

Chapter 2

Bayesian Decision Theory



Decision Theory

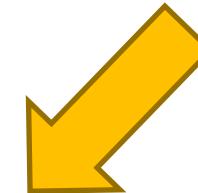
Decision

Make choice under uncertainty



Pattern Recognition

Pattern → 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 (贝叶斯定理)



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)

A Specific Example

State of Nature (自然状态)

- 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 ω 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 (先验概率)

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 ω with c possible states of nature $\{\omega_1, \omega_2, \dots, \omega_c\}$, such that:

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

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  $P(\omega_1) = 2/3; P(\omega_2) = 1/3$



Decision Before Observation

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 ω_1 if $P(\omega_1) > P(\omega_2)$; otherwise decide ω_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 \mathbf{R} , such that:

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

- 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 (类条件概率密度)

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

$$p(x|\omega)$$

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

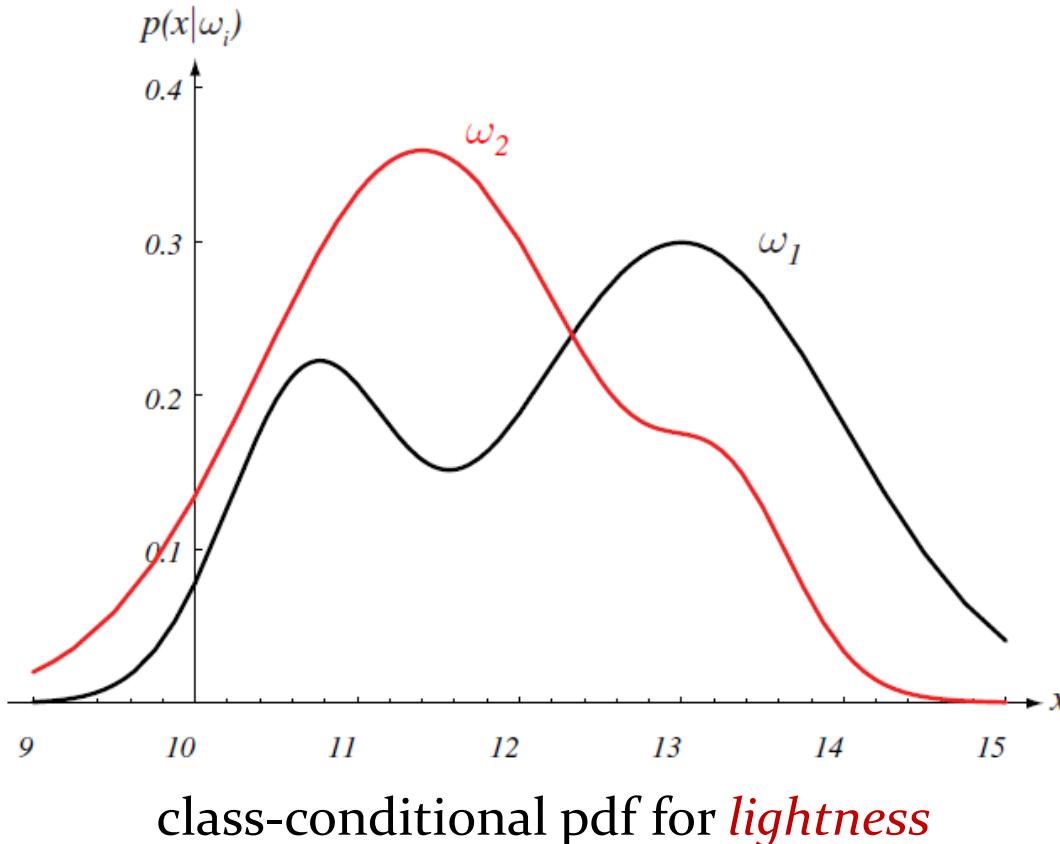
- The *class-conditional* pdf describes the difference in the distribution of observations under different classes

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



Class-Conditional PDF

An illustrative example



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

Known

Prior probability

$$P(\omega_j) \quad (1 \leq j \leq c)$$

Class-conditional
pdf

$$p(x|\omega_j) \quad (1 \leq j \leq c)$$

Observation for
test example

$$x^* \quad (\text{e.g.: fish lightness})$$

Unknown

The quantity which we want to use
in decision naturally (by exploiting
observation information)

Bayes
Formula

Posterior probability

$$P(\omega_j|x^*) \quad (1 \leq j \leq c)$$

Convert the prior probability $P(\omega_j)$
to the posterior probability $P(\omega_j|x^*)$

Bayes Formula Revisited

Joint probability density function (联合分布) $p(\omega, x)$

Marginal distribution (边缘分布) $P(\omega)$ $p(x)$

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

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



Law of total probability (全概率公式) [ref. pp.615]

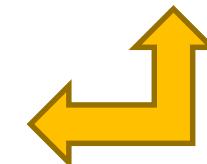
$$p(\omega, x) = P(\omega|x) \cdot p(x)$$

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



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

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



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

[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.)

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

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

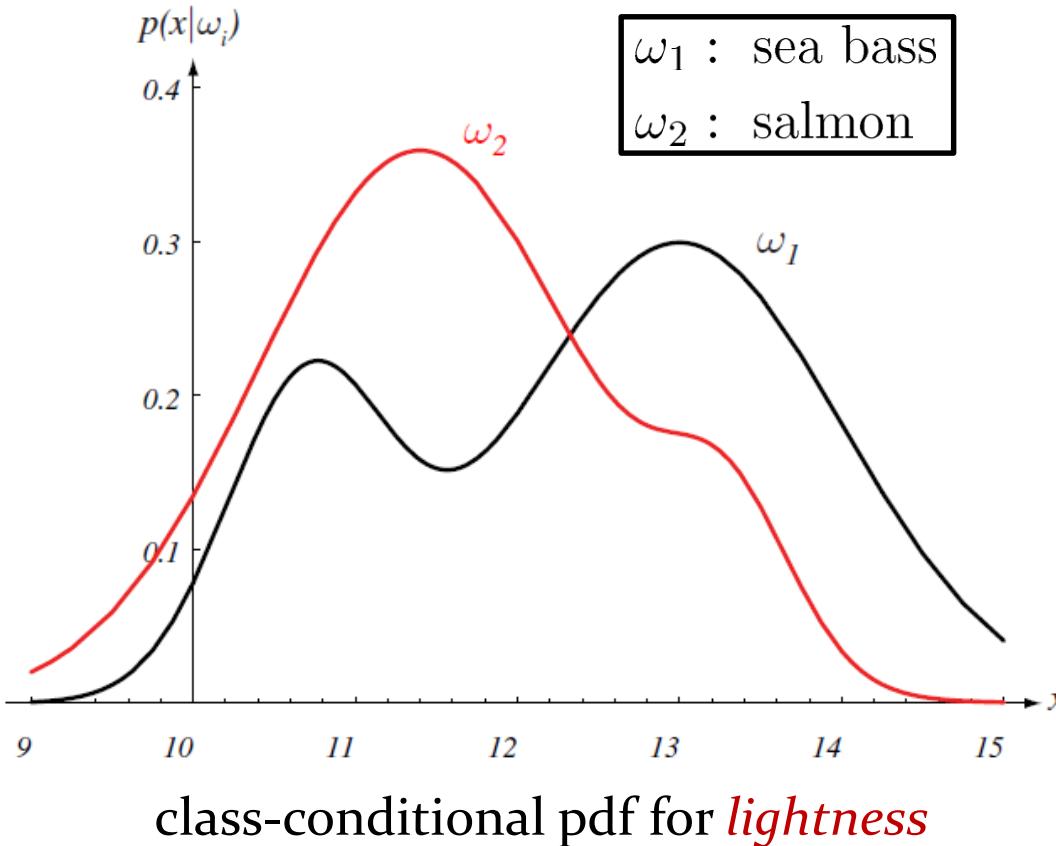
$$p(x|\omega_1) = p(x|\omega_2) = \cdots = p(x|\omega_c)$$


Degenerate to naive decision rule

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

Bayes Formula Revisited (Cont.)

An illustrative example



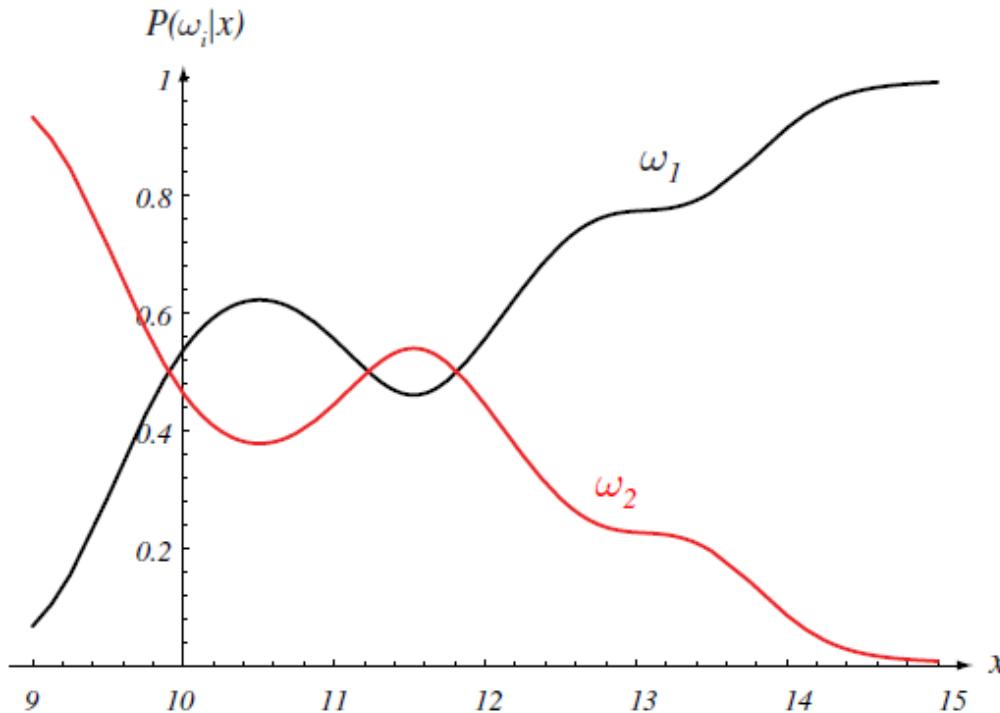
$$P(\omega_1) = \frac{2}{3}$$

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

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.)

ω_1 : cancer

ω_2 : no cancer

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

$$P(\omega_1) = 0.008$$

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

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

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

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

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

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

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

$P(\omega_2 | +) > P(\omega_1 | +)$
No cancer!



Feasibility of Bayes Formula

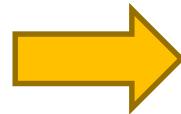
$$P(\omega|x) = \frac{p(x|\omega) \cdot P(\omega)}{p(x)} \quad (\text{Bayes Formula})$$

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 (相对频率)
- An advanced solution: Conduct density estimation (概率密度估计)

A Further Example

Problem statement

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

ω_1 : price > \$50,000

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

ω_2 : price \leq \$50,000

?

x : height of car

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

Quantities to know:

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



Counting relative frequencies via collected samples

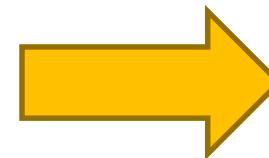
A Further Example (Cont.)

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 ω_1 : 221



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

cars in ω_2 : 988

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

A 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 ω_1 , # cars in I_x
is 46

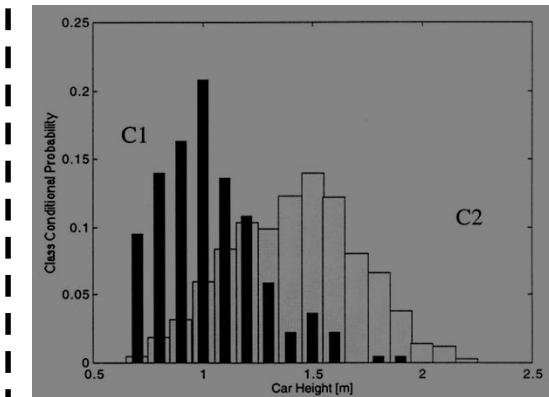
For ω_2 , # cars in I_x
is 59

$$p(x = 1.05|\omega_1)$$

$$= \frac{46}{221} = 0.2081$$

$$p(x = 1.05|\omega_2)$$

$$= \frac{59}{988} = 0.0597$$



A Further Example (Cont.)

Question

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

Estimated quantities

$$P(\omega_1) = 0.183$$

$$P(\omega_2) = 0.817$$

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

$$\begin{aligned} \frac{P(\omega_2 \mid x = 1.05)}{P(\omega_1 \mid x = 1.05)} &= \frac{P(\omega_2) \cdot p(x = 1.05 \mid \omega_2)}{p(x = 1.05)} \Bigg/ \frac{P(\omega_1) \cdot p(x = 1.05 \mid \omega_1)}{p(x = 1.05)} \\ &= \frac{P(\omega_2) \cdot p(x = 1.05 \mid \omega_2)}{P(\omega_1) \cdot p(x = 1.05 \mid \omega_1)} \\ &= \frac{0.817 \times 0.0597}{0.183 \times 0.2081} = 1.280 \end{aligned}$$

$P(\omega_2 \mid x) > P(\omega_1 \mid x)$
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 ω_1 ; Otherwise ω_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

Bayes Decision Rule – 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

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

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

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

Note : $c \neq a$

Bayes Decision Rule – The General Case (Cont.)

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

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

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



For ease of reference,
usually written as:
 $\lambda(\alpha_i | \omega_j)$

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.)

The problem

Given a particular \mathbf{x} , we have to decide which action to take



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

true state of nature is ω_j



the action being taken is α_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 (Cont.)

Expected loss (期望损失)



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

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



The incurred loss of taking action α_i in case of true state of nature being ω_j

The probability of ω_j being the true state of nature

The expected loss is also named as (*conditional*) **risk** (条件风险)

Bayes Decision Rule – The General Case (Cont.)

Suppose we have:

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

For a particular \mathbf{x} :

$$P(\omega_1 | \mathbf{x}) = 0.01$$
$$P(\omega_2 | \mathbf{x}) = 0.99$$

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

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



Bayes Decision Rule – 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 $\alpha_1, \dots, \alpha_a$

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*



Bayes Decision Rule – 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 \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)

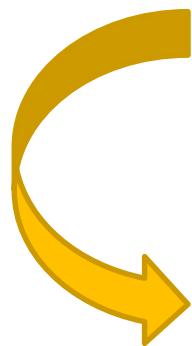
$$\begin{aligned}\alpha(\mathbf{x}) &= \arg \min_{\alpha_i \in \mathcal{A}} R(\alpha_i \mid \mathbf{x}) \\ &= \arg \min_{\alpha_i \in \mathcal{A}} \sum_{j=1}^c \lambda(\alpha_i \mid \omega_j) \cdot P(\omega_j \mid \mathbf{x})\end{aligned}$$

- 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

λ_{11}	λ_{21}
λ_{12}	λ_{22}



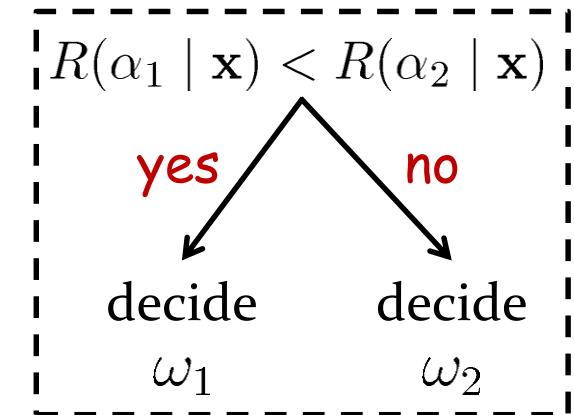
- $\Omega = \{\omega_1, \omega_2\}$ (two states of nature)
- $\mathcal{A} = \{\alpha_1, \alpha_2\}$ (α_1 = decide ω_1 ; α_2 = decide ω_2)

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

The conditional risk:

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

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



Two-Category Classification (Cont.)

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

by definition



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

by re-arrangement



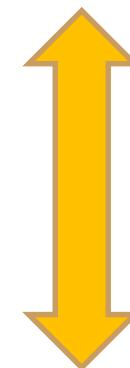
$$\boxed{(\lambda_{21} - \lambda_{11})P(\omega_1 \mid \mathbf{x}) > (\lambda_{12} - \lambda_{22})P(\omega_2 \mid \mathbf{x})}$$

likelihood ratio



constant θ
independent of \mathbf{x}

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



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

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

by Bayes theorem



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

Minimum-Error-Rate Classification

Classification setting

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

Zero-one (symmetrical) loss function

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

- 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.)

$$\begin{aligned} R(\alpha_i \mid \mathbf{x}) &= \sum_{j=1}^c \lambda(\alpha_i \mid \omega_j) \cdot P(\omega_j \mid \mathbf{x}) \\ &= \sum_{j \neq i} \lambda(\alpha_i \mid \omega_j) \cdot P(\omega_j \mid \mathbf{x}) + \lambda(\alpha_i \mid \omega_i) \cdot P(\omega_i \mid \mathbf{x}) \\ &= \sum_{j \neq i} P(\omega_j \mid \mathbf{x}) \quad \text{error rate (误差率/错误率)} \\ &= (1 - P(\omega_i \mid \mathbf{x})) \quad \text{the probability that action } \alpha_i \text{ (decide } \omega_i) \text{ is wrong} \end{aligned}$$

Minimum error rate

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



Minimax Criterion

Generally, we assume that the prior probabilities over the states of nature $\Omega = \{\omega_1, \omega_2, \dots, \omega_c\}$ are fixed

Nonetheless, in some cases we need to design classifiers which can perform well under **varying prior probabilities**

e.g. the prior probabilities of catching a sea bass or salmon fish might vary in different regions

Varying prior probabilities leads to varying overall risk



The minimax criterion (极小化极大准则) aims to find the classifier which can **minimize the worst overall risk** for any value of the priors

Minimax Criterion (Cont.)

Two-category classification

λ_{11}	λ_{21}
λ_{12}	λ_{22}



- $\Omega = \{\omega_1, \omega_2\}$ (two states of nature)
- $\mathcal{A} = \{\alpha_1, \alpha_2\}$ (α_1 = decide ω_1 ; α_2 = decide ω_2)

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

Suppose the two-category classifier $\alpha(\cdot)$ decides ω_1 in region \mathcal{R}_1 and decides ω_2 in region \mathcal{R}_2 . Here, $\mathcal{R}_1 \cup \mathcal{R}_2 = \mathbf{R}^d$ and $\mathcal{R}_1 \cap \mathcal{R}_2 = \emptyset$.

The overall risk:

$$\begin{aligned} R &= \int R(\alpha(\mathbf{x}) | \mathbf{x})) \cdot p(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathcal{R}_1} R(\alpha_1 | \mathbf{x})) \cdot p(\mathbf{x}) d\mathbf{x} + \int_{\mathcal{R}_2} R(\alpha_2 | \mathbf{x})) \cdot p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

Minimax Criterion (Cont.)

$$R = \int_{\mathcal{R}_1} R(\alpha_1 \mid \mathbf{x})) \cdot p(\mathbf{x}) d\mathbf{x} + \int_{\mathcal{R}_2} R(\alpha_2 \mid \mathbf{x})) \cdot p(\mathbf{x}) d\mathbf{x}$$

$$\begin{aligned} & \int_{\mathcal{R}_1} R(\alpha_1 \mid \mathbf{x})) \cdot p(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathcal{R}_1} \sum_{j=1}^2 R(\alpha_1 \mid \omega_j) \cdot P(\omega_j \mid \mathbf{x}) \cdot p(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathcal{R}_1} \sum_{j=1}^2 \lambda_{1j} \cdot P(\omega_j) \cdot p(\mathbf{x} \mid \omega_j) d\mathbf{x} \\ &= \int_{\mathcal{R}_1} [\lambda_{11} \cdot P(\omega_1) \cdot p(\mathbf{x} \mid \omega_1) + \lambda_{12} \cdot P(\omega_2) \cdot p(\mathbf{x} \mid \omega_2)] d\mathbf{x} \end{aligned}$$

$$\begin{aligned} & \int_{\mathcal{R}_2} R(\alpha_2 \mid \mathbf{x})) \cdot p(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathcal{R}_2} [\lambda_{21} \cdot P(\omega_1) \cdot p(\mathbf{x} \mid \omega_1) + \lambda_{22} \cdot P(\omega_2) \cdot p(\mathbf{x} \mid \omega_2)] d\mathbf{x} \end{aligned}$$

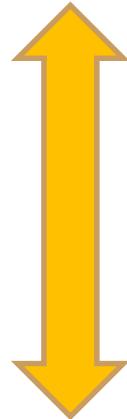
Eq.22 [pp.28]



Minimax Criterion (Cont.)

$$R = \int_{\mathcal{R}_1} [\lambda_{11} \cdot P(\omega_1) \cdot p(\mathbf{x} \mid \omega_1) + \lambda_{12} \cdot P(\omega_2) \cdot p(\mathbf{x} \mid \omega_2)] d\mathbf{x}$$

$$+ \int_{\mathcal{R}_2} [\lambda_{21} \cdot P(\omega_1) \cdot p(\mathbf{x} \mid \omega_1) + \lambda_{22} \cdot P(\omega_2) \cdot p(\mathbf{x} \mid \omega_2)] d\mathbf{x}$$



Rewrite the overall risk R as a function of $P(\omega_1)$ via:

- $P(\omega_1) = 1 - P(\omega_2)$
- $\int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_1) d\mathbf{x} = 1 - \int_{\mathcal{R}_2} p(\mathbf{x} \mid \omega_1) d\mathbf{x}$

$$R = \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x}$$
$$+ P(\omega_1) \left[(\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(\mathbf{x} \mid \omega_1) d\mathbf{x} - (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x} \right]$$

Minimax Criterion (Cont.)

$$R_{mm} = \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x}$$

$$= R_{mm}, \text{ minimax risk} \quad = \lambda_{11} + (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(\mathbf{x} \mid \omega_1) d\mathbf{x}$$

$$R = \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x} \\ + P(\omega_1) \left[(\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(\mathbf{x} \mid \omega_1) d\mathbf{x} - (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x} \right]$$

=0 for minimax solution

A linear function of $P(\omega_1)$, which can also be expressed as a linear function of $P(\omega_2)$ in similar way.



Discriminant Function (判別函数)

Classification

Pattern → Category

actions ←→ decide categories

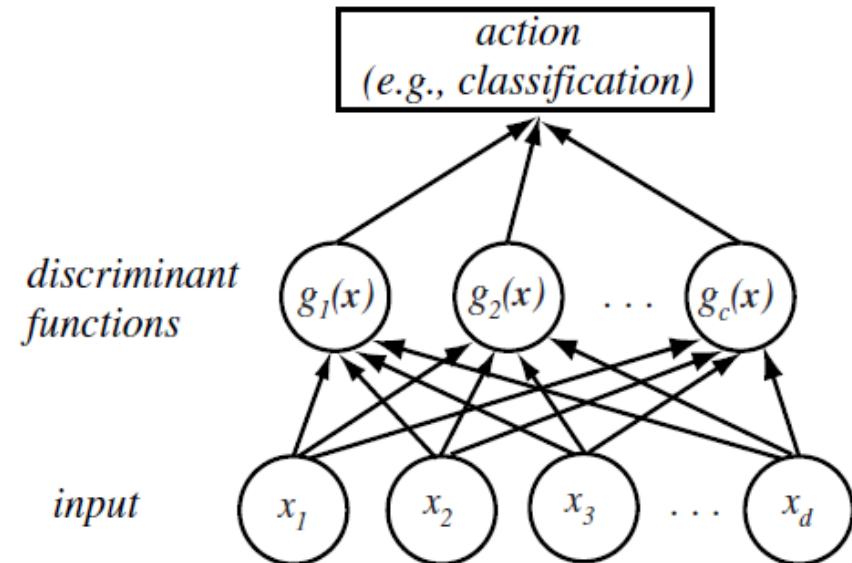
Discriminant functions

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

- Useful way to represent classifiers
- One function per category

Decide ω_i

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



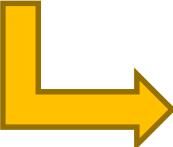
Discriminant Function (Cont.)

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

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

Various
discriminant functions  Identical
classification results

$f(\cdot)$ is a *monotonically increasing function* (单调递增函数)

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

e.g.:

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

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



Discriminant Function (Cont.)

Decision region (决策区域)

c discriminant functions

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



c decision regions

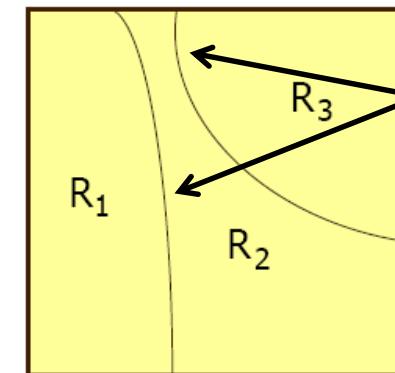
$$\mathcal{R}_i \subset \mathbf{R}^d \quad (1 \leq i \leq c)$$

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

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

Decision boundary (决策边界)

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



decision
boundary

Expected Value

Expected value (数学期望), a.k.a. *expectation, mean* or *average* of a random variable x

Discrete case

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

$$x \sim P(\cdot)$$

(\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 (Cont.)



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

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

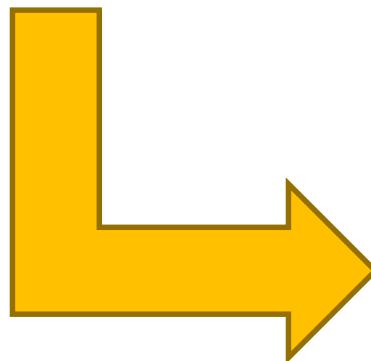
Continuous case: $\text{Var}[x] = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot p(x) dx$

Notation: $\sigma^2 = \text{Var}[x]$ (σ : standard deviation (标准偏差))

Gaussian Density – Univariate Case

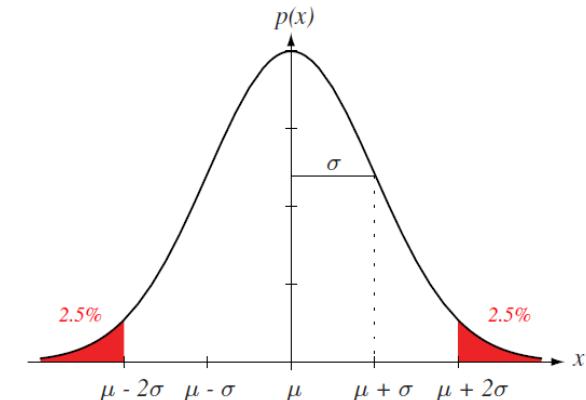
Gaussian density (高斯密度函数), a.k.a. *normal density* (正态密度函数), for continuous random variable

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



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

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



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

Vector Random Variables (随机向量)

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

$$\boxed{\begin{aligned} \mathbf{x} &\sim p(\mathbf{x}) = p(x_1, x_2, \dots, x_d) && \text{(joint pdf)} \\ p(\mathbf{x}_1) &= \int p(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_2 && \text{(marginal pdf)} \\ (\mathbf{x}_1 \cap \mathbf{x}_2 &= \emptyset; \mathbf{x}_1 \cup \mathbf{x}_2 = \mathbf{x}) \end{aligned}}$$

Expected vector

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

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

Notation:

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

 marginal pdf on
the i -th component



Vector Random Variables (Cont.)

Covariance matrix (协方差矩阵)

Properties of Σ

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

- symmetric**
(对称矩阵)
- Positive semidefinite**
(半正定矩阵)

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

Appendix A.4.9 [pp.617]

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

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

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



Gaussian Density – Multivariate Case

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

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

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

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

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

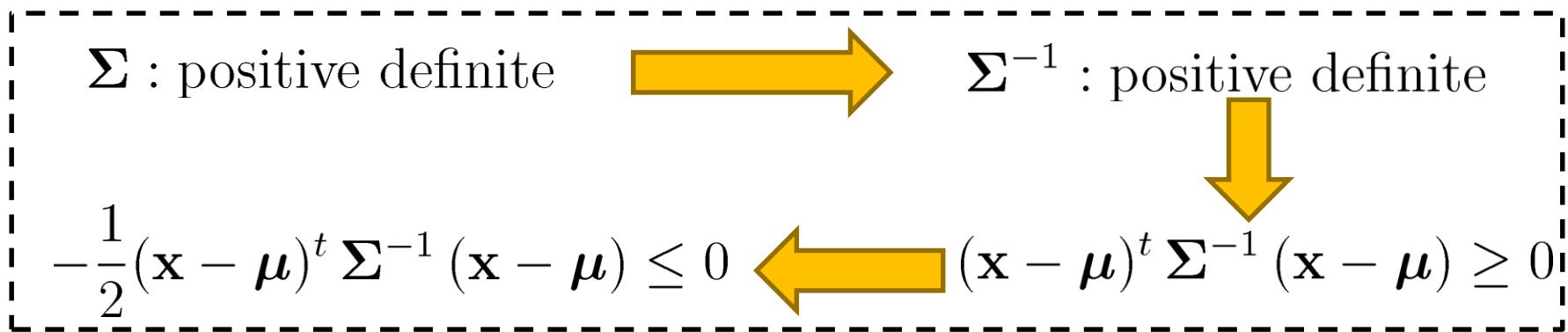
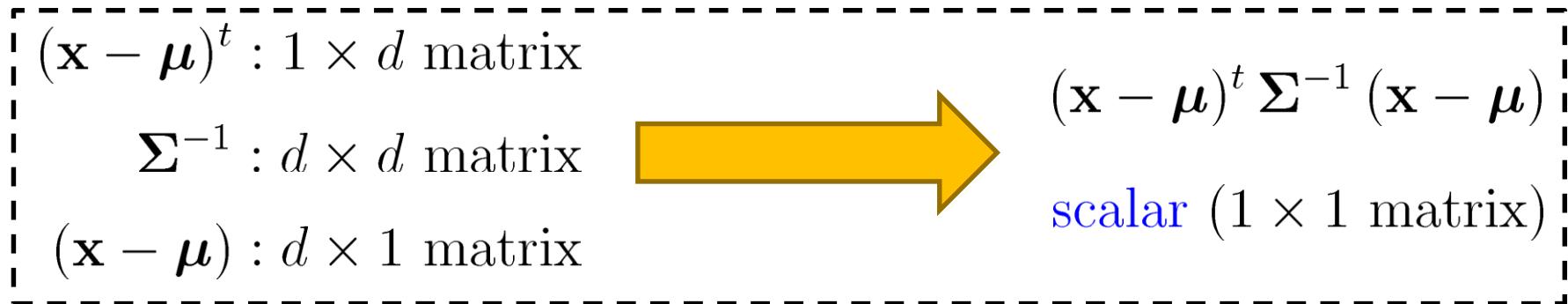
$$\Sigma = [\sigma_{ij}]_{1 \leq i,j \leq d} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_{dd} \end{pmatrix}$$

d × d covariance matrix
 $|\Sigma|$: determinant
 Σ^{-1} : inverse



Gaussian Density – Multivariate Case (Cont.)

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



Discriminant Functions for Gaussian Density

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: σ^2 times the identity matrix \mathbf{I}

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

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



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

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



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

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

Linear discriminant functions (线性判别函数)

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

$$\mathbf{w}_i = \frac{1}{\sigma^2} \boldsymbol{\mu}_i \quad \text{weight vector (权值向量)}$$

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

Case II: $\Sigma_i = \Sigma$

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

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

Covariance matrix: *identical* for all classes

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

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

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



P. C. Mahalanobis
(1893-1972)

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

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



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

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

Linear discriminant functions

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

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

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

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 (二次判别函数)

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

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

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

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



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 → 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 \leftrightarrow 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
 - 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

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

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