# Bayesian Decision Theory – Continuous Features

**Chapter 02**

Ref. Book: Pattern Recognition – S. Theodoridis

Rakib Mahmud
Lecturer, BUBT

# Bayesian Classifier Examples

## Example 01

Given:

- A doctor knows that meningitis causes stiff neck 50% of the time
- Prior probability of any patient having meningitis is 1/50000
- Prior probability of any patient having stiff neck is 1/20

Question: If a patient has stiff neck, does he/she has meningitis?

# Bayesian Classifier Examples

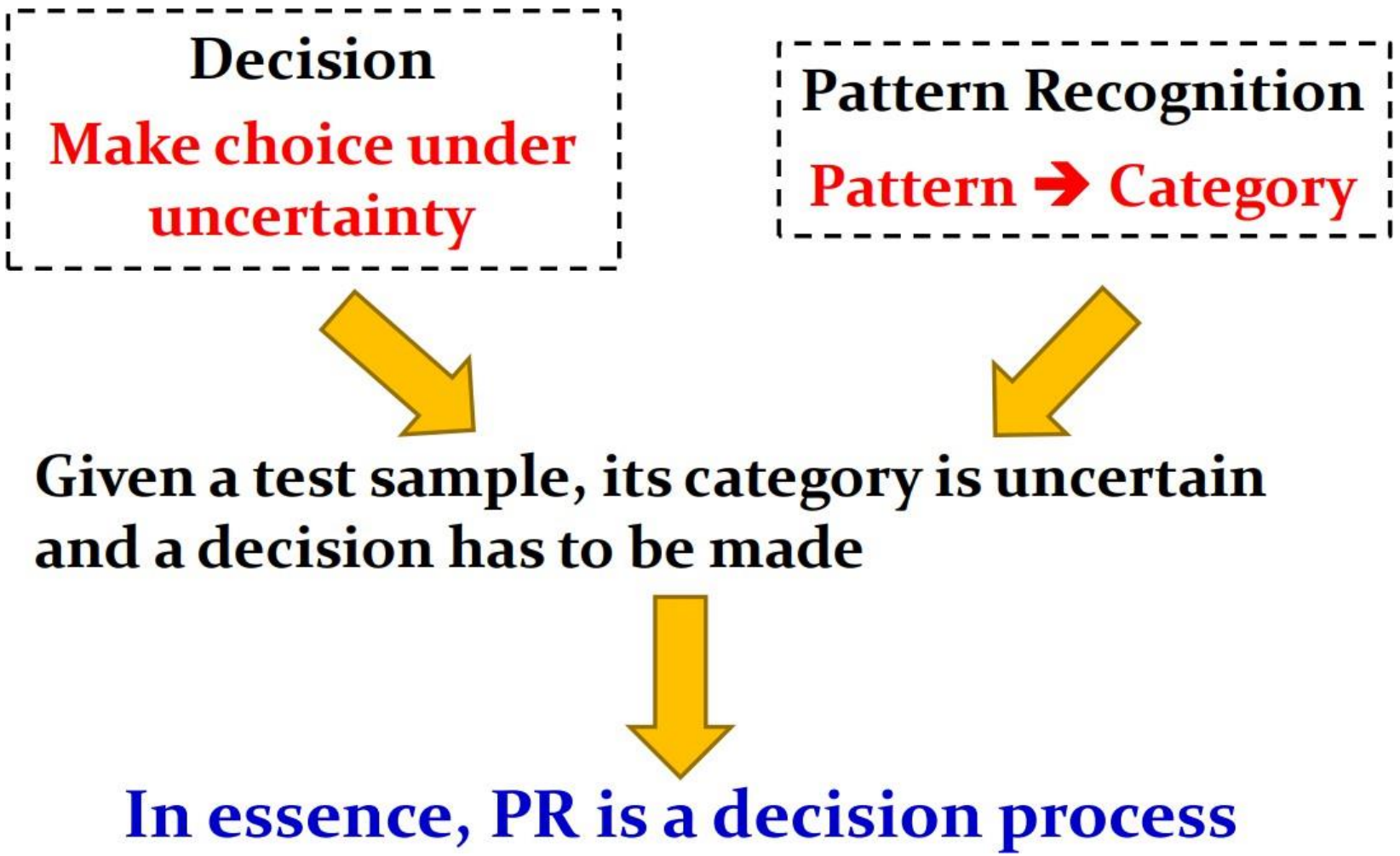
## Example 02

The sea bass/salmon example

- We now the previous counts of salmon/sea bass
- Can we predict which fish is coming in the conveyor?

State of nature is random variable
The catch of salmon and sea bass is equiprobable
- P ($\omega_1$) = P ($\omega_2$) (uniform priors)
- P ($\omega_1$) + P ($\omega_2$) = 1 (exclusivity and exhaustivity)

## Decision Theory



# 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**

# Bayes Theorem

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

$X$: 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** (先验概率) that $H$ holds (*e.g. the probability of catching a salmon*)

$P(X|H)$: the **likelihood** (似然度) 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** (后验概率) that $H$ holds given $X$ (e.g. *the probability of X being salmon given its length is 3-inch*)

**Thomas Bayes (1702-1761)**

## State of Nature

- Future events that might occur

  *e.g. the next fish arriving along the conveyor belt*

- State of nature is unpredictable

  *e.g. it is hard to predict what type will emerge next*

From statistical/probabilistic point of view, the state of nature should be favorably regarded as a random variable

*e.g. let $\omega$ denote the (discrete) random variable representing the state of nature (class) of fish types*

$\omega = \omega_1$ : sea bass

$\omega = \omega_2$ : salmon

## Prior Probability

Prior probability is the probability distribution which reflects one's prior knowledge on the random variable

**Probability distribution (for discrete random variable)**

Let $P(\cdot)$ be the probability distribution on the random variable $\omega$ with $c$ possible states of nature $\{\omega_1, \omega_2, \ldots, \omega_c\}$, such that:

$$P(\omega_i) \geq 0 \ (non\text{-}negativity) \qquad \sum_{i=1}^{c} P(\omega_i) = 1 \ (normalization)$$

*the catch produced as much sea bass as salmon* ⇒ $P(\omega_1) = P(\omega_2) = 1/2$

*the catch produced more sea bass than salmon* ⇒ $P(\omega_1) = 2/3; P(\omega_2) = 1/3$

# Why Bayes?

- Decision rule with only the prior information
  - Decide $\omega_1$ if P ($\omega_1$)>P ($\omega_2$), otherwise decide $\omega_2$
  - Misclassify many fishes
- Use of the class-conditional information
  - Use lightness
- $P(x|\omega_1)$ and $P(x|\omega_2)$ describe lightness in salmon and seabass

# Why Bayes (Cont.)

## The Problem

To make a decision on the type of fish arriving next, where 1) prior probability is known; 2) no observation is allowed

## Naive Decision Rule

Decide $\omega_1$ if $P(\omega_1) > P(\omega_2)$; otherwise decide $\omega_2$

- This is the *best* we can do without observation
- Fixed prior probabilities ➔ Same decisions all the time

Incorporate observations into decision!

*Good when* $P(\omega_1)$ *is much greater (smaller) than* $P(\omega_2)$

*Poor when* $P(\omega_1)$ *is close to* $P(\omega_2)$

[only 50% chance of being right if $P(\omega_1) = P(\omega_2)$ ]

# Probability Density Function (PDF)

**Probability density function (pdf)**
**(for continuous random variable)**

Let $p(\cdot)$ be the probability density function on the continuous random variable $x$ taking values in **R**, such that:

$$p(x) \geq 0 \ (non\text{-}negativity) \qquad \int_{-\infty}^{\infty} p(x)dx = 1 \ (normalization)$$

- For continuous random variable, it no longer makes sense to talk about the probability that *x* has a particular value (almost always be zero)
- We instead talk about the probability of *x* falling into a region *R*, say *R*=(*a*,*b*), which could be computed with the pdf:

$$\Pr[x \in R] = \int_{x \in R} p(x)dx = \int_{a}^{b} p(x)dx$$

# Incorporate Observations

## The Problem

Suppose the fish *lightness measurement x* is observed, how could we incorporate this knowledge into usage?

## Class-conditional probability density function

- It is a probability density function (pdf) for $x$ given that the state of nature (class) is $\omega$ , i.e.:

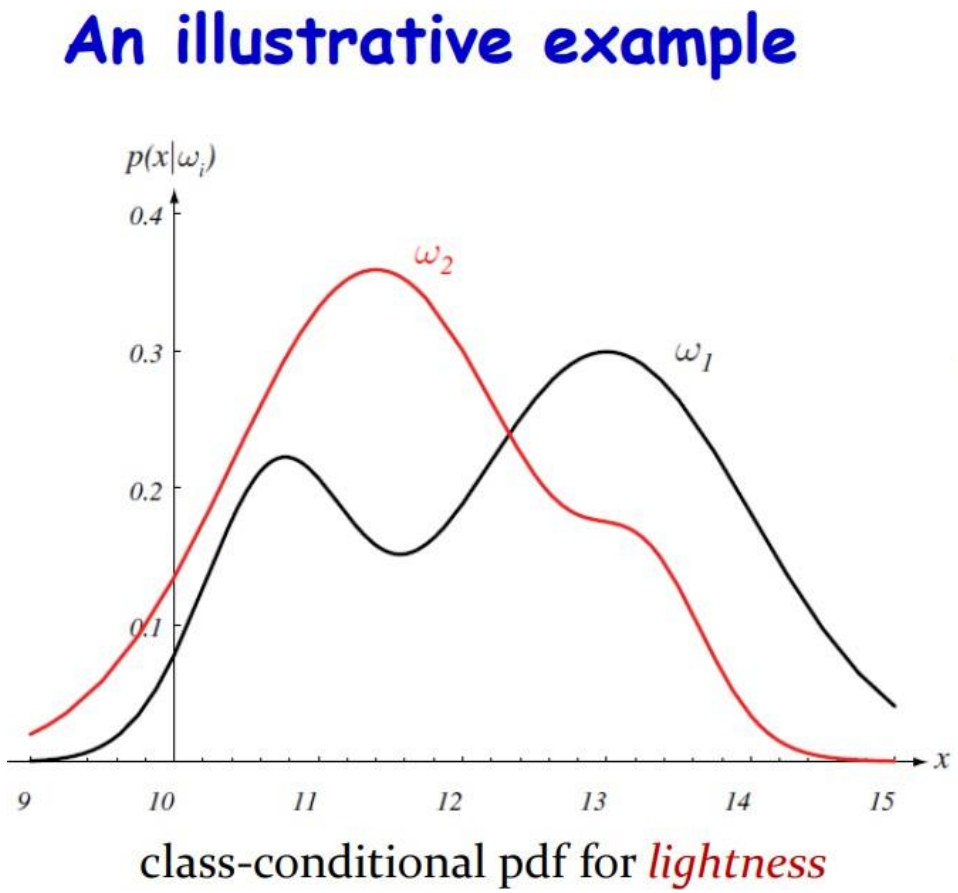
$p(x|\omega)$

$$p(x|\omega) \geq 0 \qquad \int_{-\infty}^{\infty} p(x|\omega)dx = 1$$

- The *class-conditional* pdf describes the difference in the distribution of observations under different classes

$p(x|\omega_1)$ should be different to $p(x|\omega_2)$

# Class Conditional PDF




h-axis: lightness of fish scales

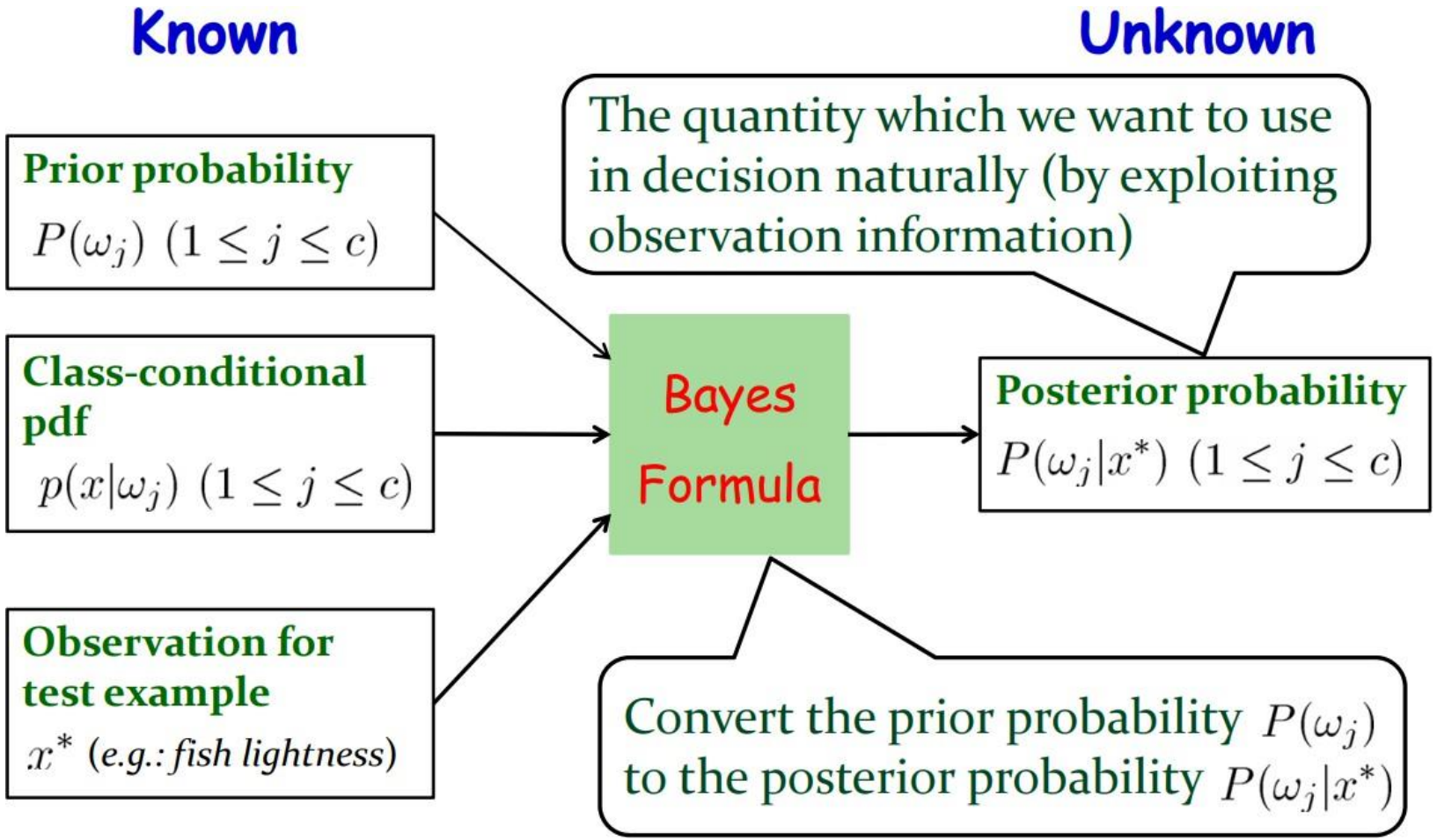
v-axis: class-conditional pdf values

black curve: sea bass

red curve: salmon

- The area under each curve is 1.0 (*normalization*)
- Sea bass is somewhat brighter than salmon

# Decision After Observation



# Bayes Formula Revisited

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

## Bayes Decision Rule

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

- $P(\omega_j)$ 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.)

$$P(\omega_j|x) = \frac{p(x|\omega_j) \cdot P(\omega_j)}{p(x)} \quad \left( posterior = \frac{likelihood \times prior}{evidence} \right)$$

**Special Case I: Equal prior probability**

$$P(\omega_1) = P(\omega_2) = \cdots = P(\omega_c) = \frac{1}{c}$$ ⇒ Depends on the likelihood $P(x|\omega_j)$

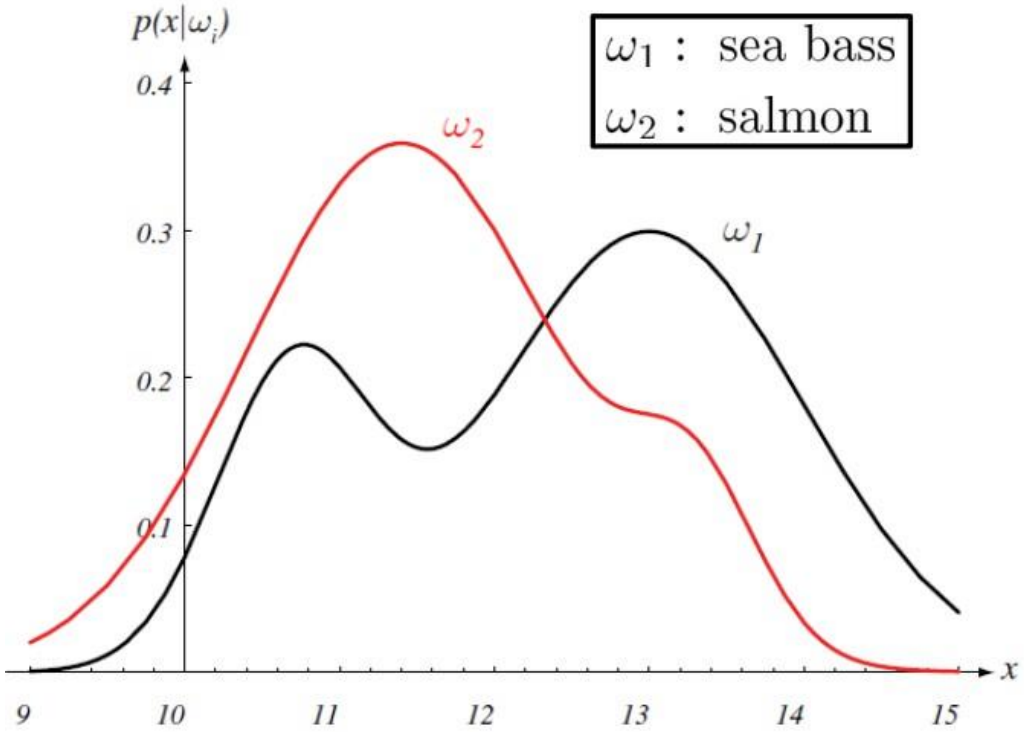
**Special Case II: Equal likelihood**

$$p(x|\omega_1) = p(x|\omega_2) = \cdots = p(x|\omega_c)$$ ⇒ Degenerate to naive decision rule

Normally, prior probability and likelihood function together in Bayesian decision process

# Bayes Formula Revisited (Cont.)

## An illustrative example




class-conditional pdf for *lightness*

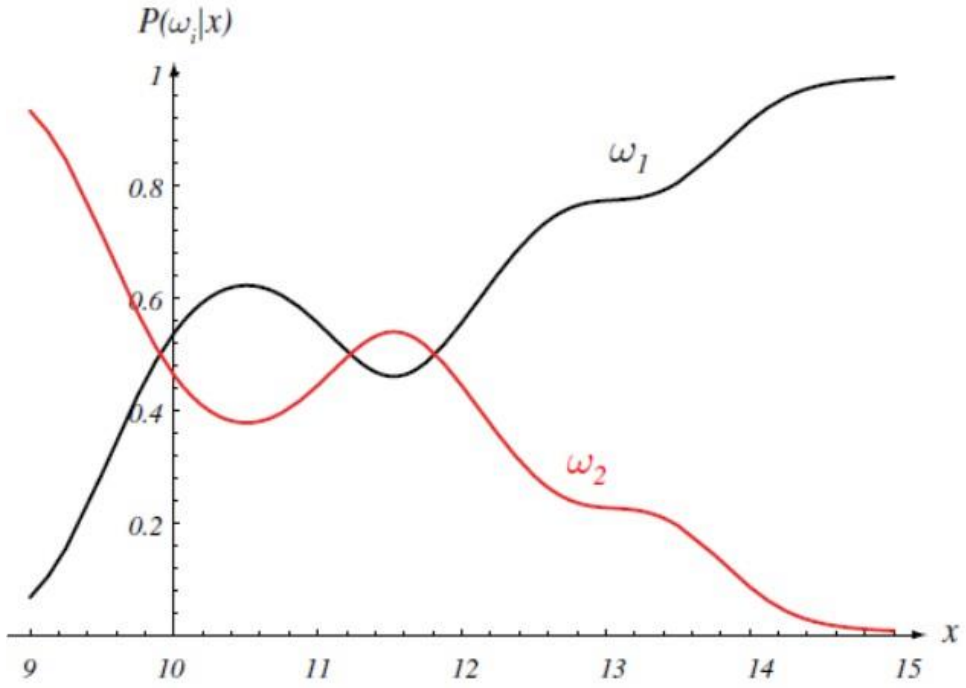
$$P(\omega_1) = \frac{2}{3}$$

$$P(\omega_2) = \frac{1}{3}$$

What will the posterior probability for either type of fish look like?

# Bayes Formula Revisited (Cont.)

## An illustrative example




posterior probability for either type of fish

h-axis: lightness of fish scales

v-axis: posterior probability for either type of fish

black curve: sea bass

red curve: salmon

- For each value of $x$, the higher curve yields the output of Bayesian decision
- For each value of $x$, the posteriors of either curve sum to 1.0

# Another Example

## 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?

# Another Example (Cont.)

$\omega_1$ : cancer $\qquad$ $\omega_2$ : no cancer $\qquad$ $x \in \{+, -\}$

$$P(\omega_1) = 0.008 \qquad P(\omega_2) = 1 - P(\omega_1) = 0.992$$

$$P(+ \mid \omega_1) = 0.98 \qquad P(- \mid \omega_1) = 1 - P(+ \mid \omega_1) = 0.02$$

$$P(- \mid \omega_2) = 0.97 \qquad P(+ \mid \omega_2) = 1 - P(- \mid \omega_2) = 0.03$$

$$P(\omega_1 \mid +) = \frac{P(\omega_1)P(+ \mid \omega_1)}{P(+)} = \frac{P(\omega_1)P(+ \mid \omega_1)}{P(\omega_1)P(+ \mid \omega_1) + P(\omega_2)P(+ \mid \omega_2)}$$

$$= \frac{0.008 \times 0.98}{0.008 \times 0.98 + 0.992 \times 0.03} = 0.2085$$

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

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

No cancer!

## Another Example (Cont.)

$$P(\omega|x) = \frac{p(x|\omega) \cdot P(\omega)}{p(x)} \quad \text{(Bayes Formula)}$$

To compute posterior probability $P(\omega|x)$ , we need to know:

Prior probability: $P(\omega)$ Likelihood: $p(x|\omega)$

How do we know these probabilities?

- A simple solution: Counting

**Relative Frequencies**

- An advanced solution: Conduct

**Density Estimation**

## Further Example

### Problem statement

Based on the height of a car in some campus, decide whether it costs more than \$50,000 or not

$\omega_1$ : price $> \$50,000$

$\omega_2$ : price $\leq \$50,000$

$x$ : height of car

$P(\omega_1|x) > P(\omega_2|x)$

?

$P(\omega_1|x) < P(\omega_2|x)$

### Quantities to know:

$P(\omega_1)$ $P(\omega_2)$ $p(x|\omega_1)$ $p(x|\omega_2)$

Counting relative frequencies via collected samples

## Further Example (Cont.)

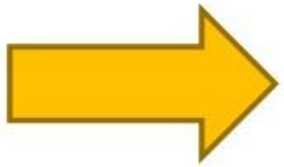
### Collecting samples

Suppose we have randomly picked 1209 cars in the campus, got prices from their owners, and measured their heights

### Compute $P(\omega_1), P(\omega_2)$ :

# cars in $\omega_1$ : 221

# cars in $\omega_2$ : 988



$$P(\omega_1) = \frac{221}{1209} = 0.183$$

$$P(\omega_2) = \frac{988}{1209} = 0.817$$

# Further Example (Cont.)

**Compute** $p(x|\omega_1), p(x|\omega_2)$ :

Discretize the height spectrum (say [0.5m, 2.5m]) into 20 intervals each with length 0.1m, and then count the number of cars falling into each interval for either class

Suppose $x = 1.05$

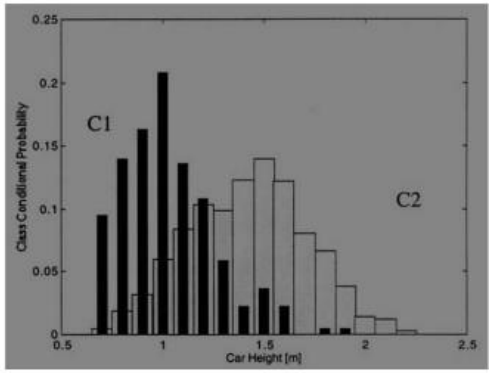
$x$ falls into interval $I_x$=[1.0m, 1.1m]

For $\omega_1$, # cars in $I_x$ is 46

For $\omega_2$, # cars in $I_x$ is 59

$$p(x = 1.05|\omega_1) = \frac{46}{221} = 0.2081$$

$$p(x = 1.05|\omega_2) = \frac{59}{988} = 0.0597$$



# Further Example (Cont.)

## Question

For a car with height 1.05m, is its price greater than \$50,000?

Estimated quantities

$P(\omega_1) = 0.183$ $\quad P(\omega_2) = 0.817$

$p(x = 1.05 \mid \omega_1) = 0.2081$ $\quad p(x = 1.05 \mid \omega_2) = 0.0597$

$$\frac{P(\omega_2 \mid x = 1.05)}{P(\omega_1 \mid x = 1.05)} = \frac{P(\omega_2) \cdot p(x = 1.05 \mid \omega_2)}{p(x = 1.05)} \Big/ \frac{P(\omega_1) \cdot p(x = 1.05 \mid \omega_1)}{p(x = 1.05)}$$

$$= \frac{P(\omega_2) \cdot p(x = 1.05 \mid \omega_2)}{P(\omega_1) \cdot p(x = 1.05 \mid \omega_1)}$$

$$= \frac{0.817 \times 0.0597}{0.183 \times 0.2081} = 1.280$$

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

$\text{price} \leq \$50,000$

# Is Bayes Decision Rule Optimal?

## 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:

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

**Under Bayes decision rule, we have**

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

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

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

# Bayes Decision Rules

- Generalization of the preceding ideas
  - Use of more than one feature
  - Use more than two states of nature
  - Allowing actions and not only decide on the state of nature
  - Introduce a loss function which is more general than the probability of error

## The General Case

- By allowing to use more than one feature

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

- By allowing more than two states of nature

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

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

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

Note : $c \neq a$

## The General Case (Cont.)

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

$$\lambda : \ \Omega \times \mathcal{A} \rightarrow \mathbf{R} \quad \text{(loss function)}$$

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

For ease of reference, usually written as:

$$\lambda(\alpha_i \mid \omega_j)$$

**A simple loss function**

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

## The General Case (Cont.)

### The problem

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

We need to know the *loss* of taking each action $\alpha_i$ $(1 \leq i \leq a)$

true state of nature is $\omega_j$

the action being taken is $\alpha_i$

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

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

**Expected (average) loss**

# The General Case (Cont.)

## Expected Loss

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

Average by enumerating over all possible states of nature

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

The expected loss is also named as ***(conditional)***

**Risk**

# The General Case (Cont.)

## Suppose we have:

| Class \ Action | $\alpha_1$ = "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 **x**:

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

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

$$R(\alpha_1 \mid \mathbf{x}) = \sum_{j=1}^{2} \lambda(\alpha_1 \mid \omega_j) \cdot P(\omega_j \mid \mathbf{x})$$

$$= \lambda(\alpha_1 \mid \omega_1) \cdot P(\omega_1 \mid \mathbf{x}) + \lambda(\alpha_1 \mid \omega_2) \cdot P(\omega_2 \mid \mathbf{x})$$

$$= 5 \times 0.01 + 60 \times 0.99 = 59.45$$

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

## The General Case (Cont.)

**The task:** *find a mapping from patterns to actions*

$$\alpha : \mathbf{R}^d \rightarrow \mathcal{A} \quad \text{(decision function)}$$

In other words, for every $\mathbf{x}$, the decision function $\alpha(\mathbf{x})$ assumes one of the $a$ actions $\alpha_1, \ldots, \alpha_a$

**Overall risk $R$**

*expected loss with decision function* $\alpha(\cdot)$

$$R = \int \underline{\underline{R(\alpha(\mathbf{x}) \mid \mathbf{x})}} \cdot \underline{\underline{p(\mathbf{x})}} d\mathbf{x}$$

*Conditional risk* for pattern $\mathbf{x}$ with action $\alpha(\mathbf{x})$

pdf for patterns

## The General Case (Cont.)

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

For every **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 **x** must be as small as possible

**Bayes decision rule** (*General case*)

$$\alpha(\mathbf{x}) = \arg\min_{\alpha_i \in \mathcal{A}} R(\alpha_i \mid \mathbf{x})$$

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

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

# Two Category Classification

## Special case

$\lambda_{11}$ $\lambda_{21}$

$\lambda_{12}$ $\lambda_{22}$

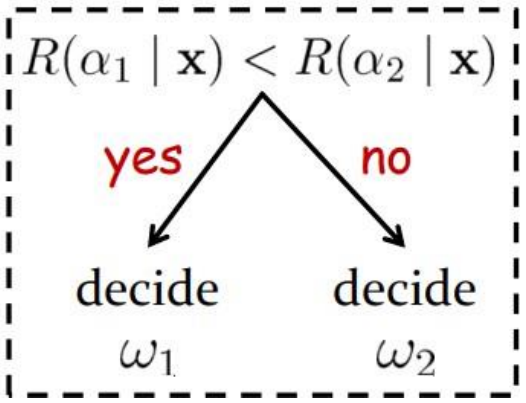
- $\Omega = \{\omega_1, \omega_2\}$ (two states of nature)
- $\mathcal{A} = \{\alpha_1, \alpha_2\}$ ($\alpha_1 =$ decide $\omega_1$; $\alpha_2 =$ decide $\omega_2$)

$\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$ : 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

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

by re-arrangement

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

by Bayes theorem

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

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

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

likelihood ratio

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

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

# Minimum Error-rate Classification

## Classification setting

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

## Zero-one (symmetrical) loss function

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

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

# Minimum Error-rate Classification (Cont.)

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

$$= \sum_{j \neq i} \lambda(\alpha_i \mid \omega_j) \cdot P(\omega_j \mid \mathbf{x}) \; + \; \lambda(\alpha_i \mid \omega_i) \cdot P(\omega_i \mid \mathbf{x})$$

$$= \sum_{j \neq i} P(\omega_j \mid \mathbf{x})$$

$$= 1 - P(\omega_i \mid \mathbf{x})$$

**Error Rate**

the probability that action $\alpha_i$ (decide $\omega_i$) is wrong

**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 Function

**Classification**

**Pattern → Category**

**actions** ⟷ **decide categories**

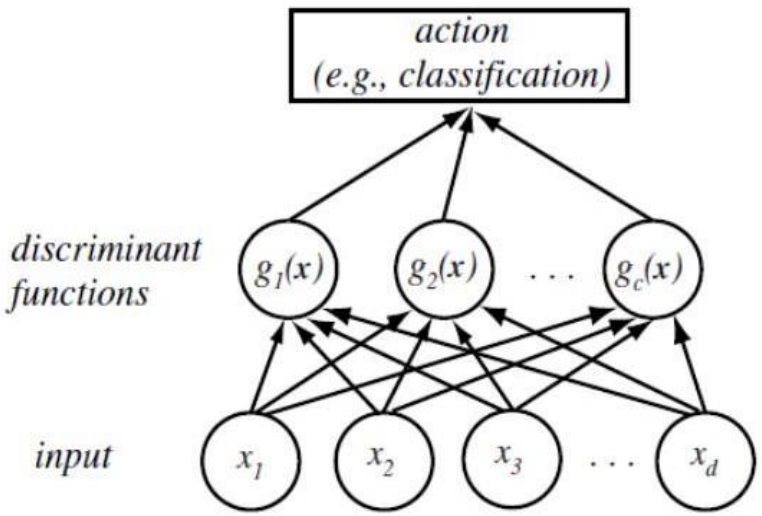
## Discriminant functions

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

- *Useful way to represent classifiers*
- *One function per category*

Decide $\omega_i$

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



## Discriminant Function (Cont.)

Minimum risk: $g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x}) \quad (1 \le i \le c)$

Minimum-error-rate: $g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) \quad (1 \le i \le c)$

**Various discriminant functions** → **Identical classification results**

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

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

e.g.:

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

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

# Discriminant Function (Cont.)

## Decision region

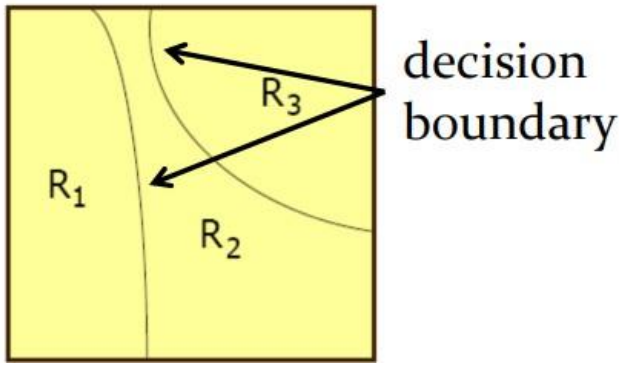
$c$ discriminant functions $g_i(\cdot)\ (1 \le i \le c)$ → $c$ decision regions $\mathcal{R}_i \subset \mathbf{R}^d\ (1 \le i \le c)$

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

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



# Expected Value

**Expected Value** , a.k.a. *expectation*, *mean* or *average* of a random variable $x$

## Discrete case

$x \in \mathcal{X} = \{x_1, x_2, \ldots, x_c\}$

$x \sim P(\cdot)$

($\sim$: "has the distribution")

$$\mathcal{E}[x] = \sum_{x \in \mathcal{X}} x \cdot P(x) = \sum_{i=1}^{c} x_i \cdot P(x_i)$$

## Continuous case

Notation: $\mu = \mathcal{E}[x]$

$x \in \mathbf{R}$

$x \sim p(\cdot)$

$$\mathcal{E}[x] = \int_{-\infty}^{\infty} x \cdot p(x)\, dx$$

# Expected Value

Given random variable $x$ and function $f(\cdot)$, what is the expected value of $f(x)$?

**Discrete case:** $\mathcal{E}[f(x)] = \sum_{x \in \mathcal{X}} f(x) \cdot P(x) = \sum_{i=1}^{c} f(x_i) \cdot P(x_i)$

**Continuous case:** $\mathcal{E}[f(x)] = \int_{-\infty}^{\infty} f(x) \cdot p(x)\, dx$

**Variance** $\mathrm{Var}[x] = \mathcal{E}[(x - \mathcal{E}[x])^2]$ (i.e. $f(x) = (x - \mu)^2$)

**Discrete case:** $\mathrm{Var}[x] = \sum_{i=1}^{c} (x_i - \mu)^2 \cdot P(x_i)$

**Continuous case:** $\mathrm{Var}[x] = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot p(x)\, dx$

Notation: $\sigma^2 = \mathrm{Var}[x]$ ($\sigma$: *standard deviation* )

# Gaussian Density- Univariate Case

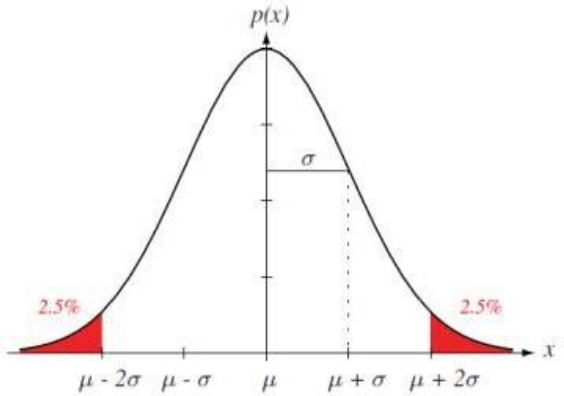
Gaussian density , a.k.a. *normal density*

for continuous random variable

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

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

$$\mathcal{E}[x] = \int_{-\infty}^{\infty} x \cdot p(x) = \mu$$




$$\mathrm{Var}[x] = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot p(x) = \sigma^2$$

# Vector Random Variable

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}$$

$$\mathbf{x} \sim p(\mathbf{x}) = p(x_1, x_2, \ldots, x_d) \quad \textbf{(joint pdf)}$$

$$p(\mathbf{x}_1) = \int p(\mathbf{x}_1, \mathbf{x}_2)\, d\mathbf{x}_2 \quad \textbf{(marginal pdf)}$$

$$(\mathbf{x}_1 \cap \mathbf{x}_2 = \emptyset;\ \mathbf{x}_1 \cup \mathbf{x}_2 = \mathbf{x})$$

## Expected vector

$$\mathcal{E}[\mathbf{x}] = \begin{pmatrix} \mathcal{E}[x_1] \\ \mathcal{E}[x_2] \\ \vdots \\ \mathcal{E}[x_d] \end{pmatrix}$$

$$\mathcal{E}[x_i] = \int_{-\infty}^{\infty} x_i \cdot p(x_i)\, dx_i \quad (1 \le i \le d)$$

**marginal pdf** on the $i$-th component

Notation:

$$\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}];\ \mu_i = \mathcal{E}[x_i]\ (1 \le i \le d)$$

# Vector Random Variable

## Covariance matrix

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

Properties of $\boldsymbol{\Sigma}$

- symmetric
- Positive semidefinite

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

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j) \cdot p(x_i, x_j)\, dx_i dx_j$$

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

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

# Gaussian Density- Multivariate Case

$$\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \mu_i = \mathcal{E}[x_i] \quad \sigma_{ij} = \sigma_{ji} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)]$$

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

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

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

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

$d \times d$ *covariance matrix*

$|\boldsymbol{\Sigma}|$ : determinant

$\boldsymbol{\Sigma}^{-1}$ : inverse

## Gaussian Density- Multivariate Case

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

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

$\boldsymbol{\Sigma}^{-1}$ : $d \times d$ matrix

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

$(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$

scalar $(1 \times 1$ matrix$)$

$\boldsymbol{\Sigma}$ : positive definite

$\boldsymbol{\Sigma}^{-1}$ : positive definite

$(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \geq 0$

$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \leq 0$

# Discriminant Function for Gaussian Density

Minimum-error-rate classification

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

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

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

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

Constant, could be ignored

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

# Discriminant Function for Gaussian Density (Cont.)

## Case I: $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}$

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

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

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

$$\boldsymbol{\Sigma}_i = \sigma^2 \cdot \begin{pmatrix} 1 & & \\ & 1 & \\ & & \ddots \\ & & & 1 \end{pmatrix} = \begin{pmatrix} \sigma^2 & & \\ & \sigma^2 & \\ & & \ddots \\ & & & \sigma^2 \end{pmatrix} \Rightarrow \quad |\boldsymbol{\Sigma}_i| = \sigma^{2d}$$

$$\boldsymbol{\Sigma}_i^{-1} = (1/\sigma^2)\,\mathbf{I}$$

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

$||\cdot||$ : *Euclidean norm*

$$||\mathbf{x}-\boldsymbol{\mu}_i||^2 = (\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i)$$

# Discriminant Function for Gaussian Density (Cont.)

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

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

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

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

Linear discriminant functions

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

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i \quad \textit{weight vector}$$

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

# Discriminant Function for Gaussian Density (Cont.)

## Case II: $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$

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

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

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* (马氏距离)

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

P. C. Mahalanobis
(1893-1972)

# Discriminant Function for Gaussian Density (Cont.)

## 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

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

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

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

# Discriminant Function for Gaussian Density (Cont.)

## Case III: $\boldsymbol{\Sigma}_i = \text{arbitrary}$

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

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

quadratic discriminant functions (二次判别函数)

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

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

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

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

# The End