Neural Networks for Signal Processing-I

Prof. Shayan Srinivasa Garani

Department of Electronic Systems Engineering

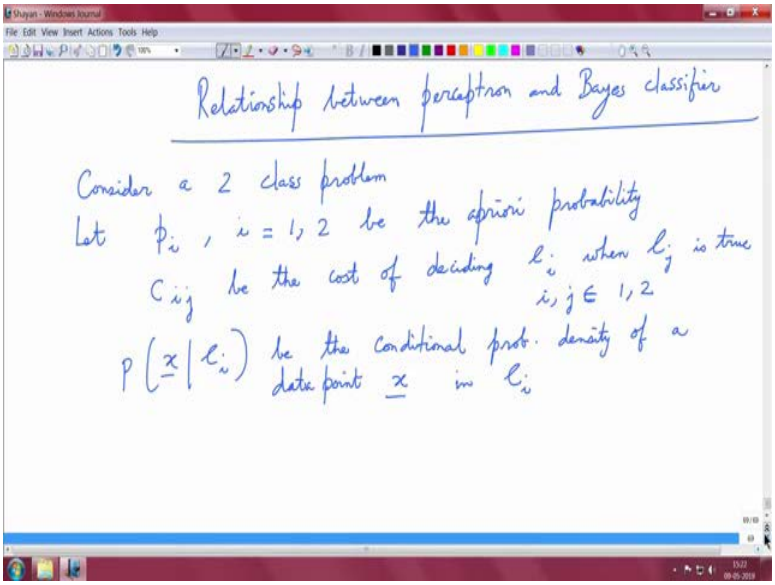
Indian Institute of Science, Bengaluru

Lecture – 11

Perception and Bayes Classifier

Having explored the perceptron algorithm and derived the method to obtain an optimal hyperplane for separating linearly separable classes, we can now delve into the relationship between the perceptron and the Bayes classifier. Let's examine this connection in detail.

(Refer Slide Time: 00:50)

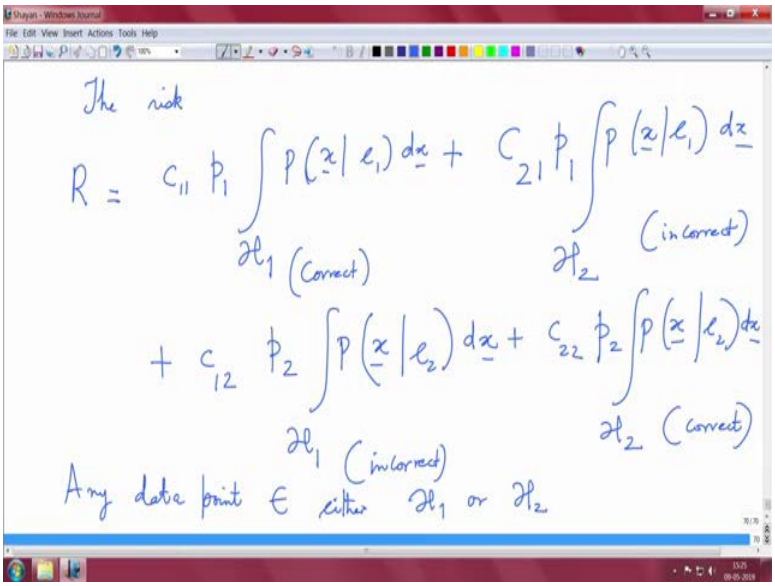


To address this, we must first establish a two-class problem. Consider a scenario involving two classes. Let p(i) for i = 1, 2 represent the a priori probability, which provides information about the likelihood of a particular data point belonging to either class 1 or class 2. Essentially, this is the a priori probability.

Let C_{ij} denote the cost of deciding that a data point belongs to class C_i when, in reality, it belongs to class C_j. This means there is a cost associated with misclassifying a data point from class C_j as class C_i. This cost applies to all combinations of i and j from the set {1, 2}. In other words, i can take values 1 and 2, and j can also take values 1 and 2.

Next, let $p(x \mid C_i)$ represent the conditional probability density of a data point x given that it belongs to class C_i. With this setup in place, we will proceed to formulate a risk function.

(Refer Slide Time: 03:09)



So, the risk R is given by the following expression:

$$R = C_{11} \cdot p_1 \int_{H_1} p(x \mid C_1)\, dx$$

Here, C_{11} is multiplied by the a priori probability p_1 and integrated over the space H_1, assuming the true class is 1 and it is correctly labeled as class 1. The a priori probability p_1 is associated with the true class 1.

Now, if there is a misclassification, where we start with class 1 (correct) but incorrectly label it as class 2, the a priori probability associated is still p_1, but we now integrate over the space H_2:

$$C_{21} \cdot p_1 \int_{H_2} p(x \mid C_1)\, dx$$

This means the true class is 1, but it is misclassified as 2, incurring a cost associated with C_{21}.

Similarly, we have:

$$C_{12} \cdot p_2 \int_{H_1} p(x \mid C_2)\, dx$$

This indicates a misclassification where the true class is 2, but it is incorrectly labeled as 1. The a priori probability associated is p_2, and the integration is over the space H_1.

Finally, if the classification is correct for class 2, we have:

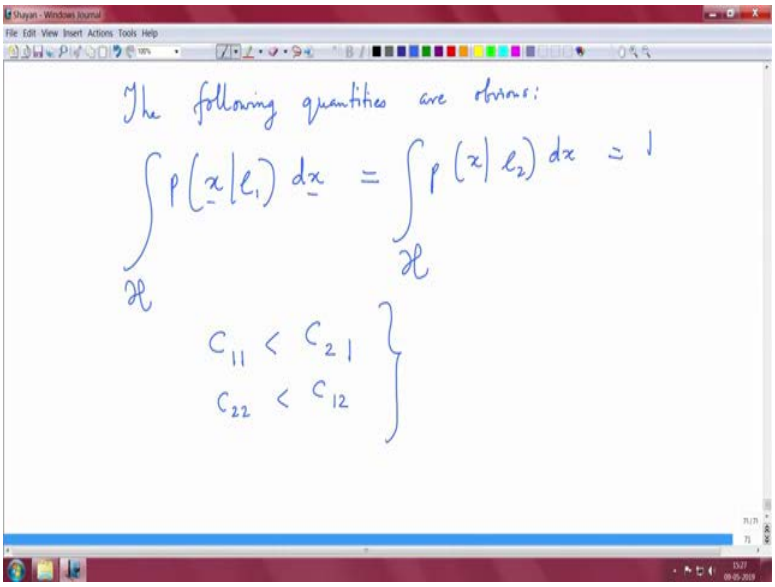
$$C_{22} \cdot p_2 \int_{H_2} p(x \mid C_2)\, dx$$

Here, the true class is 2, it is correctly labeled as 2, and the a priori probability associated is p_2. The integration is over the space H_2.

Each of these integrals involves the conditional probability density $p(x \mid C_i)$ over their respective spaces H_1 or H_2.

Now, it's essential to note that any vector or data point belongs to either the space H_1 or the space H_2. With this understanding, let's proceed to the next step.

(Refer Slide Time: 06:24)



The following quantities are evident. First, the integral of the conditional distribution $p(x \mid C_1)$ over the entire observation space is equal to 1:

$$\int p(x \mid C_1)\, dx = 1$$

Similarly, this holds true for the conditional density $p(x \mid C_2)$:

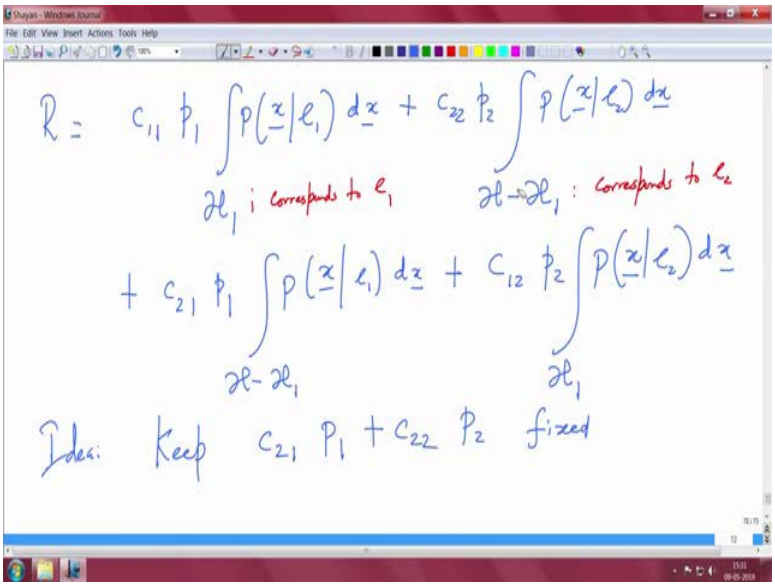
$$\int p(x \mid C_2)\, dx = 1$$

This result is intuitive and aligns with practical experience.

The cost associated with labeling a data point correctly as class 1, given it is indeed class 1, should be less than the cost associated with misclassifying it as class 2 when it actually belongs to class 1. This is because a higher penalty is incurred with misclassification, which reflects a higher cost. This principle is also supported by practical experience and should be clear.

With these understandings, we will simplify the risk associated with the two-class problem.

(Refer Slide Time: 07:56)



Let's rewrite this risk R. Recall that C_{ij} represents the cost where j is the true class label, and i is the class you are assigning to the data point.

So, the risk R is expressed as:

$$R = C_{11} \cdot p_1 \int_{H_1} p(x \mid C_1)\, dx + C_{22} \cdot p_2 \int_{H_2} p(x \mid C_2)\, dx$$

Here, H_1 corresponds to the observation space for class 1, and H_2 (which can be considered as H - H_1) corresponds to class 2.

Next, we add the terms involving misclassification costs:

$$R = C_{21} \cdot p_1 \int_{H_2} p(x \mid C_1)\, dx + C_{12} \cdot p_2 \int_{H_1} p(x \mid C_2)\, dx$$

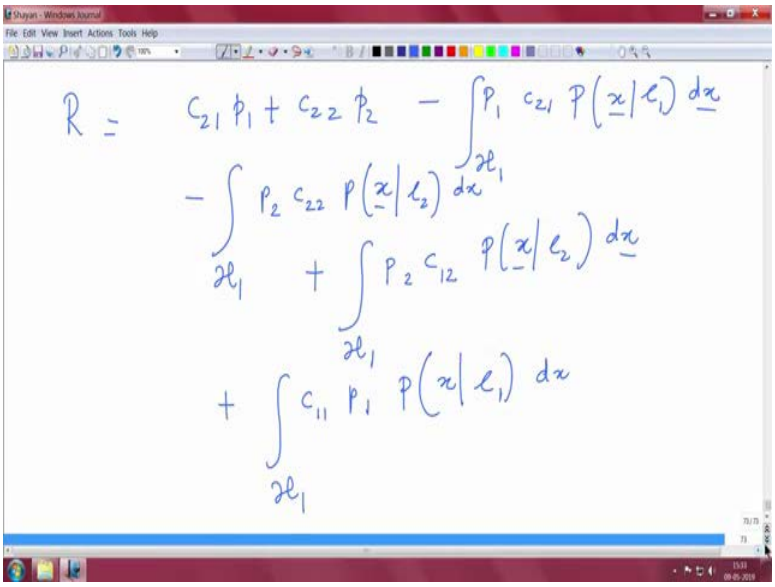
Here, H_2 corresponds to the space where we misclassify class 1 as class 2, and H_1 is the space where we misclassify class 2 as class 1.

To further simplify this, we aim to keep the term $C_{21} \cdot p_1 + C_{22} \cdot p_2$ fixed. Let's assume we want to examine the cost of labeling a data point as class 2. This cost can arise from correctly labeling class 2 or from misclassifying class 1 as class 2. We need to fix this cost and minimize the remaining quantities.

Thus, keeping the cost of labeling as class 2 fixed, we proceed to minimize the other components. Now, let's simplify this expression further:

$$R = C_{11} \cdot p_1 \int_{H_1} p(x \mid C_1)\, dx + C_{22} \cdot p_2 \int_{H_2} p(x \mid C_2)\, dx + C_{21} \cdot p_1 \int_{H_2} p(x \mid C_1)\, dx + C_{12} \cdot p_2 \int_{H_1} p(x \mid C_2)\, dx$$

(Refer Slide Time: 10:51)



By fixing the cost terms, we can focus on minimizing the integrals associated with misclassification. This approach will simplify the risk calculation for the two-class problem.

So, we start with the risk R expressed as $C_{21}p_1 + C_{22}p_2$, and we fix this value. Now, if we examine the integral, we see C_{21} p_1 integrated over the space H_2. The probability measure π associated with H_1 and H_2 must sum to 1. Thus, the probability measure for H_2 is 1 - π for H_1.

Therefore, we need to adjust the integral by subtracting the appropriate quantity. Specifically, we subtract $\int_{H_2} p(x \mid C_1)\ dx$ for C_{21} p_1, giving us:

$$\int_{H_2} p_1 C_{21} p(x \mid C_1)\ dx$$

Similarly, for C_{22}, we need to adjust the integral. Here, we have p_2 C_{22} integrated over H_2, and since we want to express this in terms of H_1, we subtract:

$$\int_{H_1} p_2 C_{22} p(x \mid C_2)\ dx$$

We also include the other terms, maintaining the integrity of the original quantities:

$$\int_{H_1} p_2 C_{12} p(x \mid C_2)\ dx + \int_{H_1} C_{11} p_1 p(x \mid C_1)\ dx$$

By converting all terms to be in terms of H_1, we simplify the expression accordingly. This careful adjustment ensures that the total probability remains consistent across the spaces H_1 and H_2. Thus, we achieve a simplified and coherent risk function:

$$R = C_{21}p_1 - \int_{H_2} p_1 C_{21} p(x \mid C_1)\ dx + C_{22}p_2 - \int_{H_1} p_2 C_{22} p(x \mid C_2)\ dx + \int_{H_1} p_2 C_{12} p(x \mid C_2)\ dx + \int_{H_1} C_{11} p_1 p(x \mid C_1)\ dx$$

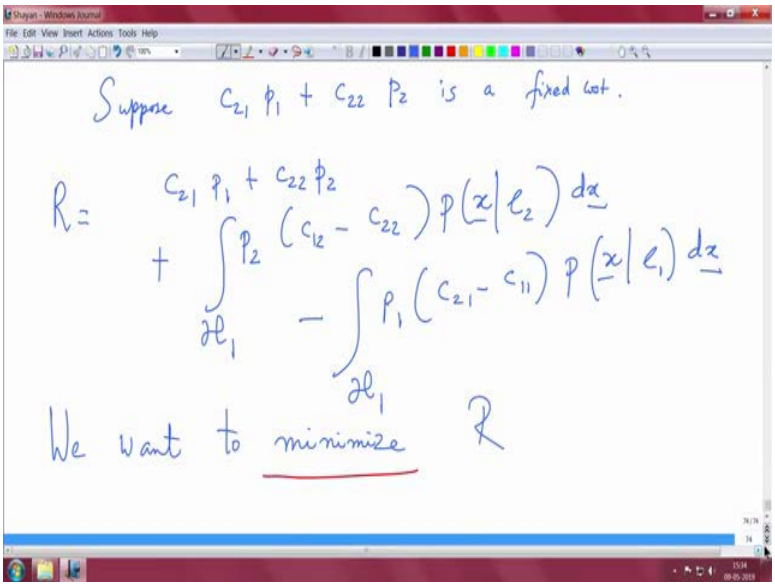
This approach ensures that we accurately account for all probabilities and costs, presenting a clear and precise risk function for the two-class problem.

Now, suppose $C_{21}p_1 + C_{22}p_2$ represents a fixed cost. We can simplify the risk R as

follows:

$$R = C_{21}p_1 + C_{22}p_2 + \int_{H_1} [p_2(C_{12} - C_{22})p(x \mid C_2)]dx$$
$$- \int_{H_1} [p_1(C_{21} - C_{11})p(x \mid C_1)]dx$$

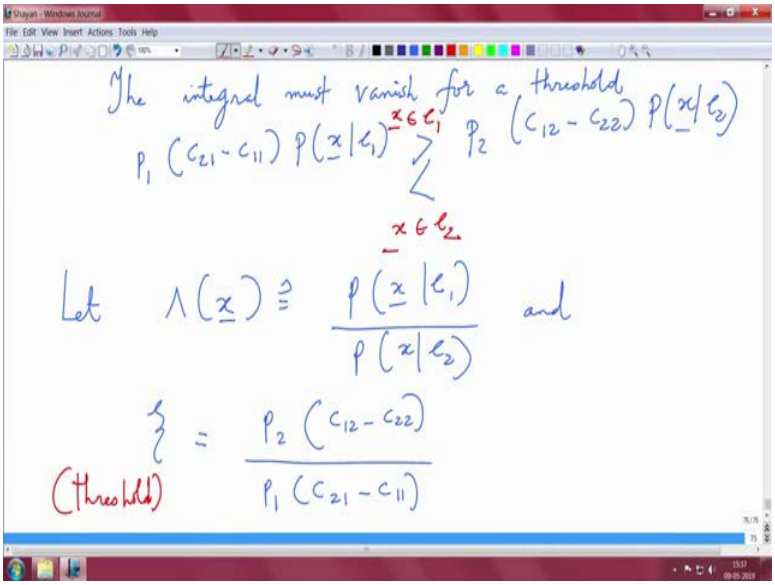
(Refer Slide Time: 13:53)



To obtain a simplified metric, we want to integrate over the space H_1. Our goal is to minimize R. It becomes apparent that if we can make the total value of these integrals equal to zero, we achieve the optimal risk. Therefore, we aim to eliminate these integrals to minimize the risk.

By addressing the integral terms and striving to nullify their contributions, we can effectively reduce the overall risk R and thus achieve the optimal solution for our two-class problem.

(Refer Slide Time: 15:45)



The integral over the space H_1 must vanish for a specific threshold. This occurs if:

$$p_1(C_{21} - C_{11})p(x \mid C_1) = p_2(C_{12} - C_{22})p(x \mid C_2)$$

When this condition holds, the integral vanishes. For the classification problem, we need to consider the inequality between these quantities. If:

$$p_1(C_{21} - C_{11})p(x \mid C_1) > p_2(C_{12} - C_{22})p(x \mid C_2)$$

we decide that x belongs to class 1. Otherwise, x belongs to class 2.

Let λ(x) be defined as the conditional likelihood ratio:

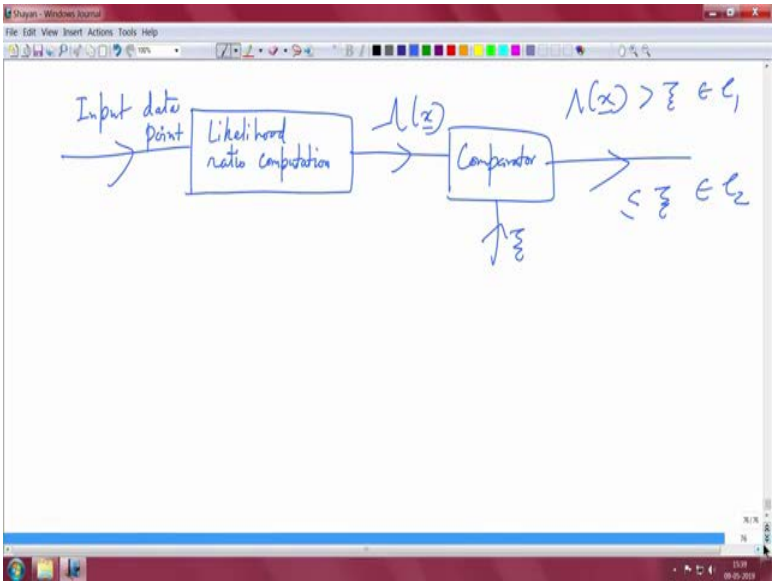
$$\lambda(x) = \frac{p(x \mid C_1)}{p(x \mid C_2)}$$

And let ζ be defined as:

$$\zeta = \frac{p_2(C_{12} - C_{22})}{p_1(C_{21} - C_{11})}$$

We use this likelihood ratio statistic and compare it against the threshold ζ. This threshold helps in making the decision between the two classes. Therefore, if λ(x) > ζ, we classify x as belonging to class 1; otherwise, we classify it as belonging to class 2. This forms the basis of our decision rule.

(Refer Slide Time: 18:34)

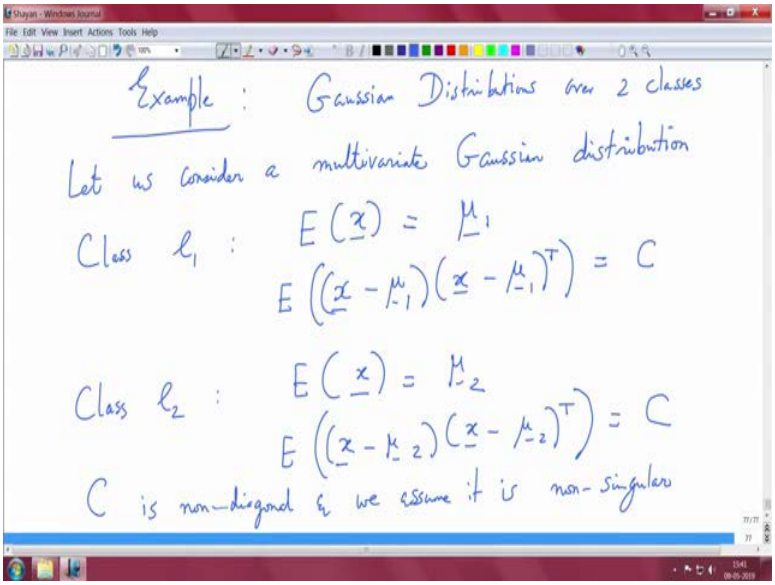


Now, if we sketch this as a schematic, we start with a data point that goes to a likelihood ratio computation unit. Here, we calculate λ(x) and compare it against ζ. The decision rule is formulated as follows: if λ(x) > ζ, we classify x as belonging to class 1; otherwise, it belongs to class 2.

There is a significant advantage in performing this likelihood operation in the log domain, which we will illustrate with an example. This is the fundamental idea behind a Bayes two-class classification problem that minimizes a certain risk.

Next, we will examine this in the context of Gaussian distributions. We will carry out a simple exercise and try to link this example to the perceptron problem, which was our original objective. So, with this setup in mind—based on the likelihood ratio computation and the comparator for making the decision in the two-class problem—let us explore the Gaussian distribution case.

(Refer Slide Time: 20:17)



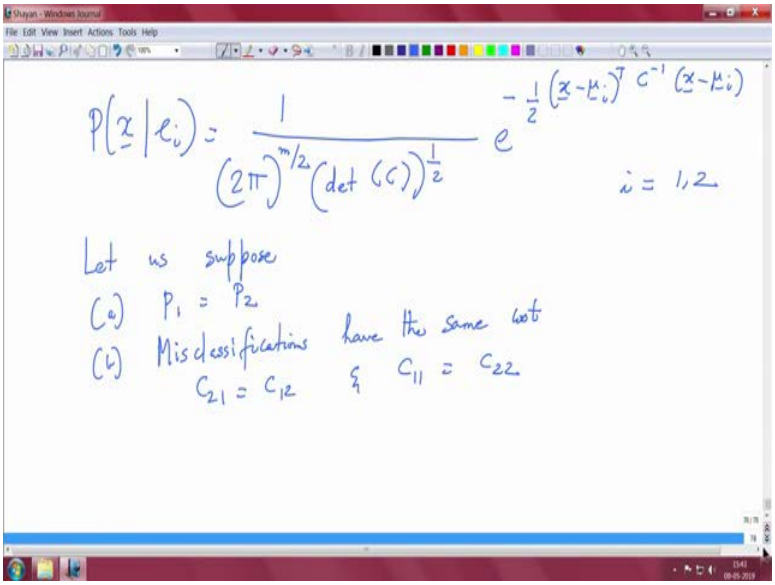
Let's start with an example. Consider a multivariate Gaussian distribution. For class C_1, we have the following quantities for its mean and covariance: the mean vector μ_1 and the covariance matrix σ, which is given by $(x - \mu_1)(x - \mu_1)^T$.

For the sake of simplification, let's assume that the covariance matrices for both classes are the same and denote it as σ. While in a more general version, we could have different covariance matrices, σ_1 and σ_2, for our example, we will assume they are identical.

Now, consider another class C_2. The mean vector for this class is μ_2, but the covariance matrix remains the same as σ. Generally, σ is a non-diagonal and non-singular matrix.

This setup allows us to explore the classification problem under the assumption of shared covariance, simplifying our calculations and focusing on the distinction between the mean vectors of the two classes.

(Refer Slide Time: 22:50)



Let's determine the conditional densities. For the multivariate Gaussian distribution, the conditional density p(x | C_i) is given by:

$$p(x \mid C_i) = \frac{1}{(2\pi)^{m/2}\ \det(\sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \sigma^{-1}(x-\mu_i)\right)$$

where m represents the number of variables in the random vector x. Here, σ is the covariance matrix. The determinant of σ should not be zero, which is why we assume σ is non-singular.

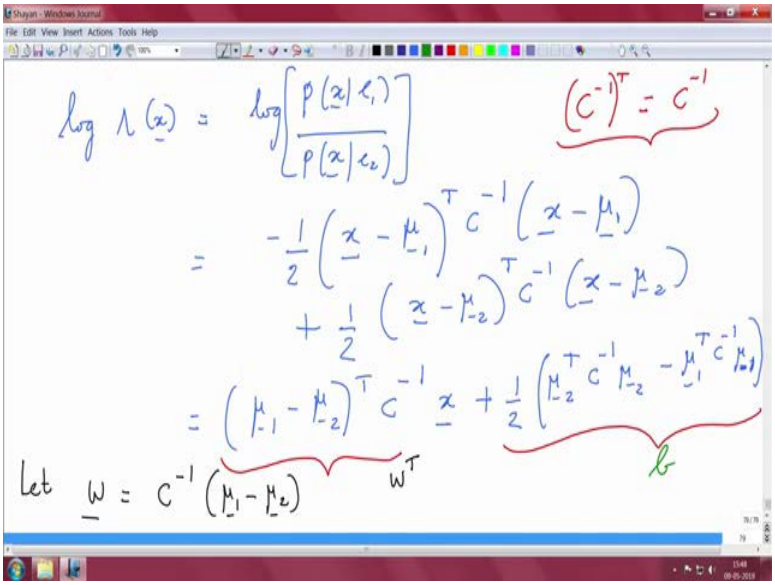
For simplicity, let's make the following assumptions:

1. $p_1 = p_2$, which simplifies our calculations, though this is not always the case in general.

2. The costs of misclassification are equal, meaning $C_{21} = C_{12}$ and $C_{11} = C_{22}$.

These assumptions will help streamline our analysis and calculations.

(Refer Slide Time: 24:48)



Now, let's analyze the log of the likelihood ratio λ(x). We define λ(x) as:

$$\lambda(x) = \frac{p(x \mid C_1)}{p(x \mid C_2)}$$

Taking the logarithm of λ(x), we get:

$$\log \lambda(x) = \log\left(\frac{p(x \mid C_1)}{p(x \mid C_2)}\right)$$

Since the normalization constants $\frac{1}{(2\pi)^{m/2}\sqrt{\det(\sigma)}}$ cancel out, we only need to consider the exponentials. Simplifying, we have:

$$\log \lambda(x) = -\frac{1}{2}(x - \mu_1)^T \sigma^{-1}(x - \mu_1) + \frac{1}{2}(x - \mu_2)^T \sigma^{-1}(x - \mu_2)$$

Upon further algebraic manipulation, this simplifies to:

$$\log \lambda(x) = (\mu_1 - \mu_2)^T \sigma^{-1} x + \frac{1}{2}[(\mu_2^T \sigma^{-1} \mu_2) - (\mu_1^T \sigma^{-1} \mu_1)]$$

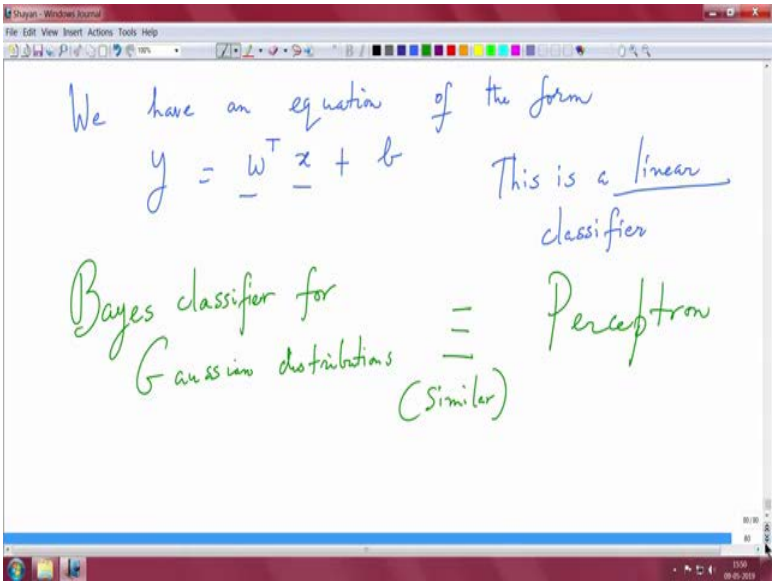
To simplify this, let us define $w = \sigma^{-1}(\mu_1 - \mu_2)$. Consequently, the expression becomes:

$$\log \lambda(x) = w^T x + \frac{1}{2}[\mu_2^T \sigma^{-1} \mu_2 - \mu_1^T \sigma^{-1} \mu_1]$$

Here, w^T represents $\sigma^{-1}(\mu_1 - \mu_2)^T$. Note that σ^{-1} is symmetric, so $(\sigma^{-1})^T = \sigma^{-1}$. This symmetry

property of the covariance matrix σ is fundamental to the simplification and is a standard result in linear algebra.

(Refer Slide Time: 29:34)



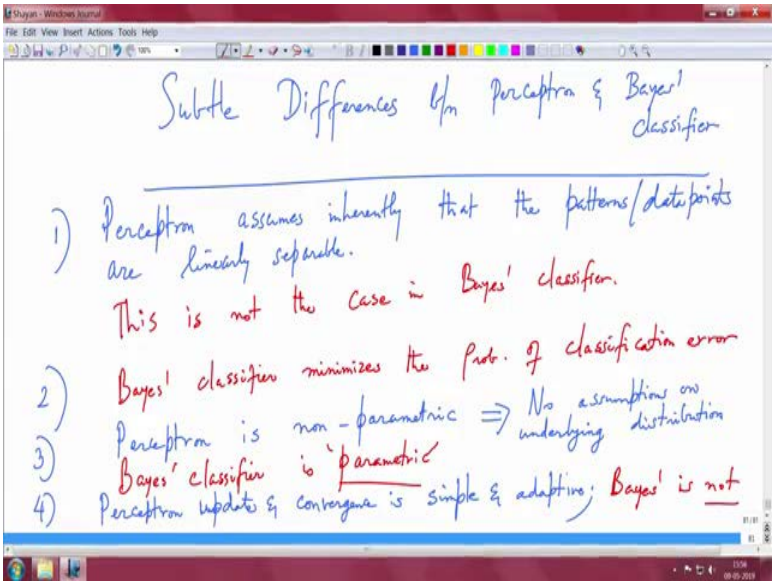
Let's denote the quantity we have as the bias b. With this, we have an equation of the form:

$$y = w^T x + b$$

This represents a linear classifier.

What we want to demonstrate is that the Bayes classifier for Gaussian distributions, assuming equal covariance matrices but different means, resembles the perceptron in that it also derives a hyperplane for classification. However, there are subtle differences between the two methods that need to be addressed. We will highlight these differences in our discussion.

(Refer Slide Time: 31:25)



Here are the subtle differences between the Perceptron and the Bayes classifier under the given constraints:

1. Linearly Separable Patterns:

- Perceptron: Operates under the assumption that patterns are linearly separable. This means that for the perceptron to work effectively, the data points or patterns must be separable by a linear boundary. If this condition is not met, the perceptron may not perform well.

- Bayes Classifier: Does not assume linear separability. It accounts for the possibility of overlapping patterns and calculates the decision boundary accordingly. This means the Bayes classifier can handle cases where patterns overlap and is capable of dealing with such complexities.

2. Error Minimization:

- Bayes Classifier: Minimizes the probability of classification error. It aims to reduce the overall probability of making an incorrect classification by accounting for the distribution of data points and their overlap.

- Perceptron: Does not explicitly minimize classification error in the probabilistic sense. Instead, it assumes that classes are separable and focuses on finding a linear boundary that separates them.

3. Parametric vs. Non-Parametric:

- Perceptron: Non-parametric. It does not make assumptions about the underlying distribution of the data points. This flexibility means it can adapt to the data without needing to estimate the distribution.

- Bayes Classifier: Parametric. It relies on assumptions about the data distribution, such as Gaussian distributions with known parameters. Changes in the distribution or covariance matrices require reevaluation of the classifier.

4. Adaptability and Convergence:

- Perceptron: Simple and adaptive. It can update weights and biases incrementally as new data points are received. It is well-suited for non-stationary environments where data statistics change over time.

- Bayes Classifier: Less adaptive in practice. If the covariance matrices or other parameters change, it requires a complete recalculation or derivation from first principles, which can be complex.

5. Handling Non-Stationary Data:

- Perceptron: Handles changes in data statistics effectively. As it receives new data points, it updates the weights and biases accordingly without needing to rederive the classifier.

- Bayes Classifier: Struggles with non-stationary data. Adapting to changes in data statistics requires recalculating the model, which can be challenging.

- These distinctions highlight the strengths and limitations of each approach. When implementing these algorithms or choosing between them, understanding these differences will help in selecting the best method for your specific application.

With this overview, we conclude this module. In the next module, we will explore the backpropagation algorithm, which extends the perceptron to multiple layers, forming a multilayer perceptron. We will discuss its architecture and applications in the upcoming session.