

# BAYES CLASSIFIER

Dr. Umarani Jayaraman  
Assistant Professor



INDIAN INSTITUTE OF INFORMATION TECHNOLOGY,  
DESIGN AND MANUFACTURING,  
KANCHEEPURAM



# Chapter 2

## Bayesian Decision Theory

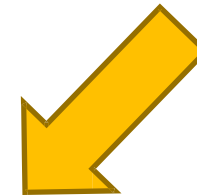
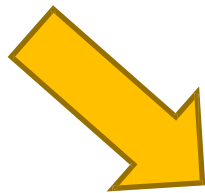
# Decision Theory

## Decision

**Make choice under  
uncertainty**

## Pattern Recognition

**Pattern  $\rightarrow$  Category**



**Given a test sample, its category is uncertain and a decision has to be made**



**In essence, PR is a decision process**

# Bayesian Decision Theory

Bayesian decision theory is a **statistical approach** to pattern recognition

*The fundamentals of most PR algorithms are rooted from Bayesian decision theory*

## Basic Assumptions

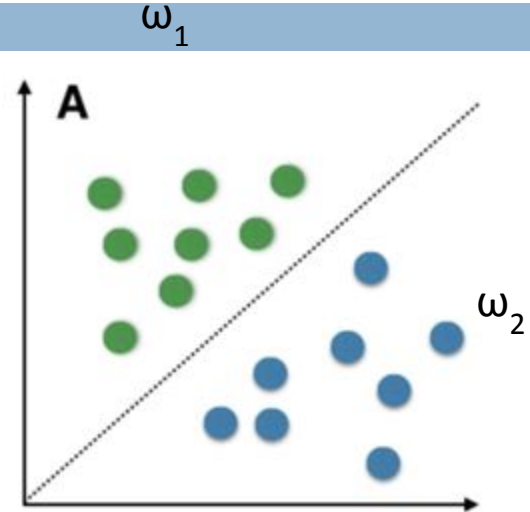
- The decision problem is posed (formalized) in **probabilistic** terms
- All the relevant probability values are known

**Key Principle**

**Bayes Theorem**

# Linear separable classes

Let  $\omega_1$  and  $\omega_2$   
**are two classes**

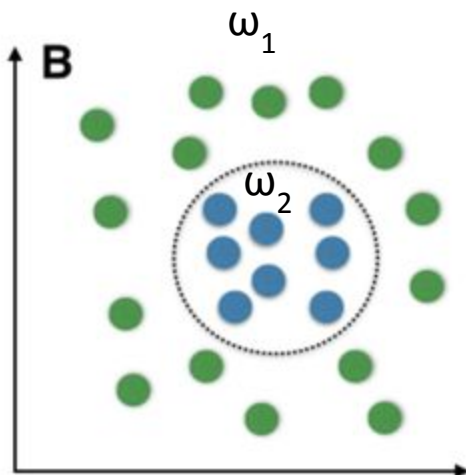


- ❑ In this case, the two classes can be separated by linear boundary, this is also known as linearly separable classes
- ❑ This is supervised learning

# Non linear separable classes

**Among these non-linear separable classes, the most common are**

- 1. Quadratic classifier**
- 2. Cubic classifier**

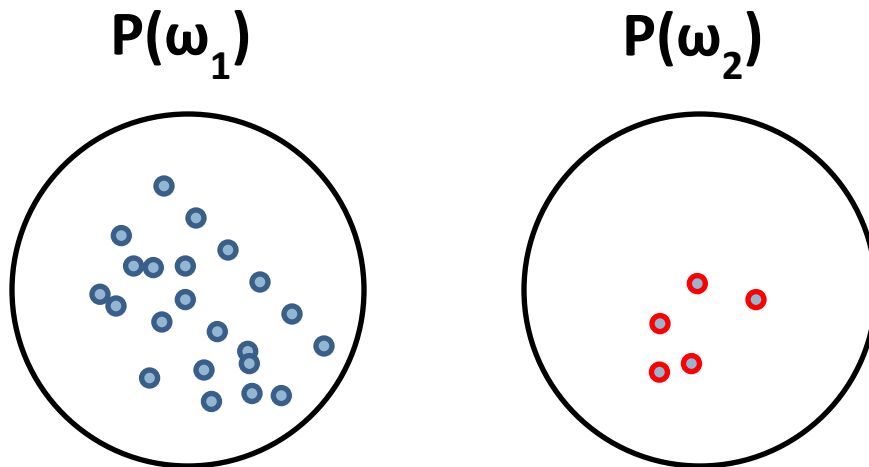


# Two class problem

- ❑ Consider any manufacturing company which produces goods for example steel plant
- ❑ Quality control department should take the decision in which of the two classes it should go
- ❑  $\omega_1$  -> accept
- ❑  $\omega_2$  -> reject

# Two class problem

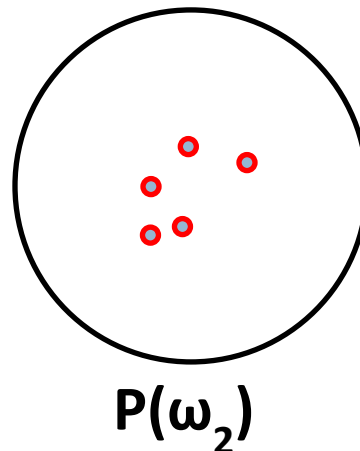
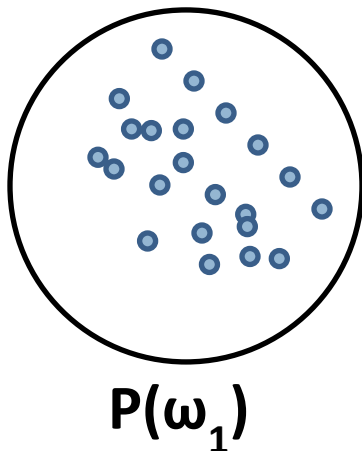
- ❑  $\omega_1 \rightarrow$  accept
- ❑  $\omega_2 \rightarrow$  reject
- ❑ We may take the previous history (i.e) how many objects are accepted and how many are rejected by the quality control department





# Two class problem

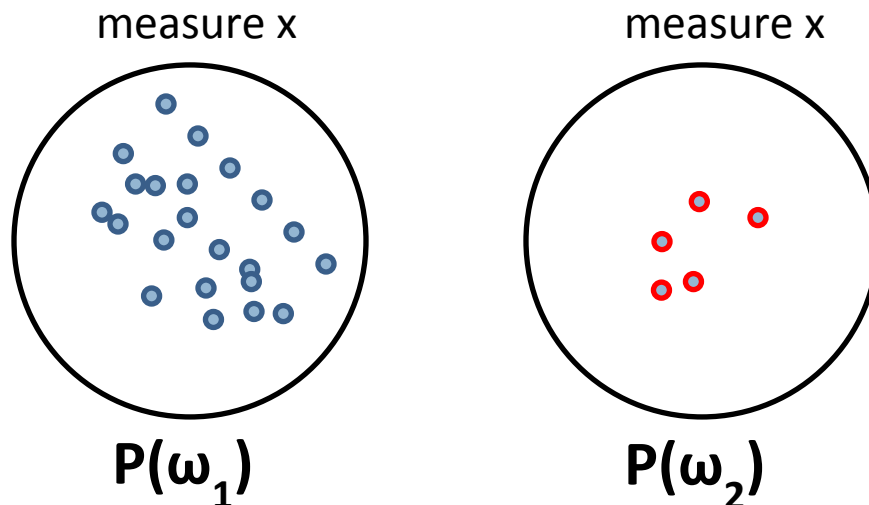
- ❑  $P(\omega_1) > P(\omega_2) \Rightarrow \omega_1$
- ❑  $P(\omega_1) < P(\omega_2) \Rightarrow \omega_2$
- ❑ But, this is not really logical because the object is always accepted or always be rejected based on a priori probability (i.e)  $p(\omega_1)$  and  $p(\omega_2)$



**Solution:**  
Incorporate  
observations  
into decision!

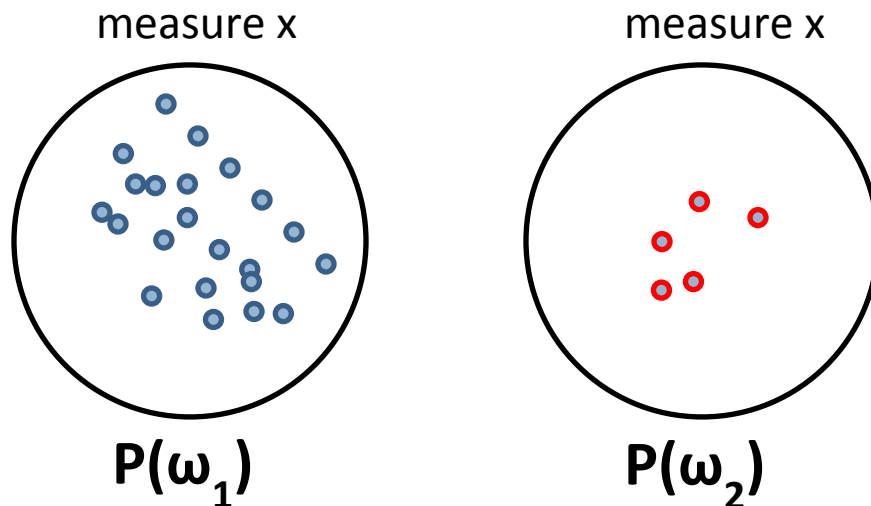
# Two class problem

- ❑ Let the observation be  $x$
- ❑ We can find out  $P(x/\omega_1)$  and  $P(x/\omega_2)$
- ❑ It is nothing but probability density function  $x$  taking the objects from class  $\omega_1$  and  $\omega_2$  respectively (**class conditional PDF**)



# Two class problem

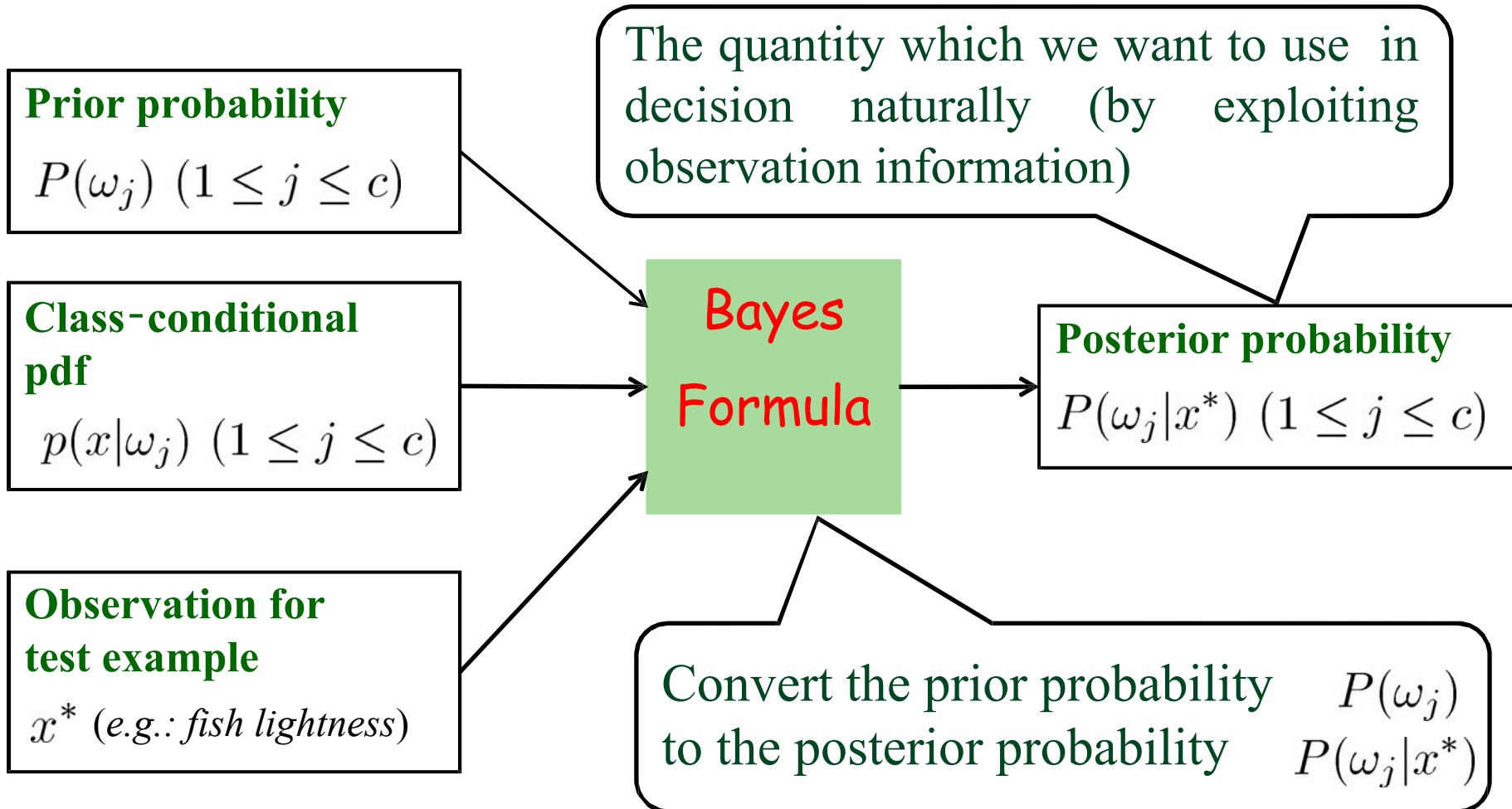
- ❑ Now the decision rule may be
- ❑  $P(\omega_1/x) > P(\omega_2/x) \Rightarrow \omega_1$ , in favor of class  $\omega_1$
- ❑  $P(\omega_1/x) < P(\omega_2/x) \Rightarrow \omega_2$ , in favor of class  $\omega_2$
- ❑ A more logical will be if this  $P(\omega_1/x)$  and  $P(\omega_2/x)$  can be combined with **a priori probability**  $P(\omega_1)$  and  $P(\omega_2)$



# Decision After Observation

**Known**

**Unknown**



# Bayes Formula Revisited

From the preliminary probability theory,

**Joint probability density function (Joint PDF)**

$$p(\omega, x)$$



**Law of total probability**

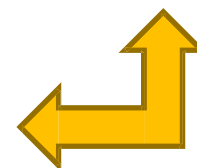
$$p(\omega, x) = P(\omega|x) \cdot p(x)$$

$$p(\omega, x) = P(\omega) \cdot p(x|\omega)$$



$$P(\omega|x) \cdot p(x) = P(\omega) \cdot p(x|\omega)$$

$$P(\omega|x) = \frac{p(x|\omega) \cdot P(\omega)}{p(x)}$$



# Bayes Formula Revisited (Cont.)

$$P(\omega_j|x) = \frac{p(x|\omega_j) \cdot P(\omega_j)}{p(x)} \quad (1 \leq j \leq c) \quad (\text{Bayes Formula})$$

## Bayes Decision Rule

$$\boxed{\text{if } P(\omega_j|x) > P(\omega_i|x), \forall i \neq j \implies \text{Decide } \omega_j}$$

- $P(\omega_j)$  and  $p(x|\omega_j)$  are **assumed to be known**
- $p(x)$  is **irrelevant** for Bayesian decision (serving as a normalization factor, not related to any state of nature)

$$p(x) = \sum_{j=1}^c p(\omega_j, x) = \sum_{j=1}^c p(x|\omega_j) \cdot P(\omega_j)$$

# Bayes Formula Revisited (Cont.)

$$P(\omega_j|x) = \frac{p(x|\omega_j) \cdot P(\omega_j)}{p(x)} = P(x/\omega_1) \cdot P(\omega_1) > P(x/\omega_2) \cdot P(\omega_2) \Rightarrow \omega_1$$

## Special Case I: Equal prior probability

$$P(\omega_1) = P(\omega_2) = \dots = P(\omega_c) = \frac{1}{c} \quad \longrightarrow \quad \begin{array}{l} \text{Depends on the} \\ \text{likelihood } P(x|\omega_j) \end{array}$$

## Special Case II: Equal likelihood

$$p(x|\omega_1) = p(x|\omega_2) = \dots = p(x|\omega_c) \quad \longrightarrow \quad \begin{array}{l} \text{Depends on a priori} \\ \text{probability } P(\omega_j) \end{array}$$

**Special Case III:** otherwise, prior probability and likelihood function together in Bayesian decision process

# Bayes Theorem

$$\text{Bayes theorem} \quad P(H|X) = \frac{P(H)P(X|H)}{P(X)}$$

$X$ : the observed sample (also called **evidence**; e.g.: *the length of a fish*)  $H$ : the hypothesis (e.g. *the fish belongs to the “salmon” category*)

$P(H)$ : the **prior probability** that  $H$  holds (e.g. *the probability of catching a salmon*)

$P(X|H)$ : the **likelihood** of observing  $X$  given that  $H$  holds (e.g. *the probability of observing a 3 -inch length fish which is salmon*)

$P(X)$ : the **evidence probability** that  $X$  is observed (e.g. *the probability of observing a fish with 3 -inch length*)

$P(H|X)$ : the **posterior probability** that  $H$  holds given  $X$  (e.g. *the probability of  $X$  being salmon given its length is 3 -inch*)



**Thomas Bayes**  
(1702-1761)



# Bayes Classifier

$$P(\omega_j|x) = \frac{p(x|\omega_j) \cdot P(\omega_j)}{p(x)} \quad (1 \leq j \leq c) \quad (\text{Bayes Formula})$$

[if  $P(\omega_j|x) > P(\omega_i|x), \forall i \neq j \implies$  Decide  $\omega_j$ ]

# Example 1

18

- Two boxes B1 and B2 contain 100 and 200 light bulbs respectively. The first box (B1) has 15 defective bulbs and the second has 5 defective bulbs
  - a) Suppose a box is selected at random and one bulb is picked out. What is the probability it is defective?
  - b) Suppose the bulb we tested was defective what is the probability it came from box 1?

# Example 1- cont.

19

	Defective	Not defective
B1	15	85
B2	5	195

# Example 1- cont.

20

- Since the box is selected at random they are equally likely
- $P(B1) = P(B2) = \frac{1}{2} = 0.5$
- $P(D/B1) = 15/100 = 0.15$
- $P(D/B2) = 5/100 = 0.05$
- $P(D) = P(D/B1) \cdot P(B1) + P(D/B2) \cdot P(B2)$
- $P(D) = (0.15 \times 0.5) + (0.05 \times 0.5) = 0.05$
- Thus there is about 5% probability a bulb is defective.

# Example 1- cont.

21

- $P(B1/D) = P(D/B1) \cdot P(B1) / P(D)$
- $P(B1/D) = 0.15 \times 0.5 / 0.0875 = 0.8571$
- $0.8571 > 0.5$
- Recall box 1 has three times more defective bulbs compared to box 2

# Example 2

22

## □ Problem statement

- A new medical test is used to detect whether a patient has a certain cancer or not, whose test result is either  $+$  (*positive*) or  $-$  (*negative*)
- For patient with this cancer, the probability of returning *positive* test result is 0.98
- For patient without this cancer, the probability of returning *negative* test result is 0.97
- The probability for any person to have this cancer is 0.008

## □ Question

- If *positive* test result is returned for some person, does he/she have this kind of cancer or not?

# Example 2- cont.

23

	Positive (+)	Negative (-)
Class $\omega_1$ Cancer	0.98	
Class $\omega_2$ No Cancer		0.97

## Question

If *positive* test result is returned for some person, does he/she have this kind of cancer or not?

Idea:

$$P(\omega_i / +) = ? \Rightarrow P(\omega_1 / +) = ? , P(\omega_2 / +) = ?$$

If  $P(\omega_1 / +) > P(\omega_2 / +) \Rightarrow \omega_1$

If  $P(\omega_1 / +) < P(\omega_2 / +) \Rightarrow \omega_2$

# Example 2 (Cont.)

$\omega_1$  : cancer

$\omega_2$  : no cancer

$x \in \{+, -\}$

$$P(\omega_1) = 0.008$$

$$P(\omega_2) = 1 - P(\omega_1) = 0.992$$

$$P(+ \mid \omega_1) = 0.98$$

$$P(- \mid \omega_1) = 1 - P(+ \mid \omega_1) = 0.02$$

$$P(- \mid \omega_2) = 0.97$$

$$P(+ \mid \omega_2) = 1 - P(- \mid \omega_2) = 0.03$$

$$\begin{aligned} P(\omega_1 \mid +) &= \frac{P(\omega_1)P(+ \mid \omega_1)}{P(+)} = \frac{P(\omega_1)P(+ \mid \omega_1)}{P(\omega_1)P(+ \mid \omega_1) + P(\omega_2)P(+ \mid \omega_2)} \\ &= \frac{0.008 \times 0.98}{0.008 \times 0.98 + 0.992 \times 0.03} = 0.2085 \end{aligned}$$

$$P(\omega_2 \mid +) = 1 - P(\omega_1 \mid +) = 0.7915$$

$$P(\omega_2 \mid +) > P(\omega_1 \mid +)$$

**No cancer!**



# Error in Bayes Classifier

**Probability of Error**

**Bayes Risk Classifier**

**Bayes Minimum Error Rate Classifier**

# Bayes Formula Revisited (Cont.)

$$P(\omega_j|x) = \frac{p(x|\omega_j) \cdot P(\omega_j)}{p(x)} \quad (1 \leq j \leq c) \quad (\text{Bayes Formula})$$

## Bayes Decision Rule

$$\boxed{\text{if } P(\omega_j|x) > P(\omega_i|x), \forall i \neq j \implies \text{Decide } \omega_j}$$

- $P(\omega_j)$  and  $p(x|\omega_j)$  are **assumed to be known**
- $p(x)$  is **irrelevant** for Bayesian decision (serving as a normalization factor, not related to any state of nature)

$$p(x) = \sum_{j=1}^c p(\omega_j, x) = \sum_{j=1}^c p(x|\omega_j) \cdot P(\omega_j)$$

# Bayes Formula Revisited

$$C_{P(\omega_j|x)} = \frac{p(x|\omega_j) \cdot P(\omega_j)}{p(x)} = P(x/\omega_1) \cdot P(\omega_1) > P(x/\omega_2) \cdot P(\omega_2) \Rightarrow \omega_1$$

## Special Case I: Equal prior probability

$$P(\omega_1) = P(\omega_2) = \dots = P(\omega_c) = \frac{1}{c}$$



Depends on the  
likelihood  $P(x|\omega_j)$

## Special Case II: Equal likelihood

$$p(x|\omega_1) = p(x|\omega_2) = \dots = p(x|\omega_c)$$



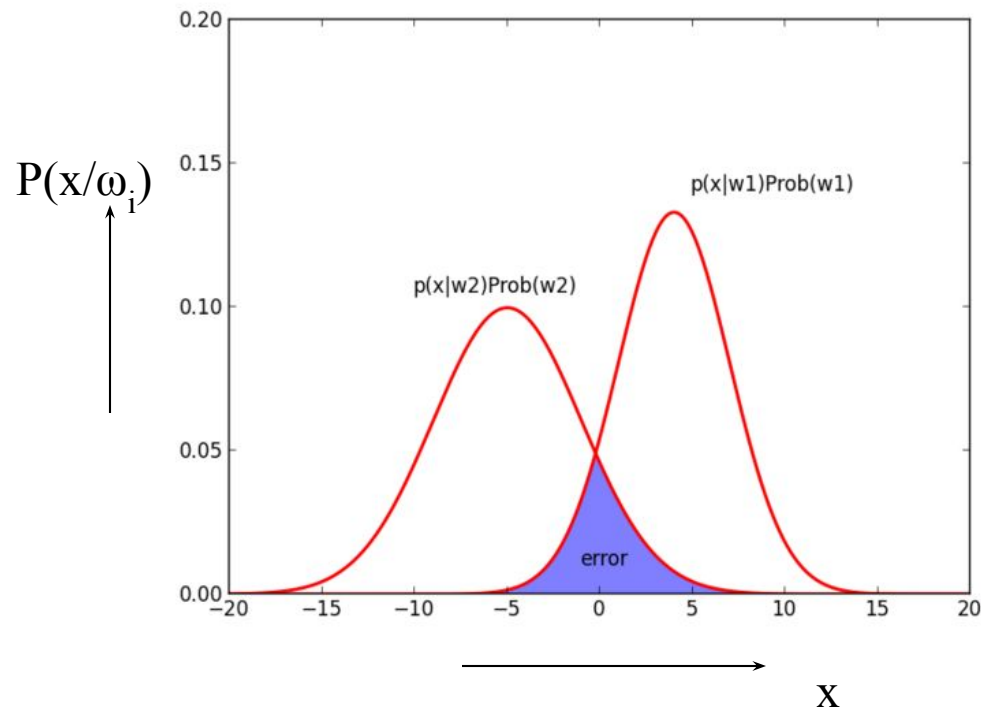
Depends on a priori  
probability  $P(\omega_j)$

**Special Case III:** otherwise, prior probability and likelihood function together in Bayesian decision process

# Error: Probability of error

28

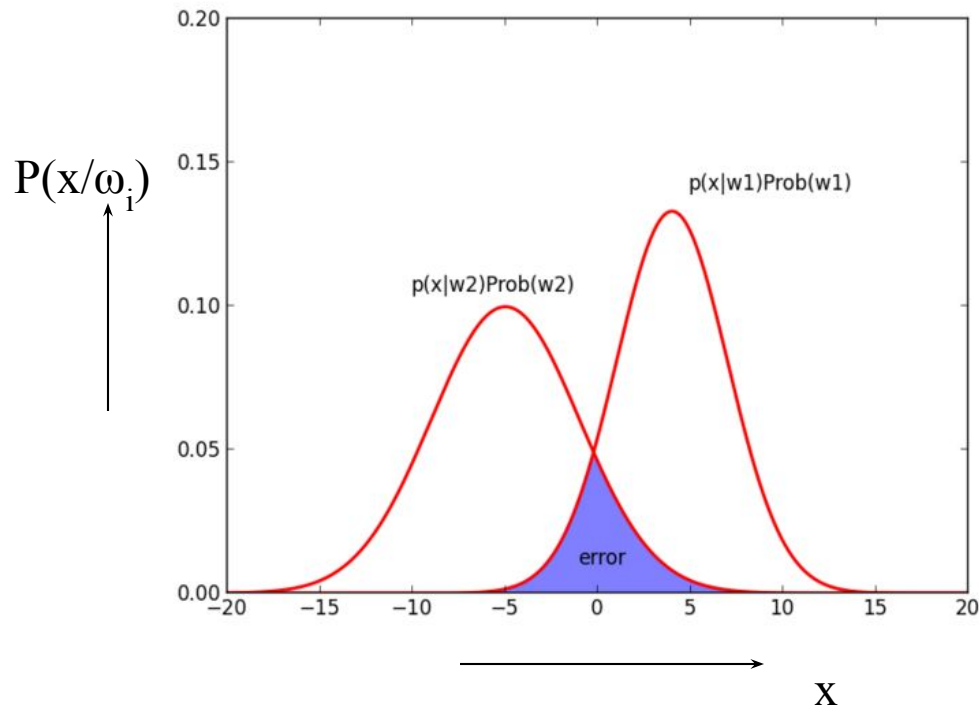
- $P(x/\omega_1) \cdot P(\omega_1) > P(x/\omega_2) \cdot P(\omega_2) \Rightarrow \omega_1$



# Error: Probability of error

29

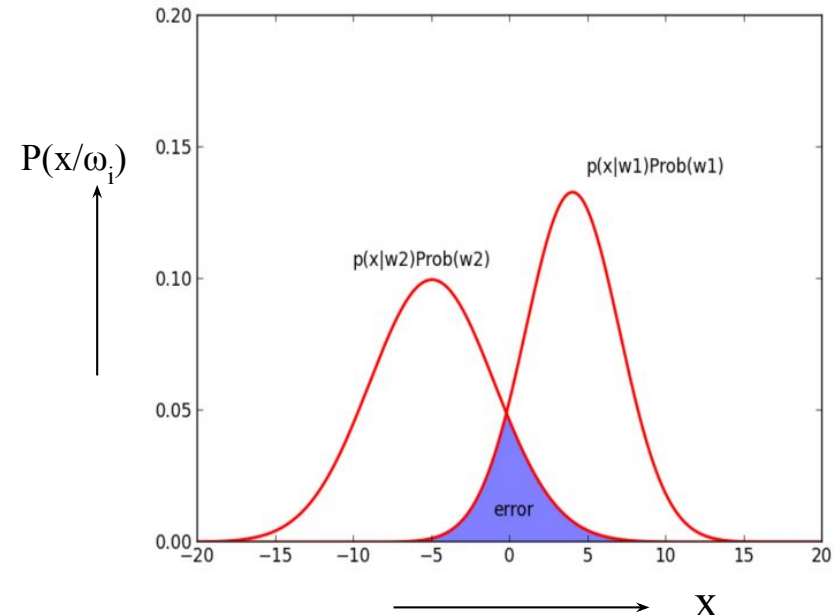
- $P(x/\omega_1) \cdot P(\omega_1) > P(x/\omega_2) \cdot P(\omega_2) \Rightarrow \omega_1$
- There is a finite probability of error
- If we decide in favor of  $\omega_1 \Rightarrow P(\omega_2/x)$
- If we decide in favor of  $\omega_2 \Rightarrow P(\omega_1/x)$



# Error: Probability of error

30

- $P(x/\omega_1) \cdot P(\omega_1) > P(x/\omega_2) \cdot P(\omega_2) \Rightarrow \omega_1$
- There is a finite probability of error
- If we decide in favor of  $\omega_1 \Rightarrow P(\omega_2/x)$
- If we decide in favor of  $\omega_2 \Rightarrow P(\omega_1/x)$
- Given, the situation like this, the total error
- $P(error | x) = \min[P(\omega_1 | x), P(\omega_2 | x)]$



# Error: Probability of error

31

## Bayes Decision Rule (In case of two classes)

if  $P(\omega_1|x) > P(\omega_2|x)$ , Decide  $\omega_1$ ; Otherwise  $\omega_2$

Whenever we observe a particular  $x$ , the **probability of error** is:

$$P(error | x) = \begin{cases} P(\omega_1 | x) & \text{if we decide } \omega_2 \\ P(\omega_2 | x) & \text{if we decide } \omega_1 \end{cases}$$

**Under Bayes decision rule, we have**

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

For every  $x$ , we ensure  
that  $P(error | x)$  is as  
small as possible



The **average probability of error**  
over all possible  $x$  must be as  
small as possible

# Is Bayes Decision Rule Optimal?

32

- For every  $x$ , Bayes classifier ensures that  $P(\text{error}/x)$  is as small as possible
- The **average probability of error** over all possible  $x$  must also be as small as possible
- The Bayes rule minimizes the expected error rate
- Minimizing the expected error rate is a pretty reasonable goal
- However, it is not always the best thing to do.



# Is Bayes Decision Rule Optimal?

33

- Consider the situation
- You are designing a pedestrian detection algorithm for an autonomous navigation system
- Your algorithm must decide whether there is a pedestrian crossing the street
- There could be two possible types of error
  - **False positive:** There is no pedestrian, but the system thinks there is a pedestrian
  - **Miss (false negative):** There is a pedestrian, but the system thinks there is not

# Is Bayes Decision Rule Optimal?

34

- In this situation, should we give equal weight to these two types of error?
- Solution: To deal with these kind of problem instead of minimizing the error rate, we minimize something called the **risk**
- First, we define the loss matrix  $L$ , which quantifies the cost of making each type of error

# Bayes Risk

35

- Element  $\lambda_{ij}$  of the loss matrix specifies the cost of taking action  $\alpha_i$  when the true state of nature is  $\omega_j$
- Typically, we set  $\lambda_{ii} = 0$  for all  $i$
- Thus a typical loss matrix for  $m=2$ , would have the form
- $$\mathbf{L} = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix} = \begin{bmatrix} 0 & \lambda_{12} \\ \lambda_{21} & 0 \end{bmatrix}$$

# Bayes Decision Rule –The General Case

36

- By allowing to use more than one feature

$$x \in \mathbf{R} \implies \mathbf{x} \in \mathbf{R}^d \text{ (} d\text{-dimensional Euclidean space)}$$

- By allowing more than two states of nature

$$\Omega = \{\omega_1, \omega_2, \dots, \omega_c\} \text{ (finite set of } c \text{ states of nature)}$$

- By allowing actions other than merely deciding the state of nature

$$\mathcal{A} = \{\alpha_1, \alpha_2, \dots, \alpha_a\} \text{ (finite set of } a \text{ possible actions)}$$

Note :  $c \neq a$

# Bayes Decision Rule – The General Case (Cont.)

37

- By introducing a loss function more general than the probability of error

$$\lambda(\alpha_i | \omega_j)$$

$\lambda(\alpha_i | \omega_j) \Rightarrow$  the loss incurred for taking action  $\alpha_i$  when the state of nature is  $\omega_j$



For ease of reference,  
usually written as:

$$\lambda_{ij}$$

## A simple loss function

Action Class	$\alpha_1 =$ “Recipe A”	$\alpha_2 =$ “Recipe B”	$\alpha_3 =$ “No Recipe”
$\omega_1 =$ “cancer”	5	50	10,000
$\omega_2 =$ “no cancer”	60	3	0

# Bayes Decision Rule – The General Case (Cont.)

38

## □ The problem

- Given a particular  $x$ , we have to decide which action to take



$\alpha_i \ (1 \leq i \leq a)$

- We need to know the *loss* of taking each action

true state of  
nature is  $\omega_j$

the action being  
taken is  $\alpha_i$



incur the loss  $\lambda(\alpha_i | \omega_j)$

**However, the true state  
of nature is uncertain**



**Expected (average) loss**

# Bayes Decision Rule – The General Case

39



Average by *enumerating* over all possible states of nature!

## Expected loss

$$R(\alpha_i | \mathbf{x}) = \sum_{j=1}^c \underbrace{\lambda(\alpha_i | \omega_j)}_{\text{The incurred loss of taking action } \alpha_i \text{ in case of true state of nature being } \omega_j} \cdot \underbrace{P(\omega_j | \mathbf{x})}_{\text{The probability of } \omega_j \text{ being the true state of nature, given the feature vector } \mathbf{x}}$$

The incurred loss of taking action  $\alpha_i$  in case of true state of nature being  $\omega_j$

The probability of  $\omega_j$  being the true state of nature, given the feature vector  $\mathbf{x}$

The expected loss is also named as *(conditional) risk or risk function*

# Bayes Decision Rule – The General Case

40

- Now, we have to choose that action  $\alpha_i$  for which the risk is minimum.
- It is also called as Bayes risk and it is the best performance that can be achieved.



# Bayes Decision Rule – The General Case (Cont.)

41

$$R = \int R(\alpha(\mathbf{x}) \mid \mathbf{x})) \cdot p(\mathbf{x}) d\mathbf{x} \quad (\text{overall risk})$$

For every  $\mathbf{x}$ , we ensure that the conditional risk  $R(\alpha(\mathbf{x}) \mid \mathbf{x})$  is as small as possible



The **overall risk** over all possible  $\mathbf{x}$  must be as small as possible

## Bayes decision rule (*General case*)

- The resulting overall risk is called the **Bayes risk** (denoted as  $R^*$ )
- The best performance achievable given  $p(\mathbf{x})$  and loss function

$$\begin{aligned} \alpha(\mathbf{x}) &= \arg \min_{\alpha_i \in \mathcal{A}} R(\alpha_i \mid \mathbf{x}) \\ &= \arg \min_{\alpha_i \in \mathcal{A}} \sum_{j=1}^c \lambda(\alpha_i \mid \omega_j) \cdot P(\omega_j \mid \mathbf{x}) \end{aligned}$$

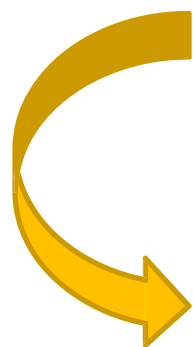
# Two-Category Classification

## Special case

$$\begin{bmatrix} \lambda_{11} & \lambda_{21} \\ \lambda_{12} & \lambda_{22} \end{bmatrix}$$

□  $\Omega = \{\omega_1, \omega_2\}$  (two states of nature)

□  $\mathcal{A} = \{\alpha_1, \alpha_2\}$  ( $\alpha_1 = \text{decide } \omega_1$ ;  $\alpha_2 = \text{decide } \omega_2$ )

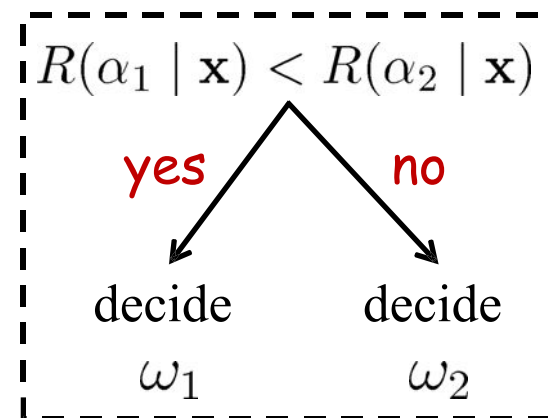


$\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$  : the loss incurred for deciding  $\omega_i$  when the true state of nature is  $\omega_j$

## The conditional risk:

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

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



# Two-Category Classification (Cont.)

$$R(\alpha_1 \mid \mathbf{x}) < R(\alpha_2 \mid \mathbf{x})$$

by  
definition



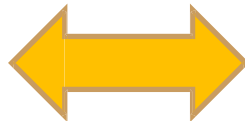
$$\begin{aligned} & \lambda_{11} \cdot P(\omega_1 \mid \mathbf{x}) + \lambda_{12} \cdot P(\omega_2 \mid \mathbf{x}) \\ & < \\ & \lambda_{21} \cdot P(\omega_1 \mid \mathbf{x}) + \lambda_{22} \cdot P(\omega_2 \mid \mathbf{x}) \end{aligned}$$

by  
re-arrangement



$$\begin{aligned} & (\lambda_{21} - \lambda_{11})P(\omega_1 \mid \mathbf{x}) \\ & > \\ & (\lambda_{12} - \lambda_{22})P(\omega_2 \mid \mathbf{x}) \end{aligned}$$

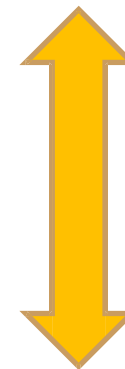
by Bayes  
theorem



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

likelihood  
ratio

constant  $\theta$   
independent of  $\mathbf{x}$



$$\lambda_{21} - \lambda_{11} > 0$$

*the loss for being  
error is ordinarily  
greater than the loss  
for being correct*

$$\begin{aligned} & (\lambda_{21} - \lambda_{11}) \cdot p(\mathbf{x} \mid \omega_1) \cdot P(\omega_1) \\ & > \\ & (\lambda_{12} - \lambda_{22}) \cdot p(\mathbf{x} \mid \omega_2) \cdot P(\omega_2) \end{aligned}$$

# Bayes Minimum Risk- Numerical Example 1

44

Suppose we have:

<b>Action</b> <b>Class</b>	$\alpha_1 =$ “Recipe A”	$\alpha_2 =$ “Recipe B”	$\alpha_3 =$ “No Recipe”
$\omega_1 =$ “cancer”	<b>5</b>	<b>50</b>	<b>10,000</b>
$\omega_2 =$ “no cancer”	<b>60</b>	<b>3</b>	<b>0</b>

For a particular  $\mathbf{x}$ :

$$P(\omega_1 \mid \mathbf{x}) = 0.01$$

$$P(\omega_2 \mid \mathbf{x}) = 0.99$$

# Bayes Minimum Risk- Numerical

## Example 1

45

calculate the risk involved for various action given in the table

<b>Action</b> <b>Class</b>	$\alpha_1 =$ “Recipe A”	$\alpha_2 =$ “Recipe B”	$\alpha_3 =$ “No Recipe”
$\omega_1 =$ “cancer”	5	50	10,000
$\omega_2 =$ “no cancer”	60	3	0

For a particular  $\mathbf{x}$ :

$$P(\omega_1 \mid \mathbf{x}) = 0.01$$

$$P(\omega_2 \mid \mathbf{x}) = 0.99$$

$$\begin{aligned} R(\alpha_1 \mid \mathbf{x}) &= \sum_{j=1}^2 \lambda(\alpha_1 \mid \omega_j) \cdot P(\omega_j \mid \mathbf{x}) \\ &= \lambda(\alpha_1 \mid \omega_1) \cdot P(\omega_1 \mid \mathbf{x}) + \lambda(\alpha_1 \mid \omega_2) \cdot P(\omega_2 \mid \mathbf{x}) \\ &= 5 \times 0.01 + 60 \times 0.99 = 59.45 \end{aligned}$$

**Similarly, we can get:**  $R(\alpha_2 \mid \mathbf{x}) = 3.47$   $R(\alpha_3 \mid \mathbf{x}) = 100$

# Bayes Minimum Risk- Numerical

## Example 2

46

### Spam Filtering: Suppose we have

<b>Action</b> <b>Class</b>	$\alpha_1 =$ Keep the mail	$\alpha_2 =$ Delete as Spam
$\omega_1$ =normal mail	<b>0</b>	<b>3</b>
$\omega_2$ =spam mail	<b>1</b>	<b>0</b>

For a particular  $\mathbf{x}$ :

$$P(\mathbf{x}/\omega_1)=0.35$$

$$P(\mathbf{x}/\omega_2)=0.65$$

$$P(\omega_1)=0.4$$

$$P(\omega_2)=0.6$$

# Bayes Minimum Risk- Numerical Example 2

47

## Spam Filtering: Suppose we have

Class \ Action	$\alpha_1 =$ Keep the mail	$\alpha_2 =$ Delete as Spam
$\omega_1$ =normal mail	0	3
$\omega_2$ =spam mail	1	0

$$R(\alpha_1 \mid \mathbf{x}) = 0.736$$

$$R(\alpha_2 \mid \mathbf{x}) = 0.792$$

Since  $R(\alpha_1 \mid \mathbf{x}) < R(\alpha_2 \mid \mathbf{x})$  we decide take action 1 and decide class 1. **Keep the mail**

For a particular  $\mathbf{x}$ :

$$P(\mathbf{x}/\omega_1) = 0.35$$

$$P(\mathbf{x}/\omega_2) = 0.65$$

$$P(\omega_1) = 0.4$$

$$P(\omega_2) = 0.6$$

# Minimum-Error-Rate Classification

## Classification setting

- $\Omega = \{\omega_1, \omega_2, \dots, \omega_c\}$  ( $c$  possible states of nature)
- $\mathcal{A} = \{\alpha_1, \alpha_2, \dots, \alpha_c\}$  ( $\alpha_i = \text{decide } \omega_i, 1 \leq i \leq c$ )

## Zero-one (symmetrical) loss function

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

- Assign no loss (i.e. 0) to a correct decision by taking action
- Assign a unit loss (i.e. 1) to any incorrect decision (**equal cost**)



# Minimum-Error-Rate Classification

## (Cont.)

49

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

**error rate**

$P(\omega_i \mid \mathbf{x})$  the probability that action  $\alpha_i$  (decide  $\omega_i$ ) is correct

### Minimum error rate

Decide  $\omega_i$  if  $P(\omega_i \mid \mathbf{x}) > P(\omega_j \mid \mathbf{x})$  for all  $j \neq i$

# Discriminant Functions

**Discriminant function for Minimum Risk,  
Minimum Error Rate classifier (Bayes  
Classifier)**

# Discriminant Function-Multi category case

## Classification

Pattern  $\mapsto$  Category

actions  $\longleftrightarrow$  decide categories

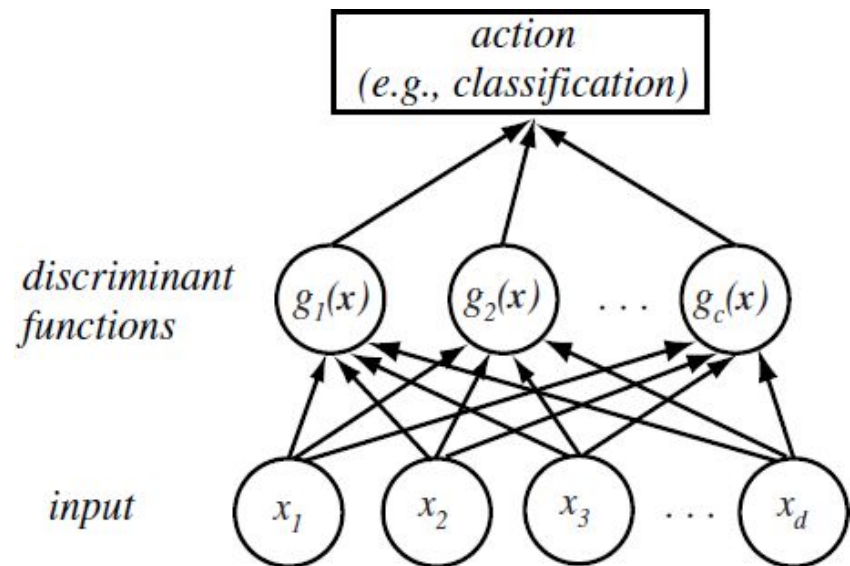
### Discriminant functions

$$g_i : \mathbf{R}^d \rightarrow \mathbf{R} \quad (1 \leq i \leq c)$$

- $\square$  Useful way to represent classifiers
- $\square$  One function per category

Decide  $\omega_i$

if  $g_i(\mathbf{x}) > g_j(\mathbf{x})$  for all  $j \neq i$



# Discriminant Function

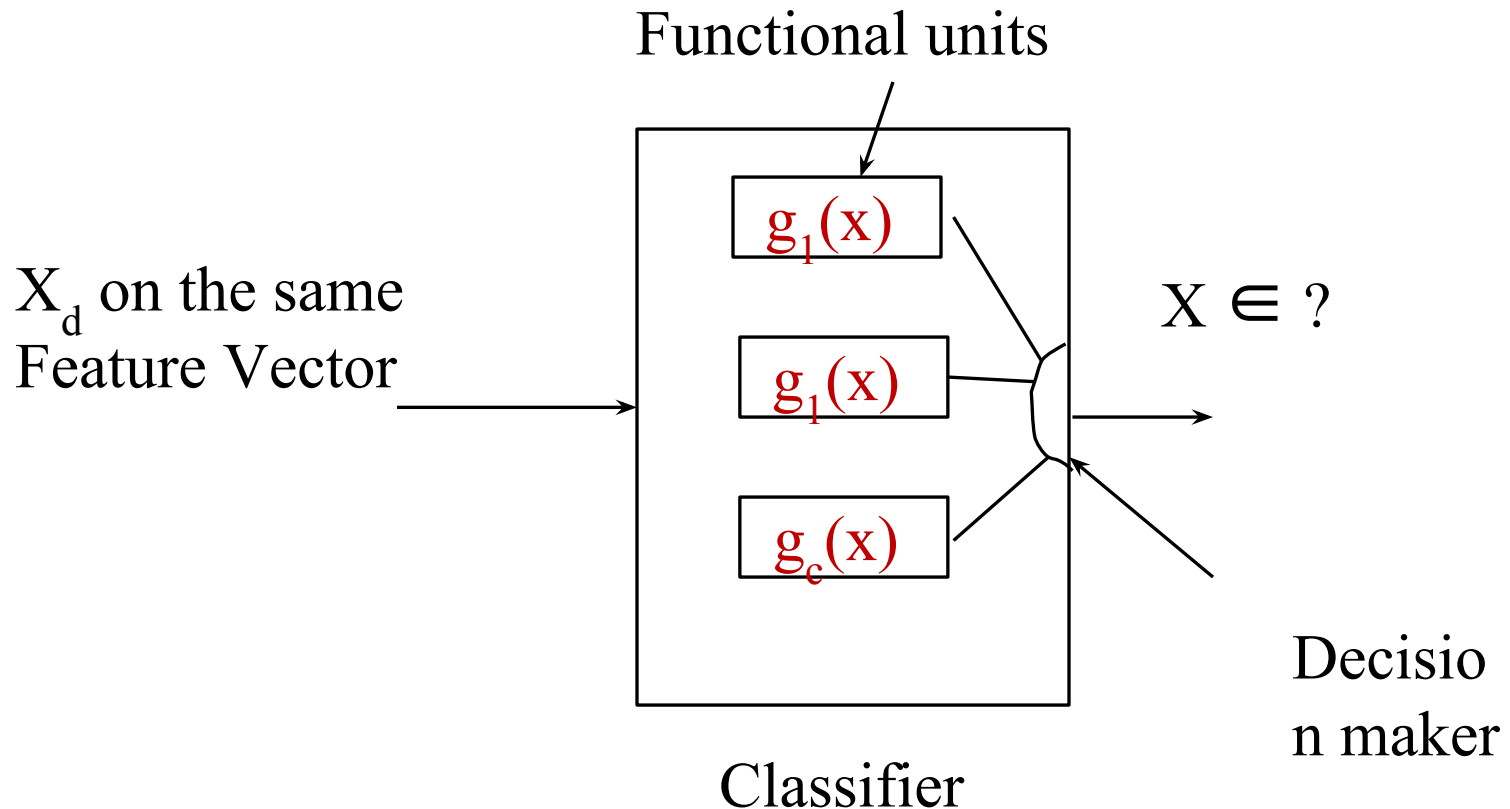
52

- The classifier is viewed as a network or machine that computes  $c$  discriminant function
- Select the category corresponding to the maximum discriminant

# Discriminant Function

53

- A network representation of a classifier is shown below



# Discriminant Function

54

- The nature of discriminant classes
- $\omega_1, \omega_2, \dots, \omega_c \Rightarrow c$  number of classes
- $g_i(x) > g_j(x)$  for all  $i \neq j \Rightarrow x \in \omega_i$

# Discriminant function under different conditions

55

Minimum risk:

$$g_i(\mathbf{x}) = -R(\alpha_i | \mathbf{x}) \quad (1 \leq i \leq c)$$

Minimum-error-rate:

$$g_i(\mathbf{x}) = P(\omega_i | \mathbf{x}) \quad (1 \leq i \leq c)$$

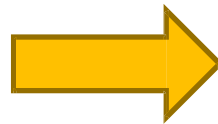
- Hence, the feature vector  $\mathbf{x}$  can assign to the class which has maximum  $g_i(\mathbf{x})$ .
- But the choice of discriminant function  $g_i(\mathbf{x})$  is **not unique**, more generally, if we replace every  $g_i(\mathbf{x})$  by  $f(g_i(\mathbf{x}))$ , where  $f(\cdot)$  is a **monotonically increasing** function, the resulting classification is **unchanged**.

# Discriminant Function

(Cont.)

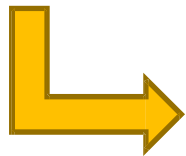
This observation can lead to significant analytical and computational simplifications

Various  
discriminant functions



Identical  
classification results


$f(\cdot)$  is a *monotonically increasing function*



$f(g_i(\mathbf{x})) \iff g_i(\mathbf{x})$  (i.e. *equivalent in decision*)

e.g.:

$f(x) = k \cdot x \ (k > 0)$    $f(g_i(\mathbf{x})) = k \cdot g_i(\mathbf{x}) \ (1 \leq i \leq c)$

$f(x) = \ln x$    $f(g_i(\mathbf{x})) = \ln g_i(\mathbf{x}) \ (1 \leq i \leq c)$



# Discriminant Function

## (Cont.)

### Decision region

$c$  discriminant functions  $c$  decision regions

$g_i(\cdot)$  ( $1 \leq i \leq c$ )



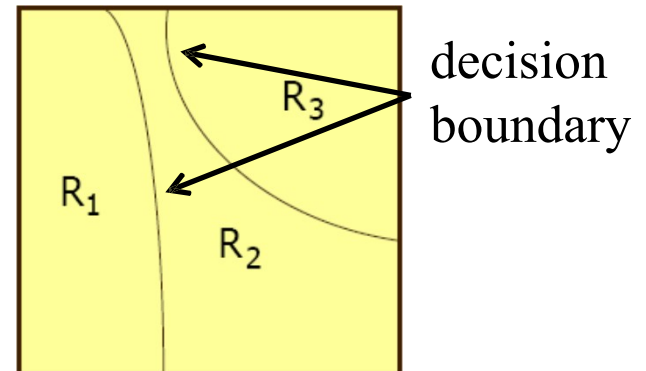
$\mathcal{R}_i \subset \mathbf{R}^d$  ( $1 \leq i \leq c$ )

$$\mathcal{R}_i = \{\mathbf{x} \mid \mathbf{x} \in \mathbf{R}^d : g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i\}$$

where  $\mathcal{R}_i \cap \mathcal{R}_j = \emptyset$  ( $i \neq j$ ) and  $\bigcup_{i=1}^c \mathcal{R}_i = \mathbf{R}^d$

### Decision boundary

surface in feature space where ties occur among several largest discriminant functions



# Dichotomizer: The two category case

58

- **Dichotomizer:** A Classifier that places a pattern in one of any two categories has a special name called a dichotomizer
- $g_1(x) = g_2(x)$
- $g(x) = g_1(x) - g_2(x)$
- Thus a dichotomizer can be viewed as a machine that computes a single discriminant function  $g(x)$ , and classifies  $x$  according to algebraic sign of the result

# Discriminant Function (Cont.)

59

- **Minimum error rate classifier**

- $g_i(x) = P(\omega_i / x)$

- $g_i(x) = P(x / \omega_i) \cdot P(\omega_i) / P(x)$

- $g_i(x) = P(x / \omega_i) \cdot P(\omega_i)$

- $f(g_i(x)) = \ln P(x / \omega_i) + \ln P(\omega_i)$

# Discriminant Function (Cont.)

60

- Minimum error rate classifier:  $g_i(x) = P(\omega_i / x)$
- **two category case**
- $g_1(x) = P(\omega_1 / x)$  ;  $g_2(x) = P(\omega_2 / x)$
- $g(x) \equiv g_1(x) - g_2(x) = 0$
- $g(x) = P(\omega_1 / x) - P(\omega_2 / x) = 0$
- $g(x) = P(x / \omega_1) \cdot P(\omega_1) - P(x / \omega_2) \cdot P(\omega_2)$
- $g(x) = \ln (P(x / \omega_1) \cdot P(\omega_1)) - \ln (P(x / \omega_2) \cdot P(\omega_2))$
- $g(x) = \ln P(x / \omega_1) + \ln P(\omega_1) - \ln P(x / \omega_2) - \ln P(\omega_2)$
- $g(x) = \ln \frac{P(X / \omega_1)}{P(X / \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$

# Discriminant function for Bayes Classifier

61

- $g_1(x) = P(\omega_1/x)$
- $g_i(x) = \ln P(\omega_i/x)$
- $g(x) = \ln P(x/\omega_i) + \ln P(\omega_i)$
- $P(x/\omega_i)$  = class conditional PDF
- $P(\omega_i)$  = PDF (a priori probability)
- Note: The structure of baye's classifier is determined by the conditional densities  $P(x/\omega_i)$  as well as prior probabilities  $P(\omega_i)$

# Discriminant function for Bayes Classifier

62

- We can have various types of probability density like i) normal density ii) poisson iii) laplacian iv) exponential and so on
- Out of these, the most common Probability Density Function (PDF) which is in use is **Normal/Gaussian density function.**

# The normal density

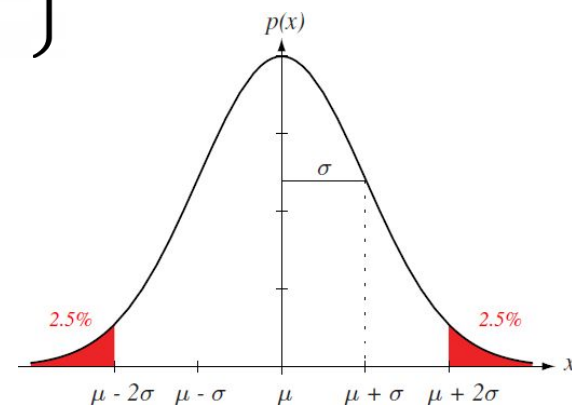
# The normal density

64

- **Univariate density**: for a single variable, the normal /Gaussian density is

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

- $P(x)$  = probability density function
- $\mu$  = mean =  $E(x)$
- $\sigma$  = standard deviation =  $E[(x - \mu)^2]$
- This particular PDF is specified by two parameters  $\mu, \sigma$
- In short  $\approx N(\mu, \sigma^2)$





# The normal density

65

- **Multivariate Probability Density function**
- Here  $X$  be a feature vector,  $X = [x_1, x_2, \dots, x_d]^T$   
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$
- $\boldsymbol{\mu} = E[X]$  ;  $\mathbf{x}$  is  $d$ -dimensional vector
  - $\boldsymbol{\mu} = [\mu_1, \mu_2, \dots, \mu_d]^T$
- $\Sigma$  = covariance matrix,  $d \times d$  matrix
  - $\Sigma = E[(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^t]$
- In short  $\approx N(\boldsymbol{\mu}, \Sigma)$

# The normal density

66

- What is expected value of individual component?
- Expected value of the  $i^{\text{th}}$  component
- $\mu_i = E[x_i]$
- $i \neq j; \sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]$
- $i = j; \sigma_{ii} = E[(x_i - \mu_i)(x_i - \mu_i)] = E[(x_i - \mu_i)^2]$

# The normal density

67

- ▣ **Bivariate Probability Density function (two variables)**
- ▣ Here  $X$  is of the form  $[x_1, x_2]^T$
- ▣ Number of dimension  $d=2$
- ▣ Assume  $x_1$  and  $x_2$  are statistically independent, and hence  $\sigma_{12}$  and  $\sigma_{21}$  are 0
- ▣ Mean  $\mu = [\mu_1, \mu_2]^T$
- ▣  $\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$

# The normal density- bivariate case

68

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

- The above multivariate normal density can be simplified to

$$P(X) = \frac{1}{2\pi |\Sigma|^{1/2}} \exp \left[ \frac{-1}{2} \left\{ \left( \frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left( \frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right\} \right]$$

# The normal density- two dimensional feature space

69

## Case 1:

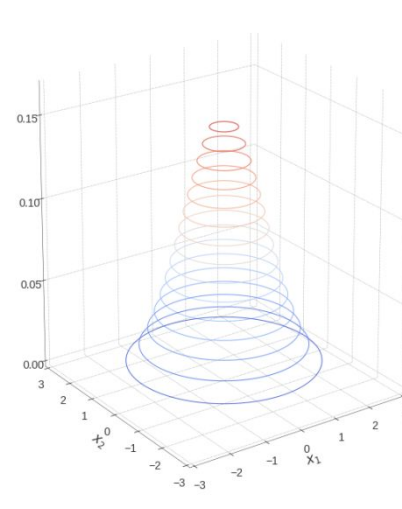
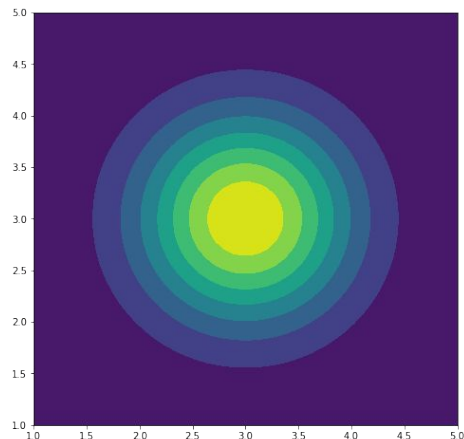
□  $\sigma_{ij}=0; i \neq j$

□  $\sigma_1^2 = \sigma_2^2; \sigma_{12} = \sigma_{21} = 0; \quad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix} = 5 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

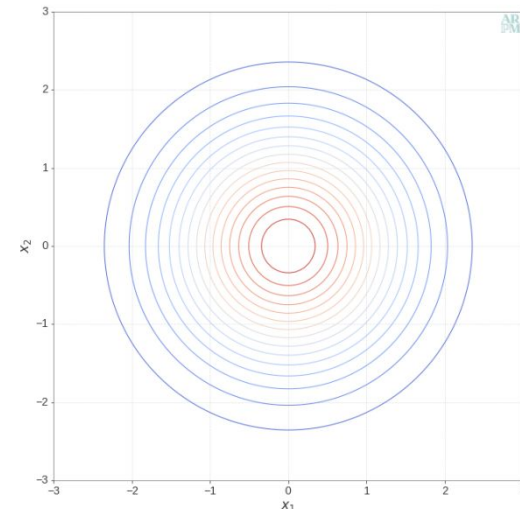
□  $x_1$  and  $x_2$  are statistically independent

□ Trace the loci of points of constant density for all value of  $x$  for which  $P(x)$  is constant.

□ Those loci of points forms circle



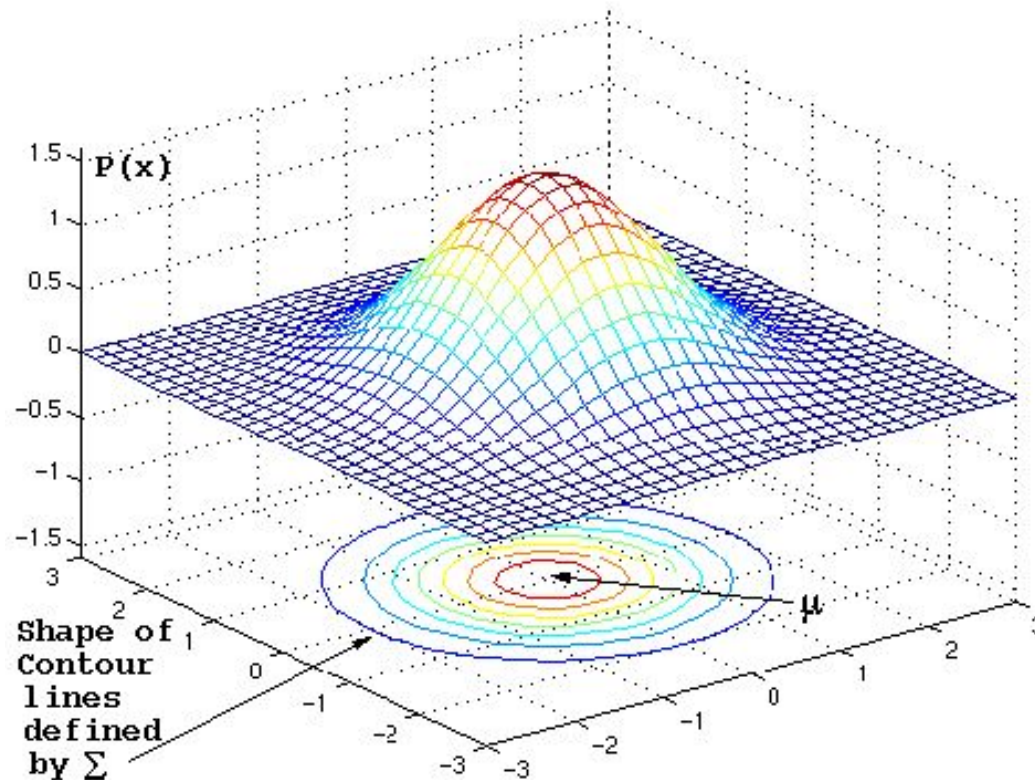
Normal pdf iso-contours



# The normal density- two dimensional feature space

70

## □ Case 1:



# The normal density- two dimensional feature space

71

□ What happens if the variants are different?

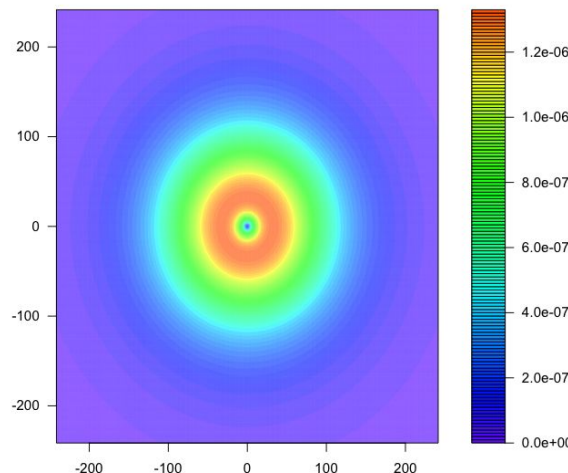
□ **Case 2:**

□  $\sigma_{ij}=0; i \neq j$

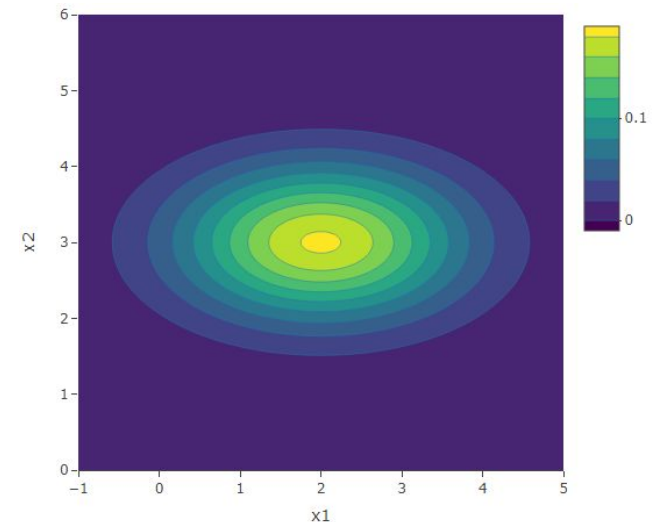
□  $\sigma_1^2 \neq \sigma_2^2; \sigma_{12} = \sigma_{21} = 0; \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 15 & 0 \\ 0 & 5 \end{bmatrix}; \sigma_1^2 > \sigma_2^2$

□  $x_1$  and  $x_2$  **are not statistically independent**. The loci of points **forms ellipse**

$$\sigma_1^2 < \sigma_2^2$$



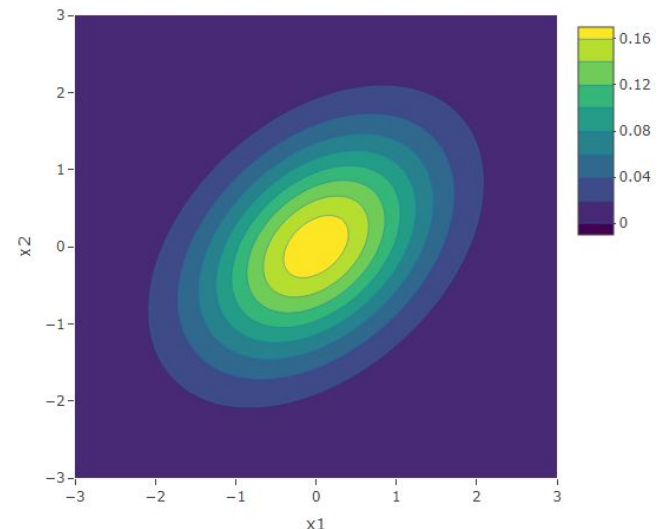
$$\sigma_1^2 > \sigma_2^2$$



# The normal density- two dimensional feature space

72

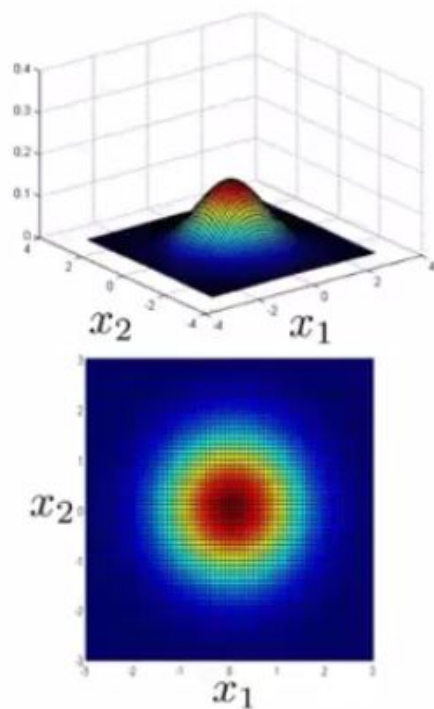
- What happens if the variants are different?
- **Case 3:**
  - $\sigma_{ij} \neq 0; i \neq j$
  - $\sigma_1^2 \neq \sigma_2^2; \sigma_{12} = \sigma_{21} = 0$
- $x_1$  and  $x_2$  are **not statistically independent**
- The direction of point distribution is determined by eigenvector of  $\Sigma$
- $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 0.4 & 0.3 \\ 0.6 & 0.7 \end{bmatrix}$



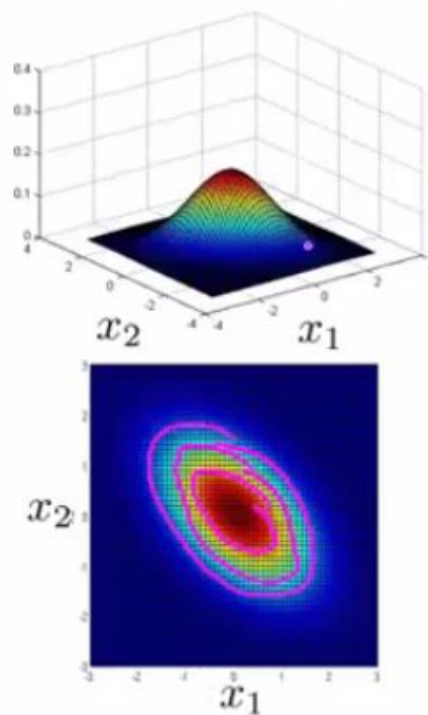


## Multivariate Gaussian (Normal) examples

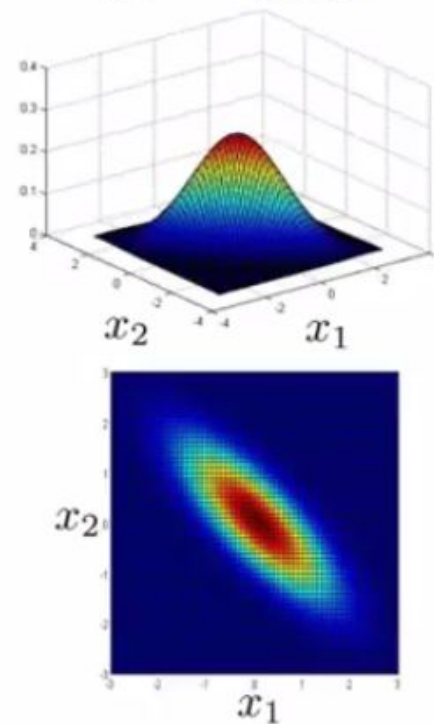
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$



## Gaussian Density–Multivariate Case

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

# Gaussian Density – Multivariate Case

$$\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boxed{\mu_i = \mathcal{E}[x_i] \quad \sigma_{ij} = \sigma_{ji} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)]}$$

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

$\mathbf{x} = (x_1, x_2, \dots, x_d)^t$  :  $d$ -dimensional *column vector*

$\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_d)^t$  :  $d$ -dimensional *mean vector*

$$\boldsymbol{\Sigma} = [\sigma_{ij}]_{1 \leq i, j \leq d} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_{dd} \end{pmatrix} \begin{array}{l} d \times d \text{ covariance} \\ \text{matrix} \\ |\boldsymbol{\Sigma}| : \text{determinant} \\ \boldsymbol{\Sigma}^{-1} : \text{inverse} \end{array}$$

# Gaussian Density – Multivariate Case

## Properties of covariance matrix

Properties of  $\Sigma$

$$\Sigma = [\sigma_{ij}]_{1 \leq i, j \leq d} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_{dd} \end{pmatrix}$$

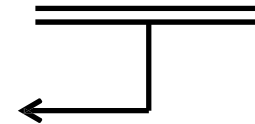
- symmetric
- Positive semidefinite

$$\sigma_{ij} = \sigma_{ji} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)]$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j) \cdot \underbrace{p(x_i, x_j)}_{\text{marginal pdf on a pair of random variables } (x_i, x_j)} dx_i dx_j$$

$$\sigma_{ii} = \text{Var}[x_i] = \sigma_i^2$$




**marginal pdf** on a pair of  
random variables  $(x_i, x_j)$



# Gaussian Density– Multivariate Case (Cont.)

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

$(\mathbf{x} - \boldsymbol{\mu})^t : 1 \times d$ matrix		
$\boldsymbol{\Sigma}^{-1} : d \times d$ matrix		$(\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$
$(\mathbf{x} - \boldsymbol{\mu}) : d \times 1$ matrix		scalar ( $1 \times 1$ matrix)

$\boldsymbol{\Sigma} : \text{positive definite}$		$\boldsymbol{\Sigma}^{-1} : \text{positive definite}$
		
$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \leq 0$		$(\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \geq 0$

# Discriminant Functions for Gaussian Density for Bayes Classifier

Bayes classification: (Minimum error rate classification)

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) \quad (1 \leq i \leq c)$$

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) \quad \longleftrightarrow \quad g_i(\mathbf{x}) = \ln P(\omega_i|\mathbf{x})$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$$

$$p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

Constant, could be ignored

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

Case I:  $\Sigma_i = \sigma^2 \mathbf{I}$

$$p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \Sigma_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)$$

Covariance matrix:  $\sigma^2$  times the identity matrix  $\mathbf{I}$

$$\Sigma_i = \sigma^2 \cdot \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} = \begin{pmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \ddots & \\ & & & \sigma^2 \end{pmatrix} \quad \longrightarrow \quad \begin{aligned} |\Sigma_i| &= \sigma^{2d} \\ \Sigma_i^{-1} &= (1/\sigma^2) \mathbf{I} \end{aligned}$$

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i) \quad \begin{aligned} \|\cdot\| &: \text{Euclidean norm} \\ \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 &= (\mathbf{x} - \boldsymbol{\mu}_i)^t (\mathbf{x} - \boldsymbol{\mu}_i) \end{aligned}$$

# Case I: $\Sigma_i = \sigma^2 \mathbf{I}$ (Cont.)

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i)$$

**Squared Euclidean distance**



$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2} [\mathbf{x}^t \mathbf{x} - 2\boldsymbol{\mu}_i^t \mathbf{x} + \boldsymbol{\mu}_i^t \boldsymbol{\mu}_i] + \ln P(\omega_i)$$

the same for all *states of nature*,  
could be ignored

**Linear discriminant functions**

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0} \Rightarrow$$

**Linear machine or linear equation**

$$\mathbf{w}_i = \frac{1}{\sigma^2} \boldsymbol{\mu}_i$$

*weight vector*

$$w_{i0} = -\frac{1}{2\sigma^2} \boldsymbol{\mu}_i^t \boldsymbol{\mu}_i + \ln P(\omega_i)$$

*threshold/bias*



## Case II: $\Sigma_i = \Sigma$

$$p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

**Covariance matrix:** *identical* for all classes

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) + \ln P(\omega_i)$$

$(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)$  : squared *Mahalanobis distance*

$\boldsymbol{\Sigma} = \mathbf{I}$   reduces to *Euclidean distance*



P. C.

~~Mahalanobis~~

(1893-1972)

## Case II: $\Sigma_i = \Sigma$ (Cont.)

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) + \ln P(\omega_i)$$



the same for all *states of nature*,  
could be ignored

$$g_i(\mathbf{x}) = -\frac{1}{2}[\mathbf{x}^t \Sigma^{-1} \mathbf{x} - 2\boldsymbol{\mu}_i^t \Sigma^{-1} \mathbf{x} + \boldsymbol{\mu}_i^t \Sigma^{-1} \boldsymbol{\mu}_i] + \ln P(\omega_i)$$

### Linear discriminant functions

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0} \Rightarrow$$

**Linear machine or  
linear equation**

$$\mathbf{w}_i = \Sigma^{-1} \boldsymbol{\mu}_i \quad \text{weight vector}$$

$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^t \Sigma^{-1} \boldsymbol{\mu}_i + \ln P(\omega_i) \quad \text{threshold/bias}$$

# Case

## III:

$\Sigma_i = \text{arbitrary}$

$$p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

Quadratic discriminant functions

$$g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + w_{i0}$$

*quadratic matrix*

$$\mathbf{W}_i = -\frac{1}{2} \boldsymbol{\Sigma}_i^{-1}$$

*weight vector*

$$\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i$$

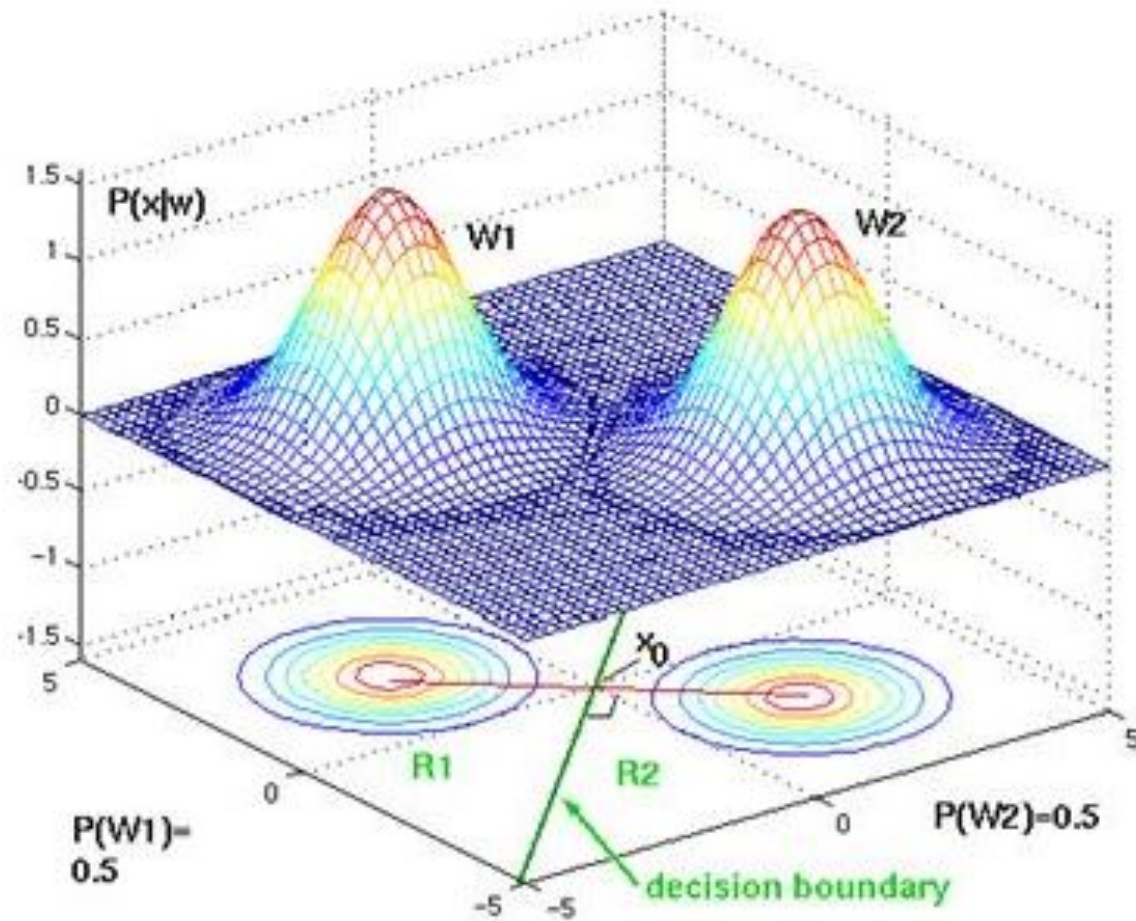
$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \quad \textit{threshold/bias}$$

# Bayes Decision Boundary for Two Classes

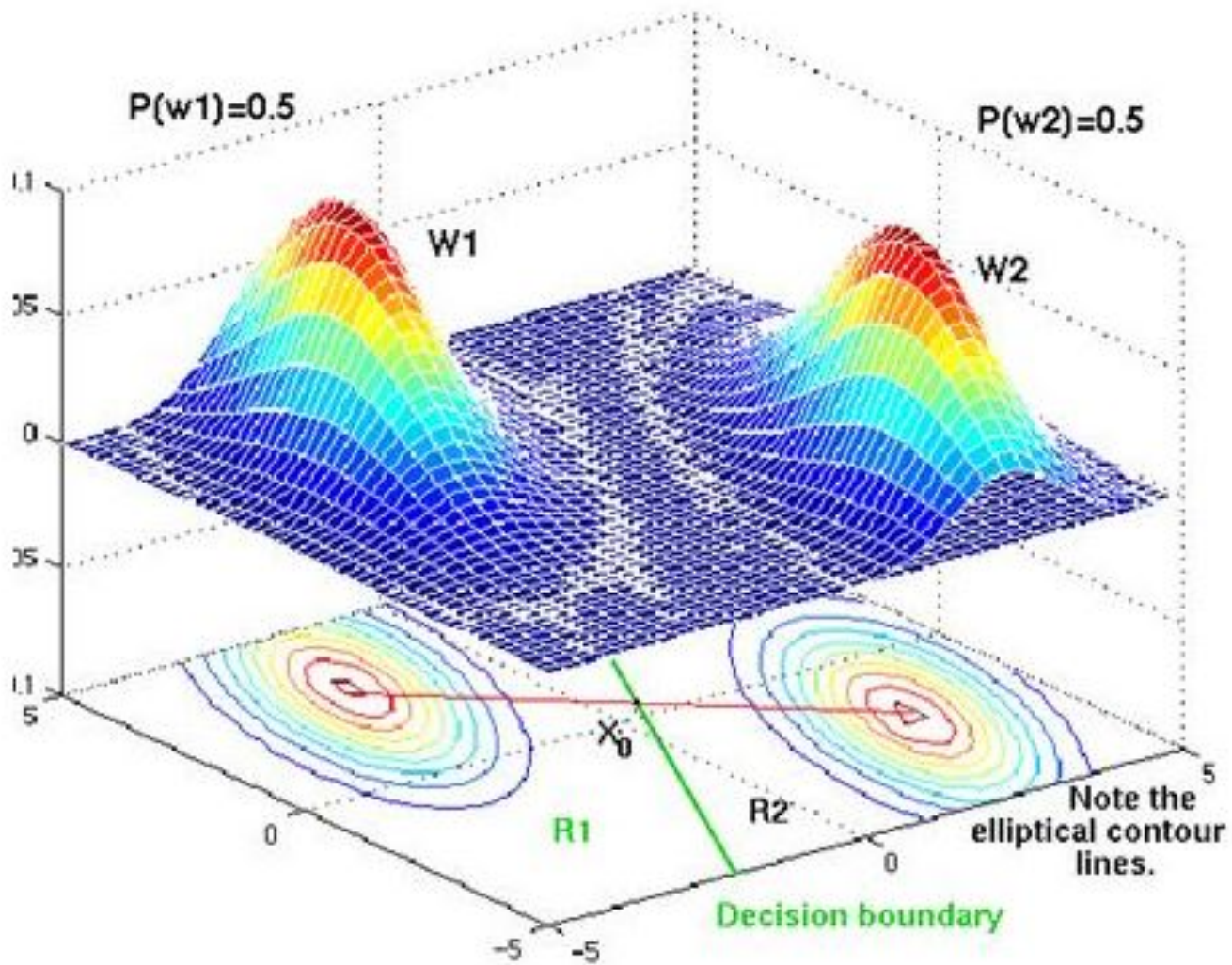
**Case 1 & 2** : Linear Separable Cases

**Case 3**: Nonlinearly Separable Cases

# Case 1

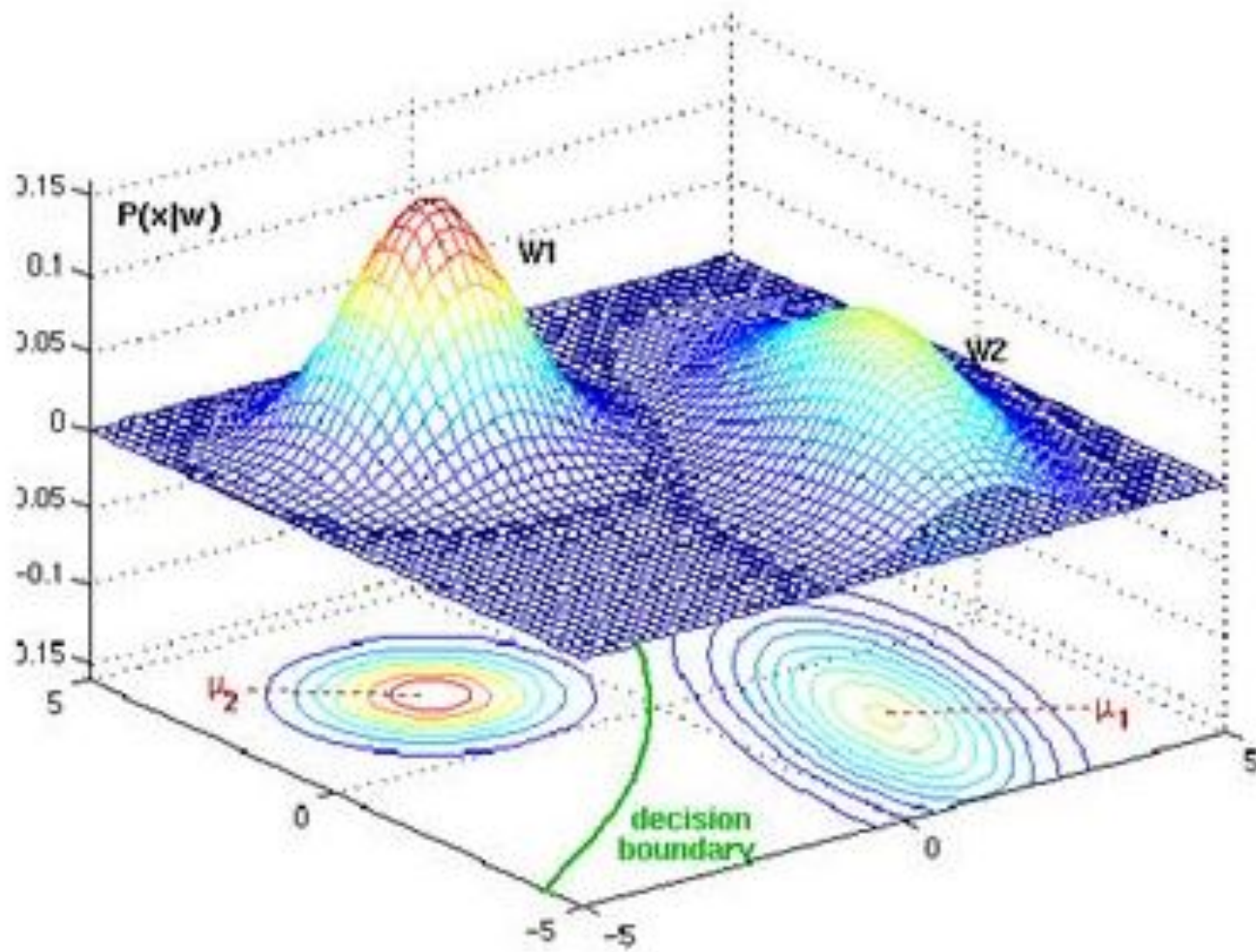


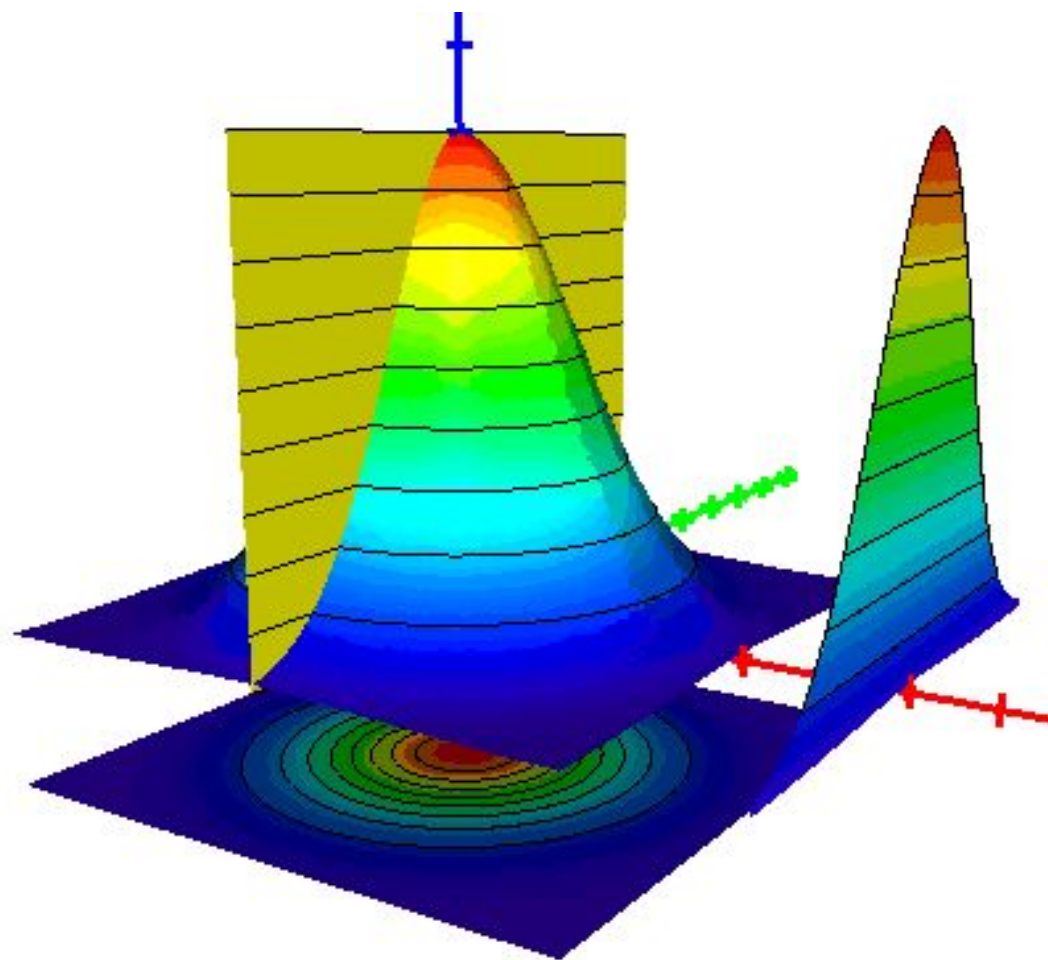
## Case 2





# Case 3

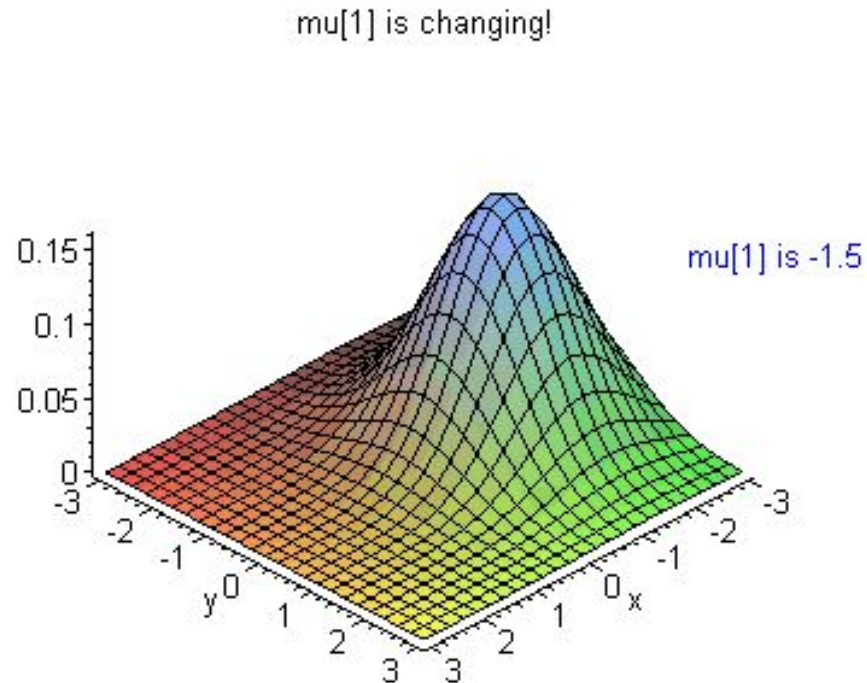






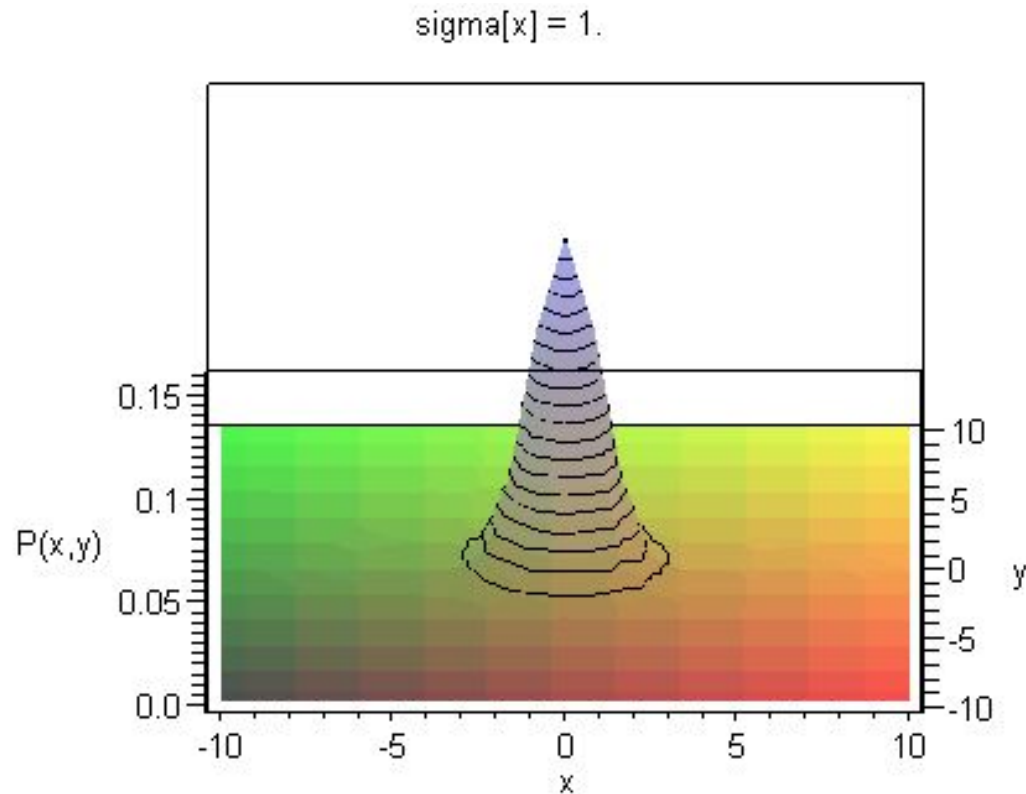
# Normal density- mean is changing

89



# Normal density- sigma is changing

90



# Summary

- Bayesian Decision Theory
  - PR: essentially a decision process
  - Basic concepts
    - States of nature
    - Probability distribution, probability density function (pdf)
    - Class-conditional pdf
    - Joint pdf, marginal distribution, law of total probability
  - Bayes theorem
    - Prior + likelihood + observation  $\rightarrow$  Posterior probability
  - Bayes decision rule
    - Decide the state of nature with maximum posterior

# Summary (Cont.)

- Feasibility of Bayes decision rule
  - Prior probability + likelihood
  - Solution I: counting relative frequencies
  - Solution II: conduct density estimation (chapters 3,4)
- Bayes decision rule: The general scenario
  - Allowing more than one feature
  - Allowing more than two states of nature
  - Allowing actions than merely deciding state of nature
  - Loss function:  $\lambda : \Omega \times \mathcal{A} \rightarrow \mathbf{R}$

# Summary (Cont.)

- Expected loss (*conditional risk*)

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

*Average by enumerating over all possible states of nature*

- General Bayes decision rule
  - Decide the action with minimum expected loss
- Minimum-error-rate classification
  - Actions □□ Decide states of nature
  - Zero-one loss function
    - Assign *no loss/unit loss* for *correct/incorrect* decisions

# Summary (Cont.)

- Discriminant functions

- General way to represent classifiers
- One function per category/class
- Induce *decision regions* and *decision boundaries*

- Gaussian/Normal density

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

- Discriminant functions for Gaussian pdf

$\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}, \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma} : \text{linear discriminant function}$

$\boldsymbol{\Sigma}_i = \text{arbitrary} : \text{quadratic discriminant function}$

Thank you

