

Computer Vision and Image Processing – Fundamentals and Applications
Professor Doctor M. K. Bhuyan
Department of Electronics and Electrical Engineering
Indian Institute of Technology, Guwahati, India
Lecture – 32
Introduction to Machine Learning

Welcome to NPTEL MOOCs course on Computer Vision and Image Processing- Fundamentals and Applications. I have been discussing about the concept of Machine Learning. Today I am going to continue the same class that is the concept of Machine Learning. So, first I will discuss the concept of Regression and after this I will discuss the concept of Bayesian Decision Theory.

(Refer Slide Time: 01:01)

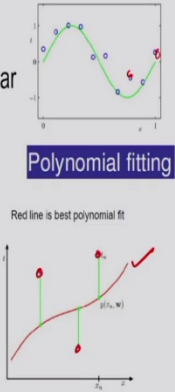
Linear Regression

Let us now consider the simplest case, i.e., simple linear regression where we are having only one dependent and one independent variable.

Given a set of points in $x - y$ plane $(x_i, y_i); i = 1, 2, \dots, n$, the linear regression attempts to find a line in 2D which best fits the points.

$$\hat{y}_i = ax_i + b$$

The error between the actual point and the line can be written as:

$$e_i = y_i - \hat{y}_i$$


Polynomial fitting

Polynomial fitting

So, what is the fundamental concept of Regression I will explain now. Linear regression is a statistical method that allows us to model the relationship between the scalar response that is the dependent variable and one or more independent variables. So, that means, I want to find the relationship between a scalar variable that is the scalar response, dependent variable and one or more independent variables.

This is done by fitting a linear equation to the observed data, the dependent variable I can say as a response or outcome and the independent variable is called as Predictor or maybe the Regressor. Suppose, if we have only one independent variable, this is called a Simple Linear

Regression. And suppose, if I consider two or more independent variables, then in this case it is the Multiple Regression.

So, this method looks for the statistical relationship between the Dependent Variable and the Independent Variable. For example, given a temperature in degree Celsius, we can find the exact value of Fahrenheit. Let us now consider the simplest case that is the Simple Linear Regression. So, where we are having only one dependent and one independent variable and simple linear regression boils down to the problem of line fitting on a 2D x-y plane.

So, suppose the given a set of points in x-y plane, x and y coordinates I am considering the linear regression attempts to find a line in 2D which is best fits the points. So, there is a concept of the linear regression. The most popular method of fitting a line is the method of least square. So, we can consider that method that is the method of Least squares, as the name suggests, at this method minimizes the sum of the squares of vertical distances from each data point to the line.

Now, the question is how to find a baseline and that is the objective of regression. Suppose, that the slope and the intercept of the required line are a and b. So, I am considering the slope is a and the b is the intercept of the line, then the equation of the line will be $y_i = ax_i + b$ that is the equation of the line. And the error between the actual point and the line it can be determined. So, I am determining the error in this figure these two figures I am showing the concept of the Polynomial fitting.

So, you can see observed data, you can see the sample points the observed data you can see, and I am fitting a curve between these observed sample points observed data and same concept I am showing in a second figure also that is the Polynomial fitting. So, mainly we want to reduce the error. So, these are my observed data you can, you can see and this is the curve, I am fitting between the observed data points. So, the objective is to minimize the error.

So, that is the objective of regression. So, you can see that I can determine the error like this $e_i = y_i - \hat{y}_i$ that I can determine.

(Refer Slide Time: 04:29)

Hence, the average error E can be computed by the following equation as:

$$E = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$
$$= \frac{1}{n} \sum_{i=1}^n (y_i - (ax_i + b))^2$$

To find the required a and b , the partial derivatives of E with respect to a and b are made equal to zero.

$$\frac{\partial E}{\partial a} = 0$$

$$\frac{\partial E}{\partial b} = 0$$

And the concept of regression is I have to minimize the error. So, you can see I can compute the error like this $E = \frac{1}{n}$. So, n number of sample points suppose, $y_i - \hat{y}_i^2$ that I am considering the average error I am determining and this is the expression for E . So, the objective is to find the slope a and the intercept b which gives minimum error so, already I have defined a line and we have the slope the slope is a and the intercept is b .

So, objective is to we have to minimize the error and corresponding to this we have to determine the slope a and the intercept b . So, to find the required values of a and b , we have to consider the partial derivative of E with respect to a and b . So, that is the partial derivative of E with respect to a that should be equal to 0 and partial derivative of E with respect to b that should be equal to 0. So, objective is to find a slope a and intercept b which gives minimum error E that is the objective.

(Refer Slide Time: 05:45)

$$\begin{aligned} \frac{\partial E}{\partial a} = 0 & \Rightarrow -\frac{2}{n} \sum_{i=1}^n (y_i - ax_i - b)x_i = 0 \\ & \Rightarrow \sum_i x_i y_i - a \sum_i x_i^2 - b \sum_i x_i = 0 \\ & a \sum_i x_i + b \sum_i x_i = \sum_i y_i = \sum_i x_i y_i \quad \checkmark \text{---(1)} \end{aligned}$$
$$\begin{aligned} \frac{\partial E}{\partial b} = 0 & \Rightarrow a \sum_i x_i + b \sum_i x_i = \sum_i y_i \quad \checkmark \text{---(2)} \end{aligned}$$

And after this you can see if I consider this equation that is the $\frac{\delta E}{\delta a} = 0$. So, based on this I can get this equation, you can do the differentiation after doing the differentiation I will be getting this 1 and similarly, if I do that $\frac{\delta E}{\delta b} = 0$ then I will be getting this equation.

So, I will be getting these two equations and this equation this equation if I consider suppose equation number 1 and equation number 2, so, I am getting two equations. So, this equation 1 and 2 are linear equations in two variables. Hence, they can be easily solved to find the values of a and b now, a means the slope and intercept is b. So, by solving these two equations, I can determine the value of a and b, a is the slope of the line and b is the intercept that I can determine.

So, in this case, I have shown only the simple case of Linear Regression. So, if you want to see the Polynomial fittings, so, you have to see the books, so, how to go for polynomial fittings between the observed data points that you have to see. So, in my discussion only I have considered the simplest regression model in which I have only 1 dependent variable and 1 independent variable. So, this is the fundamental concept of regression after this I will consider,

(Refer Slide Time: 07:16)

The slide features a dark blue header with the text "Bayes Theorem" in white. Below the header, there is a list item "• Bayes Theorem :". The first formula is $P(\omega_j | x) = \frac{p(x | \omega_j)P(\omega_j)}{P(x)}$, with red checkmarks above the likelihood and prior terms and a red underline under the posterior term. The second formula is $\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$, with red underlines under "posterior" and "evidence". Below this is the text "In the case of two categories". The final formula is $P(x) = \sum_{j=1}^{j=2} P(x | \omega_j)P(\omega_j)$, with a red underline under the entire expression.

The Bayesian decision theory. So, already I have explained the concept of the Bayes Theorem. So, here you can see this posterior density you can see these the posterior density is equal to likelihood into prior divided by evidence and if I considered two classes, then the $P(x)$ will be like this. And in this the Bayesian decision theory, this $P(x)$ is the normalizing factor. So, it has no role in classification, so, that we can neglect. So, $P(x)$ is the normalizing factor that is the evidence.

So, it has no role in classification. So, only we have to consider the likelihood and the prior. So, for a particular Feature vector, the Feature vector is x , we have to determine the class. So, the probability of obtaining a particular class given the Feature vector, the Feature vector is x . So, that we have to determine that is the objective of Bayes Decision Making.

(Refer Slide Time: 08:20)

Decision based on posterior probabilities

- Decision given the posterior probabilities

x is an observation for which:

if $P(\omega_1|x) > P(\omega_2|x)$ True state of nature = ω_1 ✓
 if $P(\omega_1|x) < P(\omega_2|x)$ True state of nature = ω_2 ✓

Therefore: whenever we observe a particular x , the probability of error is :

$P(\text{error}|x) = P(\omega_1|x)$ if we decide ω_2
 $P(\text{error}|x) = P(\omega_2|x)$ if we decide ω_1

And you can see the Decision theory will be like this. So, x is the Feature vector suppose, so, if I consider two classes that classes suppose w_1 and w_2 two classes I am considering. If the $P(w_1|x) > P(w_2|x)$, then in this case, I have to consider the class that class is w_1 and similarly, if I consider the $P(w_1|x) < P(w_2|x)$, then in this case the corresponding class will be w_2 .

So, based on this principle I can do the classification. Suppose, if I have suppose this condition, the probability 1, $P(w_1|x) = P(w_2|x)$. Then in this case, we have to see the prior probabilities. Suppose, this is a condition the $P(w_1|x) = P(w_2|x)$, then in this case, we have to see the probability there is a prior probability we have to see $P(w_1)$ and the $P(w_2)$ we have to see and based on this we can take a classification decision.

So, how to do the classification now? So, you can see, suppose, I have the Feature vector, the Feature vector is x suppose, that is the input and I can determine this values $P(w_1)P(x|w_1)$, $P(w_2)P(x|w_2)$, $P(w_c)P(x|w_c)$. So, I am determining this and I will be getting $P(w_1|x)$, I will be getting $P(w_2|x)$, $P(w_c|x)$, I will be getting. I have to pick the largest one.

So, out of these, what I have to do pick the largest. So, out of this I have to pick the largest. So, based on this I can do the classification. So, my input is x so, I can draw the x here suppose, so,

my input is x and I have to determine these values the probability of $P(w_1 \vee x)$, the probability of $P(w_2 \vee x)$ like this I have to determine and out of this I have to pick the largest and based on this I can do the classification decision.

And also I can determine the probability of error for a given Feature vector. So, the what is the probability of error? $P(error \vee x) = P(w_1 \vee x)$ that is, if we decide the class, the class is w_2 . So, if I decide the class w_2 then that is the $P(error \vee x) = P(w_1 \vee x)$. And the $P(error \vee x) = P(w_2 \vee x)$, if we decide the class the class is w_1 . So, like this I can define the probability of error.

(Refer Slide Time: 12:22)

Decision based on posterior probabilities

- Minimizing the probability of error

Decide ω_1 if $P(\omega_1 | x) > P(\omega_2 | x)$; otherwise
decide ω_2

Therefore:

$$P(error | x) = \min [P(\omega_1 | x), P(\omega_2 | x)]$$

(Bayes decision)

So, we have to minimize the probability of error. So, how to minimize this decide the class w_1 if $P(w_1 \vee x) > P(w_2 \vee x)$ otherwise, I have to decide the class the class is w_2 . So, we have to minimize the probability of error and based on this principle, I can minimize the probability of error I have to minimize the error. So, I have to select the appropriate class, that class you can select by these probabilities the $P(w_1 \text{ given } x)$ and $P(w_2 \text{ given } x)$.

(Refer Slide Time: 12:58)

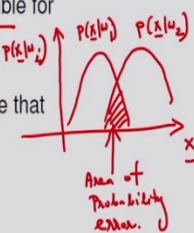
Probability of error

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}, x) dx = \int_{-\infty}^{\infty} P(\text{error} | x) p(x) dx$$

- We want $P(\text{error} | x)$ to be as small as possible for every value of x

The Bayes classifier scheme strives to achieve that

$$P(\text{error}) = \begin{cases} P(u_1 | x) & \text{if } x \in u_2 \\ P(u_2 | x) & \text{if } x \in u_1 \end{cases}$$



And you can see the average probability of error I can determine like this. So, for all the feature vectors, I can determine the average probability of error and we have to minimize this error. the probability of error we have to minimize. So, how to minimize this error. So, this probability of it will be $P(w_1|x)$ if x is assigned to this particular class. This is the Probability of Error. The probability of error is equal to probability of $P(w_1|x)$ if x is assigned to this particular class is equal to probability of $P(w_2|x)$ if x is assigned to that class, that class is w_1 .

So, we have to minimize this probability of error. So, we want probability of error given x to be as small as possible for every value of x . So, for all the feature vectors, we have to minimize this error, I can show here. So, suppose this is my x , x is a feature vector and I am considering the probability of $P(x|w_i)$. I am considering. So, this is suppose the probability of $P(x|w_1)$ or maybe I can consider suppose two classes if I consider it will be w_1 and suppose this is the probability of $P(x|w_2)$. So, this portion if you see, this is the area of, area of probability error.

This is the area of Probability Error. So, this is the concept of the Probability of Error. So, based on the probability of error, we can take a classification decision.

(Refer Slide Time: 15:15)

Bayesian classification framework for high dimensional features and more classes

Let $\{\omega_1, \omega_2, \dots, \omega_C\}$ be the set of "C" states of nature (or "categories" / "classes")

Assume, that for an unknown pattern, a d dimensional feature vector \mathbf{x} is constructed:

From Bayes rule

$$P(\omega_j | \mathbf{x}) = \frac{p(\mathbf{x} | \omega_j) P(\omega_j)}{P(\mathbf{x})}$$

We compute the posterior probability of the pattern with respect to each of the "c" classes.

In the decision making step, we assign the pattern to the class for which the posterior probability is greatest.

And suppose I have C number of classes. So, I am considering C number of classes w_1, w_2, \dots, w_C and assume that we have the d dimensional feature vector. So, I am considering the d dimensional feature vector I am considering and if you consider this is the Bayes rule, $P(w_j | \mathbf{x})$, \mathbf{x} is the feature vector that is a d dimensional feature vector is equal to $P(\mathbf{x} | w_j)$ that is the likelihood into $P(w_j)$ that is the prior probability divided by $P(\mathbf{x})$ that is the evidence.

So, we have to compute the posterior probability corresponding to the feature vector, the feature vector \mathbf{x} and for a decision making what do we have to do so, we assigned a pattern to the class for which the posterior probability is the greatest, that we can determine.

(Refer Slide Time: 16:12)

Bayesian classification framework for high dimensional features and more classes

$$\omega_{test} = \arg \max_j P(\omega_j | \mathbf{x}) = \frac{p(\mathbf{x} | \omega_j)P(\omega_j)}{P(\mathbf{x})} \quad j = 1, 2, \dots, C$$
$$P(\mathbf{x}) = \sum_{j=1}^C p(\mathbf{x} | \omega_j)P(\omega_j)$$

Evidence acts as a normalization factorterm same for all Classes.

ω_{test} is the class for which the posterior probability is highest.

The pattern is assigned to this class.

So, you can see the same thing I am showing here. So, $w_{test} = \operatorname{argmax} P(w_j \vee x)$ and I am considering this is the posterior probability is equal to this and evidence already I have explained it is nothing but the normalization factor. So, it is same for all the classes. So, that means, it has no role in classification. So, how to do the classification? So, w_{test} is the class for which the posterior probability is the highest.

So, based on this we can do the classification, that means, the pattern is assigned to this particular class. So, we can determine the posterior probability and we have to determine the highest value the highest posterior probability we have to determine and based on this we can take a classification decision. The pattern is assigned to this particular class.

(Refer Slide Time: 17:05)

Risk minimization framework

Let $\{\omega_1, \omega_2, \dots, \omega_C\}$ be the set of "C" states of nature (or "categories")

Let $\{\alpha_1, \alpha_2, \dots, \alpha_d\}$ be the set of possible "a" actions

Let $\lambda(\alpha_i | \omega_j)$ be the loss incurred for taking action α_i when the state of nature is ω_j

The diagram includes handwritten annotations: 'class' and 'Actions' in red above the left tree; 'class' and 'Action' in red above the right tree; 'True state of nature' in red below the right tree; and the loss function equation $Loss = \lambda(\alpha_i | \omega_j) = \lambda_{ij}$ in red at the bottom right.

After this another technique of decision making is by considering the Risks. So, by considering the risks, we can also take a classification decision. Suppose, I have C number of classes. So, I am considering the classes w_1, w_2 like this w_c number of classes. I am considering some actions, I am considering the actions are α_1, α_2 like this, I have number of actions and based on this I can consider the definition of loss.

So, what is the loss? I will explain, suppose x is the feature vector and x may belongs to the particular class I am considering the classes like this, suppose w_1 suppose w_2 . So, w_2 is the true class, true state of nature. So, the feature vector x may be assigned to the class, the class is w_1 or maybe w_2 like this and corresponding to this x belongs to w_1 , I am taking some actions some actions I am taking.

So, what are my actions? Action is α_i I am taking the action α_i I am taking and corresponding to this I can determine the loss. So, what is my loss? Loss is equal to $\lambda(\alpha_i \vee w_j)$ then I can determine. The loss is $\lambda(\alpha_i \vee w_j)$. So, particular action is considered the action is α_i corresponding to the class the class is w_j . So, there is a loss I can define like this.

So, loss is nothing but λ_{ij} that means, action α_i is considered for a class that class is w_j . So, I can show this again suppose, I have the classes w_1, w_2 like this, I suppose w_k , So, K number of

classes, and I am taking the actions like this $\alpha_0, \alpha_1, \dots, \alpha_k$ I am taking. So, these are my classes and I am taking some actions, actions are like this α_0, α_1 like this.

So, corresponding to w_1 I can take the action α_0 and similarly, corresponding to w_1 I can take this action also. So, action is α_k . So, corresponding to this what will be my loss the loss will be $\lambda_k' 1$ that is the loss I can determine. So, this suppose action is the reject option suppose, this action is the suppose reject option. So, in patent classification reject option is very important that means a particular feature vector may not be assigned to any 1 of the classes, then I can consider the option, the option is the reject option.

Suppose, in alphabet recognition, so, alphabet suppose A B C D like this, suppose, I am writing one alphabet something like this English alphabet. So, that means, I have to consider the option the option is the reject option I have to consider, because it does not belong to any of the classes that means, the alphabets. So, this is the concept of the loss. So, I can define the loss like this. So, action α_i is the considered for the class, the classes w_j and corresponding to this the loss is λ_{ij} . So, you can see.

(Refer Slide Time: 20:55)

Risk minimization framework

The expected loss: $R(\alpha_i) = \sum_{j=1}^c \lambda(\alpha_i | \omega_j) P(\omega_j)$ ✓

Given an observation with vector \mathbf{x} , the conditional risk is:

risk $R(\alpha_i | \mathbf{x}) = \sum_{j=1}^c \lambda(\alpha_i | \omega_j) P(\omega_j | \mathbf{x})$

At every \mathbf{x} , a decision is made: $\alpha(\mathbf{x})$, by minimizing the expected loss.

Our final goal is to minimize the total risk over all \mathbf{x} .

$$\int R(\alpha(\mathbf{x}) | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

And from this you can determine the expected loss and also the conditional risk also we can determine that is the actual α_i I am considering a for a feature vector, but feature vector is \mathbf{x} and

that is nothing but the $R(\alpha_i \vee x)$ that I am considering that is the risks I am determining this is λ_{ij}
 $\alpha_i \vee w_j P(w_j \vee x)$. So, this is nothing but λ_{ij} this is nothing but λ_{ij} .

At every x a decision is made and we have to minimize the expected loss, that concept is we have to minimize the expected loss. So, final goal is to minimize the total risks for all the feature vectors. So, we have to minimize these risks that is the objective of the risk minimization. So, by considering this, we can do the classification.

(Refer Slide Time: 21:58)

Risk minimization framework

- Two-category classification
 - α_1 : deciding ω_1
 - α_2 : deciding ω_2 ✓
 - $\lambda_{ij} = \lambda(\alpha_i | \omega_j)$ is loss incurred for deciding α_i when the true state of nature is ω_j }

λ_{11} λ_{12}
 λ_{21} λ_{22}

Conditional risk:

$$R(\alpha_1 | \mathbf{x}) = \lambda_{11}P(\omega_1 | \mathbf{x}) + \lambda_{12}P(\omega_2 | \mathbf{x})$$

$$R(\alpha_2 | \mathbf{x}) = \lambda_{21}P(\omega_1 | \mathbf{x}) + \lambda_{22}P(\omega_2 | \mathbf{x})$$

So, for Two category classification, what we can consider suppose, I am considering the action α_1 corresponding to the class the class is w_1 and α_2 I am considering, corresponding to the class w_2 and based on this I can define the loss function. So, loss function is λ_{ij} so, that means, action α_i I am considering for a class, the class is w_j . And from this we can determine the conditional risk.

So, you can see pictorially I can show you like this suppose w_1, w_2 these are the classes and I am considering the actions α_1, α_2 like this and I am taking the actions like this I am determining the actions. So, what is λ_{11} , λ_{11} , is nothing but λ_{α_1} action I am taking corresponding to the class w_1 . And similarly, I can consider λ_1 to λ_{21} and λ_{22} , I can consider like this. So, I can determine this, after this we can determine the condition risks.

(Refer Slide Time: 23:10)

Risk minimization framework

Our rule is the following:

if $R(\alpha_1 | \mathbf{x}) < R(\alpha_2 | \mathbf{x})$
action α_1 : "decide ω_1 " is taken

This results in the equivalent rule :

Decide ω_1 if:

$$\frac{(\lambda_{21} - \lambda_{11}) p(\mathbf{x} | \omega_1) P(\omega_1)}{(\lambda_{12} - \lambda_{22}) p(\mathbf{x} | \omega_2) P(\omega_2)} >$$

and decide ω_2 otherwise

After this we have to minimize the risks. So, what will be our decision rule? So, if the risks $\alpha_1 \vee \mathbf{x}$ is less than risks $\alpha_2 \vee \mathbf{x}$ that is a conditional risk, then in this case, we have to consider the action the action α_1 I have to consider that means, we have to consider the class the class w_1 we have to decide, and this is equivalent to this is equivalent to this because, from this just you can put these values you will be getting this one.

So, we can decide that class w_1 if this condition is satisfied, so, this is the condition. So, we can decide the class w_1 if this condition is satisfied, otherwise, we have to consider the class w_2 . So, that is my Decision Rule.

(Refer Slide Time: 24:08)

The slide is titled "Likelihood ratio" in a blue header. Below the title, it states: "The preceding rule is equivalent to the following rule:". The decision rule is presented as a fraction:
$$\text{if } \frac{p(\mathbf{x} | \omega_1)}{p(\mathbf{x} | \omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$
 The right-hand side of this inequality is circled in red. Below the equation, the text reads: "Then take action α_1 (decide ω_1)" and "Otherwise take action α_2 (decide ω_2)".

So, this the previous rule is equivalent to the following rule. So, that means, we can consider this one. So, from the previous equation you will be getting this one the ratio $\frac{P(x \vee w1)}{P(x \vee w2)}$ that ratio we have to determine. If it is greater than this one. So, that means, you pay consider this one then based on this, we can take a classification decision.

If this condition is satisfied, then we have to consider the actual α_1 and what is the corresponding class? The corresponding classes w1. Otherwise we have to take the action, the action is α_2 and what is the corresponding class? The class will be w2. So, based on this decision rule, we can do the classification. So, we can select either w1 or w2.

(Refer Slide Time: 25:04)

Likelihood ratio

- Regions of decision and zero-one loss function, therefore:
$$\text{Let } \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)} = \theta_\lambda$$

then decide ω_1 if: $\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \theta_\lambda$

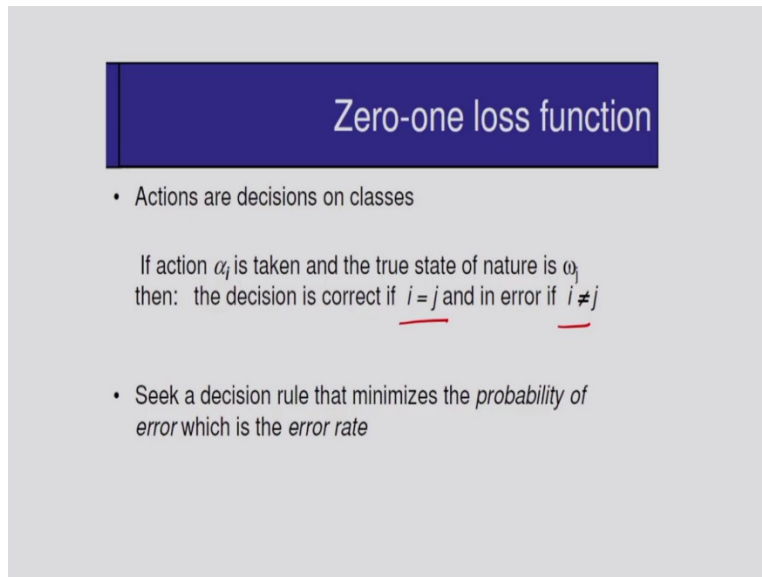
Optimal decision property ↑ Likelihood ratio ω_1

“If the likelihood ratio exceeds a threshold value independent of the input pattern \mathbf{x} , we can take optimal actions”

And we can also consider this ratio. So, this ratio is called the Likelihood Ratio, this is called the Likelihood Ratio we can determine. So, from a previous slide you can see, so, I am considering suppose this is equal to θ_k and if the likelihood ratio is greater than this threshold the threshold is suppose θ_k then based on this we can take a classification decision. So, if the likelihood ratio is greater than a particular threshold, then in this case we have to consider w_1 otherwise, we have to consider w_2 .

So, you can see this likelihood ratio is independent of the feature vector that it is likelihood ratio is independent of \mathbf{x} . So, we can determine the likelihood ratio and based on this likelihood ratio we can take a classification decision. So, you can see that based on these risks, we can take a classification decision.

(Refer Slide Time: 26:10)



Zero-one loss function

- Actions are decisions on classes

If action α_i is taken and the true state of nature is ω_j
then: the decision is correct if $i = j$ and in error if $i \neq j$

- Seek a decision rule that minimizes the *probability of error* which is the *error rate*

Now, I am defining one function that is called the Zero-one loss function. So, suppose actions are decision on classes, so, we are taking some actions for decision making. So, suppose if action α_i is taken and the true state of nature is ω_j , then the decision is correct, if i is equal to j and the decision is not correct, if it is not equal to j , we have to consider this zero to one loss function and based on this we can take a classification decision. So, objective is to minimize the probability of error that is the objective. So, we have to minimize the probability of error.

(Refer Slide Time: 26:47)

Zero-one loss function

- Introduction of the zero-one loss function:

$$\lambda(\alpha_i | \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

Therefore, the conditional risk is: All errors are equally costly.

$$\begin{aligned} R(\alpha_i | \mathbf{x}) &= \sum_{j=1}^{j=c} \lambda(\alpha_i | \omega_j) P(\omega_j | \mathbf{x}) \\ &= \sum_{j \neq i} P(\omega_j | \mathbf{x}) = 1 - P(\omega_i | \mathbf{x}) \end{aligned}$$

Select maxⁿ P($\omega_i | \mathbf{x}$)
Decide ω_i if
 $P(\omega_i | \mathbf{x}) > P(\omega_j | \mathbf{x})$ for $i \neq j$
Minimum Error rate classification.

"The risk corresponding to this loss function is the average probability error"

Now, I am defining the Zero-one loss function. So, $\lambda(\alpha_i | \omega_j)$ I am considering that is equal to 0 if i is equal to j so, that means, the loss will be 0 if i is equal to j . Otherwise, if i is not equal to j , then the zero-one loss function will be 1. So, I am defining this function zero to one loss function I am defining. After this I am considering the conditional risks, so, conditional risks already you know, so, this is the, the formula for the conditional risk and from this you can see, I am getting this one, but you can see at this point that j is not equal to i , because, when j is not equal to i , the value of zero-one loss function is 1.

So, that is why this value will be 1, if j is not equal to i or if i is not equal to j , then value will be equal to 1 and corresponding to this it is nothing but 1 minus probability of ω_i given \mathbf{x} . So, that is the meaning of this. So, to minimize the risks, what I have to consider, I have to select maximum probability of ω_i given \mathbf{x} that is the posterior probability I have to select the maximum value of this I have to select, that means the decision rule will be like this.

So, decide ω_i if probability of ω_i given \mathbf{x} is greater than probability of ω_j , given \mathbf{x} for i is not equal to j . So, this is my decision rule. And this classification technique is called the Minimum Error rate classification, because I have to minimize the error. So, this is called the Minimum Error rate classification technique.

(Refer Slide Time: 29:08)

Zero-one loss function

- Minimize the risk requires maximizing $P(\omega_i | \mathbf{x})$

$$\text{(since } R(\alpha_i | \mathbf{x}) = 1 - P(\omega_i | \mathbf{x})\text{)}$$

- For Minimum error rate

$$\text{– Decide } \omega_i \text{ if } P(\omega_i | \mathbf{x}) > P(\omega_j | \mathbf{x}) \forall j \neq i$$

So, for minimization of the risks, we have to maximize this probability since the conditional risk is equal to 1 minus probability of ω_i given \mathbf{x} and I have to minimize the error. So, what is my classification rule? Decide ω_i if probability of ω_i given \mathbf{x} is greater than probability of ω_j given \mathbf{x} . So, that is my classification rule and this is the concept of the Minimum Error rate classification.

(Refer Slide Time: 29:35)

Classifiers, Discriminant Functions and Decision Surfaces

- The multi-category case

$$\text{– Set of discriminant functions } g_i(\mathbf{x}), i = 1, \dots, c$$

– The classifier assigns a feature vector \mathbf{x} to class ω_i

if:

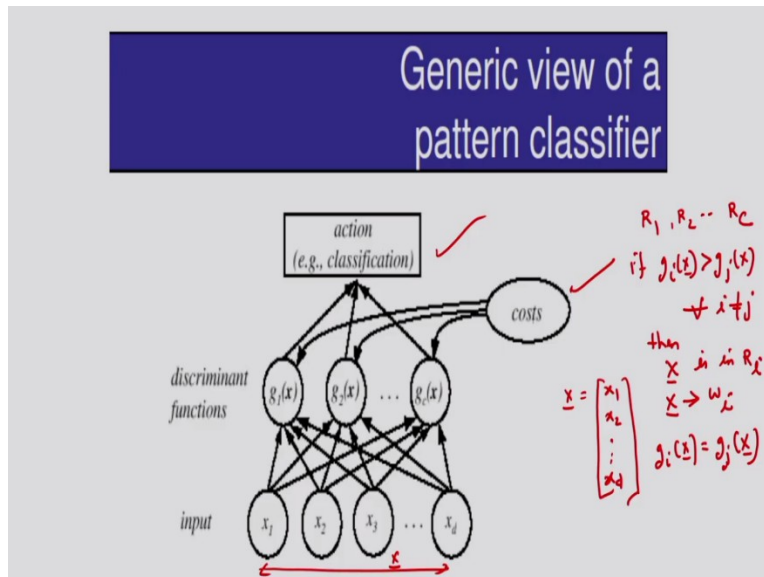
$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \forall j \neq i$$

And based on the Discriminant function, I can see the different types of decision surfaces. So, this concept I am going to explain so, what is the Discriminant function and from this how to

determine their decision boundaries, the decision surfaces. So, let us consider the multi category case. So, multiple clusters I am considering and for this I am considering the discriminant function that discriminant function is $g_i(x)$ I am considering and for each and every class, I have to determine that discriminate function.

So, I have to determine $g_1(x)$, $g_2(x)$, $g_3(x)$ like this for all the classes I have to determine the discriminant function. The classifier assigns a feature vector x to a particular class that class is supposed w_i if $g_i(x)$ is greater than $g_j(x)$ for j is not equal to i . So, based on the discriminant function, I can take a classification decision that means, if $g_i(x)$ is greater than $g_j(x)$, then based on this I can decide the class the class is w_i that means, the feature vector x will be assigned to the class the classes w_i .

(Refer Slide Time: 30:56)



And here you can see I am considering the input feature vector the input feature vector is the D dimensional feature vector. So, this is my x that is the feature vector. So, this is the D dimensional feature vector I am considering that means, x_1, x_2 like this. So, this is the D dimensional feature vector I am considering and after this you can see I am determining the discriminant function for all the classes.

So, I am considering C number of classes. So, I am determining $g_1(x), g_2(x)$ like this I am determining and for classification what we have to consider I have to find a maximum discriminant function I have to determine. So, out of this $g_1(x), g_2(x), g_c(x)$ which one is the maximum I have to determine for classification. So, you can see so, I am determining the cost that means, I have to find a maximum discriminant function and based on this I can take a classification decision.

So, what is the function of the discriminant function? It divides the features space into C decision regions that decision regions are R_1, R_2 like this, these are the decision regions R_c and if $g_i(x)$ is greater than $g_j(x)$ for i is not equal to j then x is in the region x is in the region R_i that means, the meaning is x is assigned to the class that class is w_i . So, that is the decision rule and what is the decision boundary?

The decision boundary is nothing but $g_i(x)$ is equal to $g_j(x)$. So, that is the equation of the decision boundary. So, for C number of classes we have to determine C number of discriminant functions.

(Refer Slide Time: 33:04)

Discriminant Functions

- Let $g_i(\mathbf{x}) = -R(\alpha_i | \mathbf{x})$ ✓
(max. discriminant corresponds to min. risk!)
- For the minimum error rate, we take

$$g_i(\mathbf{x}) = P(\omega_i | \mathbf{x}) \rightarrow = \frac{P(\mathbf{x} | \omega_i) P(\omega_i)}{P(\mathbf{x})}$$
- (max. discrimination corresponds to max. posterior!)
- $$g_i(\mathbf{x}) = \ln p(\mathbf{x} | \omega_i) P(\omega_i)$$
 ✓
- $$g_i(\mathbf{x}) = \ln p(\mathbf{x} | \omega_i) + \ln P(\omega_i)$$
- (ln: natural logarithm!)

So, you can see I have to determine the maximum value of the discriminant function that corresponds to the minimum risks. So, maximum discriminant function corresponds to minimum risks. So, that means, for minimum error rate $g_i(x)$ should be equal to $P(\omega_i | x)$ that is the posterior probability. So, that means, the maximum discriminant function corresponds to maximum posterior probability.

So, I am repeating this the maximum discriminant function corresponds to maximum posterior probability and what is the posterior probability if you see? This is nothing but it is equal to $P(x | \omega_i)$, $P(\omega_i)$, $P(x)$ and that I can write like this. So, promise I can write like this, because, the evidence has no role in classification. So, I can write like this, after this I can take the natural logarithm, because the multiplication is converted into addition by considering the logarithm. So, I have that this discriminant function $g_i(x)$ is equal to this.

(Refer Slide Time: 34:26)

Discriminant functions

- Discriminant functions do not change the decision, when scaled by some positive constant 'k'.

$g_1(x) \leftarrow k g_1(x)$
 $\text{if } g_1(x) > g_2(x) \text{ then } x \rightarrow w_1$
- The decision is not affected when a constant is added to all discriminant functions.

$g_1(x) = g_2(x)$
 $g_1(x) - g_2(x) = 0$
 $g(x) = 0$
 $\leftarrow \text{Decision Boundary}$

So, discriminant function does not change the decision when scale by some positive constant. So, if it is scaled by some positive constant, the discriminant function does not change the decision. The decision is not affected when a constant is added to all the discriminant function. So, that is a concept of the discriminant function. So, if I consider two classes suppose, so, for first class suppose the discriminant function is $g_1(x)$.

For the second class suppose the discriminant function is $g_2(x)$, then what will be my classification rule if $g_1(x)$ is greater than $g_2(x)$ that means their meaning is x will be assigned to the class the class is w_1 . And what will be my decision boundary? The decision boundary will be $g_1(x)$ is equal to $g_2(x)$ so, that is the equation of the decision boundary. So, this is my decision boundary.

So, this equation that means, $g_1(x) - g_2(x) = 0$ that is the equation of the decision boundary. So, I can write like this that means $g(x)$ is equal to 0. So, that is the equation of the curve that is the equation of the curve. So, if I consider these two classes, so, if I consider this the feature space so, I have two regions. The region is R_1 another region is R_2 and what will be the equation of the decision boundary?

The equation of the decision boundary is $g(x)$ equal to 0 and that is the equation of the curve or maybe in this case I am considering the line that is the decision boundary I am considering the

line I am considering two regions R1 and R2 and in the region R1 $g(x)$ is greater than 0 and in the region R2 $g(x)$ is less than 0. So, $g(x)$ is greater than 0 means, I am considering the class w_1 and $g(x)$ is less than 0 that means, I am considering the class w_2 .

(Refer Slide Time: 36:43)

Discriminant Functions

- Feature space divided into c decision regions

$$\text{if } \underline{g_i(\mathbf{x})} > \underline{g_j(\mathbf{x})} \quad \forall j \neq i \text{ then } x \text{ is in } \underline{R_i}$$

(R_i means assign x to ω_i)
- The two-category case
 - A classifier is a “dichotomizer” that has two discriminant functions $\underline{g_1}$ and $\underline{g_2}$

Let $g(\mathbf{x}) \equiv g_1(\mathbf{x}) - g_2(\mathbf{x})$

Decide $\underline{\omega_1}$ if $\underline{g(\mathbf{x})} > 0$; Otherwise decide ω_2

The feature space is divided into c number of regions, c decision regions and if $g_i(x)$ is greater than $g_j(x)$ for j is not equal to i then x is in R_i that means the feature vector will be in the region R_i that means, the feature vector x will be assigned to the class that class is w_i and for two category case that is for two classes I have to determine $g(1)$ and $g(2)$ and $g(x)$ is equal to $g_1(x) - g_2(x)$ and we can take a classification decision based on this condition. So, decide w_1 if $g(x)$ is greater than 0 otherwise decide w_2 that we have to consider.

(Refer Slide Time: 37:35)

Dichotomizer

$g_i(\mathbf{x}) = \ln p(\mathbf{x} | \omega_i) + \ln P(\omega_i)$ ✓
 (ln: natural logarithm!)

– The computation of $g(\mathbf{x})$ for dichotomizer

or

$$g(x) = \ln \frac{P(\omega_1|x)}{P(\omega_2|x)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

$g_1(x) - g_2(x)$
 $P(\omega_1|x) = \frac{P(x|\omega_1)P(\omega_1)}{P(x)}$
 $P(\omega_2|x) = ?$

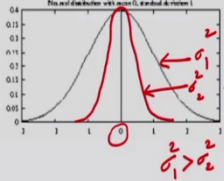
And, and this $g_i(x)$ already I have this equation the equation of the discriminant function and corresponding to this you can see this $g(x)$ is nothing but $P(\omega_1|x) - P(\omega_2|x)$ that is the $g(x)$. $g(x)$ is nothing but $g_1(x) - g_2(x)$. And from this if you can put this value because this what is $P(\omega_1|x)$ what is $P(\omega_1)$ given x that is nothing but $P(x|\omega_1)$ and $P(\omega_1)$ and the evidence is suppose $P(x)$. So, if I put these below in this equation, then you will be getting this one so, you can get this one.

So, we have to determine $g(x)$ and similarly also you have another $1 - P(\omega_2|x)$. So, $P(\omega_2)$ also x we can determine. So, if I put these two values you will be getting $g(x)$ so, the $g(x)$ is nothing but $g_1(x) - g_2(x)$. So, this is nothing but $g_1(x) - g_2(x)$.

(Refer Slide Time: 38:54)

Normal /Gaussian Distribution

A bell-shaped distribution defined by the probability density function



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \checkmark$$

- Expected, or mean, value of X is $E[X] = \int_{-\infty}^{\infty} xp(x)dx = \mu \quad \checkmark$
- Variance of X is $Var(x) = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = \sigma^2 \quad \checkmark$

Now, I will discuss the concept of Normal Distribution. So, already you know what is a Normal Distribution. So, here you see I am showing the density function corresponding to the normal

distribution $P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$. So, that is the normal distribution and corresponding to this you can see I have the bell-shaped distribution.

So, this is the bell-shaped distribution corresponding to that PDF that PDF is $P(x)$ and corresponding to this I have the, the probability density function and this is the normal distribution with mean 0. So, here you can see the mean is 0 and from this you can determine the expected value or the mean value of x . So, I can determine the mean of this random variable. So, $E[x]$ I can determine also I can determine the variance of x .

So, in this case I am considering suppose these variances suppose σ_1^2 and suppose I am considering another Gaussian function something like this then in this case suppose the variance is σ_2^2 . So, in this case the σ_1 is, $\sigma_1^2 > \sigma_2^2$. So, you can see that the variance determines the spread of the Gaussian function. So, this is about the normal and the Gaussian distribution.

(Refer Slide Time: 40:38)

Multivariate Gaussian Distribution

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$\mathbf{x} = [x_1, x_2, \dots, x_d]^T$
 $N(\boldsymbol{\mu}, \Sigma)$
 $\boldsymbol{\mu} = [\mu_1, \mu_2, \mu_3, \dots, \mu_d]^T$
 $\Sigma = E\left[\sum_{d=1}^d [x_d - \mu_d]^2\right]^T$

$$\boldsymbol{\mu} = E[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x}, \quad \Sigma = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})']$$

Where $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}$ and $\boldsymbol{\mu} = E[\mathbf{x}] = \begin{bmatrix} E[x_1] \\ E[x_2] \\ \vdots \\ E[x_d] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix}$

$\Sigma = d \times d$ Covariance matrix

$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$
 $N(\mu, \sigma^2)$

Now, let us consider the Multivariate Gaussian Distribution. So, I am considering the vector, vector is the D dimensional vector so, x is the D dimensional vector x1, x 2 like this. So, this is the D dimensional vector corresponding to this I can define the density function, the density function is Px and this is for the Multivariate Gaussian distribution. So, in this case I have two parameters one is the Mean vector and another one is the Covariance matrix.

So, I have two parameters one is the mean vector another one is the covariance. I can determine the mean vector from the input vector the input vector is x. So, the mean vector is nothing but $\boldsymbol{\mu} = \mu_1, \mu_2, \mu_3, \dots$ and since I am considering the D dimensional vector, so, this is the mean vector that means, I can determine the that means, I can determine the expected value x1, x 2 like this I can determine.

So, mean vector I can determine and also, I can determine the covariance matrix so, this is my covariance matrix. So, this is expected below $E(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})'$. So, you can see I can determine the covariance matrix and if I consider the dimension is suppose 1. So, suppose the dimension of the feature vector is 1 then in this case I will be getting the Univariate density.

So, what is the Univariate density that is already I have defined. The univariate density is twice by sigma so, I have, I will get the univariate density corresponding to D is equal to 1. So, in this

case I have two parameters one is the mean another one is the variance. So, for univariate case I have two parameters one is the mean another one is variance. So, in this case you can see I am considering the D dimensional feature vector and from this I can determine the mean vector and also, I can determine the d by d covariance matrix.

(Refer Slide Time: 43:30)

Similarly, the *covariance matrix* Σ is defined as the (square) matrix whose ij^{th} element σ_{ij} is the covariance of x_i and x_j :

$$\sigma_{ij} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)] \quad i, j = 1 \dots d,$$

Handwritten notes: $|\Sigma| = \text{determinant}$, Σ^{-1} inverse

$$\Sigma = \begin{bmatrix} \mathcal{E}[(x_1 - \mu_1)(x_1 - \mu_1)] & \mathcal{E}[(x_1 - \mu_1)(x_2 - \mu_2)] & \dots & \mathcal{E}[(x_1 - \mu_1)(x_d - \mu_d)] \\ \mathcal{E}[(x_2 - \mu_2)(x_1 - \mu_1)] & \mathcal{E}[(x_2 - \mu_2)(x_2 - \mu_2)] & \dots & \mathcal{E}[(x_2 - \mu_2)(x_d - \mu_d)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{E}[(x_d - \mu_d)(x_1 - \mu_1)] & \mathcal{E}[(x_d - \mu_d)(x_2 - \mu_2)] & \dots & \mathcal{E}[(x_d - \mu_d)(x_d - \mu_d)] \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_d^2 \end{bmatrix}$$

Handwritten notes: $\sigma_i^2 \rightarrow$ variance of respective x_i , $\sigma_{ij} \rightarrow$ co-variance betⁿ x_i & x_j , $\sigma_{ij} = 0 \Rightarrow$ univariate maximal density

Handwritten notes: $\hat{x} = \frac{(x - \mu)^T \Sigma^{-1} (x - \mu)}{\|x - \mu\|}$ - Mahalanobis distance

So, in this slide you can see how to determine the covariance matrix. So, the covariance matrix you can see expected below of this expected below $E[(x_i - \mu_i)(x_j - \mu_j)]$ that I can determine. So, that means the covariance between x_i and x_j I can determine. So, this is my covariance matrix and the diagonal elements σ_{ij} are the variances of their respective x_i . So, if I see that that if I see the diagonal elements that is nothing but the variance of respective x_i .

So, that means, σ_1^2 that corresponds to the variance variances of respective x_i and if I consider the off-diagonal elements that means, these off diagonal elements so, off diagonal elements are σ_{ij} . So, these are the covariance, covariance between covariance between x_i and x_j . So, that is a covariance between x_i and x_j and if I considered suppose, if x_i and the x_j are statistically independent, what is the meaning of this?

That means σ_{ij} is equal to 0 that is the covariance is equal to 0 that means, x_i and x_j are statistically independent then corresponding to this, this $P(x)$ that already I have defined that will

be the univariate density, univariate, univariate normal density. So, that will be the univariate normal density. So, that means, if x_i and x_j are statistically independent that corresponds to σ_{ij} is equal to 0 and that corresponds to $P(x)$ will be the univariate normal density.

Now, I can define a distance. So, what is the distance you can see, I am considering the distance $(x - \mu)' \Sigma^{-1} (x - \mu)$ so, I have this distance R square is equal to $(x - \mu)' \Sigma^{-1} (x - \mu)$. So, in this case you can see this corresponds to the determinant and this corresponds to the inverse, inverse of the covariance matrix.

So, this corresponds to inverse of the covariance matrix so, I am defining a distance that is R square is equal to $(x - \mu)' \Sigma^{-1} (x - \mu)$ that is the inverse of the covariance matrix x minus μ and that distance is called that is a very important distance this is called the Squared Mahalanobis Distance. The Mahalanobis distance, Squared Mahalanobis distance, already you know that what is the Euclidean distance.

The Euclidean distance is nothing but the distance between the vector x and μ i suppose, so, this is the Euclidean norm. So, this is the equilibrium norm and I have shown that this is the Mahalanobis distance. So, in case of the multivariate density the center that means, in this case if I consider a cluster suppose if I consider a cluster suppose a cluster I am considering some sample points I am considering.

So, these are the, these the cluster the center of the cluster is determined by the mean vector and the shape of the cluster, the shape of the cluster is determined by the covariance matrix. So, that is the importance of the mean and the covariance. So, if I consider a cluster I am considering some sample points. So, the center of the cluster is determined by the mean vector and the scope of the cluster is determined by the covariance matrix. Now, I am discussing the concept of the Bayesian classification for normal distribution.

(Refer Slide Time: 48:33)

Bayesian classification for normal distribution

$$P(w_i | x) = \frac{P(x | w_i) P(w_i)}{P(x)}$$

L - dimensional feature vector

$$P(x | w_i) = \frac{1}{(2\pi)^L |\Sigma_i|^{L/2}} \exp \left[-\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right]$$

$i = 1, 2, \dots, C$

$\Sigma_i \rightarrow L \times L$ $\mu_i = E[x]$


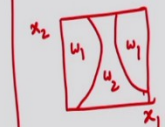
$$g_i(x) = \ln P(x | w_i) + \ln P(w_i)$$

$$= -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) + \ln P(w_i) + C_i \quad \text{--- (1)}$$

$$C_i = -\left(\frac{L}{2}\right) \ln 2\pi - \frac{1}{2} \ln |\Sigma_i|$$

$$g_i(x) = -\frac{1}{2} x^T \Sigma_i^{-1} x + \frac{1}{2} x^T \Sigma_i^{-1} \mu_i + \frac{1}{2} \mu_i^T \Sigma_i^{-1} \mu_i - \frac{1}{2} \mu_i^T \Sigma_i^{-1} \mu_i + \ln P(w_i) + C_i \quad \text{--- (2)}$$

$g_i(x) - g_j(x) = 0$
 Quadratics \Rightarrow ellipsoid, parabolas, hyperbolas, ...
 Bayesian \Rightarrow Quadratic classifier

So, let us discuss about the concept of the Bayesian classification, Bayesian classification for Normal Distribution. So, let us discuss about this concept that Bayesian classification for normal distribution. So, in case of the Bayes classifier, so, I have to determine this. So, this is nothing but $P(w_j | x)$ is equal to $P(x | w_j)$, $P(w_j)$ and $P(x)$. So, this is the Bayes law. So, for determining the probability of w_j given x , I have to consider this, that is the likelihood or the class conditional density I have to consider.

Suppose, the class conditional density follows the normal distribution. So, that means the likelihood function w_i which with respect to x in L dimensional feature space follow the general multivariate normal density. So, that means the probability of $(x | w_j)$, I am considering the density is suppose the normal density and in this case I am considering multivariate normal density. So, that means the probability of x given w_j is a multivariate normal distribution and I am considering the feature vector that is the L dimensional feature vector.

So, L dimensional feature vector I am considering. So, twice by to the power 1 by 2 sigma i to the power half an exponential minus half x minus μ_i transpose so, I am considering the multivariate normal distribution and I am considering c number of classes. So, that covariance matrix for a particular class, so, that means, I am considering the covariance matrix for class i .

So, that is the dimension is 1 by 1 covariance matrix I will be getting also I can determine the mean vector. So, μ_i also I can determine that is the mean vector I can determine that is nothing but the expected value of x_i can determine. Now, you know the discriminant function $g_i(x)$ is equal to \log you know this. So, in this case I think I better I should write w_i that is x given w_i I am writing in place of w_j I am writing just w_i . So, this is a Discriminant function.

So, if I consider this multivariate normal distribution so, just I have to put this below here. So, I will be getting $-\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) + c_i$. So, suppose this equation number 1. The c_i is a constant to what is the value of this constant? $\frac{1}{2 \ln 2 \pi} |\Sigma_i|^{-1/2} \exp(-\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i))$ so, this is c_i so, I can expand $g_i(x)$ that is a discriminant function.

So, if I expand $g_i(x)$ it will be something like this $-\frac{1}{2} x^T \Sigma_i^{-1} x + \mu_i^T \Sigma_i^{-1} x - \frac{1}{2} \mu_i^T \Sigma_i^{-1} \mu_i + c_i$ just I am expanding suppose this is the equation number 2. This is the equation number 2. So, I am getting the, the expression for a discriminant function. And in this case, what I am considering I am considering the probability of x given w_i that follows the multivariate normal distribution. So, based on this, I am calculating the discriminant function, the discriminant function is $g_i(x)$.

What will be my decision boundary? So, my decision boundary or decision curves the decision curves will be $g_i(x) - g_j(x) = 0$. So, that is the Decision Boundary. So, the decision boundary maybe the quadric decision boundary because this is the quadratic equation. So, if you see this one, so, this is a quadratic equation. So, that is why this is called a Quadric classifier.

Bayesian classifier is also called Quadric classifier because it is a Quadric equation. So, decision boundary will be quadrics. So, decision boundary may be like this Ellipsoid or maybe the Parabolas, Parabolas or maybe the Hyper Parabolas or maybe the pairs of lines I can consider like this, I have the Quadrics Decision boundary. So, this equation number 2 is the quadratic equation.

So, that is why the Bayesian classifier, the Bayesian classifier, the Bayesian classifier is also called the Quadric classifier. So, if I considered the 2-dimensional case, so, decision boundaries maybe something like this. So, this will be one class and this will be another class or maybe the

nonlinear maybe the decision boundary is maybe something like this. So, the suppose w_1, w_2 and w_3 so, the x_1 and x_2 .

So, 2-dimensional feature vector I am considering and for this I may have this type of decision boundaries. So, if I consider high dimensional case, then in this case I will be getting hyper ellipsoid, hyper parabolas like this I will be getting.

(Refer Slide Time: 58:19)

$\Sigma^{-1} x^T \Sigma^{-1} x \rightarrow$ Same for all the classes
 $\Sigma_i = \Sigma$
 $g_i(x) = w_i^T x + w_{i0}$ ← ③
 $w_i = \Sigma^{-1} \mu_i$ (Weight vector)
 $w_{i0} = \ln P(\mu_i) - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i$ (Bias or threshold)
 Case I: $\Sigma = \sigma^2 I$ (L-dimensional identity matrix)
 ③ $\Rightarrow g_i(x) = \frac{1}{\sigma^2} \mu_i^T x + w_{i0}$
 $\Sigma^{-1} = \left(\frac{1}{\sigma^2}\right) I$

Decision Hyperplanes
 $g_{ij}(x) = g_i(x) - g_j(x)$
 $= w_i^T (x - x_0)$ → $w_i^T x + w_{i0} = w_j^T x + w_{j0}$
 $\Rightarrow w^T (x_1 - x_2) = 0$
 $g_{ij}(x) = w_{ij}^T (x - x_0)$
 $w = \mu_i - \mu_j$
 $x_0 = \frac{1}{2} (\mu_i + \mu_j) - \sigma^2 \ln \left(\frac{P(\mu_i)}{P(\mu_j)} \right) \frac{\mu_i - \mu_j}{\|\mu_i - \mu_j\|^2}$
 If $P(\mu_i) = P(\mu_j)$, $x_0 = \frac{1}{2} (\mu_i + \mu_j)$
 If $P(\mu_i) < P(\mu_j) \rightarrow$ closer μ_i
 If $P(\mu_i) > P(\mu_j) \rightarrow$ closer μ_j
 σ^2 is small w.r.t $\|\mu_i - \mu_j\|^2$

Now, what about the Decision Hyper planes So, if I consider the equation number 2, you can see the equation number 2 in equation number 2 you will be having this term $x^T \Sigma_i^{-1} x$ this in equation number 2, it is same for all the discriminant function it is same for all the discriminant function. So, that means, it has no role in classification. So, that means I can neglect this one.

And suppose if I consider this case that covariance matrix is same for all the classes, the covariance matrix is same for all the classes, but it is arbitrary. So, that I am considering now. So, this, this quadratic term is same for all the classes, same for all the classes. So, that means, I can neglect this one it has no role in classification. And I am considering this case that is the covariance matrix is same for all the classes, but it is arbitrary. So, that is why I can write the discriminant function something like this $g_i(x)$ is equal to $w_i^T x + w_{i0}$.

So, this w_i is this is called the Weight vector. This is called the Weight vector and this is I can consider as a Bias or Threshold. So, Bias or Threshold I can consider like this. So, what do what is that weight vector? w_i is equal to $\sigma^2 \mu_i$. So, that is the weight vector and what is the bias? The bias is nothing but w_{i0} that is equal to $\log P(w_i) - \frac{1}{2} \mu_i^T \Sigma_i^{-1} \mu_i$.

So, here you can see if I see this expression if I see this expression that is the discriminant function $g_i(x)$ is a linear function of x and in this case the decision surfaces will be hyper planes. so, I am repeating this, that is the discriminant function $g_i(x)$ is a linear function of x , x is the input feature vector and corresponding to this the decision surfaces will be hyper planes. Now, let us consider two cases. Case number 1, I am considering the covariance matrix is same for all the classes and it is a diagonal covariance matrix.

So, that means, what I am considering the diagonal covariance matrix with equal elements. So, what is the meaning of this the feature vector is mutually uncorrelated and of and of some variance. So, that means, I am considering a diagonal covariance matrix with equal elements the meaning is feature vector is mutually uncorrelated and of same variance. And in this case, you can see what is the I ? I is the identity matrix. So, it is the 1 dimensional, 1 dimensional identity matrix.

So, I is the 1 dimensional identity matrix and suppose this equation is suppose I am considering the equation is 3 suppose. So, from 3 equation number 3 what I will be getting now, $g_i(x)$ will

be equal to $\frac{1}{\sigma^2} \mu_i^T x + w_i$, so, I will be getting w_{i0} that is the bias. And in this case what is σ^{-1} ?

That is the inverse of the covariance matrix 1 by σ^2 . I is the identity matrix.

So, what will be the decision hyper planes? Decision hyper planes, decision hyper planes will be $g_i(x)$ is equal to $g_i(x) - g_j(x)$. So, that is nothing but $w^T x - x_0$. This you can verify this you can verify like this suppose $w^T x_1 + w_0$ is equal to $w^T x_2 + w_0$. So, corresponding to this what I will be getting, what I will be getting $w^T x_1 - x_2$ is equal to 0 . So, like this you can verify that one.

So, this is the equation and that is the equation what is the equation of the decision hyper planes? $g_{ij}(x)$ is equal to $w^T(x - x_0)$ that is the weight vector $x - x_0$. So, this is the equation of the Decision Hyper Planes and in this case what is the weight vector? The weight vector is nothing but $\mu_i - \mu_j$ that is the weight vector, $\mu_i - \mu_j$. And the x_0 is equal to $\frac{1}{2}(\mu_i + \mu_j) - \sigma^2$. So, that is x_0 .

So, the decision surface is a hyper plane which will pass through the point, the point is x_0 . So, I am repeating this in this case the decision surface is a hyper plane passing through the point the point is x naught. And suppose, if I consider if the probability of w_i suppose, the probability of w_i is equal to probability of w_j then in this case corresponding to this, the point x_0 will be

$$\frac{1}{2}[\mu_i + \mu_j].$$

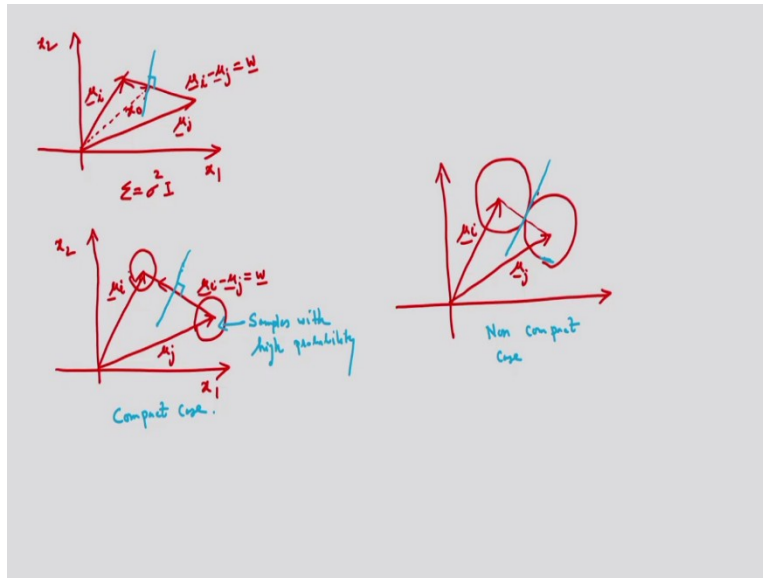
So, that is the decision surface and the point will be x_0 is equal to $\frac{1}{2}[\mu_i + \mu_j]$. That means, the hyper plane passes through the mean of $\mu_i \wedge \mu_j$. So, in this case if the probability of a particular class that class is w_i is equal to probability of the another class and the class is w_j and in this case i is not equal to j then x_0 is equal to $\frac{1}{2}[\mu_i + \mu_j]$. So, that means, the hyper plane will pass through the mean of $\mu_i \wedge \mu_j$.

And also the hyper plane will be orthogonal to the vector, the vector is w is equal to $\mu_i - \mu_j$, so, that I will show pictorially. And suppose, if the probability of w_i is less than probability of w_j then what will happen? The hyper plane will be located closer to μ_i that is the hyper plane will be located a closer to the mean vector the mean vector is μ_i and also if the probability of w_i is greater than probability of w_j , then what will happen in this case the hyper plane will be located closer to μ_j that is the mean w_j .

So, this is the case and if the variance, variance is σ^2 is small. The variance is small with respect to, with respect to the difference in the mean. So, I am just finding the Euclidean distance between these two means. The location of the hyper plane is rather insensitive to the values of $P(w_i)$ and $P(w_j)$. So, I am repeating this if the variance is small with respect to the

difference in the mean, the location of the hyper plane is insensitive to the values of the probability, the probabilities probability of w_i and the probability of w_j . So, this condition I will be showing pictorially, so, what will be my decision boundary?

(Refer Slide Time: 01:09:37)



So, I can show the decision boundary like this. So, suppose if I considered a 2-dimensional feature space x_1 and x_2 . So, suppose this is my mean vector suppose this mean vector is μ_i and, and this is μ_j suppose this vector. And I can find the vector $\mu_i - \mu_j$ that is nothing but that weight vector that is nothing but that weight vector. And also, you can see, I can determine the point the point is nothing but x_0 I can determine.

So, from that equations you can determine the point x_0 . So, that means, the hyper plane will pass through the point the point is x_0 and if I consider the probability of w_i is equal to probability of w_j then x_0 is equal to $\frac{1}{2}[\mu_i + \mu_j]$ that already I have calculated. So, that means, the hyper plane will pass through the mean of $\mu_i \wedge \mu_j$. and also the hyper plane is orthogonal to the vector w .

So, that means up consider hyper plane so, hyperplane is this suppose, so, this is my hyper plane So, that will be orthogonal. So, maybe I can consider some another color that you can understand. So, this is my hyper plane so, hyper plane is orthogonal so, it is orthogonal to the vector the vector is w . The hyper plane will be orthogonal to the vector $\mu_i - \mu_j$. So, this is $\mu_i - \mu_j$.

And in this case what I am considering the sigma is equal to sigma square I that means, the covariance I am considering the diagonal covariance matrix with equal elements. So, diagonal

covariance matrix I am considering and if I consider these two cases, one is the high variance another one is the low variance the small variance. So, what will happen you can see so, this is mean of the cluster suppose if I consider this the cluster so, the mean of the cluster is μ_i and another cluster I am considering.

So, this is another cluster so, this is music and from this you can see already I have shown you so, you can determine the vector the vector is $\mu_i - \mu_j$ that you can determine that is nothing but the weight vector level with vector and after this you can draw a decision boundary. So, this is my decision boundary you can draw that is orthogonal to the vector w , w and it will pass through the point x_{naught} .

So, in this case, I am considering the compact case this is a compact, compact compactness that means, I am considering the samples with high probability, the samples with high probability samples with high probability I can consider another case that is a non-compact case. So, in a non-compact case also I can draw the decision boundary. So, in case of a non-compact case again like the previous case, I can draw the mean vector μ_i and the μ_j and if I consider this is a cluster suppose, the μ_j is something like this.

So, this is μ_j and you can determine $\mu_i - \mu_j$ and in this case also you can determine the decision boundary. So, you can find a decision boundary so, decision boundary will be something like this. So, this is the non-compact case that means, sigma square is large in the previous case the sigma square is small with respect to $\mu_i - \mu_j$. So, that means, in the second case, the location of the decision hyper plane is much more critical as compared to the, the first case, the first case is the compact case.

So, in the compact case you can easily draw the decision boundary between these two clusters. But in case of the non-compact case, so non compact case I am showing in the second case, so, it is very difficult to draw the decision boundary. So, location of the decision hyper plane is much more critical as compared to first case, this is about the representation of the decision boundary.

(Refer Slide Time: 01:14:52)

Case 2 Non diagonal covariance matrix

$$\Sigma_i = \Sigma$$

$$g_i(x) = w^T (x - x_0) = 0$$

$$w = \Sigma^{-1} (\mu_i - \mu_j)$$

$$x_0 = \frac{1}{2} (\mu_i + \mu_j) - \ln \left(\frac{P(\mu_i)}{P(\mu_j)} \right) \frac{\mu_i - \mu_j}{\|\mu_i - \mu_j\|^2 \Sigma^{-1}}$$

Minimum distance classifier

$$g_i(x) = -\frac{1}{2} (x - \mu_i)^T \Sigma^{-1} (x - \mu_i)$$

* Σ

$$g_i(x) \rightarrow \Sigma^{-1}$$

$$d_n = \left((x - \mu_i)^T \Sigma^{-1} (x - \mu_i) \right)^{1/2}$$

$$d_n = c$$

Now, I am considering the case number 2 case 2. So, I am considering the non-diagonal covariance matrix. So, I am considering the non-diagonal covariance matrix these are case number 2 that means the Σ_i is equal to Σ . The covariance matrix is same for all the classes, in case of a case number 1 what I am considering diagonal covariance matrix, but in this case, I am considering the non-diagonal covariance matrix.

So, corresponding to these the decision boundary, the decision boundary will be $g_i(x)$ is equal to $w^T(x - x_0)$ is equal to 0 and what will be my weight vector? The weight vector will be Σ^{-1} that is the inverse of the covariance matrix $\mu_i - \mu_j$ and the x_0 the decision boundary will pass through the

point the point is $x_0 = \frac{1}{2} [\mu_i + \mu_j]$ minus \ln so, decision boundary I can find like this.

So, I can calculate the weight vector and also, I can determine the point the point is x naught. So, the decision boundary will pass through the point x_0 , but in this case the hyper plane is no longer orthogonal to the vector w that is the case for a non-diagonal covariance matrix. Now, let us consider the concept of the minimum distance classifier. So, what is the Minimum Distance Classifier?

Minimum Distance Classifier let us consider So, what is the Minimum Distance Classifier? So, from the equation number 1 you have seen that the discriminant function is equal to minus $\frac{-1}{2} - \mu_i^T x$ and σ^{-1} that is the inverse of the covariance matrix. So, you notice so, suppose, suppose the case like this, So, sigma is equal to sigma square into I so, I use the identity matrix.

So, in this case I have to determine the maximum discriminant function. So, for c number of classes I have c number of discriminant function and I have to take the decision based on this so, I have to find the maximum discriminant function and based on this I have to take the classification decision. So, that means the maximum discriminant function means minimum Euclidean distance between the respective mean points.

So, what is the Euclidean distance, the Euclidean distance dE I can write like this, the Euclidean distance between the vector x and the mean mu i so, that I have to find? So, that means the maximum discriminant function corresponds to Minimum Euclidean distance. So, I have to find a Euclidean distance and what is the, what is the classification decision? I have to find a maximum discriminant function that corresponds to minimum Euclidean distance from the respective mean points.

And suppose, if I consider the Euclidean distance is equal to constant and then in this case I will be getting the curves of the circle I will be getting so, maybe I can get the curves of the circle in case of a 2 dimensional case and if I consider the high dimensional case I will be getting hyper sphere. So, that is the Euclidean, Euclidean distance contour I can determine. So, corresponding to this I can draw this.

So, these are my contours that means, contour of equal Euclidean distance. So, two classes I can consider the contour of equal Euclidean distance and you can see these vector is the weight vector. So, suppose these is a class, class 1 and this is a class 2 and I can draw the decision boundary. So, these are decision boundary between these two classes, this is the decision boundary.

So, if I consider the 2-dimensional case then in this case I will be getting the curves of circle and if I consider the high dimensional case then I will be getting the hyper sphere. So, these are

nothing but the contours of equal Euclidean distance, the Euclidean distances are the contours of equal Euclidean distance. The second case what I am considering the second case is non diagonal covariance matrix, I am considering the non-diagonal covariance matrix.

So, that means, in this case also I have to maximize the discriminant function. So, maximizing the discriminant function is equivalent to minimizing the covariance matrix norm. So, that means, I have to maximize the discriminant function $g_i(x)$ that corresponds to the minimization of sigma to the power minus 1 that I have to minimize. So, for this I can determine the Mahalanobis distance.

So, already I have defined a Mahalanobis distance, so, $x - \mu_i^T x - \frac{1}{2}$, so, I can determine the Mahalanobis distance. So, the minimum distance corresponds to the maximum discriminant function and if I consider d_m is equal to c , then in this case, I will be getting the contours. So, maybe I can get the curves of ellipse. So, for the high dimensional case, I can consider the hyper ellipsoid and for the 2-dimensional case I will be getting the ellipse the contour of the ellipse. So, maybe something like this I will be getting.

In the previous case I, I have the curves of the circles. Now, I will be getting the curves of the ellipse, the ellipse will be like this, this is for the clusters 1 and similarly, for the cluster 2 also I will be getting the curves of the ellipse and this is the vector $\mu_i - \mu_j$ because I have to determine that this μ_i and μ_j . So, this is μ_i and μ_j and in this case already I have defined that your decision boundary, the decision boundary will not be orthogonal it will not be orthogonal to this vector, this vector is nothing but $\mu_i - \mu_j$, this vector is $\mu_i - \mu_j$.

So, you can see, I have considered the minimum distance classifier based on these two distances one is the Euclidean distance another one is the Mahalanobis distance. In this class I discussed the concept of the Bayesian classification first I discussed the concept of Probability of Error and after this I discussed the Concept of Risks. So, by considering the probability of error, I can take a classification decision.

Similarly, by considering their risks, I can also take a classification decision after this I discussed the concept of Discriminant Function. So, for c number of classes, I have c number of

discriminant functions and based on these discriminant function, I can take a classification decision. After this I consider the Normal Distribution one is the Univariate Distribution another one is the Multivariate Normal Distribution.

After this I determine the discriminant function, the discriminant function is $g_i(x)$ and after this I consider two cases one is the Diagonal Covariance Matrix and another one is the Non Diagonal Covariance Matrix, for this I determined the decision boundary in one case the decision boundary will be orthogonal to the vector w that is a Weight vector.

And in the second case, the decision boundary is not orthogonal to the weight vector, the weight vector is w the weight vector is $\mu_i - \mu_j$. And after this I discussed the concept of the Minimum Distance Classifier. So, let me stop here today. Thank you.