

(Refer Slide Time: 15:01)



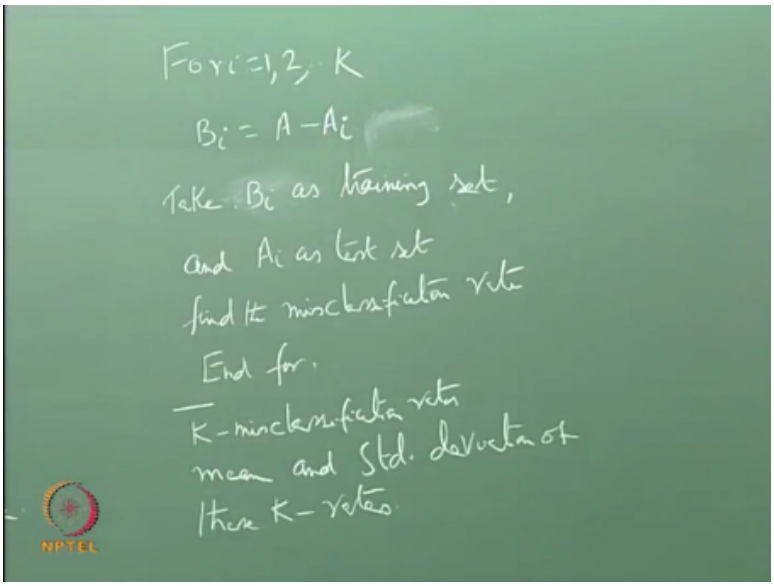
k fold cross-validation k fold cross-validation well what is the meaning of this K fold cross-validation here what we will do is our original set is all this whole formulation is there that means this is your number of points θ I divorce the label of the class I number of classes C θ i belongs to one to c so these are all there so I think okay let me just write them no problem A = x1 to xn subset of RN is your given point set θi denotes the class Of xi so number of classes is C number of classes is C so θi belongs to 1,2 up to c right θi belongs to 1, 2 up to c right.

For all I here, what we are going to do is we will take a partition of A let A1, A2, AK be such that Ai ≠ ϕ okay then Ai intersection Aj = ϕ if i ≠ j and union Ai 1,2 k = A so it is a partition of A x K subsets but there is one more condition this is the fourth condition is the sizes of Ai are more or less same let me expand on this what is the meaning of sizes of Ai is being more or less same suppose the value of k = 10 okay suppose the value of k = 10 and say small n = 100 then 100 / 10 is 10 okay.

Then every set has exactly 10 elements but on the other hand suppose small n = 101 then there will be nine sets with 10 elements and the10 the set will have 11 elements 102 then there will be 8 sets with 10 elements one set will two sets will have 11 elements like this have you understood what I wanted to say so there may be some difference but the difference should not exceed 1 okay.

Basically we would like to put the sizes to be same and if it is not possible well in some cases it would not be possible then you would like to make it more or less same now so once you do this thing.

(Refer Slide Time: 20:17)



Now you have a for loop for i = 1, 2 up to k right now you are let me just write Bi = A – Ai okay these are necessary Bi = A - Ai then what we will do is that take A- AI that is Bi as training set take Bi as training set and what is this Ai as test set take Bi as training set and Ai as test set okay and find the Miss classification rate find the Miss classification rate then you will get k such miss classification rates here so you will get k such miss classification rates k miss classification rates you are going to get.

Then you obtain mean and standard deviation mean and standard deviation of these k rates and when you are giving the result you should give the result you should give the mean and you should also give the standard deviation okay and you should mention how many folds you are taking you have taken 5 folds 10 fold, 20 folds you should mention that if you see the literature if you read articles you are going to find that the results of classifiers are given using these things they will tell you the number of folds of trans valuation.

And they will give you the mean value you can have miss classification rate as you can have correct classification rate that is fine does not matter okay give miss classification rate and give the standard deviation of this k values okay now so if you have I think I need to mention a few more points here suppose I my classifier is normal distribution based I would like to get the

mean and covariance matrix and then I will assume some prior probabilities and then I do the classification okay.

Now if I take B1 as my training set I may get some means and some covariance matrices but when I take B2 as the training set I may not get the same means and I may not get the same covariance matrices right then you see the classification scheme is same but the exact classifier there is a difference are you understanding what I am trying to say the classification scheme is same but exact classifier there is a difference because the means have changed the covariance matrices have changed.

Similarly if I use something like some nearest neighbor rule and if I take Bi as my training set and Ai as my test set I will B1 training set a one test set then I will get some classification but when I make the training set as B2 since the training sets have changed okay since the training sets have changed the performance of the classification scheme I mean the classifier is same but the exact classification rule there is a difference since the training sets have become different are you understanding and the same thing happens even with other classification schemes.

So here basically what we are doing is that we are trying to compare the performance of different classification schemes not the exact what μ you have got what σ you have got they may be different but classification schemes we are trying to compare here you are trying to compare the classification schemes herein then leave one out method note that every point every point in that data set some time Martha is going to become a point in the test set.

The same thing happens here also in k fold classic cross-validation also every point in the data set some time more than is going to become a point in the test set but there is a difference between these two methods in leave-one-out method it is an exhaustive method at a time you are leaving out one point and that you are doing it for every point in the data set but if you look at k fold cross-validation there are several if is and buts let me tell you the if is and but is I wrote here these three things $A_i \cap A_j = \phi$ $i \neq j$ union $A_i = A$.

Partition right let us just say I have 100 points and my k is 10 I know that A1 should have 10 elements A2 should have 10 elements like that A10 will have 10 elements but the actual question is how are you going to get those 10 elements right are you going to do them selectively that means no randomness involved or you are going to do it randomly how are you going to do it

how are you going to get those A1 A2 up to A10 suppose I get 1 A1, A2, A 10 and he gets another A1,A2, A 10.

But they need not be same are you understanding what I am trying to say in leave-one-out method for the same data set whatever I do or whatever he does the results are going to be the same but in k fold cross-validation since it depends on this A1, A2, and those 10 sets and they can be different for different persons so there is a problem here naturally the more and more folds you do it the better it is going to be okay the more and more folds you do it the better it is going to be and suppose you make the value of K = n then what is going to happen you will get this is it clear.

You make the value of small k = n then you will get the leave-one-out method so basically the main problem with leave one out method is that for a very high value of small n you really do not know how to implement it so in order to take care of this people are using K fold cross-validation in order to take care of that if the value of small n is really high then you can use this but then people also understand that it depends on how those A1, A2, AK are chosen how those A1, A2 AK are chosen.

So what many people do is that that also they do it randomly that means this A1, A2 AK this particular thing they do it may be 40 times are 20 times and take the average over this so that they would like to make this thing independent of this choice of A1, A2, AK are you understanding what I am trying to say is this clear no say I get 1 A1, A2, AK you get 1 A1,A2, AK he will get another A1, A2, AK like this you get some 40 or 50 such, A1, A2, AK so that you try to make it independent of the choice of this a AK okay your results this also people do it okay.

This also people do it to make it independent of the choice of A1, A2, AK so that then whatever results that you are going to show in the article they will make it independent of the choice of this the sets the K folds independent of the choice of the K fold this for some such thing had become necessary because it is difficult for you to implement the leave-one-out method directly over the whole data set so you try to do this in that way you are preserving the basic feeling of the leave-one-out method and also you are trying to make it independent of A1, A2, AK.

So this is one thing that people have been doing it if anyone can suggest something better than this and people who will take it I am not saying that this is the best way of doing naturally I mean I have told you the limitations of doing this thing maybe you can do a better choice I mean you can come out with a better scheme then whatever is existing any questions I will stop here you.

End of
Module 06 – Lecture 01

Online Video Editing / Post Production
M. Karthikeyan
M. V. Ramachandran
P. Baskar

Camera
G. Ramesh
K. Athaullah
K. R. Mahendrababu
K. Vidhya
S. Pradeepa
D. Sabapathi
Soju Francis
S. Subash
Selvam
Sridharan

Studio Assistants
Linuselvan
Krishnakumar
A. Saravanan

Additional Post – Production
Kannan Krishnamurty & Team

Animations
Dvijavanthi

NPTEL Web & Faculty Assistance Team
Allen Jacob Dinesh
Ashok Kumar
Banu. P
Deepa Venkatraman
Dinesh Babu. K.M
Karthick. B

Karthikeyan. A
Lavanya. K
Manikandan. A
Manikandasivam. G
Nandakumar. L
Prasanna Kumar. G
Pradeep Valan. G
Rekha. C
Salomi. J
Santosh Kumar Singh. P
Saravanakumar. P
Saravanakumar. R
Satishkumar. G
Senthilmurugan. K
Shobana. S
Sivakumar. S
Soundhar Raja Pandian. R
Suman Dominic. J
Udayakumar. C
Vijaya. K.R
Vijayalakshmi
Vinolin Antony Joans

Administrative Assistant
K.S. Janakiraman

Principal Project Officer
Usha Nagarajan

Video Producers
K.R. Ravindranath
Kannan Krishnamurty

IIT Madras Production

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

www.nptel.ac.in