M L
D M

Chapter 2
Bayesian Decision Theory

Bayesian Decision Theory

MIMA

- Bayesian decision theory is a statistical approach to data mining/pattern recognition

- Mathematical foundation for decision making

- Using probabilistic approach to help making decision so as to minimize the risk (cost).

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

Preliminaries and Notations

MIMA

$\omega_i \in \{\omega_1, \omega_2, \dots, \omega_c\}$: a state of nature

$P(\omega_i)$: prior probability

$\mathbf{x}$: feature vector

$p(\mathbf{x})$: evidence probability

$p(\mathbf{x} \mid \omega_i)$: class-conditional density / likelihood

$P(\omega_i \mid \mathbf{x})$: posterior probability

Decision Before Observation

- The Problem
  - To make a decision where
    - Prior probability is known
    - No observation is allowed
- Naïve Decision Rule

$$\text{Decide } \omega_1 \text{ if } P(\omega_1) > P(\omega_2), \text{otherwise } \omega_2$$

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

Bayes Theorem

MIMA

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i) P(\omega_i)}{p(\mathbf{x})}$$

$$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x} \mid \omega_i) P(\omega_i)$$

Thomas Bayes
(1702-1761)

Decision After Observation

MIMA

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{p(\mathbf{x})}$$

unimportant in making decision

$$\mathcal{D}(\mathbf{x}) = \arg\max_{\omega_i} P(\omega_i \mid \mathbf{x})$$

Decision After Observation

MIMA

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i) P(\omega_i)}{p(\mathbf{x})}$$

Bayes Formula

Known

Prior probability
Class-conditional pdf
Observation

Unknown

Posterior probability

$$P(\omega_i \mid \mathbf{x}):$$

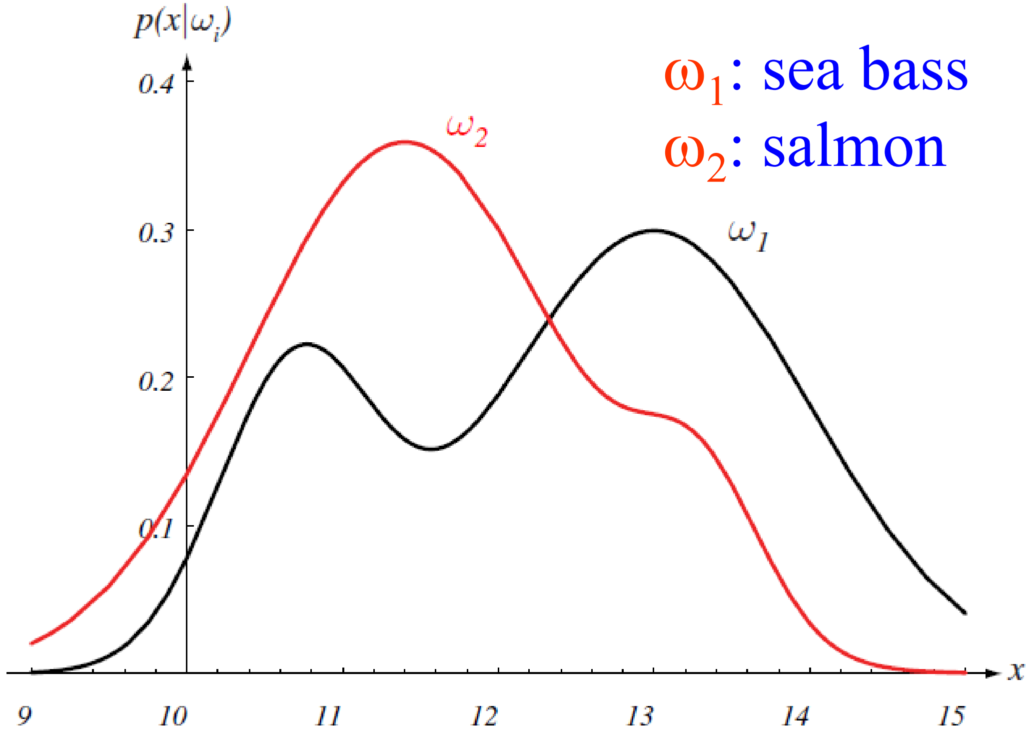
Special Cases

MIMA

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i) P(\omega_i)}{p(\mathbf{x})} \quad \left(\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \right)$$

- Case I: Equal prior probability
  - $P(\omega_1)=P(\omega_2)=\cdots=P(\omega_c)=1/c$
  - Depends on the likelihood $p(\mathbf{x}|\omega_j)$
- Case II: Equal likelihood
  - $p(\mathbf{x}|\omega_1)=p(\mathbf{x}|\omega_2)=\cdots=p(\mathbf{x}|\omega_c)$
  - Degenerate to naïve decision rule
- Normally, prior probability and likelihood function together in Bayesian decision process

An example



ω_1: sea bass
ω_2: salmon

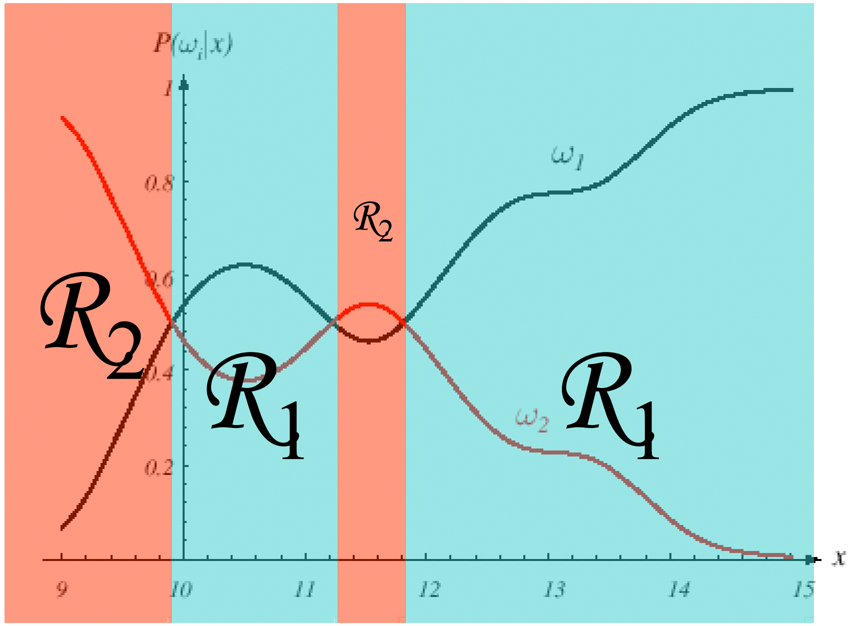
$P(\omega_1)=2/3$
$P(\omega_2)=1/3$

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

class-conditional pdf for *lightness*

Decide ω_1 if $p(\mathbf{x}|\omega_1)P(\omega_1) > p(\mathbf{x}|\omega_2)P(\omega_2)$; otherwise decide ω_2

An example



posterior probability for either type of fish

h-axis: lightness of fish scales
v-axis: posterior probability for each 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

MIMA

- 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, does she/he have cancer?

Another Example (Cont.)

MIMA

ω_1 : cancer $\qquad \omega_2$: no cancer $\qquad x \in \{+, -\}$

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

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

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

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

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

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

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

No cancer!

Feasibility of Bayes Formula

MIMA

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i) P(\omega_i)}{p(\mathbf{x})} \quad \left(\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \right)$$

- To compute posterior probability, we need to know prior probability and likelihood

How do we know these probabilities ?

- A simple solution: Counting Relative frequencies
- An advanced solution: Conduct Density estimation

A Further Example

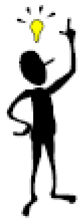
MIMA

- Problem
  - 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
ω_2 : price <=$ 50,000
x : height of a car

Decide ω_1 if $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$;
otherwise decide ω_2

Quantities to know:
$P(\omega_1)$ $P(\omega_2)$ $P(\mathrm{x}|\omega_1)$ $P(\mathrm{x}|\omega_2)$
How to get them?



Counting relative frequencies via collected samples

A Further Example (Cont.)

MIMA

- Collecting samples
  - Suppose we have randomly picked 1209 cars in the campus, got prices from their owners, and measured their heights
- Compute $P(\omega_1)$ and $P(\omega_2)$

\# cars in ω_1 : 221

\# cars in ω_2 : 988

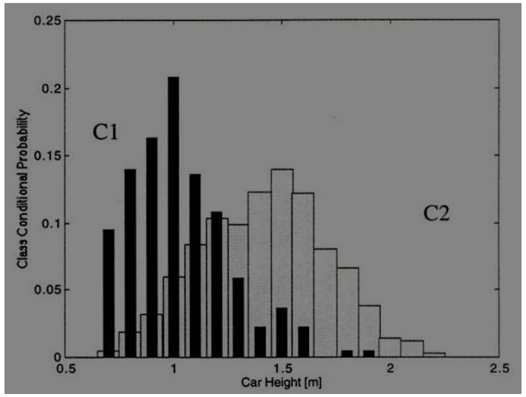
$$P(\omega_1) = \frac{221}{1209} = 0.183$$

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

A Further Example (Cont.)

MIMA

- 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 , which means that 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.)

MIMA

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

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

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

$$P(x = 1.05 \mid \omega_1) = \frac{46}{221} = 0.2081$$

$$P(x = 1.05 \mid \omega_2) = \frac{59}{988} = 0.0597$$

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

$$= \frac{P(\omega_2)P(x = 1.05 \mid \omega_2)}{P(\omega_1)P(x = 1.05 \mid \omega_1)} = \frac{0.817 \times 0.0597}{0.183 \times 0.2081}$$

$\Rightarrow$ $P(\omega_1|\mathbf{x}) < P(\omega_2|\mathbf{x})$, price<=\$50,000

Is Bayes Decision Rule Optimal

MIMA

- Consider two categories

Decide ω_1 if $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$; otherwise decide ω_2

- When we observe x, the probability of error is:

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

Thus, under Bayes decision rule, we have

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

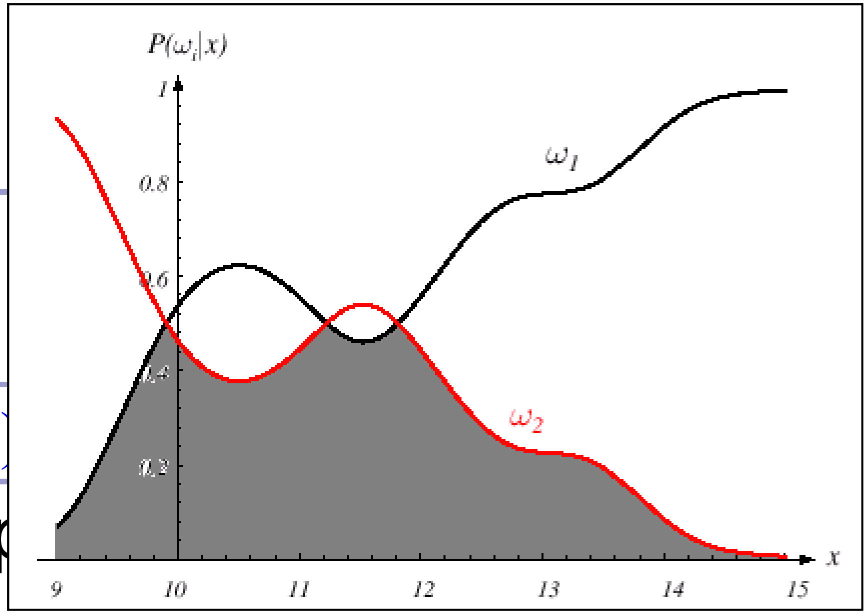
For every x, we ensure that P(error|x) is as small as possible

Is Bayes Decision

- Consider two categories

Decide ω_1 if $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$

- When we observe x, the p



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

Thus, under Bayes decision rule, we have

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

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

Generalized Bayes Decision Rule

- Allowing to use more than one feature

$$x \in R \Rightarrow x \in R^d :$$ d-dimensional Euclidean Space

- Allowing more than two states of nature

$\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$: a set of c states of nature

- Allowing actions other than merely deciding the state of nature

$A = \{\alpha_1, \alpha_2, \ldots, \alpha_a\}$: a set of a possible *actions*

Note that $c \neq a$

Generalized Bayes Decision Rule (cont.)

MIMA

- Introducing a loss function more general than the probability of error

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

$$\lambda_{ij} = \lambda(\omega_j, \alpha_i) : \text{the loss incurred for taking action } \alpha_\text{i} \text{ when the state of nature is } \omega_\text{j}$$

For ease of reference, it is usually written as:

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

We want to $\underbrace{\textit{minimize the expected loss}}_{\text{Risk}}$ *in making decision.*

Generalized Bayes Decision Rule (cont.)

MIMA

- Introd[...] [...] than the pr[...]

A simple loss function

λ

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

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

For ease of reference, it is usually written as: $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$:

We want to $\underbrace{\textit{minimize the expected loss}}_{\text{Risk}}$ *in making decision.*

Generalized Bayes Decision Rule (cont.)

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

- Problem
  - Given a particular x, we have to decide which action to take
  - To do this, we need to know the loss of taking each action $\alpha_i \ (1 \le i \le a)$

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

The action being taken α_i

True state of nature ω_j

However, the true state of nature is uncertain

Expected (average) loss

We want to minimize the expected loss in making decision.

$\underbrace{\text{minimize the expected loss}}_{\text{Risk}}$

Generalized Bayes Decision Rule (cont.)

MIMA

- Expected loss

Given $\mathbf{x}$, the expected loss (risk) associated with taking action α_i.

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

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"

Generalized Bayes Decision Rule (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:
$P(\omega_1|\mathbf{x}) = 0.01$
$P(\omega_2|\mathbf{x}) = 0.99$

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

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

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

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

Generalized Bayes Decision Rule (cont.)

MIMA

- 0/1 Loss Function

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

$$\lambda(\alpha_i \mid \omega_j) = \begin{cases} 0 & \alpha_i \text{ is a correct decision assiciated with } \omega_j \\ 1 & \text{otherwise} \end{cases}$$

$$\Longrightarrow \quad R(\alpha_i \mid \mathbf{x}) = P(error \mid \mathbf{x})$$

Generalized Bayes Decision Rule (cont.)

MIMA

- Bayes decision rule (general case)

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

- Overall risk

$$R = \int R(\underbrace{\alpha(\mathbf{x})}_{\textit{Decision function}} \mid x) \cdot p(x) dx$$

For every x, we ensure that the conditional risk R(a(x)|x) is as small as possible; Thus, the overall risk over all possible x must be as small as possible.

The optimal one to minimize the overall risk
Its resulting overall risk is called the Bayesian risk

General Case: Two-Category

MIMA

$$\Omega = \{\omega_1, \omega_2\}$$

$$A = \{\alpha_1, \alpha_2\}$$

Loss Function

State of Nature

Action		ω_1	ω_2
	α_1	λ_{11}	λ_{12}
	α_2	λ_{21}	λ_{22}

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

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

General Case: Two-Category

Perform α_1 if $\underbrace{R(\alpha_2|\mathbf{x}) > R(\alpha_1|\mathbf{x})}$; otherwise perform α_2

$$\Rightarrow \quad \lambda_{21}P(\omega_1|\mathbf{x})+\lambda_{22}P(\omega_2|\mathbf{x}) > \lambda_{11}P(\omega_1|\mathbf{x})+\lambda_{12}P(\omega_2|\mathbf{x})$$

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

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

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

General Case: Two-Category

MIMA

Perform α_1 if $\underbrace{R(\alpha_2|\mathbf{x}) > R(\alpha_1|\mathbf{x})}$; otherwise perform α_2

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

$$\Rightarrow \quad \underbrace{(\lambda_{21} - \lambda_{11})}_{positive}P(\omega_1 \mid \mathbf{x}) > \underbrace{(\lambda_{12} - \lambda_{22})}_{positive}P(\omega_2 \mid \mathbf{x})$$

Posterior probabilities are scaled before comparison.

General Case: Two-Category

MIMA

Perform α_1 if $\underbrace{R(\alpha_2|\mathbf{x}) > R(\alpha_1|\mathbf{x})}$; otherwise perform α_2

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

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

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

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

General Case: Two-Category

MIMA

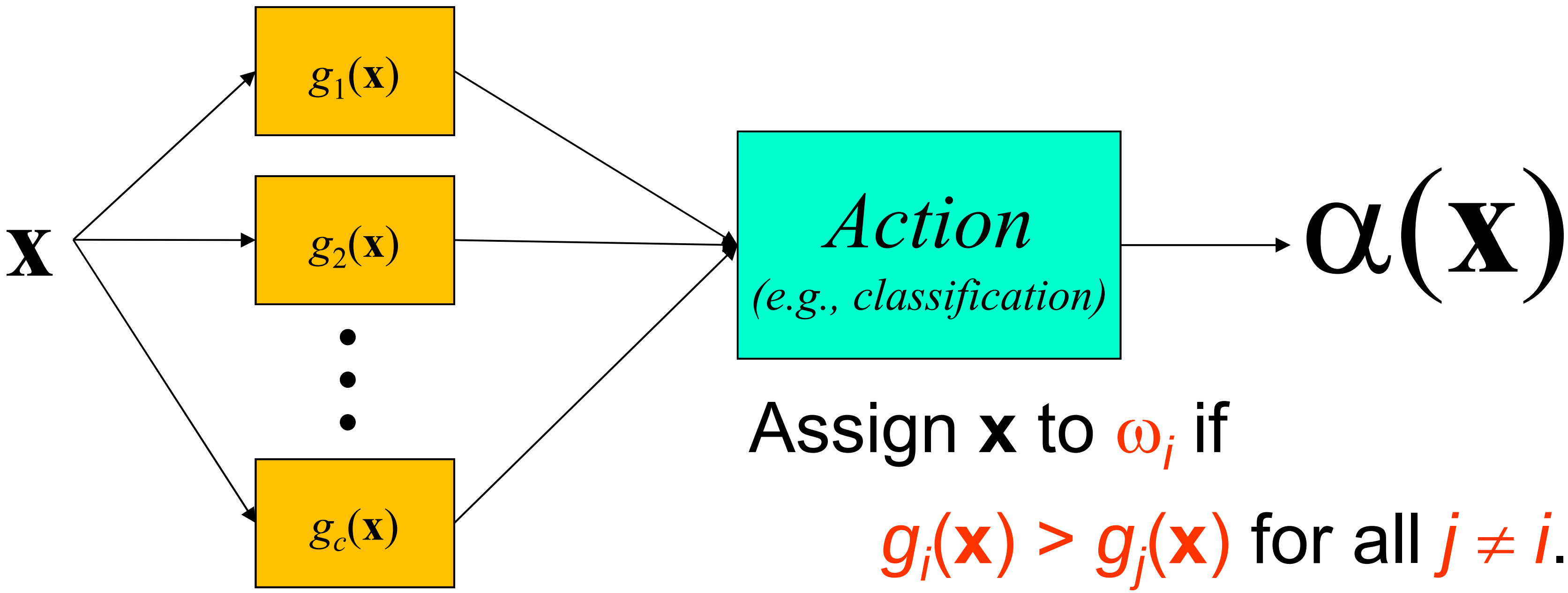
$$\Rightarrow \text{Perform } \alpha_1 \text{ if } \underbrace{\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)}}_{\textit{Likelihood Ratio}} > \underbrace{\frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \frac{P(\omega_2)}{P(\omega_1)}}_{\textit{Threshold}}$$

Discriminant Function

- Discriminant functions for multicategory

$$g_i(x): R^d \rightarrow R \quad (1 \leq i \leq c)$$

- One function per category



Assign $\mathbf{x}$ to ω_i if

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

Discriminant Function

- Minimum Risk Case:

$$g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x})$$

- Minimum Error-Rate Case:

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

$$g_i(\mathbf{x}) = p(\mathbf{x} \mid \omega_i)P(\omega_i)$$

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

Discriminant Function

MIMA

- Relationship between minimum risk and minimum error rate

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

error rate (误差率/错误率)

the probability that action α_i (decide ω_i) is wrong

Discriminant Function

- Various discriminant function
- Identical classification results

If $f(.\ )$ is a *monotonically increasing* function,
then $f(g_i(.\ ))$'s are also be discriminant functions.

- Example

$$f(x) = k \cdot x \ \ (k > 0) \Rightarrow f(g_i(x)) = k \cdot g_i(x) \ \ (1 \le i \le c)$$

$$f(x) = \ln x \Rightarrow f(g_i(x)) = \ln g_i(x) \ \ (1 \le i \le c)$$

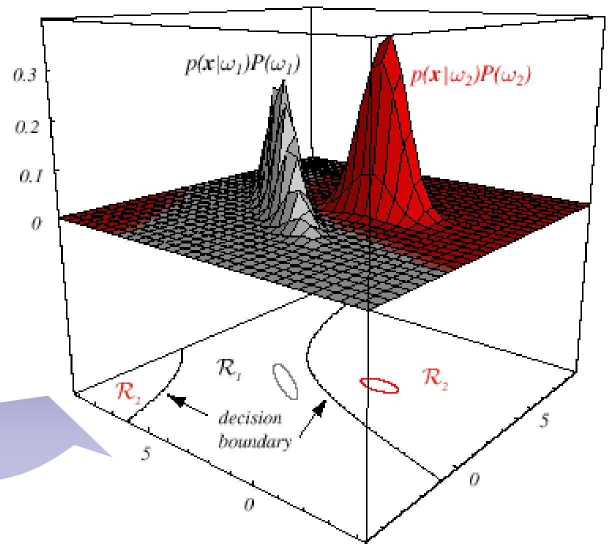
Decision Regions

- c discriminant functions result in c decision regions.

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

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

- Decision boundary
  - Decision regions are separated by decision boundaries



Two-category example

The Normal Distribution

MIMA

Discrete random variable (*X*) — Assume integer

Probability mass function (pmf): $p(x) = P(X = x)$

Cumulative distribution function (cdf): $F(x) = P(X \le x) = \sum_{t=-\infty}^{x} p(t)$

Continuous random variable (*X*)

Probability density function (pdf): $p(x)$ or $f(x)$ not a probability

Cumulative distribution function (cdf): $F(x) = P(X \le x) = \int_{-\infty}^{x} p(t)dt$

Expectations

MIMA

- a.k.a. expected value, mean or average of a random variable
- x is a random variable, the expectation of x

$$E[x] = \begin{cases} \sum_{x=-\infty}^{\infty} xp(x) & x \text{ is discrete} \\ \int_{-\infty}^{\infty} xp(x)dx & x \text{ is continuous} \end{cases}$$

The k^{th} moment $E[X^k]$

The 1^{st} moment $\mu_X = E[X]$

The k^{th} central moment $E[(X - \mu_X)^k]$

Important Expectations

MIMA

- Mean

$$\mu_X = E[X] = \begin{cases} \sum_{x=-\infty}^{\infty} x p(x) & X \text{ is discrete} \\ \int_{-\infty}^{\infty} x p(x) dx & X \text{ is continuous} \end{cases}$$

- Variance

$$\sigma_X^2 = Var[X] = E[(X-\mu_X)^2] = \begin{cases} \sum_{x=-\infty}^{\infty} (x-\mu_X)^2 p(x) & X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu_X)^2 p(x) dx & X \text{ is continuous} \end{cases}$$

Notation: $\sigma^2 = Var[x]$ (σ: standard deviation ?)

Fact: $\sigma^2 = Var[x] = E[x^2] - (E[x])^2$

Entropy

- The *entropy* measures the fundamental *uncertainty* in the value of points selected randomly from a distribution.

$$H[X] = \begin{cases} -\sum_{x=-\infty}^{\infty} p(x)\log p(x) & X \text{ is discrete} \\ -\int_{-\infty}^{\infty} p(x)\log p(x)dx & X \text{ is continuous} \end{cases}$$

Univariate Gaussian Distribution

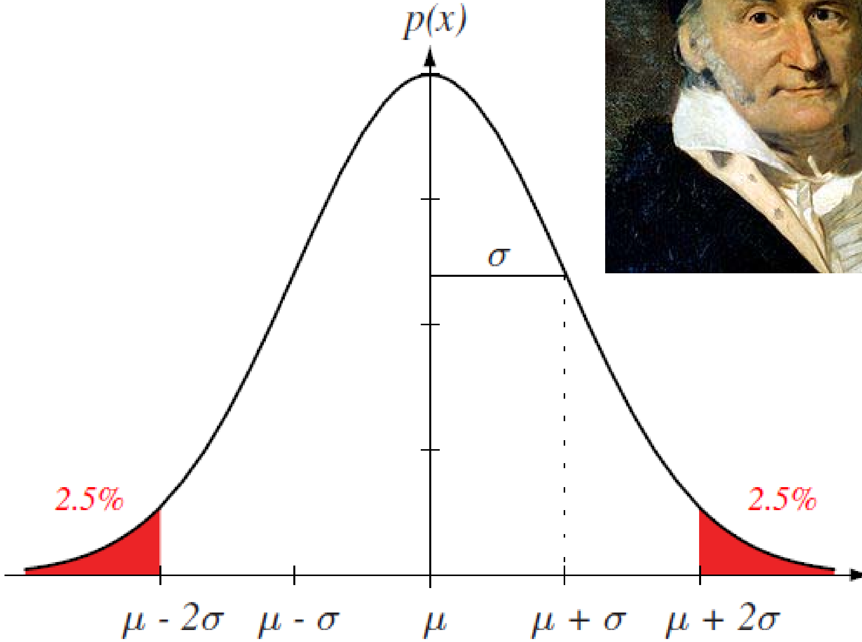
- Gaussian distribution, a.k.a. Gaussian density, normal density.

$X \sim N(\mu, \sigma^2)$

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

$E[X] = \mu$

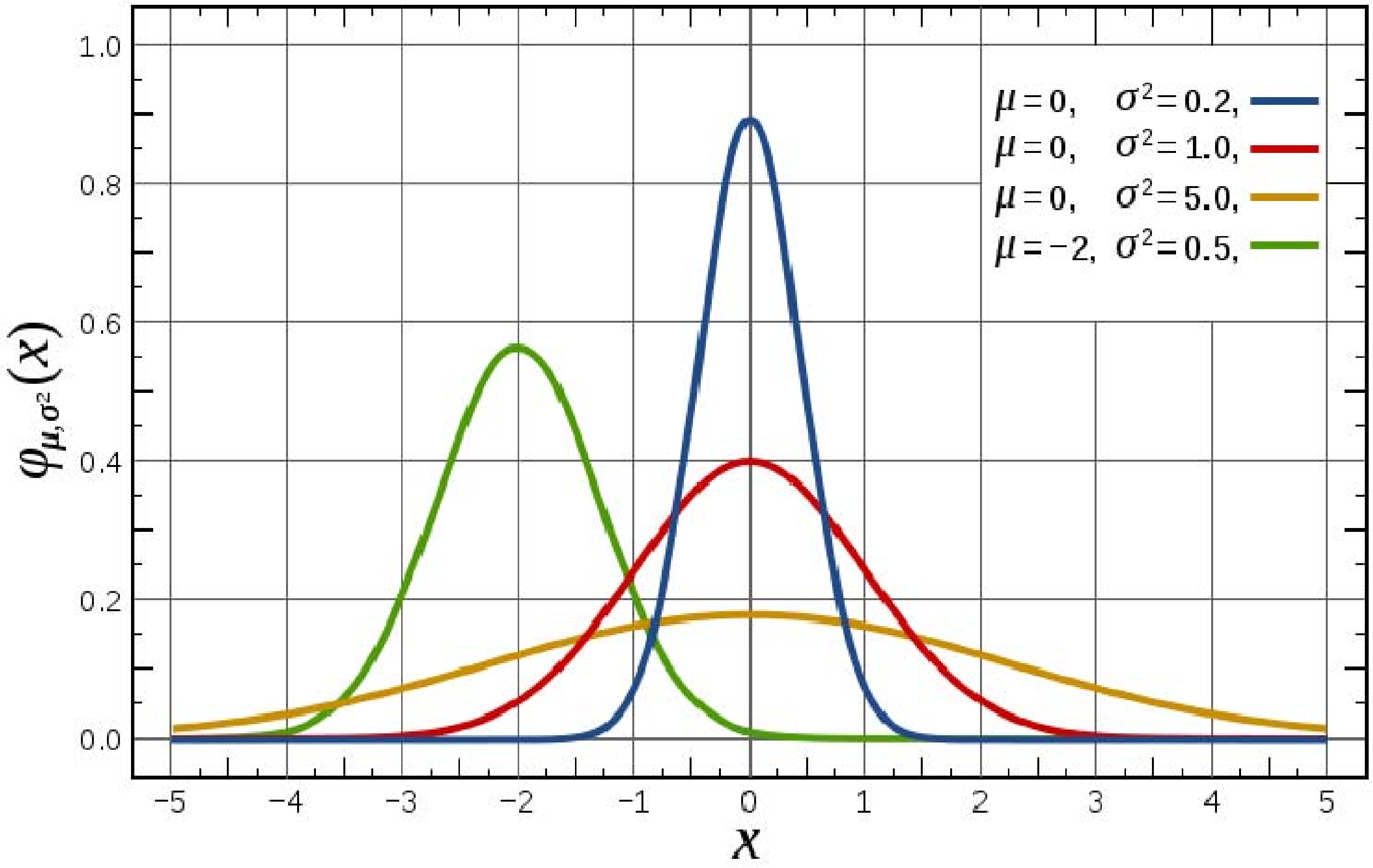
$Var[X] = \sigma^2$

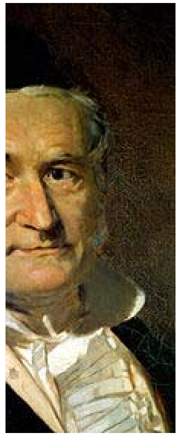


Univariate Gaussian Distribution

MIMA

- Ga ',
  nc





X

p

E

5%

x

$$Var[X] = \sigma^2$$

Random Vectors

- A d-dimensional random vector is:

$$\mathbf{X} = (x_1, x_2, \ldots, x_d)^T \quad \mathbf{X} : \Omega \to R^d$$

$$X \sim p(X) = p(x_1, x_2, \ldots, x_d) \quad \text{(joint pdf)}$$

- Expected vector

$$E[\mathbf{X}] = \begin{pmatrix} E[x_1] \\ E[x_2] \\ \vdots \\ E[x_d] \end{pmatrix} \qquad E[x_i] = \int_{-\infty}^{+\infty} x_i \underline{p(x_i)} dx_i \quad (1 \le i \le d)$$

Marginal pdf on the ith component.

$$\boldsymbol{\mu} = E[\mathbf{X}] = (\mu_1, \mu_2, \ldots, \mu_d)^T$$

Random Vectors

MIMA

- Covariance matrix

$$\mathbf{\Sigma} = E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \Lambda & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \Lambda & \sigma_{2d} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ \sigma_{d1} & \sigma_{d2} & \Lambda & \sigma_d^2 \end{pmatrix}$$

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

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

Properties:
Symmetric, Positive semidefinite

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

Multivariate Gaussian Distribution

MIMA

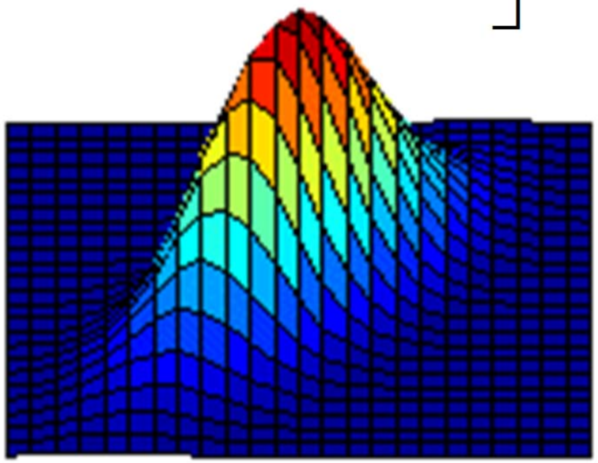
- X is a d-dimensional random vector

$$X \sim N(\mu, \Sigma)$$

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

$$E[X] = \mu$$

$$E[(X - \mu)(X - \mu)^T] = \Sigma$$

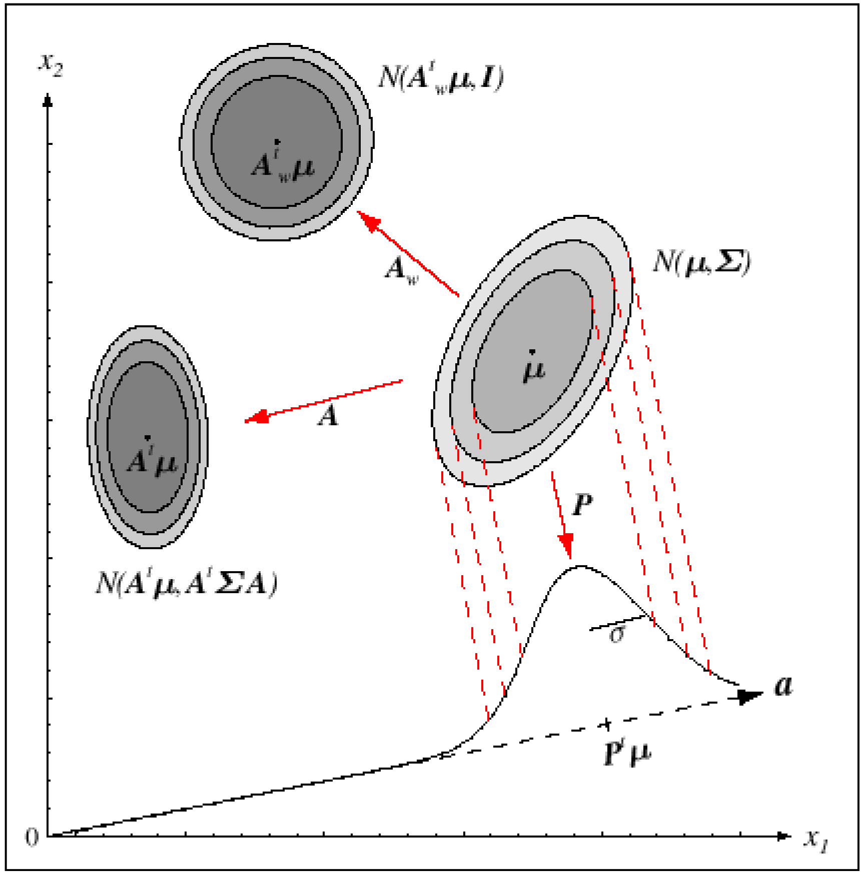


Properties of *N*(μ,Σ)

- X is a d-dimensional random vector, and

$$X \sim N(\mu, \Sigma)$$

- If $Y=A^T X$, where **A** is a $d \times k$ matrix, then



$$\mathbf{Y} \sim N(\mathbf{A}^T\boldsymbol{\mu}, \mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A})$$

On Covariance Matrix

MIMA

- As mentioned before, Σ is *symmetric* and *positive semidefinite*.

$$\boldsymbol{\Sigma} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Phi}^T$$

Φ: *orthonormal* matrix, whose columns are *eigenvectors* of Σ.

Λ: *diagonal* matrix (*eigenvalues*).

- Thus,

$$\boldsymbol{\Sigma} = (\boldsymbol{\Phi}\boldsymbol{\Lambda}^{1/2})(\boldsymbol{\Phi}\boldsymbol{\Lambda}^{1/2})^T$$

Mahalanobis Distance

- Mahalanobis distance

$$r^2 = (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})$$

$$X \sim N(\mu, \Sigma)$$

P.C. Mahalanobis (1894-1972)

$$\boxed{\underbrace{p(\mathbf{x})}_{\text{depends on the value of } r^2} = \underbrace{\frac{1}{(2\pi)^{d/2} \, |\boldsymbol{\Sigma}|^{1/2}}}_{\text{constant}} \exp\left[-\frac{1}{2} \underbrace{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}_{r^2} \right]}$$

Discriminant Functions for Gaussian Density

- Minimum-error-rate classification

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

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

-

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

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

Constant, could be ignored

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

Discriminant Functions for Gaussian Density

MIMA

- Three cases
  - Case 1 $\mathbf{\Sigma}_i = \sigma^2 \mathbf{I}$
    - Classes are centered at different mean, and their feature components are pairwisely independent have the same variance.
  - Case 2 $\mathbf{\Sigma}_i = \mathbf{\Sigma}$
    - Classes are centered at different mean, but have the same variation.
  - Case 3 $\mathbf{\Sigma}_i \neq \mathbf{\Sigma}_j$
    - Arbitrary

Case 1: $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{I}$

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

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

$$= -\frac{1}{2\sigma^2}(\underbrace{\mathbf{x}^T\mathbf{x}}_{\textit{irrelevant}} - 2\boldsymbol{\mu}_i^T\mathbf{x} + \boldsymbol{\mu}_i^T\boldsymbol{\mu}_i) + \ln P(\omega_i)$$

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

Case 1: $\Sigma_i = \sigma^2 \mathbf{I}$

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

- It is a linear discriminant function

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

- where
  - Weight vector

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

  - Threshold/bias

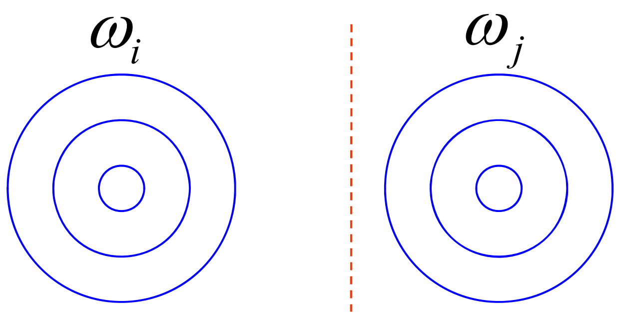
$$w_{i0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T \boldsymbol{\mu}_i + \ln P(\omega_i)$$

Case 1: $\Sigma_i = \sigma^2\mathbf{I}$

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

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

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



$$\mathbf{w}_i^T\mathbf{x} + w_{i0} = \mathbf{w}_j^T\mathbf{x} + w_{j0}$$

$$g_i(\mathbf{x}) = g_j(\mathbf{x})$$

Boundary btw.
ω_i and ω_j

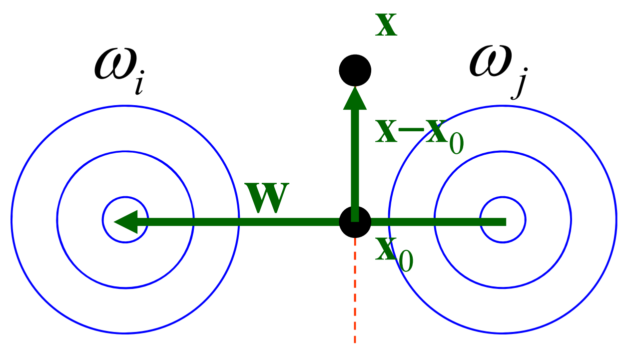
$$(\mathbf{w}_i^T - \mathbf{w}_j^T)\mathbf{x} = w_{j0} - w_{i0}$$

$$(\boldsymbol{\mu}_i^T - \boldsymbol{\mu}_j^T)\mathbf{x} = \tfrac{1}{2}(\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i - \boldsymbol{\mu}_j^T\boldsymbol{\mu}_j) - \sigma^2 \ln\frac{P(\omega_i)}{P(\omega_j)}$$

$$(\boldsymbol{\mu}_i^T - \boldsymbol{\mu}_j^T)\mathbf{x} = \tfrac{1}{2}(\boldsymbol{\mu}_i^T - \boldsymbol{\mu}_j^T)(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \sigma^2\frac{(\boldsymbol{\mu}_i^T - \boldsymbol{\mu}_j^T)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}$$

Case 1: $\Sigma_i = \sigma^2 \mathbf{I}$

- The decision boundary will be a hyperplane perpendicular to the line btw. the means at somewhere.



$$\mathbf{w}^T(\mathbf{x}-\mathbf{x}_0)=0$$

$$\mathbf{w}=\boldsymbol{\mu}_i-\boldsymbol{\mu}_j$$

$$\mathbf{x}_0=\underbrace{\tfrac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)}_{\text{midpoint}}-\underbrace{\frac{\sigma^2}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}_{0 \text{ if } P(\omega_i)=P(\omega_j)}$$

$g_i(\mathbf{x})=g_j(\mathbf{x})$

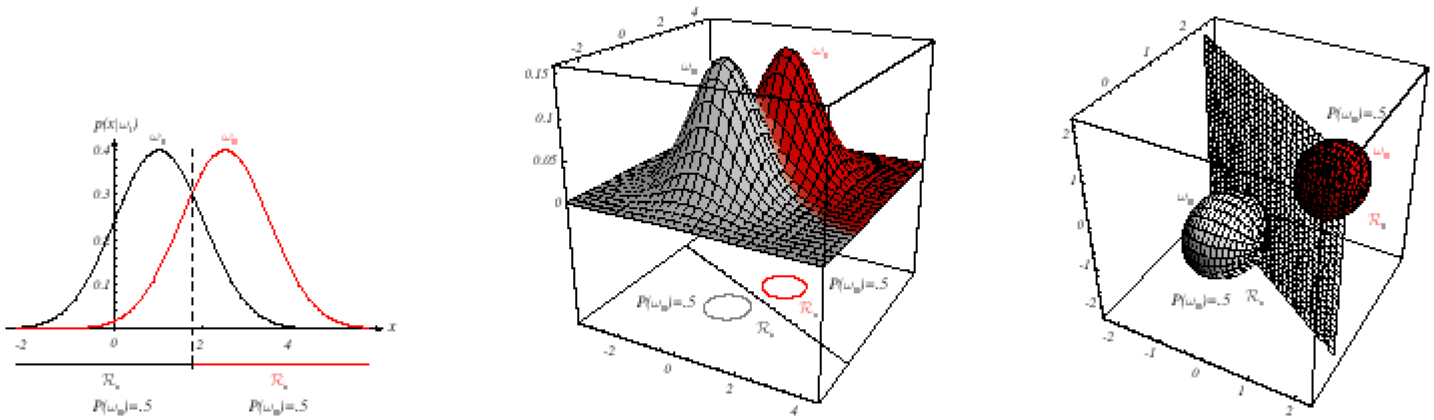
Boundary btw. ω_i and ω_j

$$\overbrace{(\boldsymbol{\mu}_i^T-\boldsymbol{\mu}_j^T)}^{\mathbf{w}^T}\mathbf{x}=\tfrac{1}{2}(\boldsymbol{\mu}_i^T-\boldsymbol{\mu}_j^T)(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\sigma^2\frac{(\boldsymbol{\mu}_i^T-\boldsymbol{\mu}_j^T)(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}$$

Case 1: $\Sigma_i = \sigma^2 \mathbf{I}$

MIMA

$$P(\omega_1) = P(\omega_2)$$

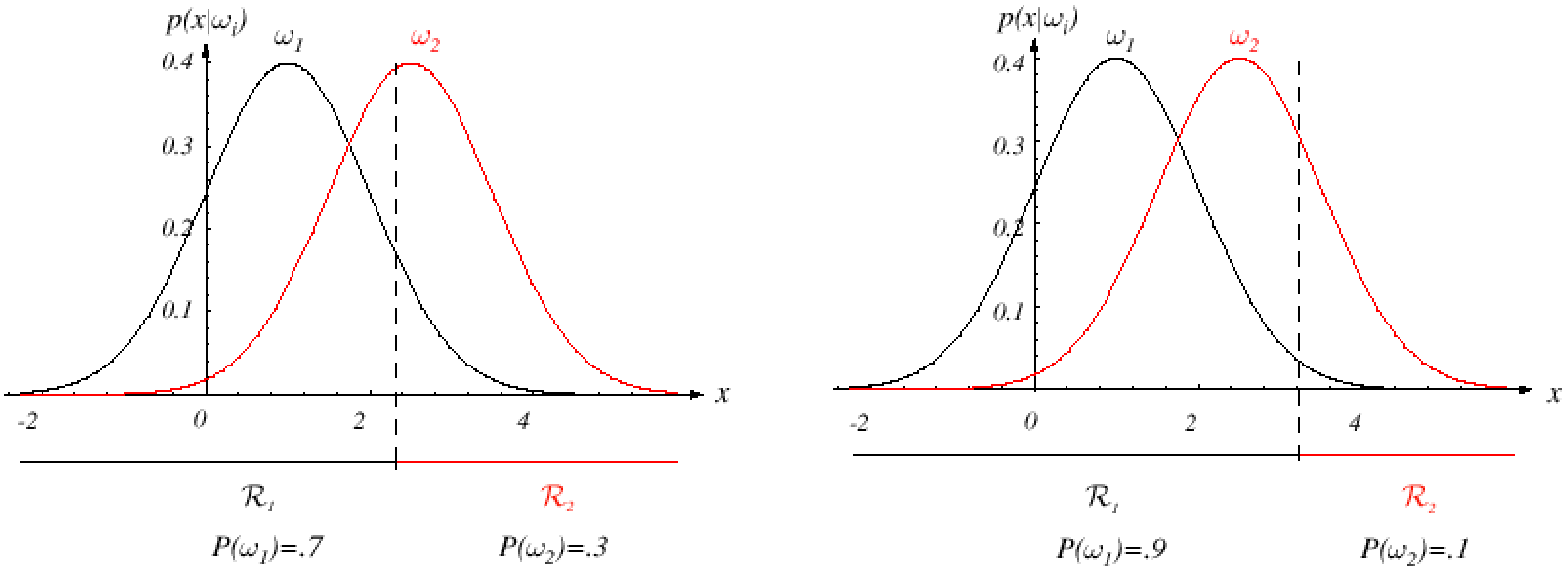


Minimum distance classifier (template matching)

Case 1: $\Sigma_i = \sigma^2 \mathbf{I}$

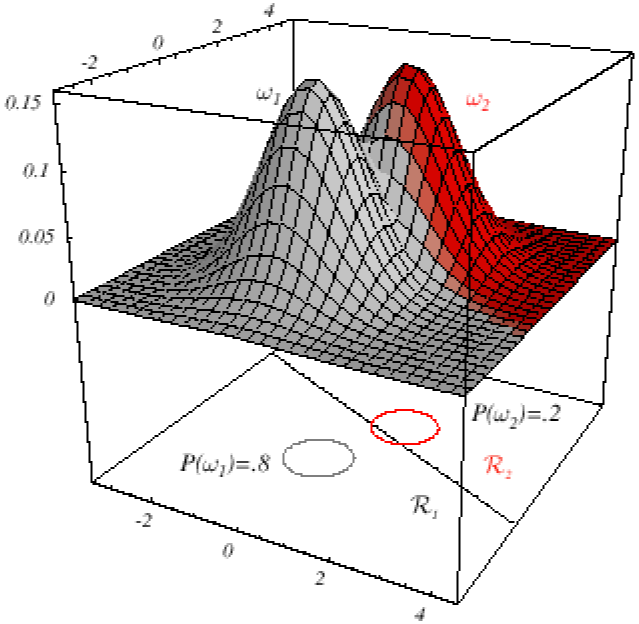
MIMA

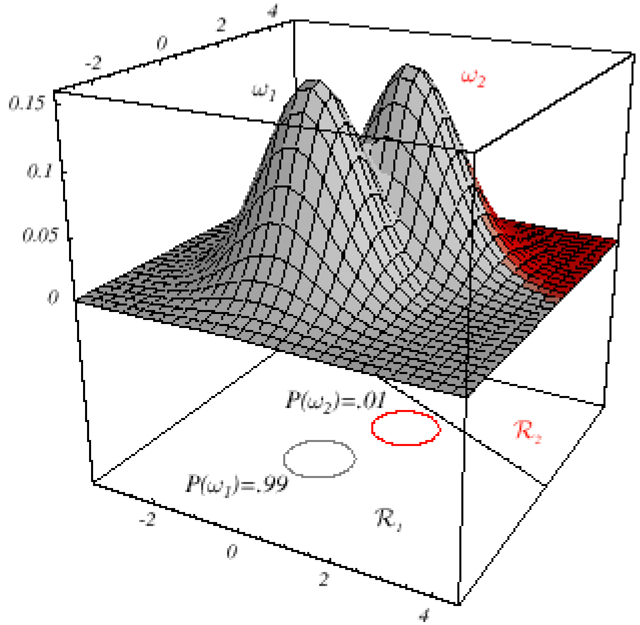
$$P(\omega_1) > P(\omega_2)$$



Case 1: $\Sigma_i = \sigma^2 \mathbf{I}$

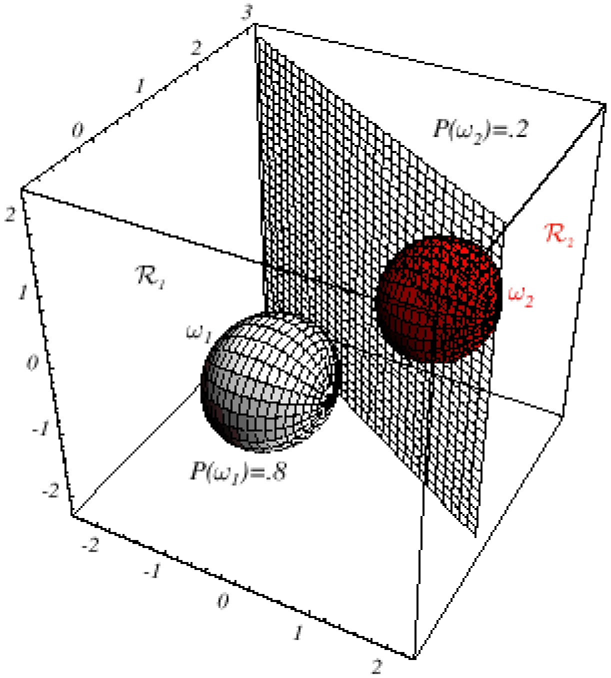
$$P(\omega_1) > P(\omega_2)$$

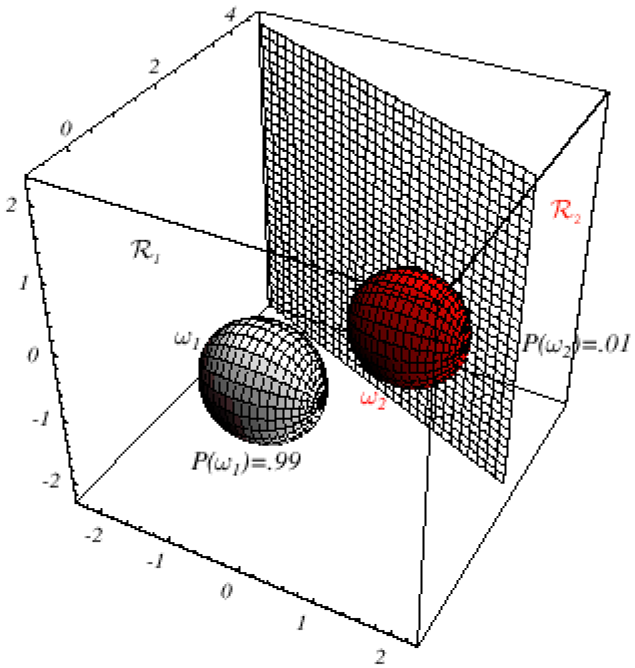




Case 1: $\Sigma_i = \sigma^2 \mathbf{I}$

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





Case 2: $\mathbf{\Sigma}_i = \mathbf{\Sigma}$

MIMA

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

$$g_i(\mathbf{x}) = -\frac{1}{2}\underbrace{(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}_{\substack{\textit{Mahalanobis}\\ \textit{Distance}}} + \underbrace{\ln P(\omega_i)}_{\substack{\textit{Irrelevant if}\\ P(\omega_i)=P(\omega_j)\ \forall i,j}}$$

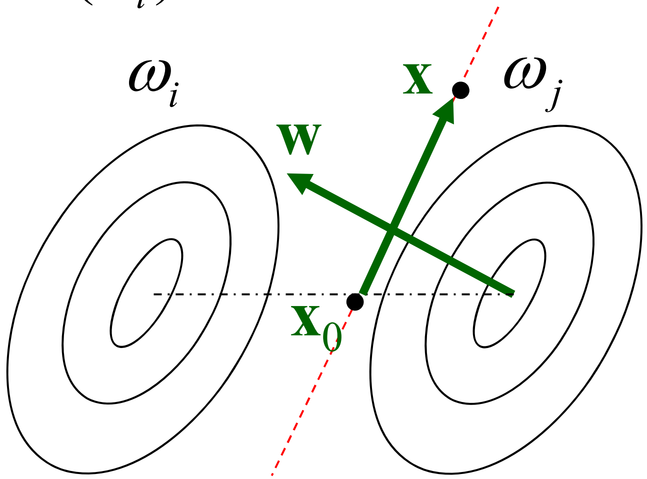
$$= -\frac{1}{2}(\underbrace{\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}}_{\textit{Irrelevant}} - 2\boldsymbol{\mu}_i^T\mathbf{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_i^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_i) + \ln P(\omega_i)$$

$$g_i(\mathbf{x}) = \mathbf{w}_i^T\mathbf{x} + w_{i0} \quad \begin{cases} \mathbf{w}_i = \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_i \\ w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i) \end{cases}$$

Case 2: $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0} \quad \begin{cases} \mathbf{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \\ w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i) \end{cases}$$

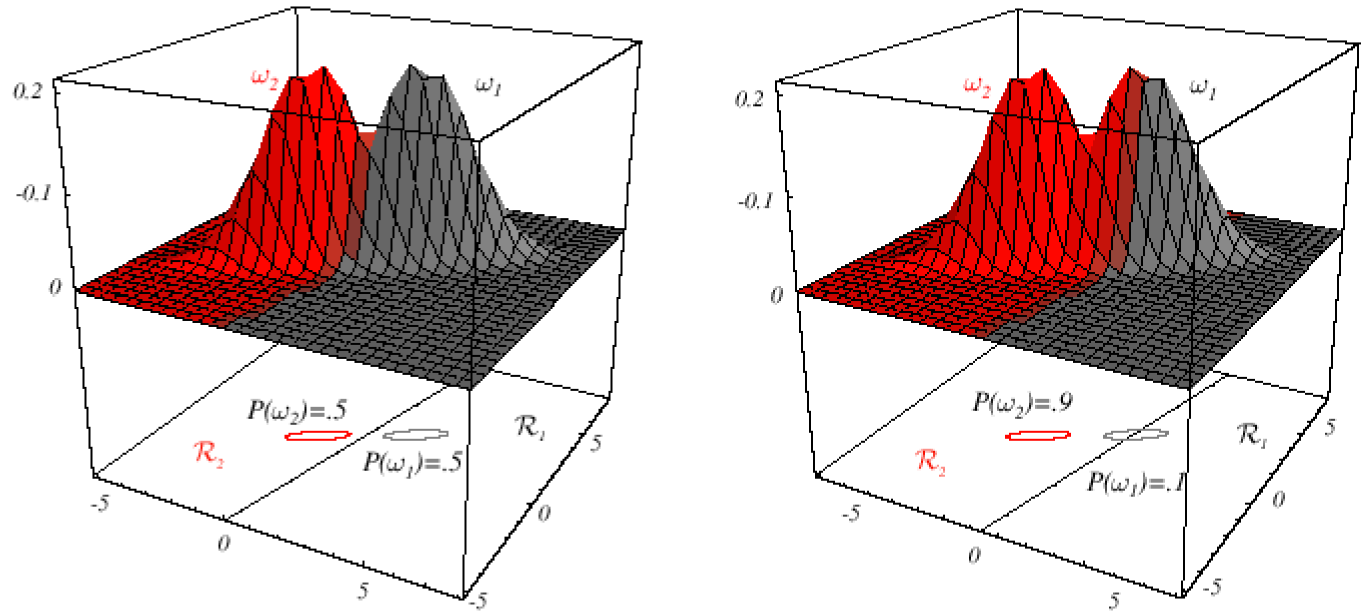
$$\boxed{\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0}$$



$$g_i(\mathbf{x}) = g_j(\mathbf{x})$$

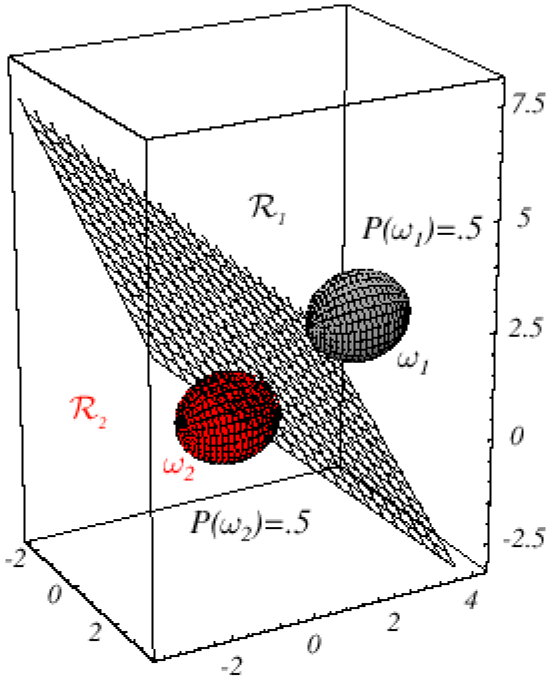
$$\begin{cases} \mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \\ \mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \dfrac{\ln[P(\omega_i)/P(\omega_j)]}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \end{cases}$$

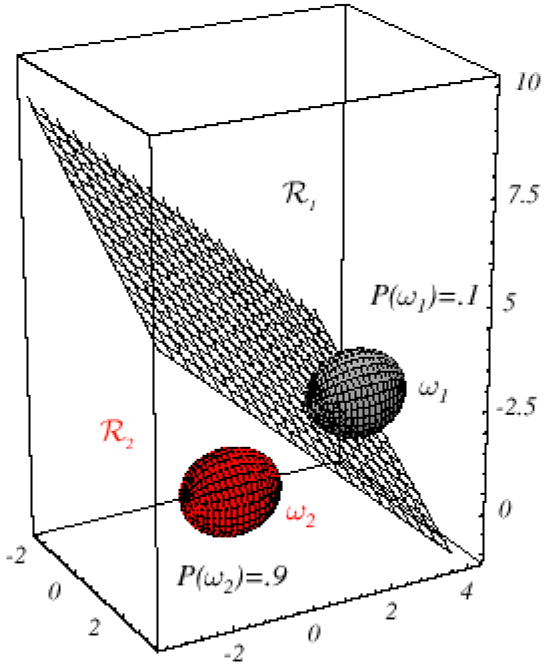
Case 2: $\Sigma_i = \Sigma$



Case 2: $\Sigma_i = \Sigma$

MIMA





Case 3: $\boldsymbol{\Sigma}_i \neq \boldsymbol{\Sigma}_j$

MIMA

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

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

$$g_i(\mathbf{x}) = \underbrace{\mathbf{x}^T \mathbf{W}_i \mathbf{x}}_{\substack{\text{Without this term} \\ \text{In Case 1 and 2}}} + \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

Without this term

In Case 1 and 2

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

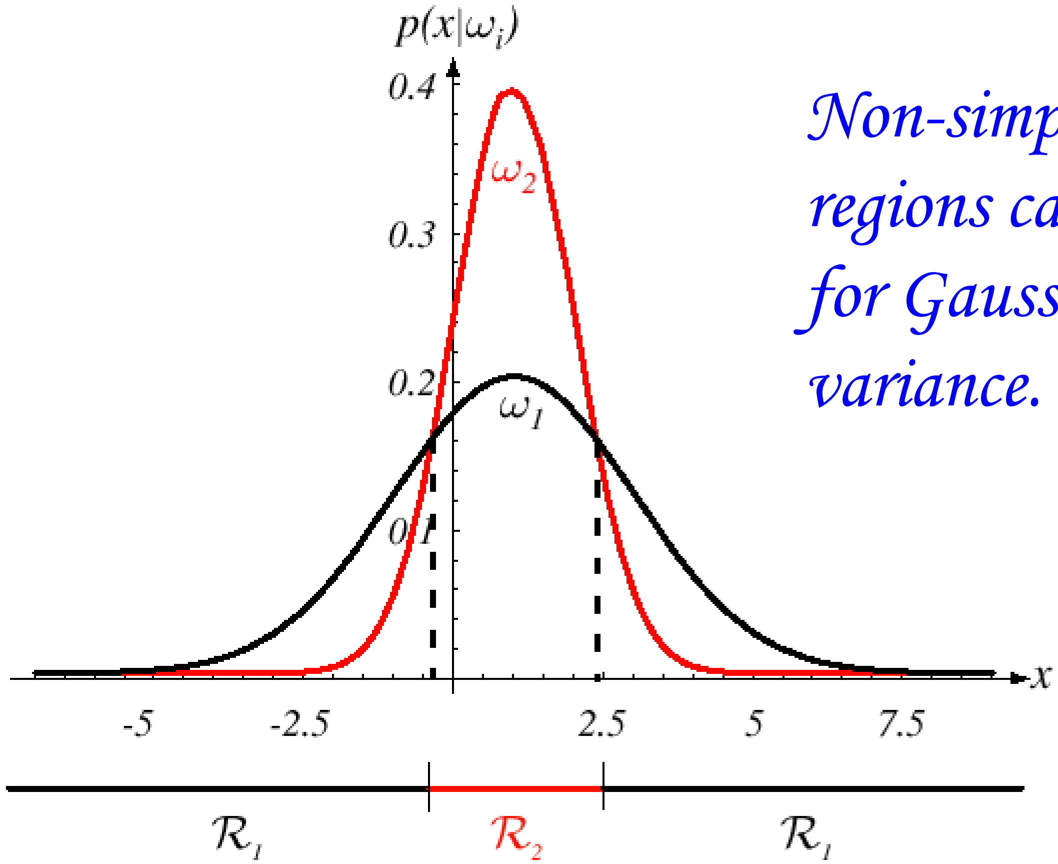
$$\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

$$w_{i0} = -\tfrac{1}{2}\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \tfrac{1}{2}\ln|\boldsymbol{\Sigma}_i^{-1}| + \ln P(\omega_i)$$

Decision surfaces are hyperquadrics, e.g.,

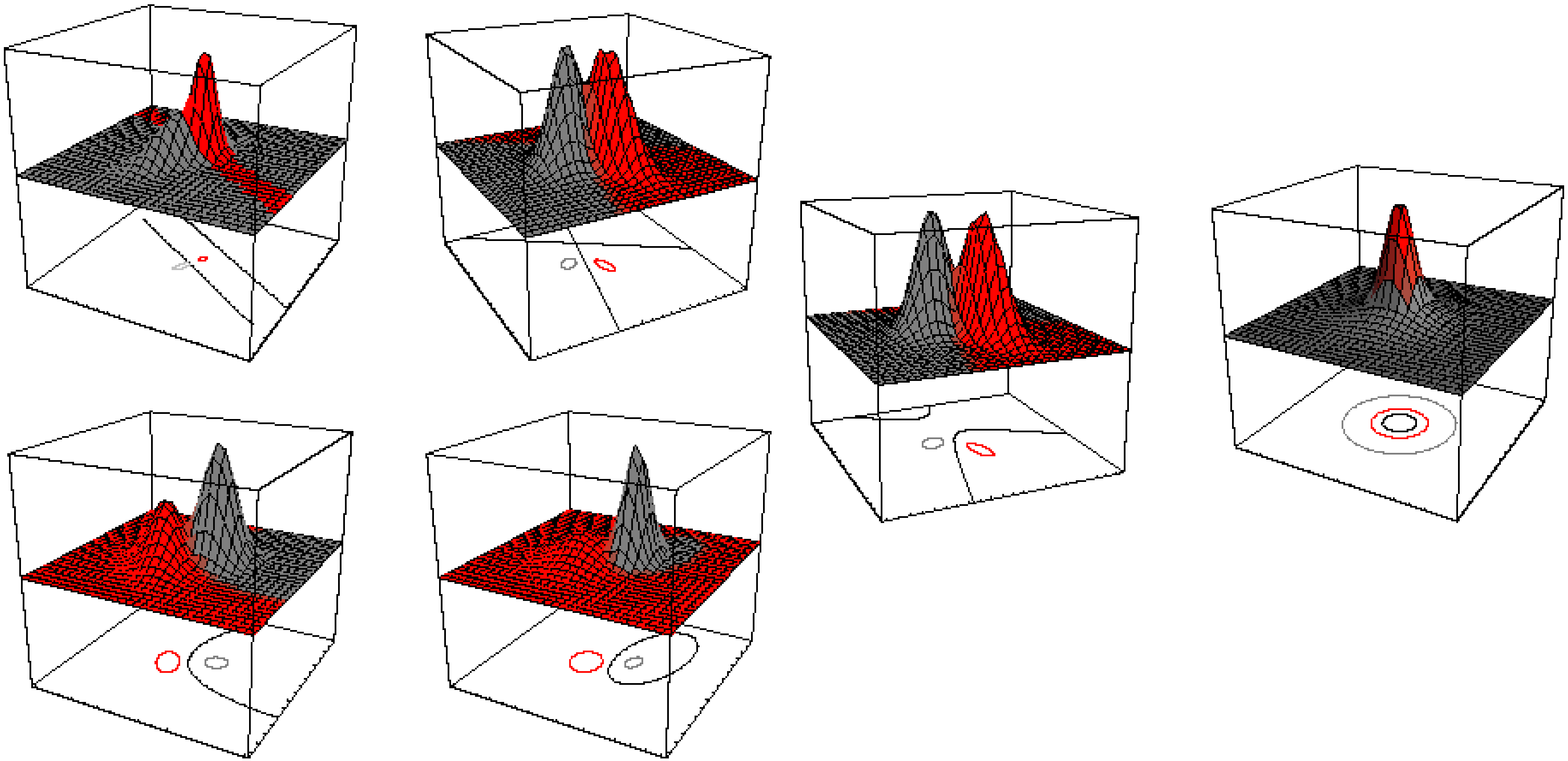
- Hyperplanes
- Hyperspheres
- Hyperellipsoids
- hyperhyperboloids

Case 3: $\Sigma_i \neq \Sigma_j$

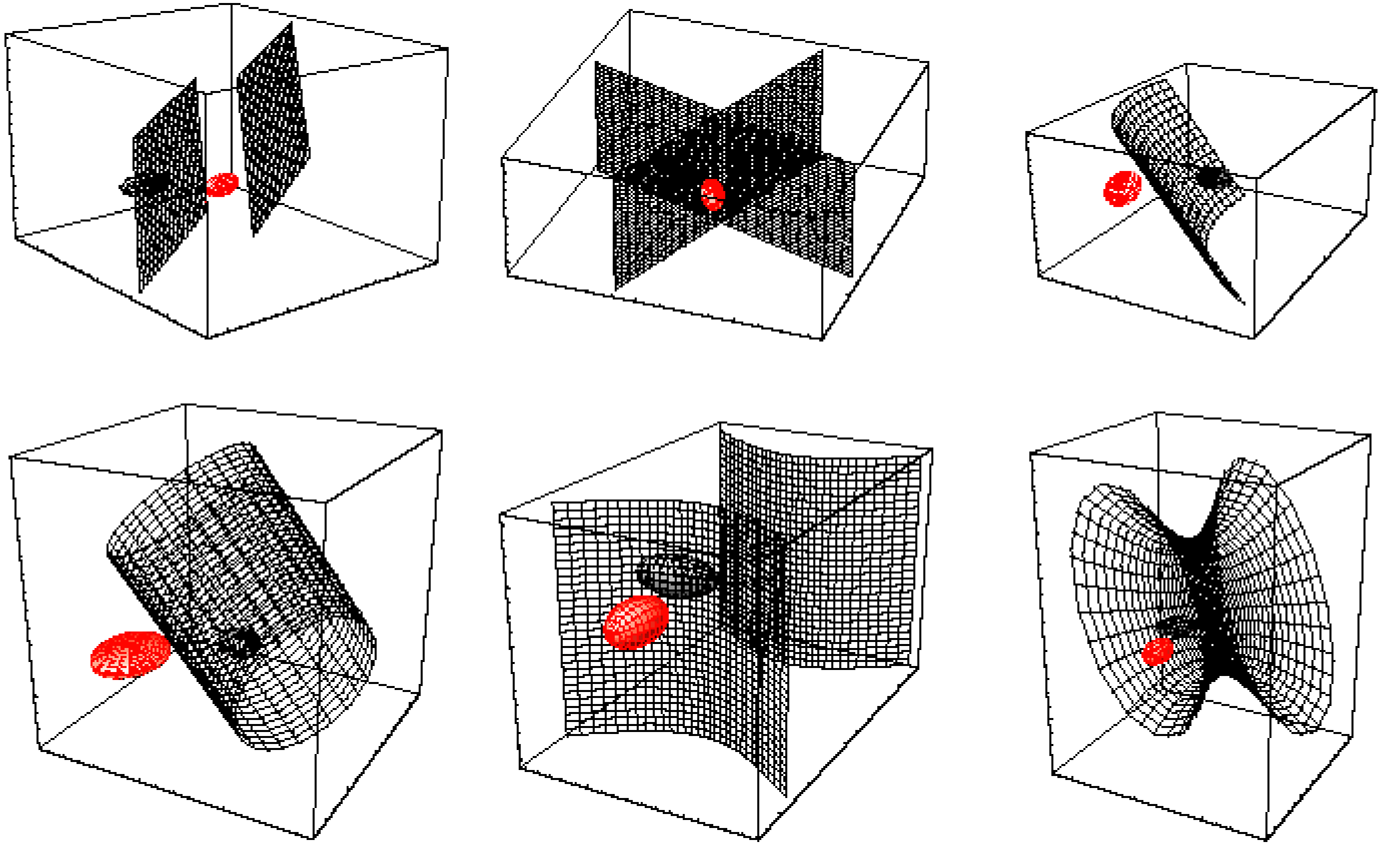


Non-simply connected decision regions can arise in one dimension for Gaussians having unequal variance.

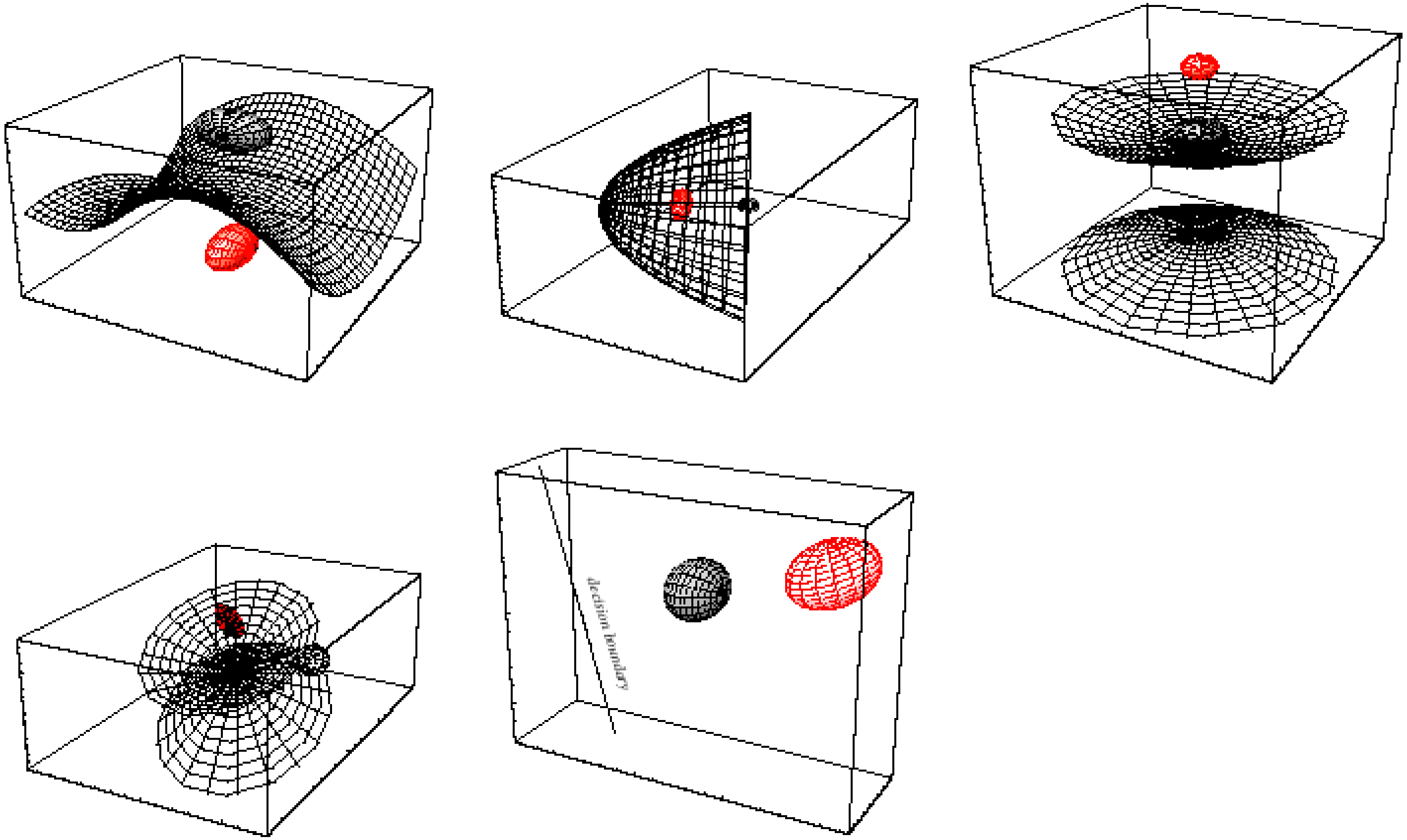
Case 3: $\Sigma_i \neq \Sigma_j$



Case 3: $\Sigma_i \neq \Sigma_j$

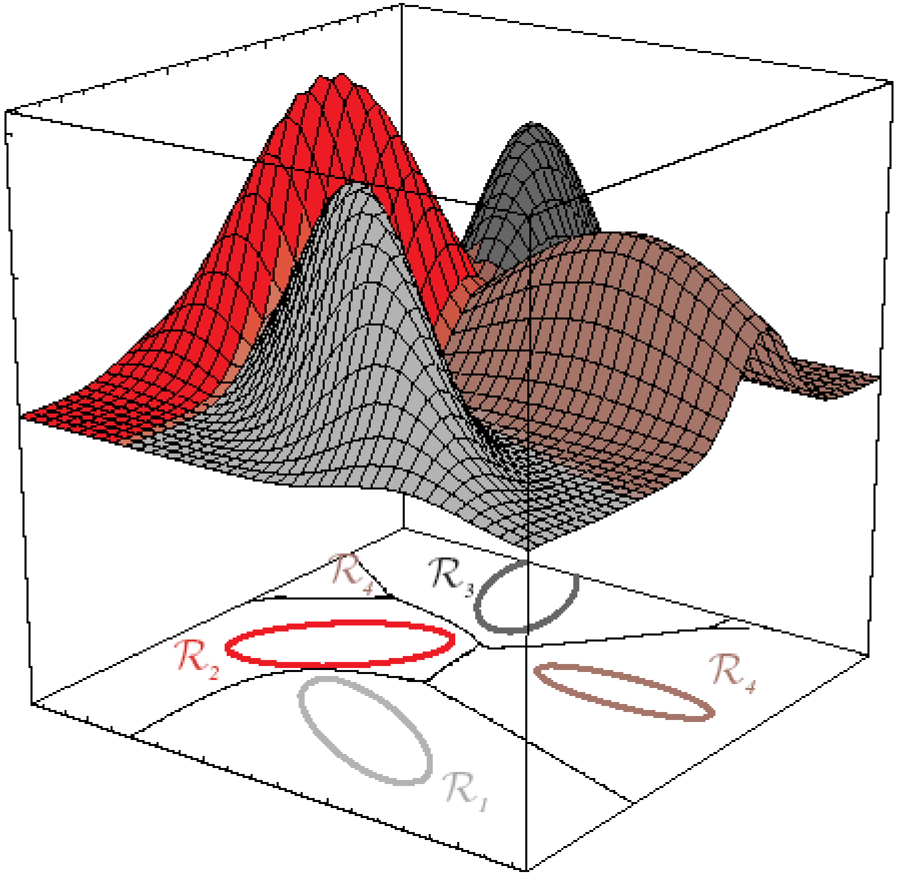


Case 3: $\Sigma_i \neq \Sigma_j$



Case 3: $\Sigma_i \neq \Sigma_j$

MIMA



Summary

MIMA

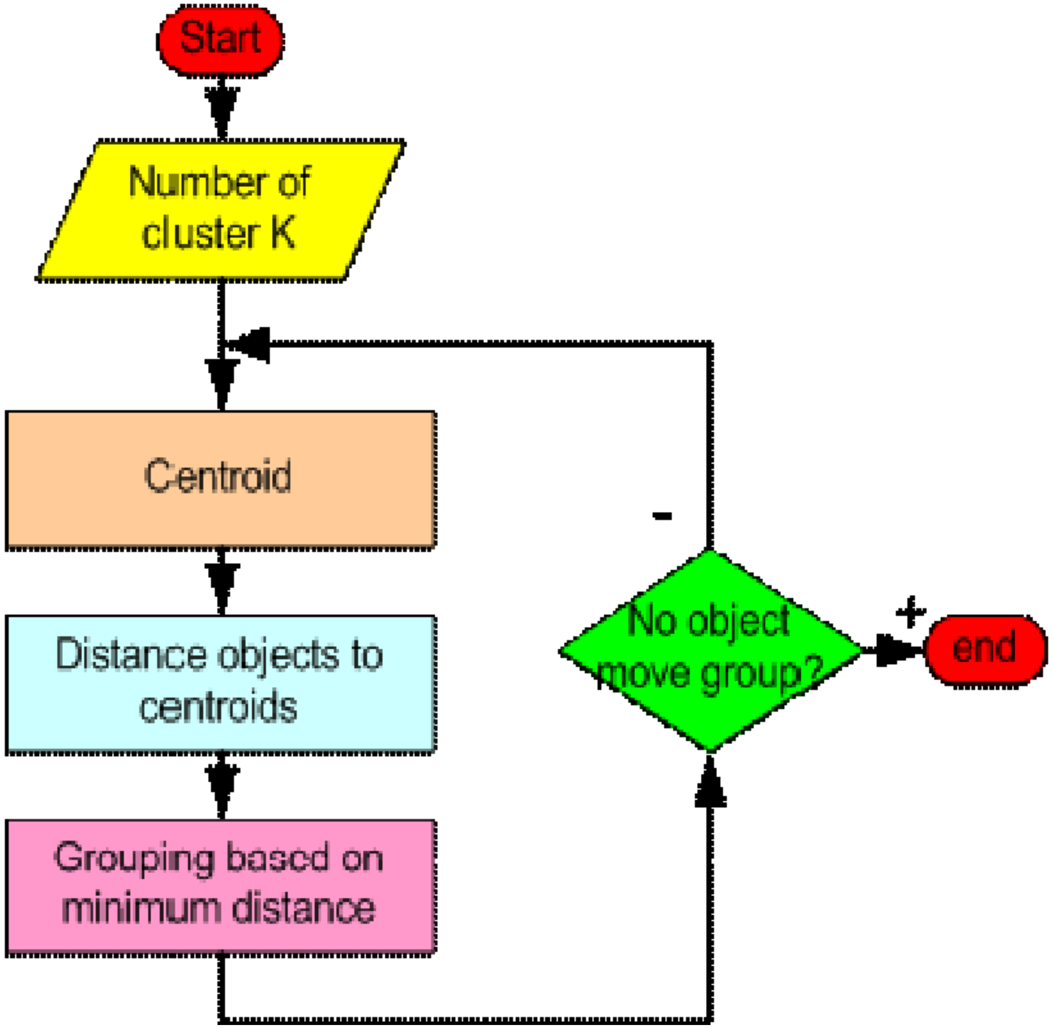
- Bayesian Decision Theory
  - Basic concepts
  - Bayes theorem
  - Bayes decision rule
- Feasibility of Bayes Decision Rule
  - Prior probability + likelihood
  - Solution I: counting relative frequencies
  - Solution II: conduct density estimation

Summary

MIMA

- 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
- Expected loss (conditional risk)
- General Bayes decision rule
- Minimum-error-rate classification
- Discriminant functions
- Gaussian density
- Discriminant functions for Gaussian pdf.

k-means


Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

【 Thank You ! 】

Any Question?