Chapter 2

Bayesian Decision Theory

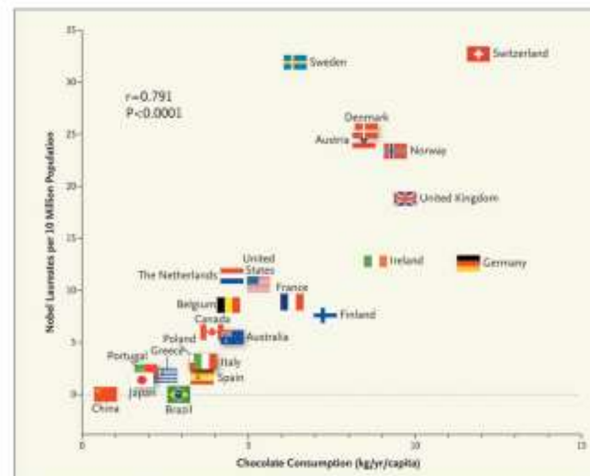
The 3W of Bayesian Decision Theory

- **W**hat is Bayesian Decision Theory?
  - What is Bayesian?
  - What is Decision?
  - What is Theory?
- **W**hy do we need Bayesian Decision Theory?
- Ho**W** to use Bayesian Decision Theory?

Decision: a tale of two sides

- Observing a photo of some animal at the farm gate, deciding whether such animal is: a sheep or a goat

- Observing the data that the chocolate consumption is highly predictive for the number of Nobel prize laureates, deciding whether to make laws for forcing people eat more chocolate

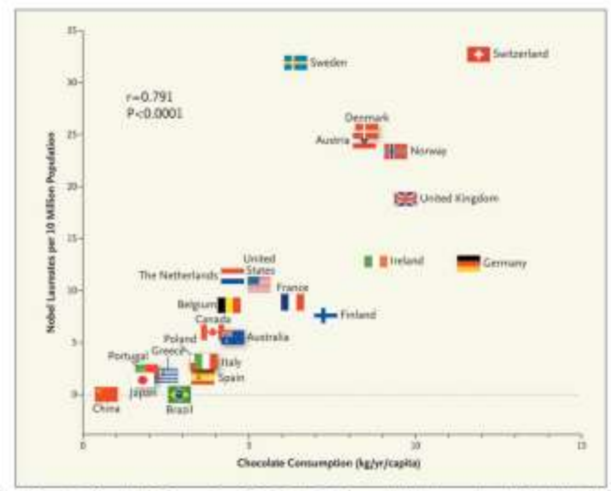


F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012

Decision: a tale of two sides (Cont.)

- Observing a photo of some animal at the farm gate, deciding whether such animal is: a sheep or a goat
  - Requires no change to the objective world
  - **Prediction, or predictive decision**

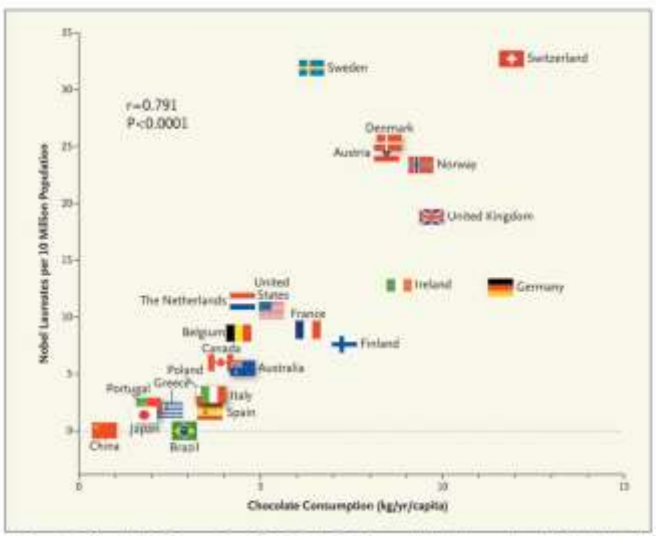
- Observing the data that the chocolate consumption is highly predictive for the number of Nobel prize laureates, deciding whether to make laws for forcing people eat more chocolate
  - Requires change to the objective world
  - **Intervention, or interventional decision**



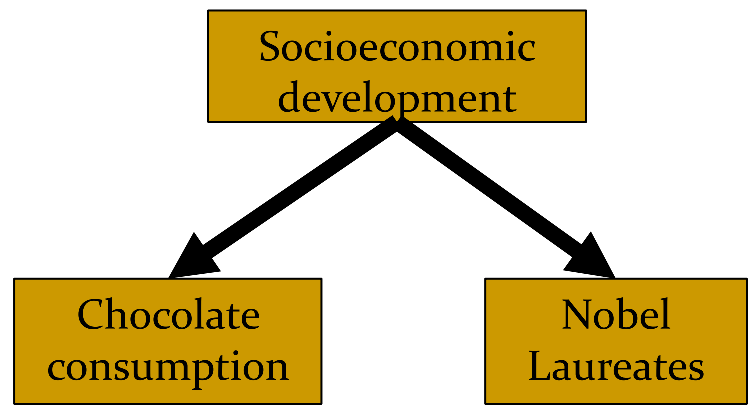
F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*. N Engl J Med 2012

Decision: a tale of two sides (Cont.)

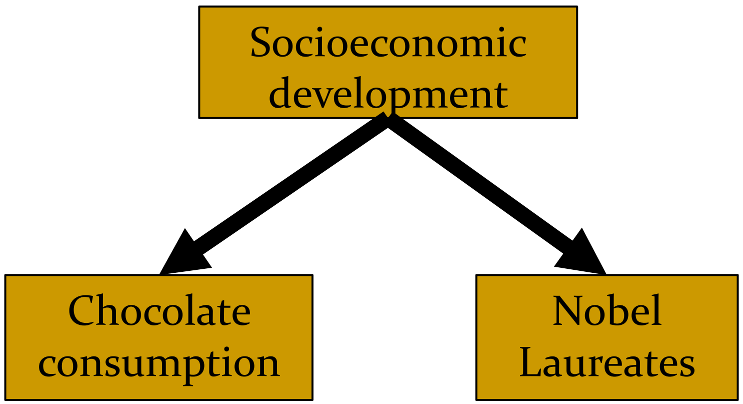
- Why increasing chocolate consumption does not affect numbers of Nobel laureates?

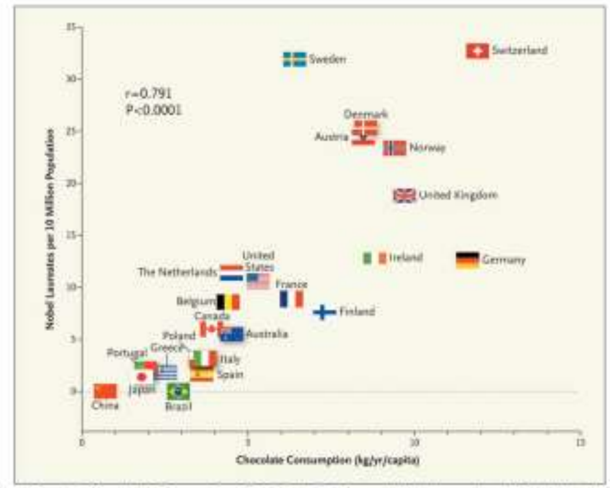


F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012



Decision: a tale of two sides (Cont.)

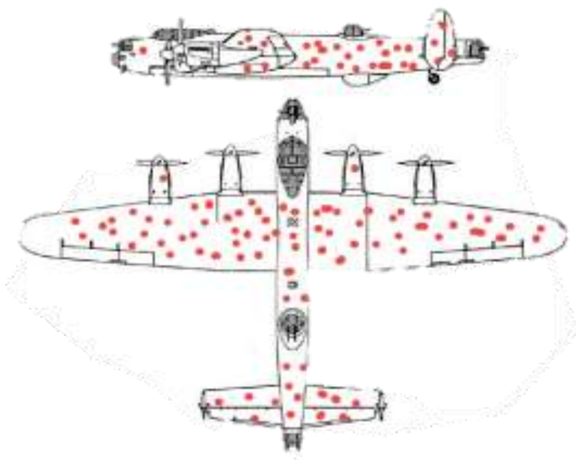


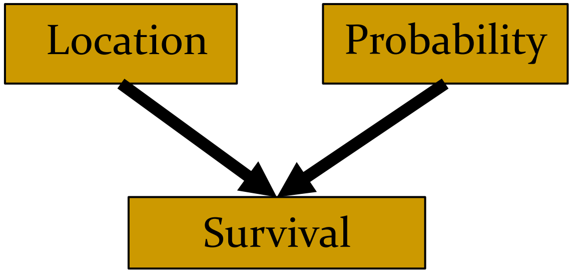


F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012

- For prediction tasks, any correlated variables can be used to predict each other.
- For intervention tasks, only the “cause” variable can be used to change the values of the “effect” variable.

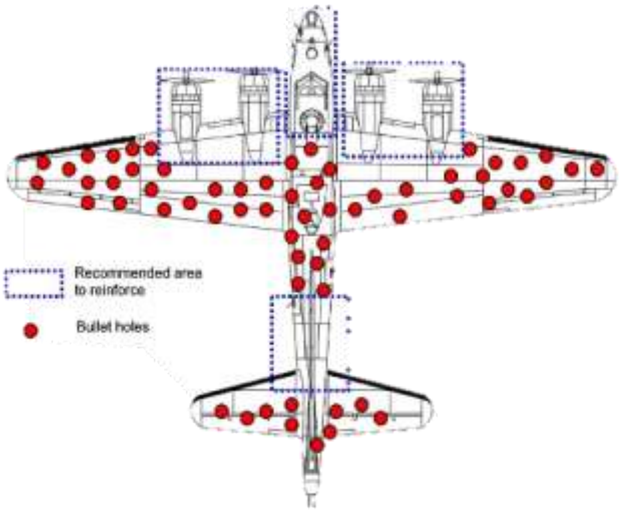
Decision: yet another perspective

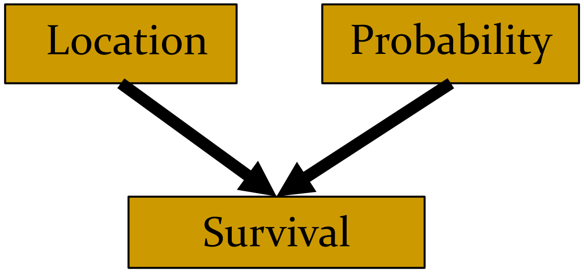




- "Location" and "probability" of bullet holes are irrelevant
- However, if we only consider planes that have survived
- Then, some locations **seems to be** less likely of being hit.

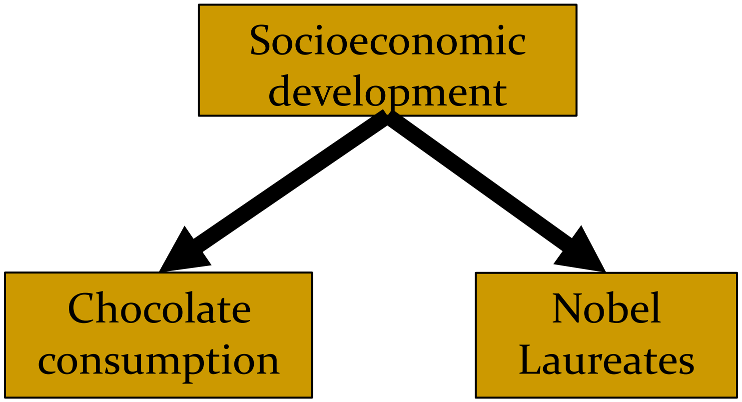
Decision: yet another perspective (cont.)

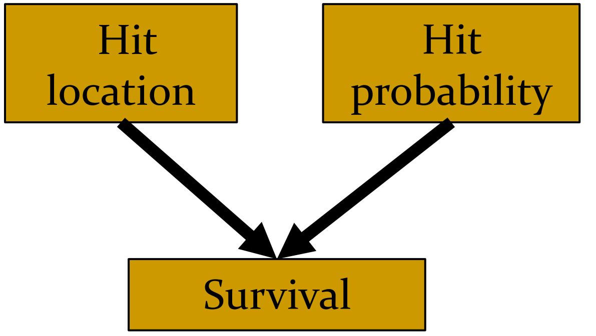




- "Location" and "probability" of bullet holes are irrelevant
- It's impossible to change either "location" or "probability"
- But we can instead improve armor!

Decision vs Prediction

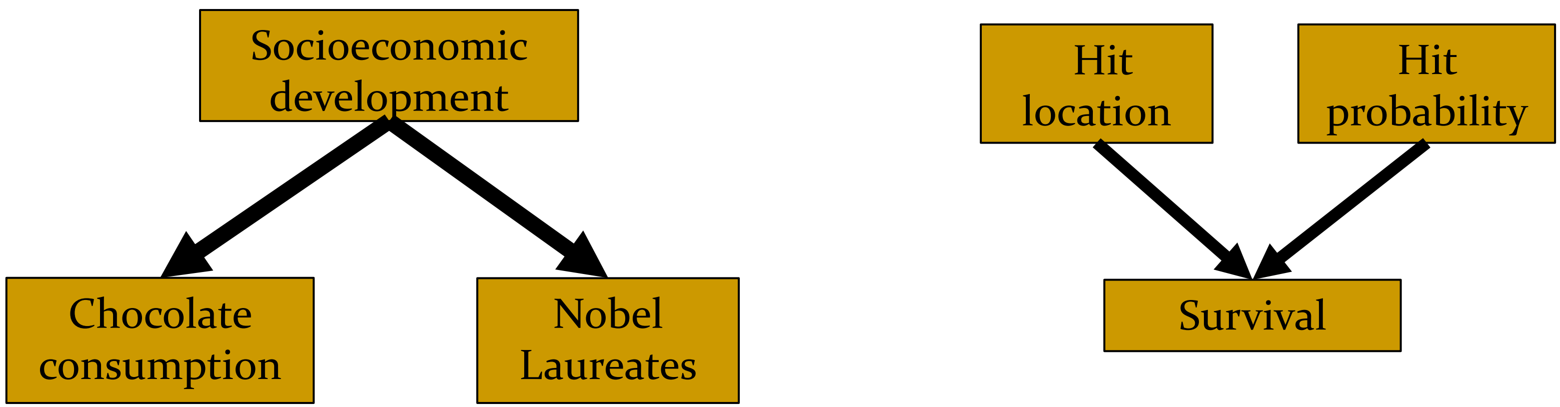




- Directed Acyclic Graph (DAG)
- Nodes represent variables, edges represents relationships
- Can be interpreted causally under certain circumstances
  - Can answer interventional questions

Most methods in this course focus on prediction.
Interventional questions requires causal methods to answer.

Decision: a tale of two sides (Cont.)



- The first example is called "spurious correlation"
- The second example is called "selection bias"

Decision: a tale of two sides (Cont.)

- Recall the goal of pattern recognition:
  - Generalization [泛化能力/推广能力]
- Selection bias may jeopardize generalization!

Training




Test


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

Which type of decision?

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 size of a goat*)

H: the hypothesis (*e.g. the sheep belongs to the "goat" category*)

$P(H)$: the **prior probability** (先验概率) that H holds (*e.g. the probability of goat and sheep*)

$P(X)$: the **evidence probability** that X is observed (*e.g. the probability of observing an animal with curly wool*)

$P(X|H)$: the **likelihood** (似然度) of observing X given that H holds (e.g. *the probability of observing a goat with very curly wool*)

$P(H|X)$: the **posterior probability** (后验概率) that H holds given X (e.g. *the probability of X being goat given its wool is curly*)

Thomas Bayes (1702-1761)

Bayesian vs Frequentist

- The Bayesian billiard game
  - Alice and Bob **can't** see the billiard table.
  - The judge rolls a ball down the table, and marks where it lands. Once this **mark** is in place, the judge rolls new balls.
  - If the ball lands to the left of the mark, Alice gets a point; if it lands to the right of the mark, Bob gets a point.
  - The first person to reach **6 points** wins the game.
  - Now say that Alice is leading with **5** points and Bob has **3** points.

What can be said about the chances of Bob to win the game?

Bayesian vs Frequentist (Cont.)

- The Frequentist Approach
  - 5 balls out of 8 balls fell on Alice's side
  - Maximum likelihood estimate of θ that balls land on Alice's side:
    - $L(\theta) = p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$
    - $\log L(\theta) = y \log \theta + (n-y) \log(1-\theta) + C$
    - $\hat{\theta} = \frac{y}{n} = 5/8$
  - Assuming this maximum likelihood probability, we can compute the probability that Bob will win, which is given by:
    - $\mathrm{P}(Bob\ Wins) = (1-0.675)^3 = 0.052734375$

Frequentist concludes that Bob got 5.2% chance of winning!

Bayesian vs Frequentist (Cont.)

- The Bayesian Approach
  - Prior distributions: $\theta \sim Uniform(0,1)$
  - $\mathbb{E}(Bob\ wins) = \int_0^1 (1-\theta)^3 P(\theta | A = 5, B = 3) d\theta$
  - $P(\theta | A = 5, B = 3) = \frac{P(\theta)P(A=5,B=3|\theta)}{\int_0^1 P(\theta)P(A=5,B=3|\theta)d\theta}$
  - $P(A = 5, B = 3|\theta) = \binom{8}{5}\theta^5(1-\theta)^3$, $P(\theta) = 1$
  - $\mathbb{E}(Bob\ wins) = \frac{\int_0^1 (1-\theta)^6 \theta^5 d\theta}{\int_0^1 (1-\theta)^3 \theta^5 d\theta} = \frac{5!6!/12!}{5!3!/9!} = \frac{1}{11}$
  - Without knowing the Bayesian probability:
    - $\mathbb{E}(Bob\ Wins) = 0.091$

$$\int_0^1 p^{m-1}(1-p)^{n-1}\, dp = \frac{\Gamma(n)\Gamma(m)}{\Gamma(m+n)}$$

$$\Gamma(n+1) = n!$$

Bayesian concludes that Bob got 9.1% chance of winning!

Bayesian vs Frequentist (Cont.)

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

Likelihood: $$\max P(X|H)$$

- **Bayesian** view
  - uses probabilities for both hypotheses and data.
  - depends on the **prior and likelihood** of observed data.
  - requires one to know or construct a prior.
  - dominated statistical practice **before** the 20th century.
- **Frequentist** view
  - never uses or gives the probability of a hypothesis.
  - depends on the **likelihood** $\boldsymbol{P(X|H)}$.
  - does not require a prior.
  - dominated statistical practice **during** the 20th century.

How about the 21st century?

Difficulties of the Bayesian approach

- Computational:
  - Almost invariably requires integrations over uncertain functions which do not have analytical solutions.
- Prior distribution:
  - Prior distributions are often unknown.
- Treating hypotheses as probability is not intuitive
  - For probabilities of non-repeatable events that are either true or false, there is no frequency interpretation.

Bayes Theorem

Bayes theorem: $$P(\omega|x) = \frac{P(\omega)P(x|\omega)}{P(x)}$$

x : the observed sample (also called **evidence**; *e.g.: the size of a goat*)

ω: the hypothesis (*e.g. the sheep belongs to the "goat" category*)

$P(\omega)$: the **prior probability** (先验概率) that ω holds (*e.g. the probability of goat and sheep*)

$P(x)$: the **evidence probability** that x is observed (*e.g. the probability of observing an animal with curly wool*)

$P(x \mid \omega)$: the **likelihood** (似然度) of observing x given that ω holds (e.g. *the probability of observing a goat with very curly wool*)

$P(\omega \mid x)$: the **posterior probability** (后验概率) that ω holds given x (e.g. *the probability of* x *being goat given its wool is curly*)

A Specific Example

State of Nature (自然状态)

- Future events that might occur

  e.g. the next animal arriving at the farm gate

- 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)

$\omega = \omega_1 :$ *sheep*

$\omega = \omega_2 :$ *goat*

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 $\{\omega_1, \omega_2, \ldots, \omega_c\}$, such that:

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

the herd has as much sheep as goat ⇨ $P(\omega_1) = P(\omega_2) = 1/2$

the herd has more sheep as goat ⇨ $P(\omega_1) = 2/3; P(\omega_2) = 1/3$

Decision Before Observation

The Problem

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

Naive Decision Rule

$$\boxed{\text{Decide } \omega_1 \text{ if } P(\omega_1) > P(\omega_2); \text{ otherwise decide } \omega_2}$$

- 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,概率密度函数)
(for continuous random variable)

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

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

- 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 *curliness 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.:

$$p(x|\omega)$$

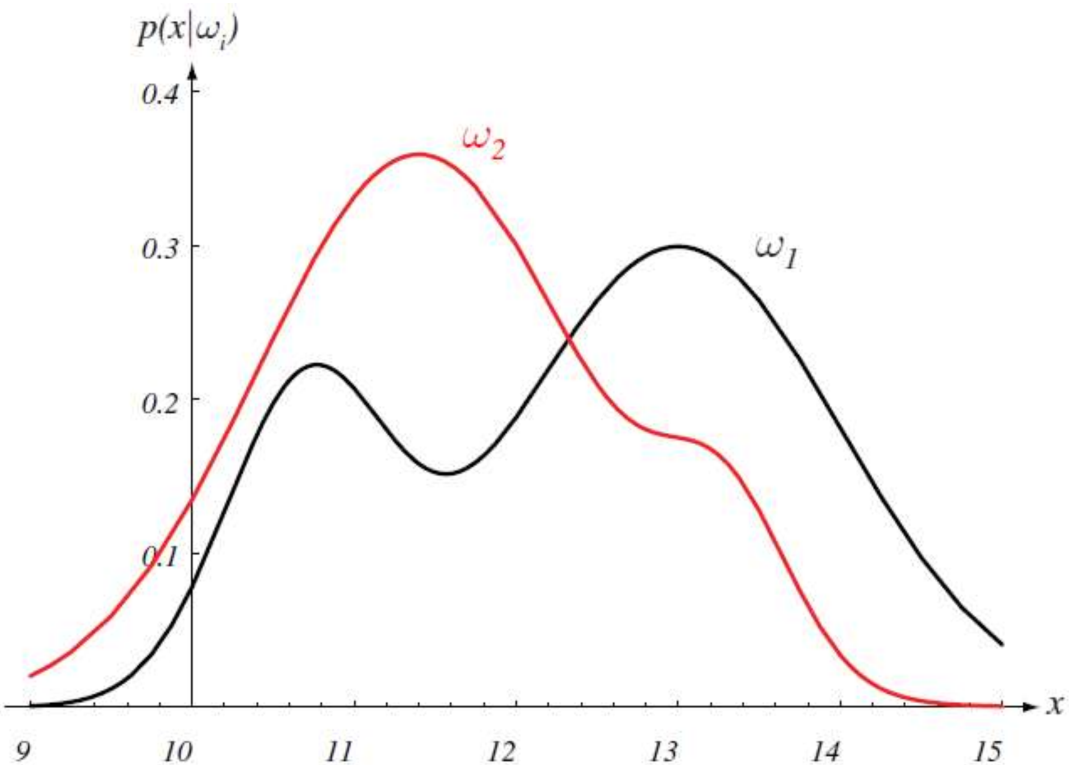
$$p(x|\omega) \geq 0 \qquad \int_{-\infty}^{\infty} p(x|\omega)dx = 1$$

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

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

Class-Conditional PDF

An illustrative example



class-conditional pdf for *curliness*

h-axis: curliness of wool

v-axis: class-conditional pdf values

black curve: sheep

red curve: goat

- The area under each curve is 1.0 (*normalization*)
- Sheep is somewhat curlier than goat

Decision After Observation

Known

Unknown

Prior probability
$P(\omega_j)\ (1 \le j \le c)$

Class-conditional pdf
$p(x|\omega_j)\ (1 \le j \le c)$

Observation for test example
x^* (*e.g.: wool curliness*)

Bayes Formula

Posterior probability
$P(\omega_j|x^*)\ (1 \le j \le c)$

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

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) $P(\omega)$ $p(x)$

$$P(\omega) = \int_{-\infty}^{\infty} p(\omega, x)dx \qquad p(x) = \sum_{j=1}^{c} p(\omega_j, x)$$

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

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

- $P(\omega_j)$ 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)

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

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

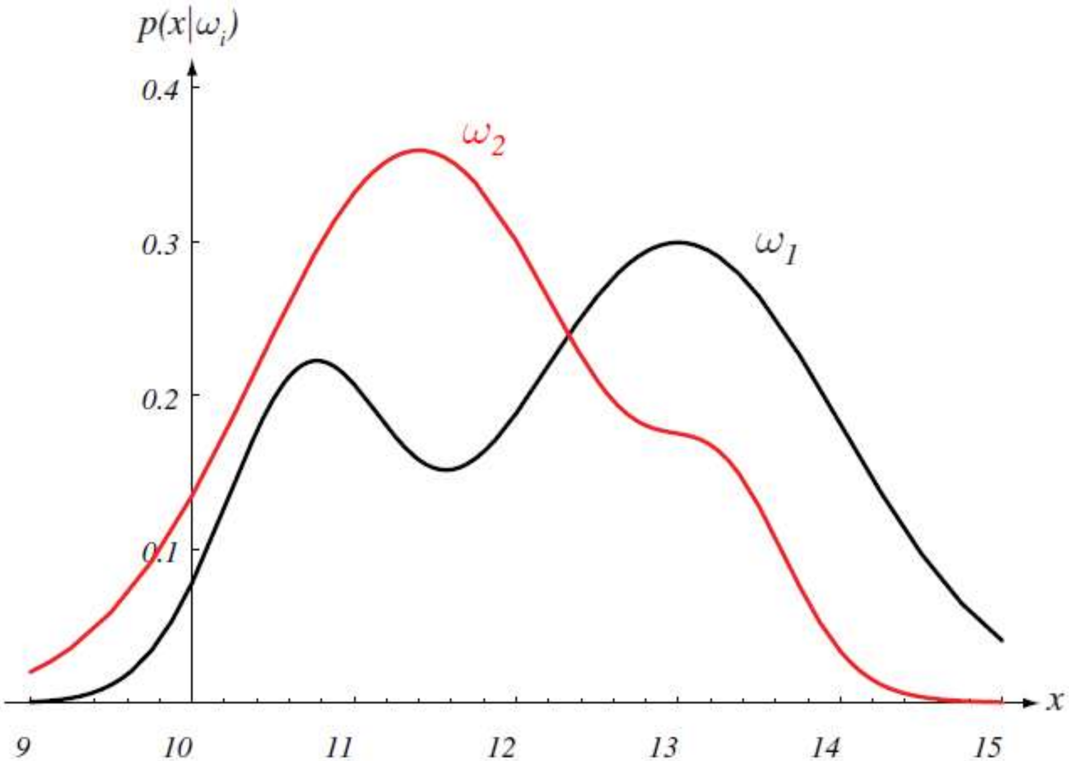
$$p(x|\omega_1) = p(x|\omega_2) = \cdots = p(x|\omega_c)$$

→ 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 *curliness*

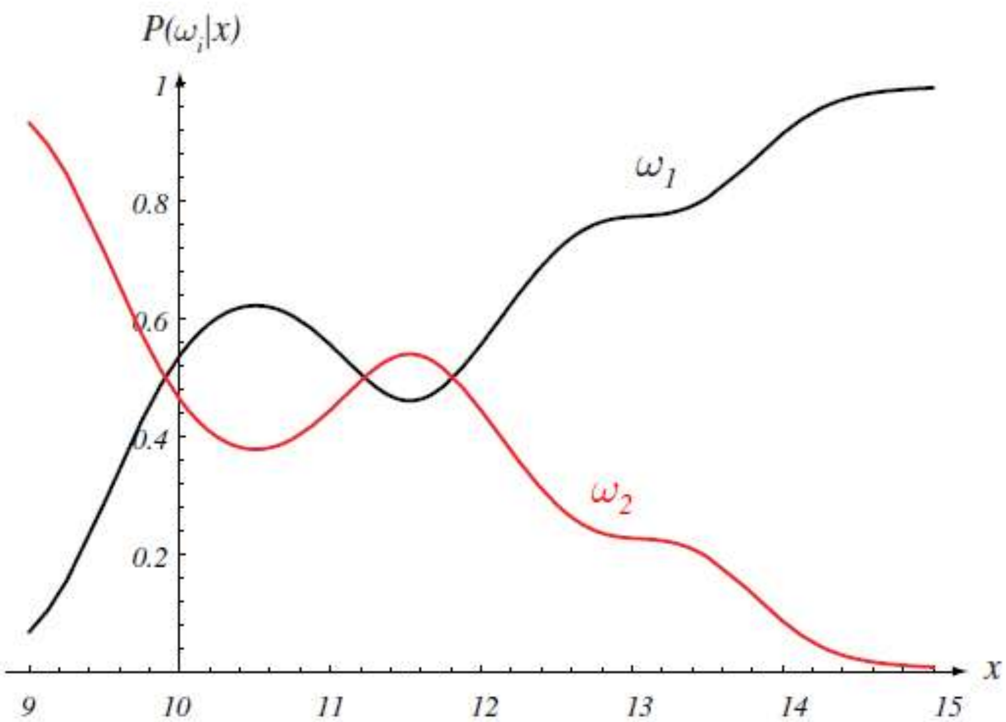
$$P(\omega_1) = \frac{2}{3}$$

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

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

Bayes Formula Revisited (Cont.)

An illustrative example



posterior probability for either type

h-axis: curliness of wools

v-axis: posterior probability for either type of sheep

black curve: sheep

red curve: goat

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

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

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

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

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

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

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

$$\boxed{P(\omega_2 \mid +) > P(\omega_1|+)}$$

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

$$\omega_1 : \text{ price} > \$50,000$$

$$\omega_2 : \text{ price} \leq \$50,000$$

$$x : \text{ height of car}$$

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

?

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

Quantities to know:

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

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

cars in ω_2 : 988

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

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

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

$x = 1.05$

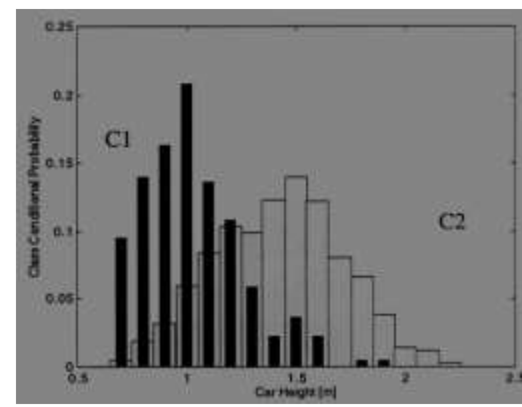
x falls into interval I_x=[1.0m, 1.1m]

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

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

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

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

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

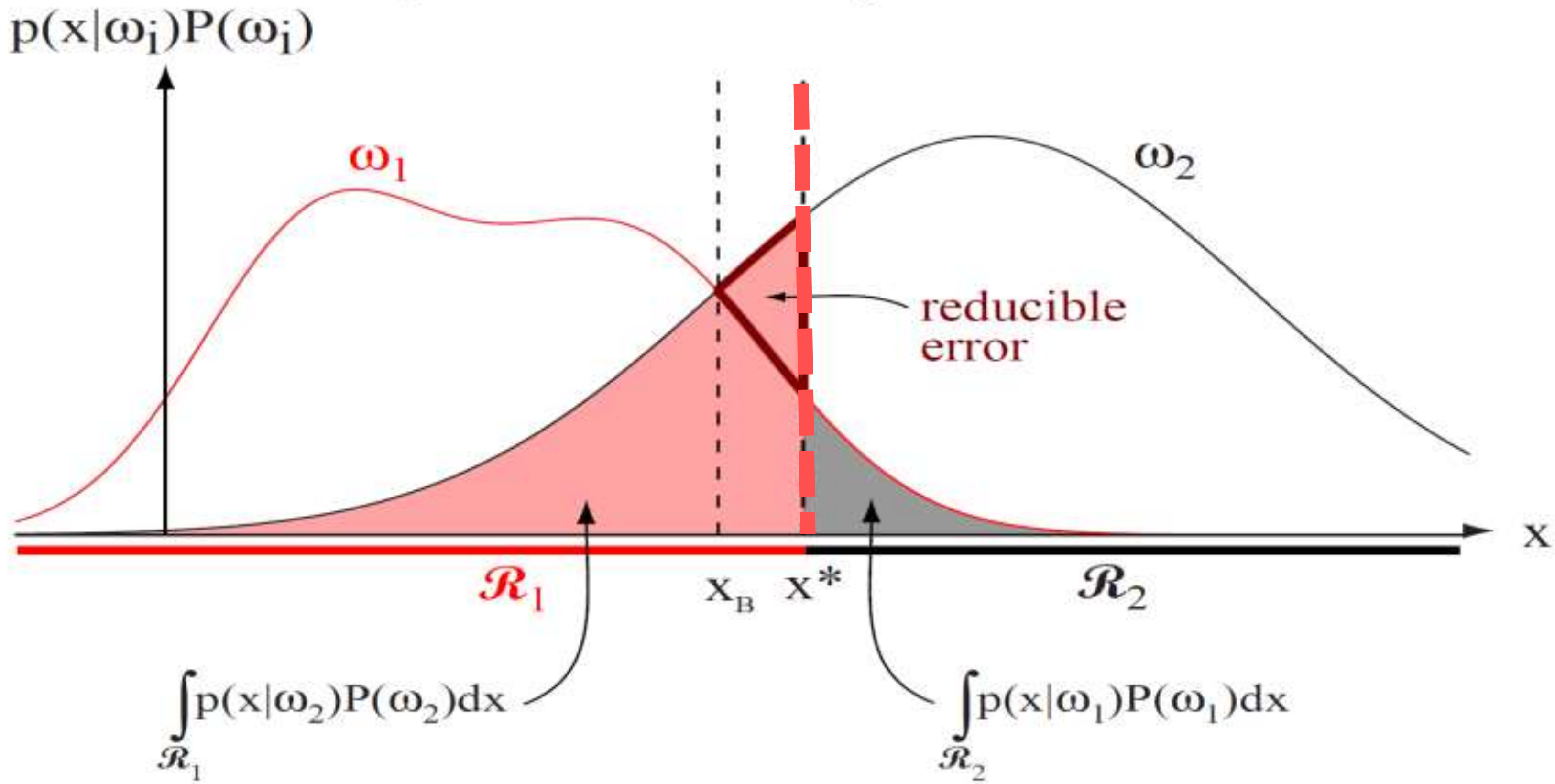
$$= \frac{0.817 \times 0.0597}{0.183 \times 0.2081} = 1.280$$

$$\boxed{\begin{array}{c} P(\omega_2 \mid x) > P(\omega_1 \mid x) \\ \text{price} \leq \$50,000 \end{array}}$$

Is Bayes Decision Rule Optimal?

Bayes Decision Rule (In case of two classes)

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



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

  $\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$ (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)

$$\text{Note} : c \neq a$$

Bayes Decision Rule – The General Case (Cont.)

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

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

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

For ease of reference, usually written as:

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

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**

Bayes Decision Rule – 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 ω_j

the action being taken is α_i

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

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!

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

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:

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 $\mathbf{x}$:

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

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

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

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

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

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

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 **x**, the (predictive) decision function $\alpha(\mathbf{x})$ assumes one of the *a* actions $\alpha_1, \ldots, \alpha_a$

Overall risk *R*
（总体风险）
expected loss with decision function $\alpha(\cdot)$

$$R = \int \underbrace{R(\alpha(\mathbf{x}) \mid \mathbf{x})}_{\text{Conditional risk for pattern } \mathbf{x} \text{ with action } \alpha(\mathbf{x})} \cdot \underbrace{p(\mathbf{x})}_{\text{pdf for patterns}} d\mathbf{x}$$

Conditional risk for pattern **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 **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*)

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

- 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

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

- $\Omega = \{\omega_1, \omega_2\}$ (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 ω_i when the true state of nature is ω_j

The conditional risk:

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

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

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

yes → decide ω_1

no → decide ω_2

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

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

likelihood ratio

constant θ independent of $\mathbf{x}$

Minimum-Error-Rate Classification

Classification setting

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

Zero-one (symmetrical) loss function

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

- 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.)

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

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

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

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

error rate (误差率/错误率)

the probability that action α_i (decide ω_i) is wrong

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 $\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$ 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 getting a sheep or a goat might vary in different farms

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

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

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

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

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

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

Minimax Criterion (Cont.)

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

Eq.22 [pp.28]

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

$$= \int_{\mathcal{R}_1} \sum_{j=1}^{2} R(\alpha_1 \mid \omega_j) \cdot P(\omega_j \mid \mathbf{x}) \cdot p(\mathbf{x}) d\mathbf{x}$$

$$= \int_{\mathcal{R}_1} \sum_{j=1}^{2} \lambda_{1j} \cdot P(\omega_j) \cdot p(\mathbf{x} \mid \omega_j) d\mathbf{x}$$

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

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

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

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:

- $P(\omega_1) + P(\omega_2) = 1$
- $\int_{R_1} p(x|\omega_1)dx + \int_{R_2} p(x|\omega_1)dx = 1$

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

Minimax Criterion (Cont.)

$$\begin{aligned} R_{mm} &= \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x} \\ &= \lambda_{11} + (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(\mathbf{x} \mid \omega_1) d\mathbf{x} \end{aligned}$$

$$\begin{aligned} R = &\overbrace{\lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x}}^{= R_{mm}, \text{ minimax risk}} \\ &+ P(\omega_1) \underbrace{\left[(\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(\mathbf{x} \mid \omega_1) d\mathbf{x} - (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x} \right]}_{=0 \text{ for minimax solution}} \end{aligned}$$

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 (判别函数)

Classification

Pattern ➔ Category

actions ⟷ **predict categories**

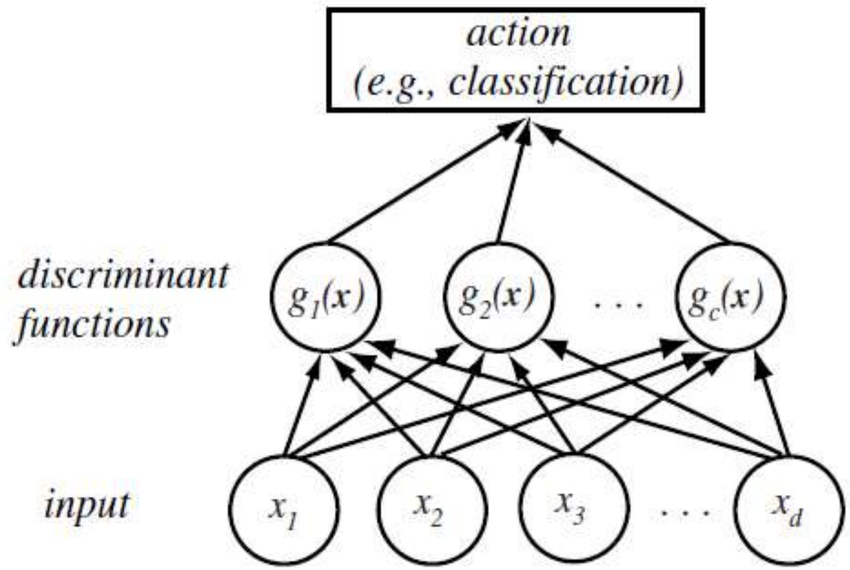
Discriminant functions

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

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

Decide ω_i

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



Discriminant Function (Cont.)

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

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

Various discriminant functions ⇒ **Identical classification results**

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

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

e.g.:

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

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

Discriminant Function (Cont.)

Decision region (决策区域)

c discriminant functions → c decision regions

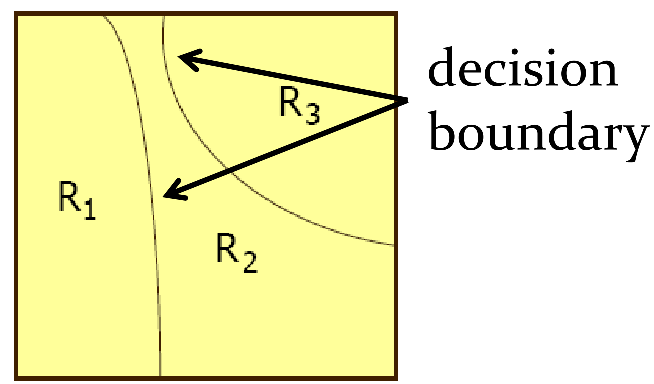
$g_i(\cdot) \ (1 \leq i \leq c)$ → $\mathcal{R}_i \subset \mathbf{R}^d \ (1 \leq i \leq c)$

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

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 (决策边界)

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



Expected Value

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

Discrete case

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

$x \sim P(\cdot)$

($\sim$: "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 (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) $\mathrm{Var}[x] = \mathcal{E}[(x - \mathcal{E}[x])^2]$ (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: $\sigma^2 = \mathrm{Var}[x]$ (σ : *standard deviation* (标准偏差))

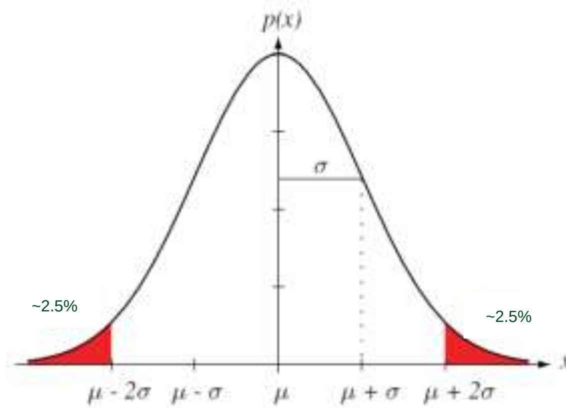
Gaussian Density – Univariate Case

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

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

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

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



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

Vector Random Variables (随机向量)

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

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

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

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

Expected vector

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

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

marginal pdf on the *i*-th component

Notation:

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

Vector Random Variables (Cont.)

Covariance matrix (协方差矩阵)

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

Properties of $\mathbf{\Sigma}$

- 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]

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

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

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

Gaussian Density – Multivariate Case

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

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

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

$\mathbf{x} = (x_1, x_2, \ldots, x_d)^t$: *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 (Cont.)

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

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

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

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

$\Rightarrow$ $(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})$ scalar (1×1 matrix)

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

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

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

Discriminant Functions 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

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

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: σ^2 times the identity matrix **I**

$$\boldsymbol{\Sigma}_i = \sigma^2 \cdot \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} = \begin{pmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \ddots & \\ & & & \sigma^2 \end{pmatrix} \Rightarrow \begin{array}{l} |\boldsymbol{\Sigma}_i| = \sigma^{2d} \\ \boldsymbol{\Sigma}_i^{-1} = (1/\sigma^2)\,\mathbf{I} \end{array}$$

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

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

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

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

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

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

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

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)

Case II: $\mathbf{\Sigma}_i = \mathbf{\Sigma}$ (Cont.)

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

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

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

Linear discriminant functions

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

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

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

Case III: $\boldsymbol{\Sigma}_i$ = 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*

Summary

- Bayesian Decision Theory
  - PR: essentially a predictive 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 ←→ 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

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

- Discriminant functions for Gaussian pdf

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

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