Indian Institute of Technology Madras
Presents

NPTEL
NATIONAL PROGRAMME ON TECHNOLOGY ENHANCED LEARNING

Pattern Recognition

Module 04

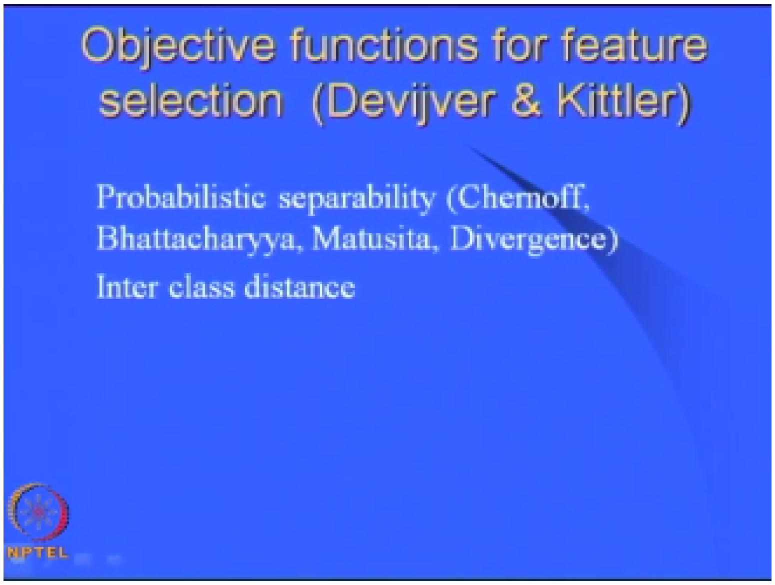
Lecture 06

Features Selection Criteria Function: Interclass Distance Based

Prof. C.A. Murthy
Machine Intelligence Unit.
Indian Statistical Institute. Kolkata

We have been discussing objective functions for feature selection there we earlier discussed the probabilistic supper ability based feature selection functions.

(Refer Slide Time: 00:25)

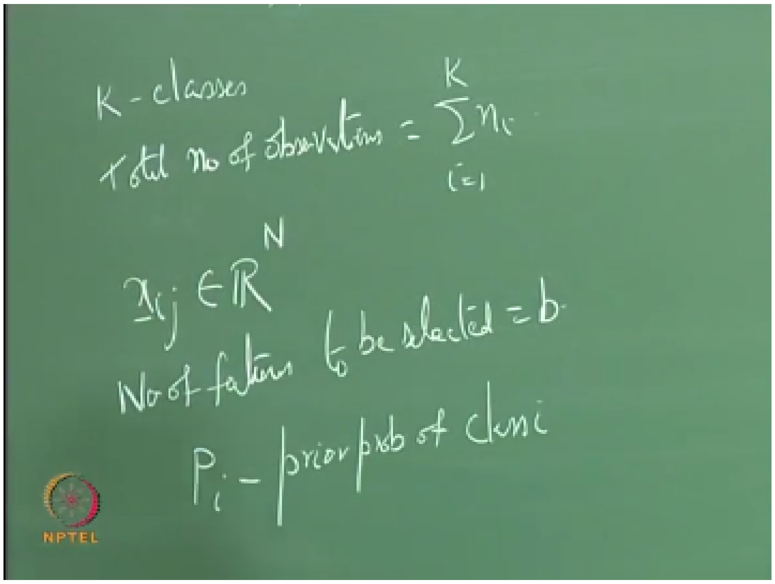


Objective functions for feature selection (Devijver & Kittler)

Probabilistic separability (Chernoff, Bhattacharyya, Matusita, Divergence)
Inter class distance

Now we are going to discuss feature selection functions where we are given a training sample search, so we are not we are going to assume that we do not have any idea on the about the density functions here, so in this setup.

(Refer Slide Time: 00:50)



K - classes

total no of observations = $\sum_{i=1}^{K} n_i$

$x_{ij} \in \mathbb{R}^N$

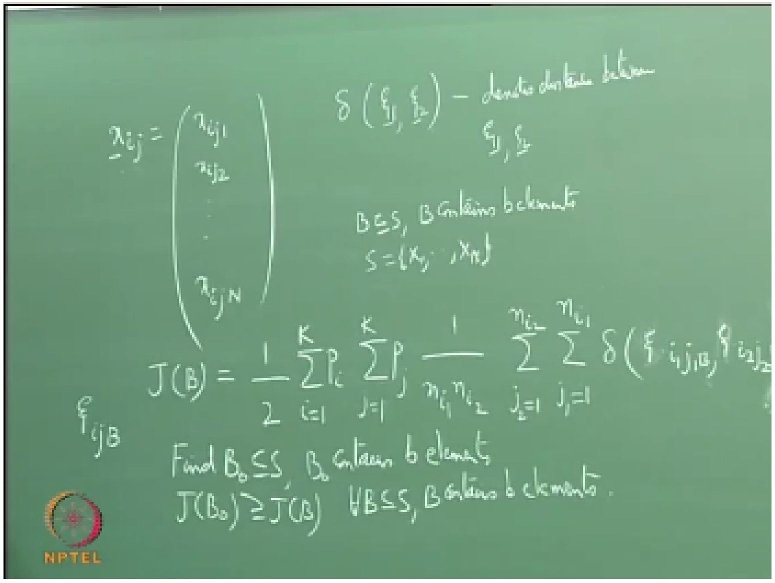
No of features to be selected = b

P_i - prior prob of class i

We have let us just say we have points I am going to represent them as xij is equal to 1, 2 up to ni i = 1 2 up to k that is we have k classes we have K classes and the number of points in the ith class is small ni the number of points in the earth classis small ni xi denotes the j observation in the ith class xij denotes the observation in the ith class right j observation in the hype class j = 1,2 of ni there are ni number of observations in the ith class and I is equal to 1 to K thus the total number of observations total number of observations is equal to summation ni.

I = 12K and each xij belongs to RN, N capital n dimensional space, so the number of features is capital and each xij belongs to RN that is the number of features is N and we are supposed to select small B number of features number of features to be selected to be selected is equal to small be totally we have N number of features we are supposed to select b number of features, so all our observations are in N dimensional space and these observations are represented by xij so each xij.

(Refer Slide Time: 03:43)



$$\underline{x}_{ij} = \begin{pmatrix} x_{ij1} \\ x_{ij2} \\ \vdots \\ x_{ijN} \end{pmatrix}$$

$\delta(\xi_1, \xi_2)$ - denotes distance between ξ_1, ξ_2

$B \subseteq S$, B contains b elements

$S = \{X_1, \cdots, X_N\}$

$$J(B) = \frac{1}{2} \sum_{i=1}^{K} P_i \sum_{j=1}^{K} P_j \frac{1}{n_{i_1} n_{i_2}} \sum_{j_2=1}^{n_{i_2}} \sum_{j_1=1}^{n_{i_1}} \delta(\xi_{i_1 j_1 B}, \xi_{i_2 j_2}$$

ξ_{ijB}

Find $B_0 \subseteq S$, B_0 contains b elements

$J(B_0) \geq J(B)$ $\forall B \subseteq S$, B contains b elements.

It is a column vector xij xij1 xij2 xij N are the notations clear are the notations clear each xij is a N dimensional vector it is represented as xij1 xij2 xijn totally there are N number of features number of classes is K and the number of observations in the ith class is small ni and total number of observations is this okay, now so now the question is somehow we have to select small B features, so the intuition here is similar to the inclusion in the probabilistic separability based measures.

There what did we do we selected features in such a way that they maximize the distance between the density functions of the respective classes here, we do not have density functions but we have some point, so we need to select features in such a way that they maximize the distance between those points it is clear we need to select features in such a way that they maximize the distance between those points now how do we put it mathematically what is a mathematical formulation of this.

So if I write δ let us say δ denotes distance between say a vector ξ1 and a vectors ξ2 this denotes distance between is ξ1 and ξ 2 so δ denotes say distance between two vectors, now using this δ how do we define distance between the classes inter-class distance, so what may be a way of defining distance here let me first assume that you have the same number of I mean the let me first define this distance in the N dimensional space, if I define distance here than for any subset of this N dimensions.

Then we can define distance correspondingly there okay, so how do I define distance now here I am writing a big formula first what I do is that suppose x_{ij1j1} x_{i2j2} I take distance between x_{ij1j1} x_{i2j2} okay and what I do is that I just do the ∑ j1 = 1 2 how much ni1 j2 = 1 2 ni2 then this will provide distance between the points in the class i1 and the class i2 write then I take average then I take average 1 /ni1 into ni2 write and suppose I have prior probabilities let us just say PI denotes prior probability of class I suppose this prior probabilities are given to us then what I will do is that I multiplied by the corresponding prior probabilities and I take the ∑ and I divide it by 2 why do I divide it by 2 the answer is simple.

The same I is appearing here and here the same class so you are measuring the same thing twice so I am dividing it by 2, so if you take some number of figures some be number of features then the corresponding vectors, if I represent it by say xij is this suppose I write ξij of a set B that means B is a set having b number of elements okay, B is a set having v number of element B is a subset of s and B contains b element and our s is x1 to xn and n features is having n number of features.

And we are taking a subset of B of and B contains small be elements with respect to those be elements you have a corresponding this b dimensional vector that I am writing it as ξ ijb is this clear to you, if it is not clear you ask me you have been hearing my lecture since morning and even other days also B is a subset containing small be elements, so corresponding to those small V elements the vector is going to be of b dimensional vector, so correspondingly for let us just say this b let us just a b= 2.

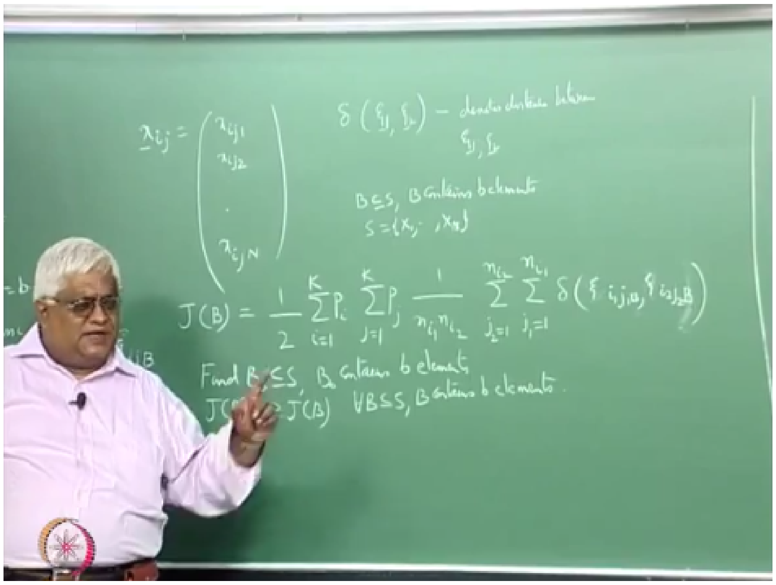
Let us just say and B is say the feature one and feature to x1 and x2 then what will be the ξ jb then this ξijb actually it will be xij value 1 xij2 get to just only these two points this corresponds to the vector containing be elements B of containing b elements corresponding to this B set for xij when you remove all the other features from xij the resultant vector is going to be ξij okay from xij we are getting this ξij corresponding to those B set then this one corresponding to those B set this will be ξijb and this will be $ξ_{i1j2B}$ and this will be jB.

From xij corresponding to those B set the resultant vector I am representing it as ξij corresponding to this set B, so naturally for i1 j1 you are going to have ξ1 j1 be 4i to date we are going to have say I to J to be, so do this thing when you are going to get jb now we want to maximize it we want to do the maximization, so well find V_0 subset of S B_0 contains b number of elements and J of B_0 is you want to do the maximization right is greater than or equal to j (B) for all be subset of S and B contains b elements.

The intuition is similar for probabilistic separability and as well as for inter-class distance-based listing we will select those features which maximize the separation, okay now there is one small point I had yeah there is one small point I have not told you what this δ is I only said that δ measures distance, but in what way it is going to measure the distance we can have several distance functions we can have several distance functions and using those distance functions you can get four different distance functions.

You are going to get different criterion functions for different distance functions you are going to get different criterion functions I shall be talking a little bit about distances in my I mean one of my next lectures, okay here what I will be doing is that I will take some particular form of the distance function, so what form shall I shall I take we take the usual form what is the usual one.

(Refer Slide Time: 16:39)



$\underline{x}_{ij} = \begin{pmatrix} x_{ij1} \\ x_{ij2} \\ \vdots \\ x_{ijN} \end{pmatrix}$

$\delta(\xi_1, \xi_2)$ – denotes distance between ξ_1, ξ_2

$B \subseteq S$, B contains b elements

$S = \{X_1, \cdots, X_N\}$

$$J(B) = \frac{1}{2} \sum_{i=1}^{K} P_i \sum_{j=1}^{K} P_j \frac{1}{n_{i_1} n_{i_2}} \sum_{j_2=1}^{n_{i_2}} \sum_{j_1=1}^{n_{i_1}} \delta(\xi_{i_1 j_1 B}, \xi_{i_2 j_2 B})$$

Find $B_0 \subseteq S$, B_0 contains b elements

$J(B_0) \geq J(B)$ $\forall B \subseteq S$, B contains b elements.

δ ξ 1 ξ 2 is equal to that is the square of the Euclidean distance is it correct it is the square of the Euclidean distance, now if you take the square of the Euclidean distance here for this jb now it gets if you take that the whole expression is going to get simplified the whole expression is going to get simplified can you tell me how it may get simplified we have any idea if possible I would like you to do this calculations in your home and maybe you can show the things to me tomorrow but I want you to do it.

I want you to do it now let us just see how it may get simplified I mean I will write down the final form I will write down the final form the final form is going to be like this let me write ξi bar as 1 over ni $\sum$j = 1 2ni ξ ij all these things are with respect to the set B all these are with respect to the set B okay, this is the mean of the IH class of the training set mean of the i^{th} class of the training set ξ ij is equal to 1 2 ni x_{ij} with respect to the set B this is ni 1by ni o okay, now let me write μ as $\sum$i = 1 2k pi in all these are with respect to the set B with respect to the set B.

Let me write μ as this μ is the overall me μ is the overall mean and this is the mean of the highest class on the basis of that sample set training sample set, now j(b) will be take the distance between $\xi_{ijB} - \xi$ I bar B this one over ni and there is a pi and there is $\sum$ i = 1 to k + $\sum$pi this is going to be this jB this is what you are going to get if you substitute in this place if you substitute in this place the δ as this corresponding to the set B and if you simplify this whole expression this is what you are going to get.

I want you to do this cloud to do the calculations on your own and you will get it now can you tell me what are these expressions let us look at this here what we are doing you are going to find the distance of a point with its class me right you are going to find the distance of a point to its class mean you are going to find all those distances and you are taking their average it is basically with in class distance this is basically with in class distance, now what about this is we take the mean of the i^{th} class and that we are taking the distance between that and the overall mean that will be what it will be between class distance.

It will be between class distance so we are actually trying to maximize the sum of the width in class distance and the between class distance when this expression was written intuitively okay, this was written just intuitively and then when we use the usual Euclidean distance as δ then this expression has boiled down to this form, where this is nothing but the sum of the width in class distance and between class distance so we are trying to maximize if you use this expression then we are trying to maximize basically.

The sum of that with in class and between class distances so probably we can suggest slightly better I should say criterion function what is that, we do not need to maximize the sum probably we would like to maximize the second part between class distance we would like to maximize and with in class distance we may want to minimize it okay, within class distance we may want to minimize it and between class distance may I want to maximize it so how does one do it we can take this divided by this.

And you maximize we can take this divided by this and try to maximize the whole thing it is clear to you because on one hand we want to maximize this and on the other hand we want to minimize this, so a slightly better criterion function would be take this by this and maximize the whole or this by this minimize the whole whichever way you take, so we can have another criterion function where we take this divided by this and we maximize so another criterion function will be we take the between class distance and divide it by with in class distance and you just maximize that.

So using this specific I should say intuition there are a few more such criterion functions where they take the width in class scatter matrix which is represented by SW between class scatter matrix which is represented by SB and then they calculate SW inverse SB okay they calculate SW inverse SB and using this SW inverses be people have selected features right, and the basic formulation of this has w inverses be again from this when you go step by step you are going to come to this SW inverses be anyone can directly talk about it.

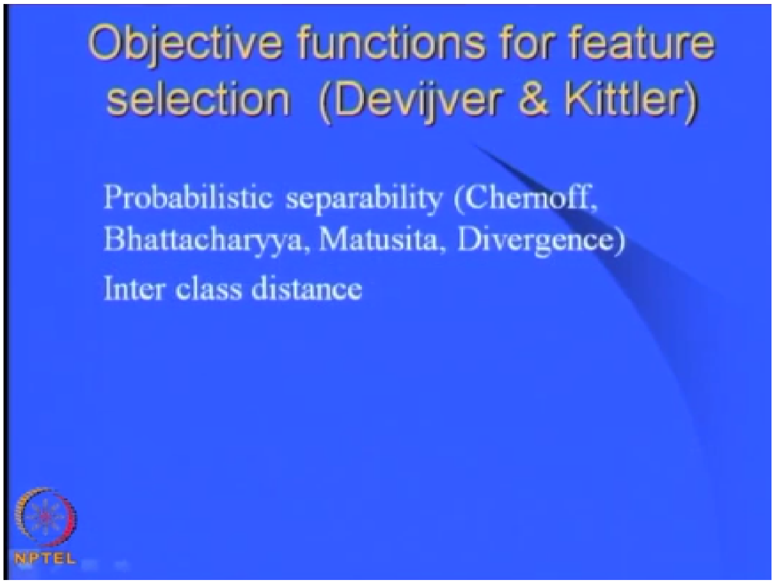
You take SW which is the width in class scatter matrix and you take SB which is the between class scatter matrix you take SW inverse SB okay and you use this SW inverse B to obtain features use the SW inverse SB to obtain features you can use the trace of the matrix do you stress you can use Eigen values Eigen vectors okay, but you can use them to obtain features about this particular aspect about SW inverse and SB I think Dr. Susan Dass will teach you these things okay that is why I am not going to the details about that.

But my part I will just about the intra class distance thing I am just stopping it here where this is the basic formulation from here you can go to that and then from there you can develop very many criterion functions, which is one of them is something based on SW inverses be okay that one is based on SW inverses B and if you go through what is known as, if you go through some of the things regarding multivariate analysis in multivariate analysis where that is in statistics where they day they deal with these matrices extensively okay.

You will be finding many ways of defining the sort of criterion functions will be finding if some more ways of defining this sort of criterion functions because, it basically statistics okay this part is basically statistics and you will be finding quite many books on this part both in statistics literature and also in pattern recognition of literature there are some formulas relating to their has something called canonical correlations, and there are also some other topics you know if you go through the multivariate analysis books.

Basically discriminant analysis there you are going to get in statistics you are going to get the different formulations of this way of selecting features, but they do not use the word feature selection but they try to tell you all these formulations there okay, so at this portion I am ending this way of taking features that is inter class distance-based feature selection, but I will be also talking something more about.

(Refer Slide Time: 32:06)



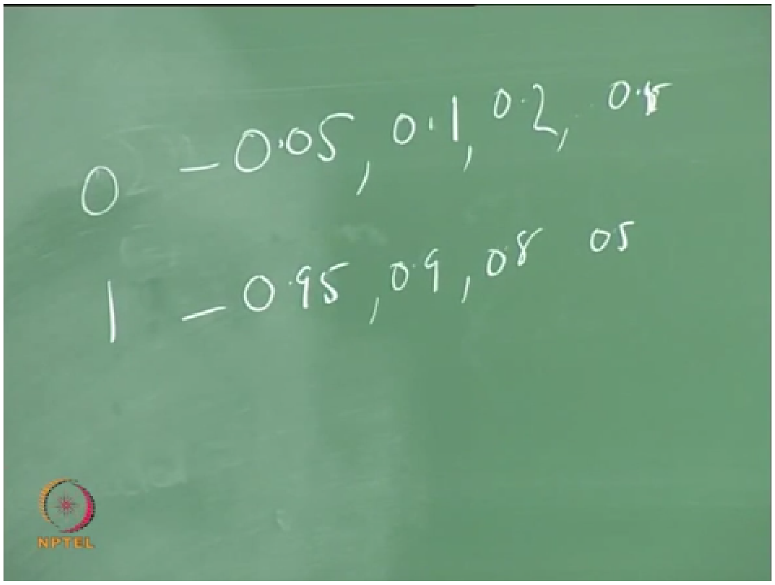
Objective functions for feature selection (Devijver & Kittler)

Probabilistic separability (Chernoff, Bhattacharyya, Matusita, Divergence)
Inter class distance

Let me just see do you know what entropy means let me talk a little bit about that entropy if the basic concept has come from physics probably you might be knowing the people who have physics background they will be knowing some laws regarding thermodynamics where seemingly the second law of thermodynamics, it deals with entropy where entropy is maximized means you have more disordered entropy less is small means you have more order entropy more means you have more disorder okay.

That is the basic I should say feeling or whatever it is about entropy but the concept of entropy that is used in our fields the main feeling has come from physics what is the meaning of how do one say that there is more disorder in the system, how does one say that there is more disorder in the system let us just say my system it has only two values.

(Refer Slide Time: 33:47)



0 – 0.05, 0.1, 0.2, 0.5

1 – 0.95, 0.9, 0.8 0.5

0 and 1 suppose 0 occurs with probability that is this a 0.5 and one occurs with 0.95 probability and you have second case 0 occurs with a 0.1 probability and say this occurs with a 0.9 probability and say this is 0.2 and this is 0.8, so somewhere you are going to get 0.5 and 0.5 now in this situation we have only two values that the variable is taking where do you think you are going to have the maximum disorder I think it will be at this place 0.5 and 0.5 in a case like this 0.95 0.05 you know that most of the times one is going to occur.

You can say that but in such a case like this you cannot say it so you have maximum disorder at the place where 0 and 1 they are occurring with equal probability okay, similarly if you have three states then each of them occurs with say probability 1/3 there you have the maximum disorder in fact the more and more number of states the disorder is going to increase and for the same number of states, if you have unequal probabilities then disorder is less equal probabilities disorder is more are you understanding.

If the more and more number of states means you have more disorder and for the same number of states, if you have different probabilities of occurrence then your disorder is less the same probability of occurrence for each of the states means the disorder is more okay, so this thing is represented mathematically by this formula suppose in a system you have n number of states and I i^{th} occurs with probability pi that means ∑pi1 to n = 1 ith state occurs with probability pi then this is said to be the entropy of the whole system.

This is the entropy of the whole system - ∑ pi log pa why the - sign is there since all the pi are lying between 0 and 1 log of that is going to be negative logarithm of that is going to be negative right log of any value between 0 and 1 is negative, so minus will make it positive minus will make it positive now one can also say this thing as the information contained in the system, now why the word information is used if you have more of an uncertainty then you have to find more about the system.

That means the information contained in the system is more that is the basic feeling that is why people also use the term information people also use the term information I think let me just give you an example suppose say today's weather in Chennai there are let us just say there are four possible states one is raining one is hot other one is say normal raining hot and what is the fourth one fourth one is cold normal weather means you have normal temperature, okay pleasant temperature let us just say hot means well it is hot okay.

And rarely means it is raining today cold means it is cold well generally we all know that in Chennai it is generally not cold okay, but today it is slightly I mean to the colder side I am I correct if you go outside it is slightly towards the colder side, so there are four possible states to the system normal weather hot cold rain, now between these four possibilities what is the usual thing that occurs here I think probably usually it is hot I do not know you people must be able to tell me tell this thing more than me usually probably it is hot.

The normal weather that is the press the pleasant weather probably it occurs only a few weeks rainy season probably that also occurs a few weeks visitor that is cold probably that also occurs have very few weeks most of the time I suppose it is hot that is why whenever people talk about

Chennai is hot, so they have this information because they know these probabilities the probability of something being hot is more here, so when people come to visit Chennai and other places like me and my family which is going to visit we are not going to we are not carrying any winter clothing.

Are you understanding so this is the consequence of these things because there is less uncertainty here entropy is less on the other hand, if you look at something like say some other places where these things are occurring probably equally likely are maybe then the entropy is may are the entropy is more than one has to be prepared for all the eventualities if you are going to that place and you do not know whether it is going to be hot or cold on one hand you will have to take the cold garments for winter.

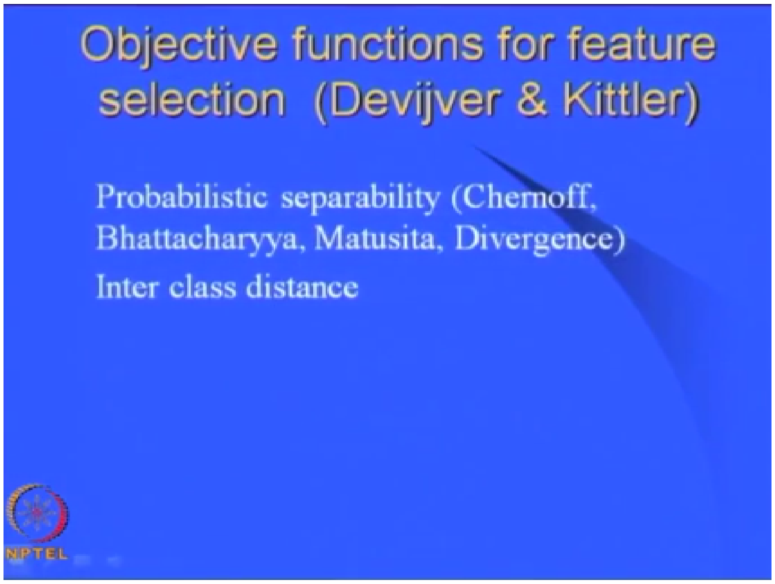
And on the other hand if it is hot I mean you must be prepared to bear the heat then the more uncertainty you have the move the things you are going to carry there are you understanding, so this is one function that tells you how much uncertain or how much entropy is present or how much information is presented this is one function that does it there are also a few more such functions.

But this is the one that is used a lot now there are some criterion functions which take which take this information theoretic measure this one and they define criterion functions based on this there are some functions there are some criterion functions which define they are created there are some criterion functions which use this definition to derive the form of the criterion functions, this is one thing that I wanted to tell you and though I dealt with only probabilistic separability inter-class distance based measures in more detail.

What I am trying to tell you that is that these are not the only ways in which you can define the criterion function you can also use information theoretic way of defining a criterion functions and in fact there are many other ways in which you can define criterion functions there are many ways based on fuzzy set theory, there are also some other ways in which you can use a mixture of these things you can use a mixture of a few of these things and in fact if you look at the literature on feature selection.

This is one of the most I mean this is an area on which you will find just too many papers quite many papers and you can come across several different criterion functions they are just too many of them they are just too many of them I have dealt with only a very few a very small number of criterion functions this is not the end this is only the beginning okay, there are many other criterion functions existing in the literature now I will show you a few of my slides.

(Refer Slide Time: 44:54)

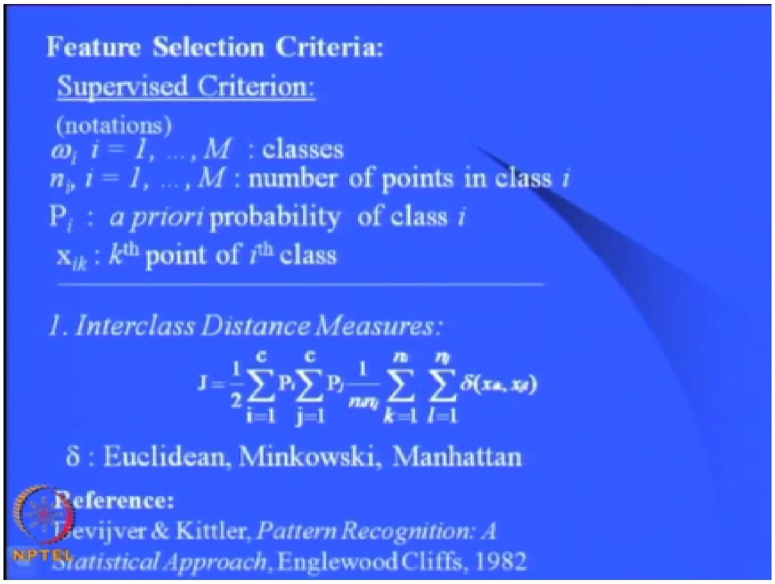


Objective functions for feature selection (Devijver & Kittler)

Probabilistic separability (Chernoff, Bhattacharyya, Matusita, Divergence)

Inter class distance

(Refer Slide Time: 44:55)



Feature Selection Criteria:

Supervised Criterion:

(notations)

ω_i $i = 1, \ldots, M$: classes

n_i, $i = 1, \ldots, M$: number of points in class i

P_i : *a priori* probability of class i

x_{ik} : k^{th} point of i^{th} class

1. Interclass Distance Measures:

$$J = \frac{1}{2}\sum_{i=1}^{c} P_i \sum_{j=1}^{c} P_j \frac{1}{n_i n_j} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \delta(x_{ik}, x_{jl})$$

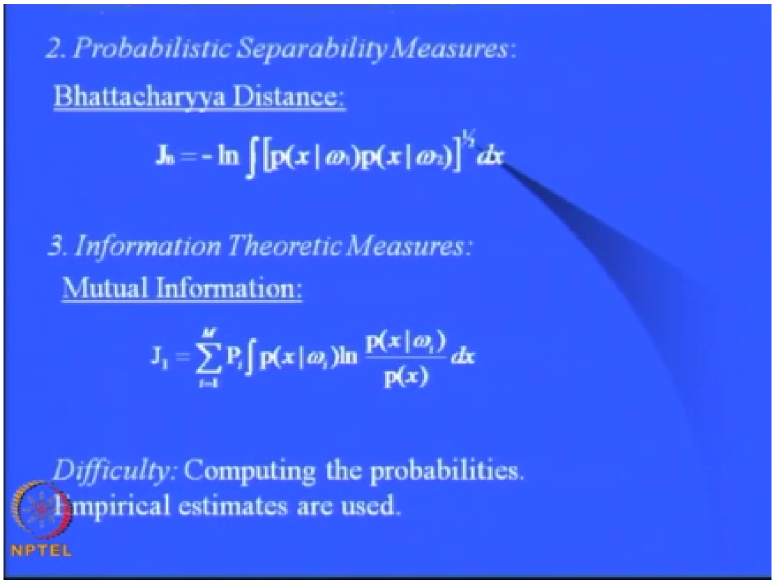
δ : Euclidean, Minkowski, Manhattan

Reference:

Devijver & Kittler, *Pattern Recognition: A Statistical Approach*, Englewood Cliffs, 1982

This intra class distance measures this one is the one that I was telling you today.

(Refer Slide Time: 54:08)



2. Probabilistic Separability Measures:

Bhattacharyya Distance:

$$J_B = -\ln \int \left[p(x|\omega_1) p(x|\omega_2) \right]^{1/2} dx$$

3. Information Theoretic Measures:

Mutual Information:

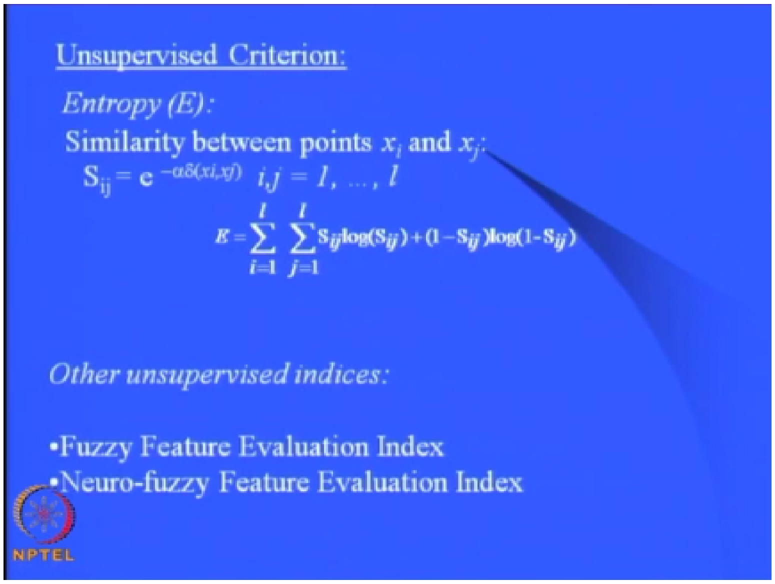
$$J_1 = \sum_{i=1}^{M} P_i \int p(x|\omega_i) \ln \frac{p(x|\omega_i)}{p(x)} dx$$

Difficulty: Computing the probabilities.

Empirical estimates are used.

Bhattacharya distance I already mentioned this is a mutual information this is something that I was telling you earlier I mean this one I did not discuss but this one is also existing this is okay.

(Refer Slide Time: 45:26)



Unsupervised Criterion:

Entropy (E):

Similarity between points x_i and x_j:

$S_{ij} = e^{-\alpha\delta(xi,xj)}$ $i,j = 1, \dots, l$

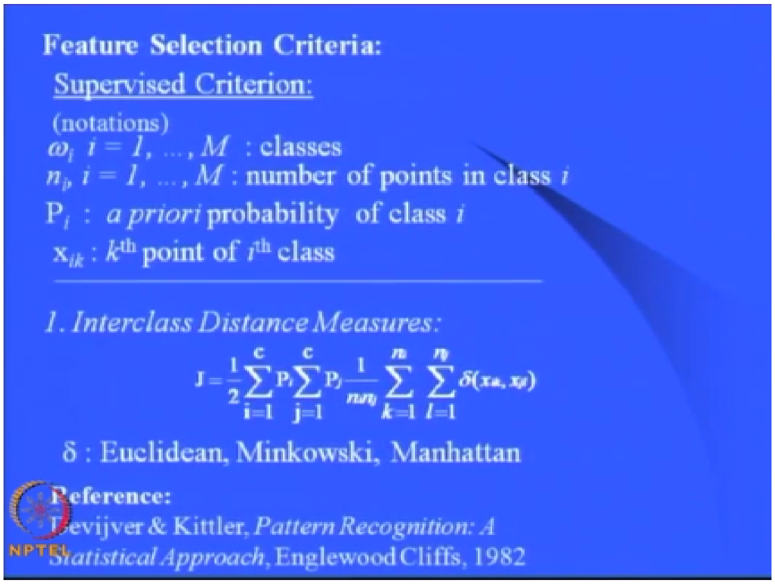
$$E = \sum_{i=1}^{l} \sum_{j=1}^{l} S_{ij}\log(S_{ij}) + (1 - S_{ij})\log(1 - S_{ij})$$

Other unsupervised indices:

•Fuzzy Feature Evaluation Index
•Neuro-fuzzy Feature Evaluation Index

And there are yes this is entropy based one this is entropy based one these are none supervise the criterion and there are some fuzzy feature evaluation indices there are also Neuro-fuzzy feature evaluation indices.

(Refer Slide Time: 45:52)



Feature Selection Criteria:

Supervised Criterion:

(notations)

ω_i $i = 1, \ldots, M$: classes

n_i, $i = 1, \ldots, M$: number of points in class i

P_i : *a priori* probability of class i

x_{ik} : k^{th} point of i^{th} class

1. Interclass Distance Measures:

$$J = \frac{1}{2}\sum_{i=1}^{c} P_i \sum_{j=1}^{c} P_j \frac{1}{n_i n_j}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j}\delta(x_{ik}, x_{jl})$$

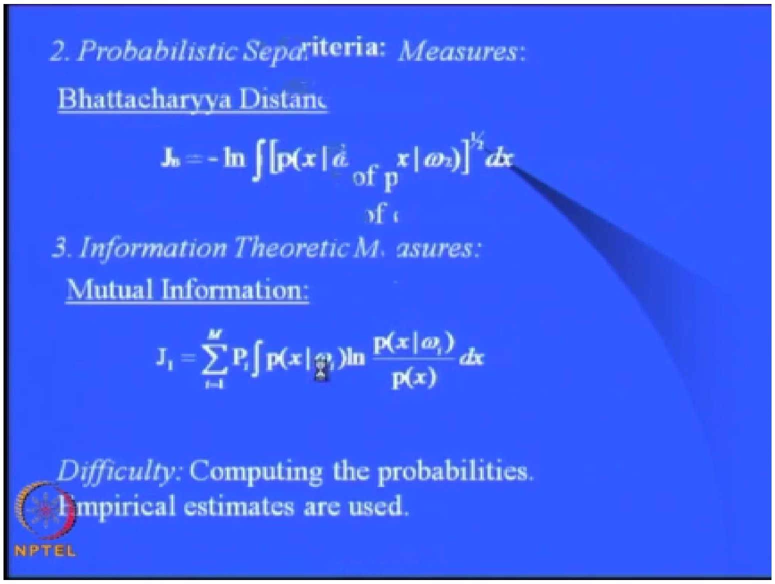
δ : Euclidean, Minkowski, Manhattan

Reference:

Devijver & Kittler, *Pattern Recognition: A Statistical Approach*, Englewood Cliffs, 1982

Yes this one is the intra class distance-based one which I discussed.

(Refer Slide Time: 45:58)



*2. Probabilistic Sepa*riteria: *Measures:*

Bhattacharyya Distanc

$$J_B = -\ln \int [p(x|\omega_1)\, p(x|\omega_2)]^{1/2} dx$$

of p

of (

3. Information Theoretic M asures:

Mutual Information:

$$J_I = \sum_{i=1}^{M} P_i \int p(x|\omega_i) \ln \frac{p(x|\omega_i)}{p(x)} dx$$

Difficulty: Computing the probabilities. Empirical estimates are used.

Bhattacharya one I discussed the third one the mutual information this I did not discuss similarity entropy based on this also I did not discuss I discussed a little bit about the just the formulation part of it and there are many others like fuzzy feature evaluation in this sense Neuro- fuzzy feature evaluation index in fact I just wrote a few of them there are simply too many other such measures they are simply too many other such measures with this I will stop okay.

End of
Module 04 – Lecture 06

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.iitm.ac.in

Copyrights Reserved