Indian Institute of Technology Madras

NPTEL
National Programme on Technology Enhanced Learning

Pattern Recognition

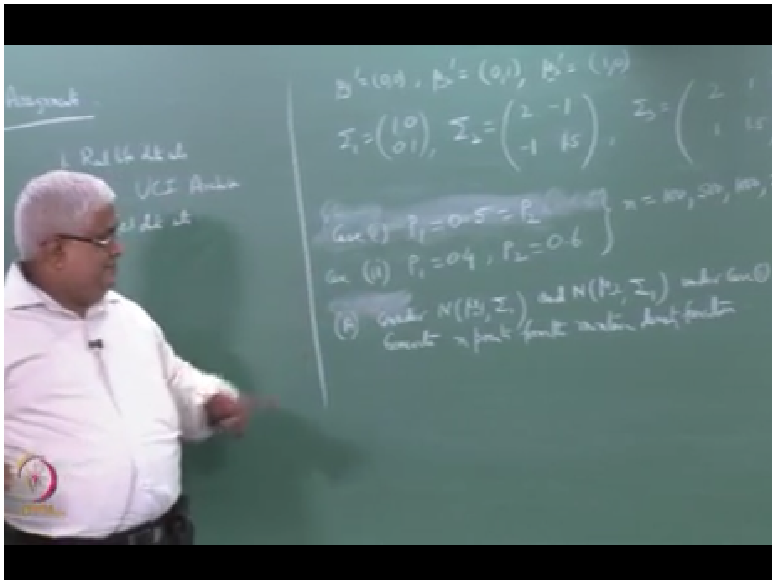
Module 02

Lecture16

Assignments

Prof. C.A.Murthy
Machine intelligence unit.
Indian statistical institute Kolkata

(Refer Slide Time: 00:16)


Assignment
UCI Archive
$\Sigma_1 = \begin{pmatrix}1&0\\0&1\end{pmatrix}$, $\Sigma_2 = \begin{pmatrix}2&-1\\-1&1.5\end{pmatrix}$
Case (i) $P_1 = 0.5 = P_2$
Case (ii) $P_1 = 0.4$, $P_2 = 0.6$
$N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_1)$

Assignments for the students can be divided in two different ways you can convert into two parts first part assignments on artificial datasets the second part is assignment on real-life datasets let me first talk about assignments on real-life datasets so if you want to really do the assignments on real-life datasets the first question is where are the real-life data sets available well these are available in UCI archive these are available in you CA archive you will get simply many datasets at least 50 if not more in fact 50 is a conservative number if not more okay, you will get datasets labeled unlabeled.

And you can do classification you can do clustering you can do feature selection all these things by downloading the datasets these datasets are freely downloadable you not have to pay a single PI for downloading them they are freely available so you can do whatever techniques that have been taught here you can apply those techniques on the respective datasets this is for real life datasets but then you will always have a problem there when you want to apply it on a real life data set the problem is that well.

You have applied many of these techniques maybe some of them are giving good result of them are not giving good results okay, the question is that how do I know the first thing is that how do I know which particular method is to be applied to which data set this is one of the questions that people have been bothered about for a long time see in order to give a partial answer to this question you need to actually look at artificial datasets in artificial datasets the modeling part we are going to do because, we are going to generate the data.

We know the model of the data so we know everything about the data set then if a method is working well whether a method is working well or not on the data set that can be found out easily if you generate the data on your own whereas if some other person generates a data set on some problem and he gives you the data set and if you say that some X method is working.

On this dataset then the person may come and ask you array I will give you ten more observations how then can you say I will give you ten more observations on the same from the same data set but these observations are unknown to you so if I give you those ten more observations which are not present in your data set then do you think your method is going to work then your answer will be you do not know where as an artificial datasets since we have we know the model of the data when we do it we can take many more points.

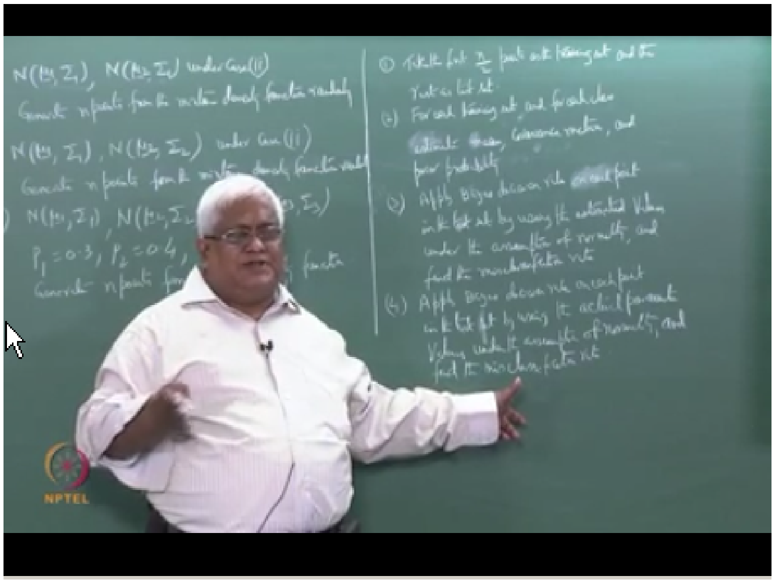
We can make the value of small n as such as possible to really see whether maybe for 1000 points one method is giving better than the other method but then for 2000 the one which was giving better results it is not giving better results though the one which was not giving better than it is giving better results now so these variations you can see because you can generate as many points as you can on artificial data sets you really know you can make your entry very large to check whether one method is better than other method or not whereas in real life data sets.

Since the number is limited whatever maybe your conclusions you are concluding only on those points only the moment a new point comes in you still have a problem of generalizing some of these methods to the next point you are not in a position many times many times you are not in a position to say the method that is working best for those many points whether it will still work better than the other methods if you include some more points whereas these problems you won't have for artificial datasets.

It is always necessary to work on artificial datasets for understanding the limitations of the method that is why now I am going to discuss about how to do the assignments on artificial datasets for this the first thing is that you need to generate the data set so what is it that you let me just give some examples synthetic artificial synthetic datasets artificial datasets so let me just give an assignment right now I will take points in r2.I have taken three two dimensional points prime means transpose so it is basically a column vector zero.

This is column vector 0 1 this is column vector 1 0 these are three covariance matrices and case 1case to case 1 p1 = 0.5 = p2 case 2 P 1 = 0.4 P2 = 0.6 ok now let me just write capital A this is the first problem first one consider normal μ 1 $\sum$1 and normal μ 2 $\sum$ 1 only normal μ 1 $\sum$ 1normal μ 2 $\sum$ 1 under case 1 that means the prior probabilities are 0.5 in point 5 okay yes and generate end points from the mixture density function mixture density function generate endpoints from the mixture density function.

What is the value of this n the value of n is we'll start with 100 the next value is 500 the next value is 1000 the next is 2,000 and the next is 5,000 okay, so each time say for example here you should generate first hundred points then next 500 points next 1000 points next 2,000 and then next 5,000 points so when you generate appoints in the mixture density function. You will get the label and also the point you will get the label and also the points this is A.
(Refer Slide Time: 10:50)


N(μ1, Σ1), N(μ2, Σ1) under Case(ii)
Generate n points from the mixture density function randomly
N(μ1, Σ1), N(μ2, Σ2) under Case(ii)
Generate n points from the mixture density function randomly
N(μ1, Σ1), N(μ2, Σ2 ... μ3, Σ3)
P1 = 0.3, P2 = 0.4,
Generate n points from ... function
(1) Take the first n/2 points as training set and the rest as test set
(2) For each training set and for each class estimate mean, covariance matrix, and prior probability
(3) Apply Bayes decision rule on each point in the test set by using the estimated values under the assumption of normality, and find the misclassification rate
(4) Apply Bayes decision rule on each point in the test set by using the actual parameter values under the assumption of normality, and find the misclassification rate
NPTEL

B normal μ 1 ∑1 normal μ 2 this is also ∑ 1 under case 1 under case - sorry here the prior probabilities are different 0.4 and point 6 generate endpoints from the mixture density function mixture density function C is note that in the first two cases the covariance matrices are same so in the third case I will take covariance matrices to be different normal μ 2 and ∑ 2 and here under case ok let me just take two so here the covariance matrices are different generate end points from the mixture distribution mixture density function.

And the last one is normal μ 1 ∑ 1normal μ 2 ∑ 2 and normal μ 3∑ 3 and here P 1 is equal to 0.3 say P 2 = 0.4 P 3 = 0.3generate endpoints generate endpoints from the mixer density function so here so data sets are generated note that when you are generating these points you are not only knowing the point but also the class label of the points after the generation now what are the works that you are going to do well the first one is that so there are totally four cases and in each one the value of n can be any one of the five so totally twenty different situations totally twenty different situations now take the first n by two points at the training set.

And the rest as test said take the first n by dew points as the training set and the rest as that just says just sit now for each training set and for each class for each training set and for each class find okay estimate mean covariance matrix estimate mean estimate covariance matrix and prior probability for each training set and for each class estimate mean estimate covariance matrix mean covariance matrix and prior probability after estimating these three so apply Bayes decision rule apply Bayes decision rule.

On apply the restriction rule on the test set on your application based on each point in the test set by using the estimated mean by using the estimated values estimated values what I mean is estimated mean estimated covariance matrix and estimated prior probability so under the assumption of normality to all these things if you assume normal distribution under the assumption of normality so assume normal distribution since you are assuming normal distribution what you need to know is the mean covalence matrix.

And then the prior probabilities for applying the base station so you have estimated those things so apply base station rule and find the misclassification rate what is the misclassification rate take a point in the test set you know already the label of the point then you do all these things suppose the label says that it should go to class 1 but suppose your rule says that it is going to class 2 then it is misclassified if the point is in class I and if you put it in class size and there is no miss classification otherwise there is miss classification.

So what is miss classification rate means number of misclassified points divided by the total number of points number of misclassified points divided by the total number of points that is the misclassification rate note that all these things you are doing on the estimated ones now you do the same thing on the actual ones you already know the actual meaning the actual covariance matrix and the actual prior probabilities apply Bayes decision rule on each point in the test set by using the actual values by using the actual parameter values not.

The estimated under the assumption of normality we already know actually the normal and find the misclassification rate so you have to miss classification rates now one by using the estimated values another by using the actual values as n increases these two Misclassification rates the difference should decrease these two rates should be very close then we know that this whole thing is fine whatever theory that has been taught we assume normal distribution we have done many things okay one of the ways.

That you can check is by doing this so this is an assignment but let me tell you a small problem that you may face in all these things where is this generate endpoint from the mixture distribution and end points in the mixture density and end points end point from the mixture density function all these things there is one thing that I need to write randomly always so randomly, randomly here also randomly this randomly word is extremely important.

Because computer cannot generate random numbers it can generate only pseudo-random numbers whatever methods that you are going to use it can generate only pseudo-random numbers it cannot generate random numbers so this is one thing that you need to keep on your mind to the extent possible I do not know how you are going to do it you should generate the points randomly okay, and if your method of random number generation is fine you are going to see this result you are going to see this result.

Let me tell you how to generate points from the mixture density function this generation for generating one point from the mixture density function it has two parts the first part II generation out of a random number in the interval 0to 1 the first part is generation of a random number in the interval 0 to 1once you generate it you should see where the random number belongs to that is in this case 1 if the random number belongs to the interval 0 to 0.5 then you should generate point from the first distribution if it belongs to 0.5 to 1then if you generate point from the second distribution.

This is what one needs to do in each one of these four cases a-b-c-d for example in this case generator run a number between 0 and 1if it belongs to 0 to 0.3 then you take it from the you should generate a number from the random vector from the first distribution if it belongs to 0.3 to 0.7then you should generate a random vector from the second distribution from 0.72 one you should generate a random number random vector from the third distribution so generating a random vector from one distribution means it has from the mixture distribution means it has two parts one is from which one of the original distributions.

You need to do it and the second part is generate from the distribution so that you are going to get the label and as well as the point you will get a point and the according and you will get the label also like that you need to generate points from you need to generate 100 points set then500 points at 1,000 points at 2,000points at 5,000 points it then so if you apply this thing then the first n by 2points will be training and the rest will be test set that means if you taken as 100 first 50 points will be training and then read that as 50 will be test.

Similarly for other values for example for this the first 2,500 will be training and the next two thousand five will be test set then you use the training set and then you estimate the mean covariance matrix and prior probability for each one of the classes then using the estimated things then for every point on the test set in the test set you see whether itis rightly classified are wrongly classified and you get the misclassification rate similarly you get the misclassification rate by using the original parameter values where is null parameter values are known µ 1 µ 2 µ3 ∑ 1 ∑ 2 ∑ 3 they are all known so you will get 2 error rates 1for the original one for the estimated ones.

So you need to check whether that the estimated ones are going to cover is null world note that the original one the misclassification rate according to the theorems as n goes to ∞ the misclassification rate by using the original parameter values it should go to the base misclassifications probability our misclassification probability of the base decision rule since we are applying Bayes decision rule somehow we need to talk about misclassification probability of the base station so what I am trying to say is that if you use the estimated ones then also for sufficiently large and as if n is increasing then the estimated misclassification rate.

It will go to the base miss classification probability it will go to the base miss classification probability so for this since we do not know the base mix classification probability even that we are doing some sort of an estimation here though we are using the original parameters still it depends on what the training what the test set is the test set is varying if I take one set of 5000 points I will get one test set and if you take something then it is going to be a different test set now the next question is that the same is the case with training also how do I know that the misclassification rate.

That we are getting whether they are really good estimates are not for that in principle you are supposed to do the same experiment in many times say for n is equal to 100 and for this one you do it let us just say ten times so hundred point sets you should generate it ten times okay so you will get 10 such hundred point sets and for each set you have an misclassification rate based on estimated values misclassification rate based on actual values misclassification rate based on estimated values misclassification rate based.

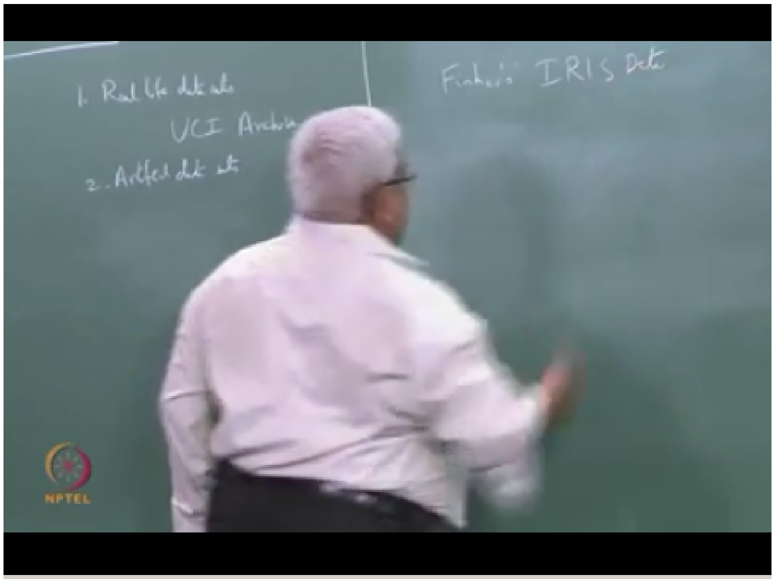
On actual value so your misclassification rate based on estimated values if you take the average it is supposed to be close to miss the average that you are going to have when you are the Miss class you can read rate based on the actual value these two averages should be closed so in principle one should do it so this also you can take it as a part of the assignment this is on base decision rule so we have been discussing about applying Bayes decision rule are in fact by base station rule the other classifiers that have been developed so far about classification.

I mean for classification they can all be used you have a training set and you have a test set using the test set first using the training set you develop the classifier and then using the test set you find the misclassification rate that is it whatever classifiers that you have for example if you have minimum distance classifier well there you have to have the means with respect to the mean you calculate the distance.

And for whichever class the distances whichever mean the distance is the minimum put it in that class so you can do the same thing by using this training and test sets okay and so like that whatever class files that have been developed you have this training data set and test set data set and you can apply those classifiers and check their performance now that is about artificial data set.

Now there are this real-life data sets about this real-life data sets I would like to make a few comments those post those are you who are aware of image processing you know that there is this llama image on which most of the persons would work probably on an image you know that there are several work that have been done similarly an unreliable data sets also there is one specific data set on which lot of work have been done that is Fisher's iris data Fisher's iris data.

(Refer Slide Time: 31: 50)



So you have four-dimensional observations you have three classes and from each class there are 50 points 50 + 50 + 50totally hundred and 50 points. So you have to generate a classifier this is one of the most widely used datasets. I do not know many persons they are just too many whatever classification methods have been developed I am more or less sure that almost all of them have been applied on this data set and if you develop a new scheme.

And if you apply it and if you want to say that your method works better than the other scheme then one of the ways in which you can say is that apply your method on Fisher Syrus data set and then you see the latest literature you will find many persons working on this thing and showing the thing.

So you say that you are one if you are one is better than that whatever you find in the latest literature then that will add plus point to your work so Fisher Syrus data is one such standard data set and like that nowhere are many more data sets before I conclude I would like to mention one thing there are some datasets available there for which the training sample sets have also been given okay so that I mean you also always have a problem of what is the size of your training samples which that is one and the second one is that once you choose the size what are the points that you are going to have in that so you will find some data sets.

There where the training samples that are also given there so in that sense for those datasets you can do the comparison very well but for many data sets there are no training sample sets they are just labeled data then since they are standard data sets you will find too many persons working on those data sets and you will find many results if your results on just say let us just say twenty to twenty-five data sets if they are better than the existing results then at least you have a contention to do you have something to say to the reviewers that your results.

Are better it is very difficult for the reviewers to reject your paper straightly even if you are saying it only experimentally because they are standard all the available data sets and you are saying that your results are better than all the existing results that is a good use of I mean this is why people like to apply their own methods on real-life data sets on where these data sets on these data sets already there are many works.

So that they are they have become benchmark data sets so I would like you people to work on these data sets and if any new method is developed you should apply it and find the veracity of your method find how good or how bad is your method on the basis of results that you are going to get on the real-life datasets thank you.

Online Video Editing /Post Production
K.R.Mahendra Babu
Soju Francis
S.Pradeepa
S.Subash

Camera
Selvam
Robert Joseph
Karthikeyan
Ram Kumar
Ramganesh
Sathiaraj

Studio Assistants
Krishankumar
Linuselvan
Saranraj

Animations

Anushree Santhosh
Pradeep Valan .S.L

NPTEL Web &Faculty Assistance Team

Allen Jacob Dinesh
Bharathi Balaji
Deepa Venkatraman
Dianis Bertin
Gayathri
Gurumoorthi
Jason Prasad
Jayanthi
Kamala Ramakrishnan
Lakshmi Priya
Malarvizhi
Manikandasivam
Mohana Sundari
Muthu Kumaran
Naveen Kumar
Palani
Salomi
Senthil
Sridharan
Suriyakumari

Administrative Assistant

Janakiraman.K.S

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