Indian Institute of Technology Madras

NPTEL
National Programme on Technology Enhanced Learning

Pattern Recognition

Module 02

Lecture 04

Training Set, Test Set

Prof. C.A. Murthy

Machine Intelligence Unit
Indian statistical institute Kolkata

Mentioning to you about the estimation of parameters given a data set and if you know that it follows normal distribution how to estimate mean and how to estimate covariance matrix these details I was giving you in the last class now this all these details are given to you on the basis of something what is that that you are given a data set and for the data set you would like to find mean and covariance matrix now this is there is a small caveat here since.

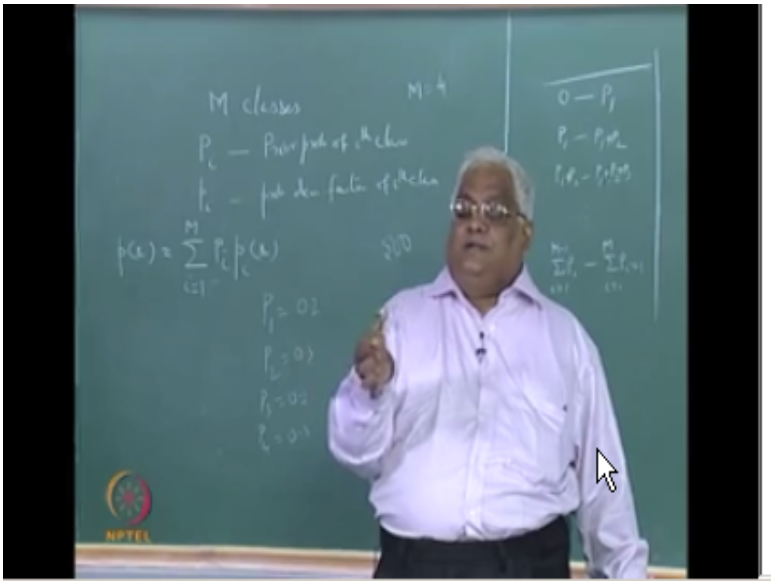
I am talking about classification problem since there is not just one class in general you have many classes from each class you need to have some set of points from each class you need to have some point so that you can estimate the mean and covariance matrix if the distribution is normal okay if the distribution is normal then from each class you need to have some point snow there was also one phrase that was used that is the independent and identically distributed that was the phrase that was used.

Now if you ask me a question is it really true that for real-life data sets the phrase independent and identically distributed whether it is valid or not then that creates a small problem many times it is not valid I will give you a few examples ,I will give you a few examples suppose you are interested in classifying pixels in satellite image data you are interested in classifying pixels from a satellite image data okay, so for that you need to have let us just say the number of classes is two you need to classify a pixel into water or land only two classes let us assume that okay so for that you need to get some pixels belonging to water class some pixels belonging to land classes

so that you can somehow get their means and mean of the water class and mean of the land class variance covariance matrix corresponding.

To water class variance covariance matrix corresponding to land class so that if you want up a normal distribution then you can help these means then using this when an unknown pixel comes in then you will classify that to either water or land okay so for this we need to have some points from water class and some points from land class now here the meaning of independent and identically distributedthe meaning is the following.

(Refer Slide Time: 03:41)



If you have M classes and prior probabilities are capital P a is prior probability of IH class and small P a this is the probability density function for either class probability density function usually it is called PDF a 5 class then what you will have is this is known as mixture distribution which I am representing it by p(x) which I am representing it by p(x) summation I is equal to 1 PM capital PI multiplied by small p IX now from this mixture density function.

You are supposed to draw points randomly from the mixture density function you are supposed to draw points randomly now what is the meaning of drawing points randomly from a Mixel density function it is like this there are this P 1 P 2capital P 1 capital P 2 capital p m.

These are basically Clausen krummel's one interval is 0 2 P 1 another one is P1 2 P 1 +P 2 the other one is P 1+ P 2 2 P 1 + P 2+ P 3etcetera and the last one is P I =1 to M -1 to I =1 to M P I

which is nothing but 1 this is actually equal to 1 these are class intervals you draw a point randomly from 0 to 1 if the point falls in this interval then you will select a point from the first class if it falls in this interval you will select a point from the second class like this if it falls in this interval you will select a point from the m-f class that is the first one.

So first you are deciding from which plus you are supposed to select a point then once you have decided the class then suppose say it is class 1 now you are supposed to draw a point following small p 1x if it is class 2 then you are supposed to draw a point following the density small P 2 etcetera like this suppose you have drawn say let us just say 500 points let us say you have drawn 500 points and let us say the number of classes is 3 let us say the number of classes is 3 and let us just say it is for.

So that I can get the numbers nicely let us say a number of classes is four and you have got five hundred you have to run five hundred points and let us say P 1 = 0.2 say P 2 = 0.3 say P 3 = 0.2 say P 4 = 0.3 okay then approximately if you draw500 points approximately 0.2 into 500what is the value of that around 100 and on 100 points will be from class 1 and on 150 points will be from class 2 and around 100 points will be from class 1class 3 and there on 150 points will be from class 2.

In that way you are getting points from the four classes this is what should have happened then we are drawing points independent and identically distributed from this mixture density function if you have 100points from class 1 150 from class 200from class 3 150 from class 4 then and if we say that they are IAD then what could have happened what must have what we want to have happened is the following is this one is this clear to you what we want to have happen is the one that I stated but had has it really happened unfortunately the answer is no why say as I was telling classification of pixel from satellite imagery you want o classify a pixel to land our water.

So you are supposed to get points from the water class randomly you are supposed to get points from the land class randomly and you are supposed to have the capital pas that means the prior probabilities also you are supposed to have them you are supposed to have those values but what is generally the case is that you do not know the prior probabilities generally the case is that you do not know the prior probabilities that is the first point now the second point is that points from the water class you are supposed to have drawn them randomly.

That means from all the available water pixels you are supposed to have taken those points randomly now are you actually doing it unfortunately the answer is no why the answer is not the answer is no because you are not it takes really lot of time and energy and money to draw points randomly lot of time energy and money to draw points randomly you have chosen somehow one particular position then you are supposed to go to that particular position okay then and first it is like this you have what a satellite imagery the is a very big area of satellite image.

(Refer Slide Time: 11:24)

These this is an area and you have got usually I mentioned Calcutta as an example because I am basically from Calcutta we have Hooghly River it just goes like this and Calcutta has very many points there are too many points in Calcutta and they are curate to many places there is hurrah bridge which is somewhere here and this bridge we call it his near diction is where there is a bridge here and there is a bridge here and there is another ability of second to play bridge and there are very many small water bodies several places in Calcutta now you are supposed to get pixels from water.

Now if I if you know that if you this is this area is say this is somewhere near duction is fair and this is second who play bridge if you take that just the frame just above this then it will go like this you go up to barrack poor and so on and so forth okay you go up to barrack for in fact the river continues this is Ganges you can go up to Himalayas also okay if you go on looking at the path of the river now your aim is to put a pixel into land class or water class only two classes

now how many satellite how many such images do you have from which you would like to make.

This thing into water class our land class how many such images you have you might be interested in classifying pixels only from this image or you may be interested in classifying pixels from forty fifty hundred may be thousand such images the whole path of Ganges River you have taken then it would be minimum 1000 images if not more the whole path of Ganges River you have taken okay and so you would have minimum 1000 such locations if not more than from these locations you take a pixel and you would like to say whether.

It is belong to water class our land class that is what is your aim okay are you might be interested in classifying only this region now let me just consider one example where you are interested in classifying pixels and only this region where a each pixel you would like to put it into land class or water that was okay now suppose I know the prior probabilities how do I know the prior probabilities how do I know that how many percentage of pixels they are in water how much percentage of pixels is in line.

If I really know it that means I have an idea of which pixel is from water which pixel is from land then only I would be knowing this ratio or I would be doing and someone else has done some survey I might be taking data from there if someone else if someone has done some survey if I take the data from there then I more or less know many things about this region I more or less know many things about this region I more or less know which pixel belongs to water which pixel belongs to land then I do not need to do the classification.

Why do you want to do the classification you want to do the classification because for something that is completely unknown to you that pixel are that particular observation you would like to put into one of the classes here is something if you know everything you do not need to do classification most of the times what happens is that you will know the prior probabilities that means you know the proportion from each class that happen only because you know everything about it and if you know everything you do not need to do the classification right so prior probabilities let us assume that they are not known to you renovation that prior probabilities are not known to you if prior probabilities are not known and I need to get pixels from some pixels from what are some pixels from Lange how do.

I get them again I need to do some sort of a survey over some small area I need to do some sort of a survey over some small area in the region under consideration and from there I need to pick up the points like this from there I need to pick up the points like this so that this information I would like to sort of generalize it to the whole area that means what you would send someone who would go to let us just say some portion here and from here he would get some pixels.

Because this is in the river area maybe from here he would get some pixels from here from deadeye about some pixels maybe from here he would get data about some pixels in that way he is getting data about the river pixels similarly you would send someone maybe to some place here then he would get data from there about the land like this land from here land like this so you need to send people to some areas but how do you know those areas see the main problem comes in because them the main problem is that most of the times.

The observations are not taken in this way whatever is suitable for us whatever is within our means we would use that to get the I should say the set corresponding the initial set which I am calling it as the set of observation is belonging to each one of the classes that is the data set that is given to us this given to us is coming if you have to perform the experiment you need to send some persons to that corresponding area and those persons somehow they get hold of the class information about different classes and they will come to you and usually.

It is not done in the aid way and it is many times impossible to do it in the aid way independent and identically distributed that is one of the problems so that is the data that is given to you are given some number of observations from each class say from the class one say from the class one here say you have around 100 observations from class to say it is 150 observation this is again100 let us just say this is 150 this is the data that is given to you let us say now you are supposed to do the classification how do you do the classification I said that there are many classifier rules of classification.

Then which rule you want to apply and why do you want to apply how do you know which rule is better and why it is better in order to do all these things what people do is that they divide this whole set into two parts the first part they call it as training such the second part they call it as test set from the training set they develop the classifier develop the classifier means if someone wants to up like a nearest neighbor rule he would have like a nearest neighbor rule on the training set.

If someone wants to apply a normal distribution issue normal distribution estimate the parameters and use Bayes decision rule somehow estimate the prior probability is also fine that can be done if someone wants to do something like a multi-layer perceptrons using the training set that can be done in that way very many classifiers sousing the training set they would develop the classifiers now once the classifiers are developed they use the classifier they use this classifier information classifier to classify points in that test set to classify points in the test set they use this classifier.

Now whichever classifier is giving better performance on the test set assuming that the performances of all these classifiers on the training set are same then that classifier is a better classifier than the other classifiers let me repeat if you have chosen four classifiers and the performance of the four class voice on the training set is more or less same then that classifier which gives better performance on the tests the data set that you take it as a better classifier than the other classifiers and this is one of this is the purpose of actually using training and test sets.

The purpose is fine but there are very many questions here the questions are here you have 500 points right the data set has 500 points the first question is how many points should be dead in the training set how many should be there in the test set okay how do you get hold of those numbers is it the case that we will have 250 in the training set 500 exactly half you take250 in the training set 250 in the test set first you have a question of numbers the second question is suppose you have those numbers how do you take those points let us just say 250 points are must be there in the training set and250 points must be there in the test set let us say somehow we have chosen this number then among.

Those 250 how many must be there from class 1 how many must be there from class 2 how many must be there from class 3 how many must be there from class 4 this is one question and again if you have chosen those numbers how to get those many points from this thing how to get those many points from here what is the basis for it what is the basis for it now these questions they do not have clear-cut answers they do not have clear-cut answers the reason is the following I will tell you again partial reasons not complete reasons.

The subject pattern recognition I should say it is was started by statisticians Bayes decision rule was by statisticians and the usual Fisher's linear discriminate f unction which of seconded us

would be teaching and other such things there are many things are by statisticians the electrical engineers came into the picture sometime in 60s electrical electronics engineers they came into the picture sometime in 60s and basically because they have got some phenomenon I mean they were doing signal processing there they are getting many observations from different classes.

So how do you do the classification the signal can be character recognition problems you might be having many such application domains in fact one of the first places where pattern recognition was applied was in character recognition problems so that is how electrical engine is selected in electronics engineers are I Triple E okay it came into the picture then once there once you started looking at the application they are just too many of them then I mean the subject improved enormous thing I mean the literature has increased enormously so you would actually see in the literature two distinct perspectives.

One is the perspective from a statistician and one is the perfect perspective from computer scientists statisticians would want more accuracy so you would get more accuracy if you take the size of the training data said to be more test data said to be less so if you look at the literature the early part of the literature then you would see in the statistics literature around 80 85 percent is the size of the training dataset sometimes 90 percent and then 15 percent is the size of that test dataset okay.

Now you see the papers from the computer scientists computer scientists they do not want to do many calculations they want to get the results fast now if the training set sizes are high you need to do too many calculations right so if you look at the literature on pattern recognition from the computer science perspective you would see the sizes often training sets to be small test sets are large are you understanding what I am saying the test set sizes have large training set sizes are small because training set size is small that implies.

You need to-do less number of calculations okay so you would see clearly these two distinct perspectives in the literature and unfortunately I belong to both the categories statistics as well as computer science I do not want to put myself in this are that I am just telling you what is there in the literature okay so this is so I do not want to tell you how to choose those numbers how many points must be there in the training set how many must be there in the test set that I do not want to make a statement on that okay.

Now once you have chosen those numbers then how to select those points that I will make some statements suppose in one class the observations.
(Refer Slide Time: 28:09)

That are given to us in one class the observers are given to us say they are like this is in one class these are the observation that are given to us now somehow you have select you have decided some number of points to be selected know which one I mean among these points let us just say you are supposed to select say some 50 points from.

This now which 50 points you would like to select that is the question now suppose I select the point only from this region then is it proper the answer is no so that means how do we select the points we select the points in such a way that it is somehow the points are spanning somehow the whole of the training set the whole of sorry the whole of the set that is given to us we need to select points in such a way that they span this whole of the region that is given to us are you understanding this is I mean this is the main thing in training such and this is also true for Test section.

This is also true for test set let us say most of the points in the test set say they are coming from this region then the test set is biased towards this region then is the test set I should say reliable your answer would-be known so test set also should have the property of spanning the whole region from which the points are given to us it should be the property of both the training set as

well as the test set then only the classifier that you have developed that would be reliable and its performance would be reliable.

If your test set is chosen nicely you might be having a fantastic classifier let us just say your class your distributions are really normal but if your training set has got points from here then your estimates of mean and covariance they will be very bad right your estimates of mean and covariance they will be very bad are if your training set is really taken from the whole thing but the test set most of the points are from here then your test set reliable there is no reliability of the test set your method may be a very good method.

But if your data is not proper you will not get good results if your method may be a very good method but if a dev sure data is not proper the way you have done the experimentation is not proper then you will not get good results okay the meaning of getting good results means you must have confidence on the results that you have got you will be lacking the confidence if that data or the processing of the data is not proper if the processing of the data is not proper then you will not get good results.

Now this particular thing about the points coming from the training the training set points are spanning this region this comment holds good even for the selection of points I will tell you the meaning suppose your whole class now I am not talking about training set I am not talking about any set now suppose this is your whole class say a class one of some particular problem this is your whole class now unfortunately the one who has collected observations from this class suppose he has spanned only this part we have sent someone to collectobservations on water and land say this is the whole of water class.

But the person who had gone to collect the observations on water he had only got points from say this region but not the whole of it then your method may be very good but the generalization capability will be very because the data is not proper that is where basically independent identically distributed comes in if they are independent and identically distributed they will span the whole of the thing which is very difficult to put it into practice because of very many reasons so once you are given a data set one of the first things that you need.

To do is how reliable is this data so you are going to spend most of your time on that data if your data itself is not reliable your time will be useless I mean that is the time that you are spending

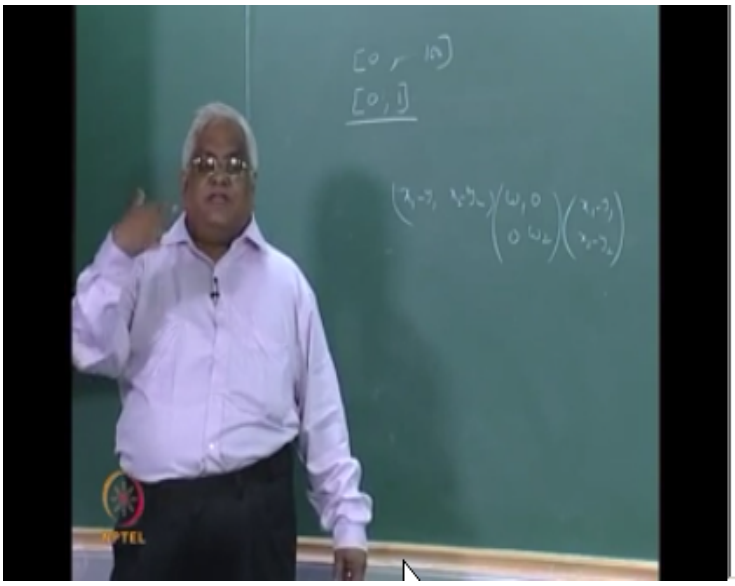
on the dataset will be useless the one who has given you the data if he has given you only for this region whereas you are supposed to get this whole thing then it is not going to be of any use it is not going to be of any use at least you must be able to say that the data is only for this region at least that thing at least you must be able to say even that many of us many times.

We are not in a position to state many of us many times we are not in a position to state that because someone else has gone and collected the data and the data is coming to you and you really do not know professor seer of many of his lectures this is the first that he mentioned this the point that I made just now once someone gives you a data forget about all the analysis and other stuff the first thing that you need to do is how reliable is that data that is something that you need to find out but how that is not known from data to data from problem to problem these things would change there are no general methods for it okay.

There are no general methods for finding how much the data is reliable so once you are satisfied with the reliability then you can go ahead doing all these things training set test set develop the classifier so on and so forth but this is very much important if someone has given you the observations only for this region it is very difficult for you to hope to get results for this whole region this is something that is very much important okay so training set we would like it to span the whole region the whole class that test set also we would like.

It to span the whole of the class this is these are the main two imperatives regarding training and test set now there is one another thing that I need to cover standardization and normalization okay suppose you are given a data set say one particular feature.

(Refer Slide Time: 37:05)



It is taking values say in the interval 0to 100 and there is another feature which is taking values say in the interval 0 to 1 one feature is taking values in the interval 0 to 200 another feature is taking values in the interval0 to 1 suppose you would like to calculate something like a distance.

Now this is 0 to 100 this is 0 to 1 since the span is more if you write something like X I -y I whole square it is likely that this is going to have more impact on the distance than this let me repeat if one figure is taking values in the interval zero to hundred and another figure is taking values in the interval 0 to 1 and this has more variation than this then when you are calculating distance between two such quantities where the first feature is this one then the second feature is this then X I X 1-the corresponding y 1 whole square.

Since the variant the variation is more it may be likely that this feature has more importance are this may be contributing more to the distance than this because here it is less than you would have a question is it really appropriate that this feature is contributing more to the distance than this sort of problem I tackled in one of the previous classes where I was talking about height and weight where I have used some x2 - right - then I have used some weights if you see one of the previous lectures.

If you see one of the previous lectures there that the problem was not about having more variation and the problem was about the unit's one is measured in centimeters another one is measured in cages then how do you get the distance okay there the problem was about the unit's here the problem was about how much variation you have in the variable values of the variable

now sometimes this variation may be important for you so that you would like to keep it sometimes it may not be important for you so that you would like to reduce this variation.

I will give you examples where this variation may be unimportant I will give you an example where the variation may be unimportant suppose you are looking at the classification of persons belonging to let us say hilly regions not eastern region of our country I do not know that you have seen people from that region or not people who are living really on hills their noses have very small length are you understanding what I am trying to say their noses have very small length compared to the people who are living in plains.

So length of the nose is a feature which differentiates between people we people who are living in plain from the hilly regions people by the way how much do you think is the variation there if you measure the length of the nose in centimeters it is very small value right the difference between our lengths and their length it is really very small that you compare it with height and weight there you are going to have big differences height and weight length of the nose will be small small one so they the small difference you would like to give more importance.

Than this one the small difference you would like to give more importance than the bigger one so but if you look at something like PCA principal components there you are looking at larger variances there you are giving more importance to larger variances so I mean it varies from problem to problem it varies from situation to situation how you would look at the variation in variables are the variance of the variables.

So now if you want to somehow make the variations to be same there are generally two methods that people have followed one people have one is called standardization another one is called normalization I think I will stop here.

End of Module 02-Lecture 04

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
Selvam
Sridharan

Studio Assistants
Linuselvan
Krishankumar
A.Saravanan

Additional Post –Production
Kannan Krishnamurthy& 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
Santhosh 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