Ch 02.

Bayesian Decision Theory

Symbols

- Random variables representing categories ω
- Class labels $\omega_i, i = 1,2,3,...$
  - Such as：

    ω_1 ： salmon；ω_2 ： sea bass
- Class Prior Probability $P(\omega_i) = P(\omega = \omega_i)$
  - When all categories are mutually exclusive and complete $\sum_{i=1}^{c} P(\omega_i) = 1$
- Class conditional probability density function

  $p(x \mid \omega_i)$

Before Observation

- Problem

  Given the prior probability of all possible categories, predict the category of the next possible pattern without observation

- Optimal Decision Rule

If $P(\omega_j) \geq P(\omega_i), \forall i \neq j$, Then the next model is predicted to be ω_j

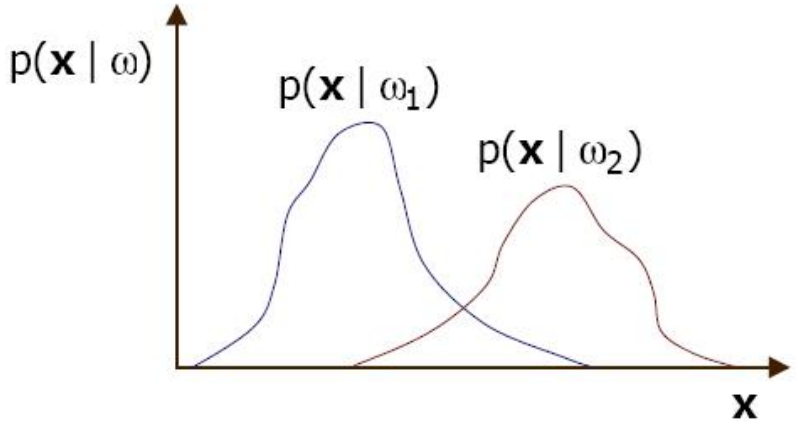
  - In the absence of any observation of the emerging pattern, the decision rule has the least probability of error and is therefore the optimal decision rule;
  - If the prior probability is constant, each prediction is the same.

- If there is more information, can better predictions be made?

After Observation

- When the class is ω_i , the probability of observing the feature x is $p(x \mid \omega_i)$
- Feature x that can be used for classification whose probability distribution should be different in different categories

$$\int_{\mathcal{R}_i} p(\mathbf{x} \mid \omega_i)\, d\mathbf{x} = 1$$



Prediction based on Observed Features

- Goal：When x are observed, the probability of class ω_i is $P(\omega_i \mid x),\ i = 1, 2, 3..., c$
- Decision Rule:

When x is observed from the sample,

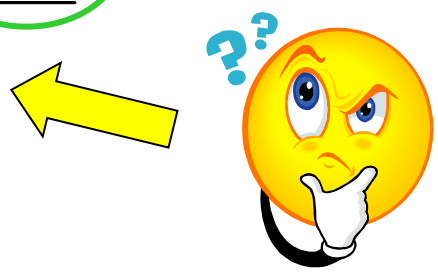
if $P(\omega_j \mid x) \geq P(\omega_i \mid x), \forall i \neq j$,

then the sample is predicted as ω_j

Bayes Formula

- Use prior probability（before observe ***x***）to calculate posterior probability （after observe ***x***）

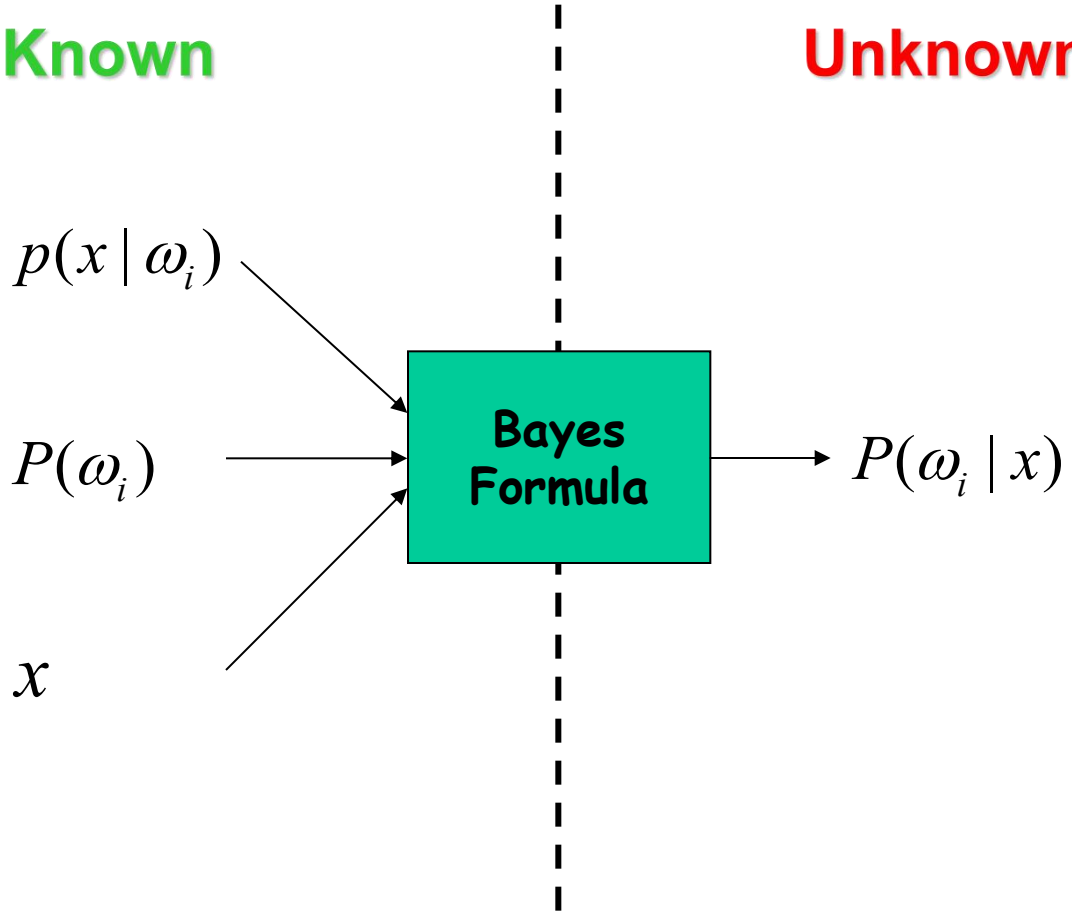
$$P(\omega_i \mid x) = \frac{p(x \mid \omega_i) P(\omega_i)}{p(x)}$$



$$p(x) = \sum_{i=1}^{c} p(x \mid \omega_i) P(\omega_i)$$ **Constant!**

$$\sum_{i=1}^{c} P(\omega_i \mid x) = 1$$

Bayes Formula



Special Case

- Case One
  - Uniform prior probability: $P(\omega_1) = P(\omega_2) = \cdots = P(\omega_c) = \frac{1}{c}$
  - Decision only depends on $p(x \mid \omega_i)$

When x is observed from the sample,

if $P(x \mid \omega_j) \geq P(x \mid \omega_i), \forall i \neq j$,

then the sample is predicted as ω_j

Special Case

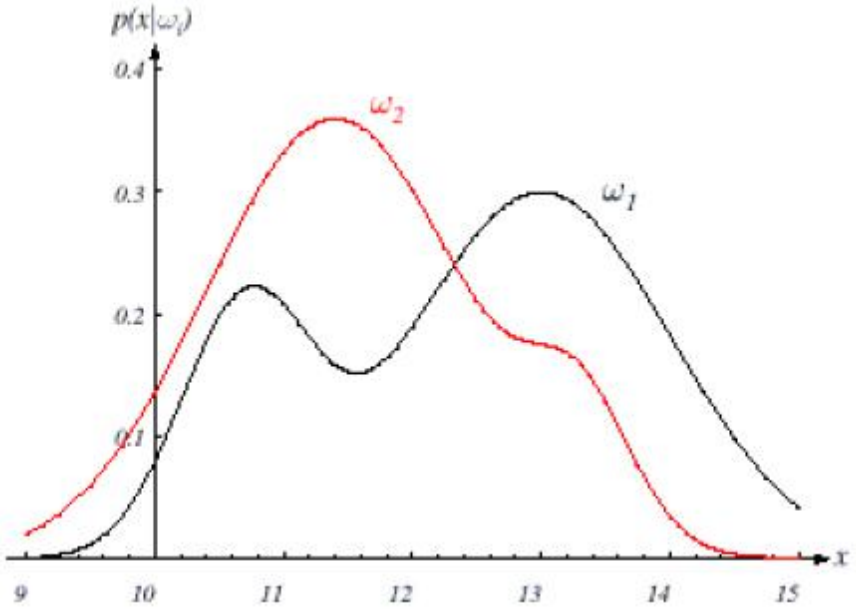
- Case Two
  - Same class conditional probability density function：

    $p(\mathbf{x} \mid \omega_1) = p(\mathbf{x} \mid \omega_2) = \cdots = p(\mathbf{x} \mid \omega_c)$

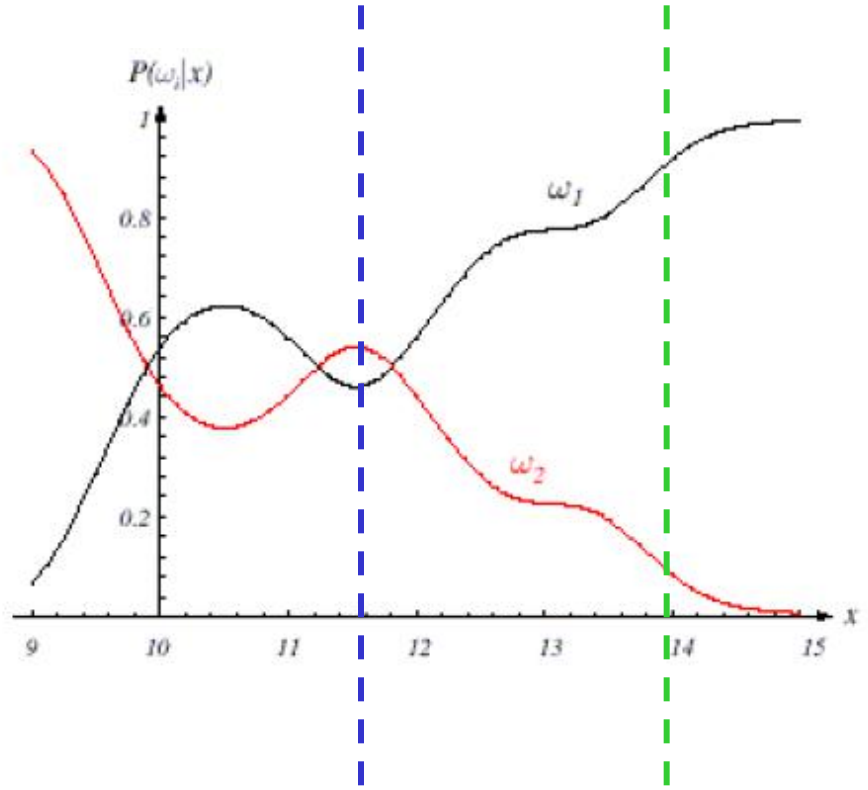
  - Decision only depends on prior probability

If $P(\omega_j) \geq P(\omega_i), \forall i \neq j$, then the sample is predicted as ω_j

Example One



Class conditional probability density function



Posterior probability

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

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

Example One

Problem statement

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

Question

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

Example One

ω_1 : cancer $\qquad$ ω_2 : no cancer $\qquad$ $x \in \{+, -\}$

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

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

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

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

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

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

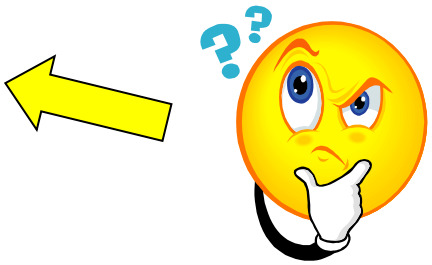
$P(\omega_2 \mid +) > P(\omega_1|+)$

No cancer!

How to Determine Probability ?

- Applying Bayesian decision rules, the following probabilities is required to be known

$p(x \mid \omega_i)$ $P(\omega_i)$



- For a specific problem, it is often necessary to calculate the relative frequency by experiments, or to use probability density estimation techniques to determine the above probability

Example Two

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 $> \$50,000$

ω_2 : price $\leq \$50,000$

x : height of car

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

?

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

Quantities to know:

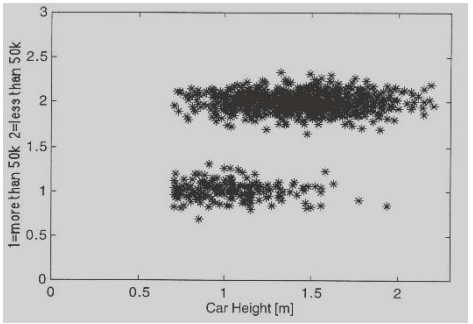
$P(\omega_1)$ $P(\omega_2)$ $p(x|\omega_1)$ $p(x|\omega_2)$

Counting relative frequencies via collected samples

Example Two

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**

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

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

Example Two

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

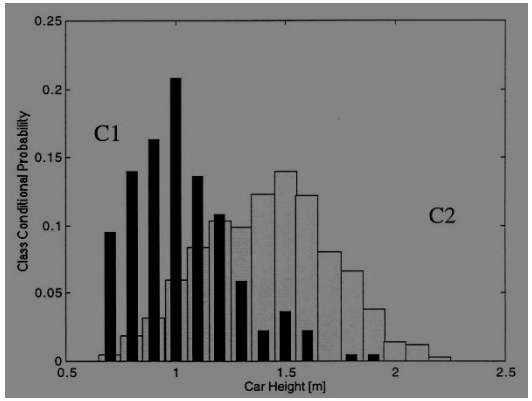
***x* falls into interval I_x=[1.0m, 1.1m]**

For ω_1, # cars in I_x is 46

For ω_2, # cars in I_x is 59

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

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



Example Two

Question

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

Estimated quantities

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

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

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

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

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

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

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

Is Bayes Decision Rule Optimal?

Bayes Decision Rule (In case of two classes)

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

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

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

Under Bayes decision rule, we have

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

For every *x*, we ensure that *P*(*error* | *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

- Generalize the problem as follows：
  - Allow for multi-class；
  - Allowing other actions and not just classification；
  - A more general loss function is introduced to replace the error probability。
- Loss Function $\lambda(\alpha_i \mid \omega_j)$
  - When the real class is ω_j， losses arising from α_i action taken
  - Allowing the cost of a certain classification error more than other classification errors

Bayes Decision Rule – The General Case

- Conditional risk (Expected loss)
  - When ***x*** are observed, the expected loss α_i action is taken

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

Optimality of Bayesian Decision

- **Decision Rule Function** $\alpha(\mathbf{x})$

  Function of mapping observed feature x to action to be taken

- **Total Risk** $R = \int R(\alpha(\mathbf{x}) \mid \mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}$

  Expected loss of a decision rule

- **Optimal Decision**
  - Decision rule that makes the total risk lowest
  - For given feature ***x***, if the action chosen by decision rule $\alpha(\mathbf{x})$ can minimize conditional risk $R(\alpha(\mathbf{x}) \mid \mathbf{x})$, total risk would be minimized

Optimality of Bayesian Decision

- ***Bayesian Decision Rule***: For all i=1,2,…,a, compute conditional risk $R(\alpha_i \mid \mathbf{x})$, choose action α_j to minimize the condition risk $R(\alpha_j \mid \mathbf{x})$

The minimum total risk obtained by Bayesian decision is called *Bayesian Risk*, which is expressed as R^*

Two-Class Classification

- Action
  - α_1 : decide ω_1
  - α_2 : decide ω_2
- Loss
  - $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$
- Conditional Risk

$$R(\alpha_1 \mid \mathbf{x}) = \lambda_{11} P(\omega_1 \mid \mathbf{x}) + \lambda_{12} P(\omega_2 \mid \mathbf{x})$$
$$R(\alpha_2 \mid \mathbf{x}) = \lambda_{21} P(\omega_1 \mid \mathbf{x}) + \lambda_{22} P(\omega_2 \mid \mathbf{x})$$

- Minimum risk decision rule

If $R(\alpha_j \mid \mathbf{x}) \le R(\alpha_i \mid \mathbf{x}), i \ne j$, decide ω_j

Two-Class Classification

- Equivalent minimum risk decision rules

$$R(\alpha_j \mid \mathbf{x}) \le R(\alpha_i \mid \mathbf{x}),\ i \ne j$$

$$(\lambda_{i1} - \lambda_{j1})\, P(\omega_1 \mid \mathbf{x}) \ge (\lambda_{j2} - \lambda_{i2})\, P(\omega_2 \mid \mathbf{x}),\ i \ne j$$

$$(\lambda_{i1} - \lambda_{j1})\, p(\mathbf{x} \mid \omega_1)\, P(\omega_1) \ge (\lambda_{j2} - \lambda_{i2})\, p(\mathbf{x} \mid \omega_2)\, P(\omega_2),\ i \ne j$$

- In general, the loss of classification errors is greater than loss of correct (often no loss when correct)

$$\lambda_{21} - \lambda_{11} > 0 \qquad \lambda_{12} - \lambda_{22} > 0$$

- Likelihood Ratio

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

independent of x, a precomputed constant θ

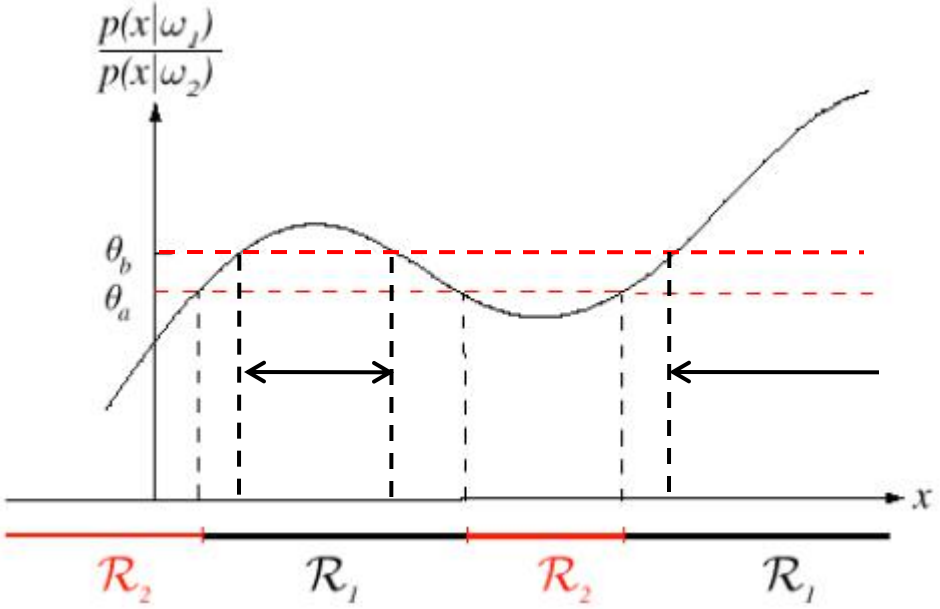
Two-Class Classification

- Bayesian decision rule based on likelihood ratio

If $\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} \geq \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$, decide ω_1

otherwise decide ω_2

Example



- Different loss functions determine different decision thresholds θ_a and θ_b
  - θ_a: "0-1" Loss
  - θ_b: $\lambda_{12} > \lambda_{21}$
- Each class of decision domains may be discontinuous

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 ←→ Decide states of nature**
  - **0-1 loss function**
    - **Assign *no loss*/*unit loss* for *correct*/*incorrect* decisions**

"0-1" Loss

- "0-1" Loss（对称损失）Function

$$\lambda(\alpha_i \mid \omega_j) = \begin{cases} 0 & i = j \text{ (correct decision)} \\ 1 & i \neq j \text{ (incorrect decision)} \end{cases} \quad i, j = 1, \ldots, c$$

  - When the decision is correct, there is no loss, and any loss of error is equal to 1, that is, all misjudgments are equivalent

- Conditional Risk

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

$$= \sum_{j \neq i} P(\omega_j \mid \mathbf{x}) = 1 - P(\omega_i \mid \mathbf{x})$$

  - Conditional risk is the error rate
  - Minimize the conditional risk $R(\alpha_i \mid \mathbf{x})$ is equal to maximize posterior probability $P(\omega_i \mid \mathbf{x})$

Minimum-Error-Rate Classification

- Minimum-error-rate classificaiton is minimum risk classification using “0-1” loss function
- Minimum-error-rate classification rule in two-class classification

If $\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} \geq \frac{P(\omega_2)}{P(\omega_1)}$, decide ω_1

otherwise decide ω_2

- Minimum-error-rate classification rule in multi-class classification

If $P(\omega_j \mid x) \geq P(\omega_i \mid x), \forall i \neq j$,

decide ω_j

Minimax Rule

- Minimum risk classifier depends on prior probability
- How to design a classifier with less risk when the prior probability is unknown?
  - **Minimize maximum possible total risk**
  - **Minimax Rule**

Minimax Rule

- Total Risk

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

For two-class classification：

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

Minimax Rule

- Denote $P(\omega_1)+P(\omega_2)=1$ and $\int_{\mathcal{R}_1} p(\mathbf{x}\mid\omega_i)\,d\mathbf{x}+\int_{\mathcal{R}_2} p(\mathbf{x}\mid\omega_i)\,d\mathbf{x}=1$, total risk is

Once $\mathcal{R}_1$ and $\mathcal{R}_2$ determined, these are constants

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

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

- R and $P(\omega_1)$ is linear $R=\mu P(\omega_1)+R_{mm}$
- Choose $\mathcal{R}_1$ and $\mathcal{R}_2$ to make $\mu=0$, then total risk is independent of $P(\omega_1)$, which is called as minimax risk

Minimax Rule

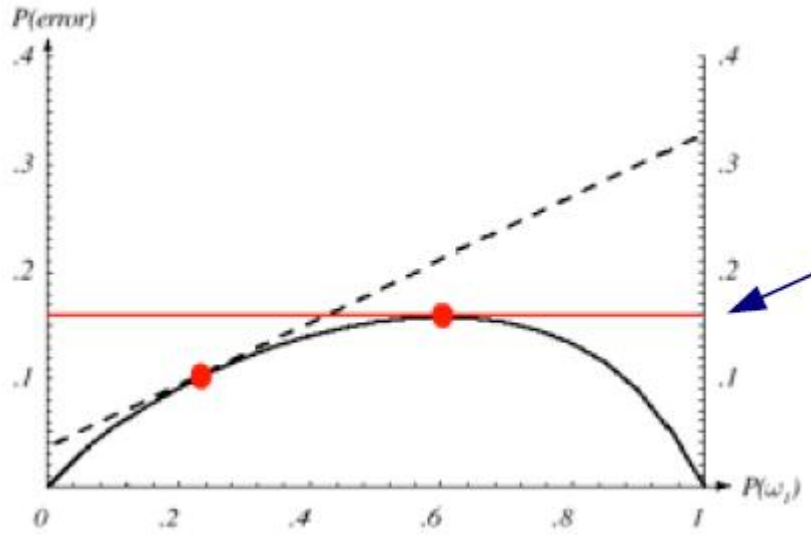
- Minimax Risk

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

Minimax risk can also denoted as

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

- Example

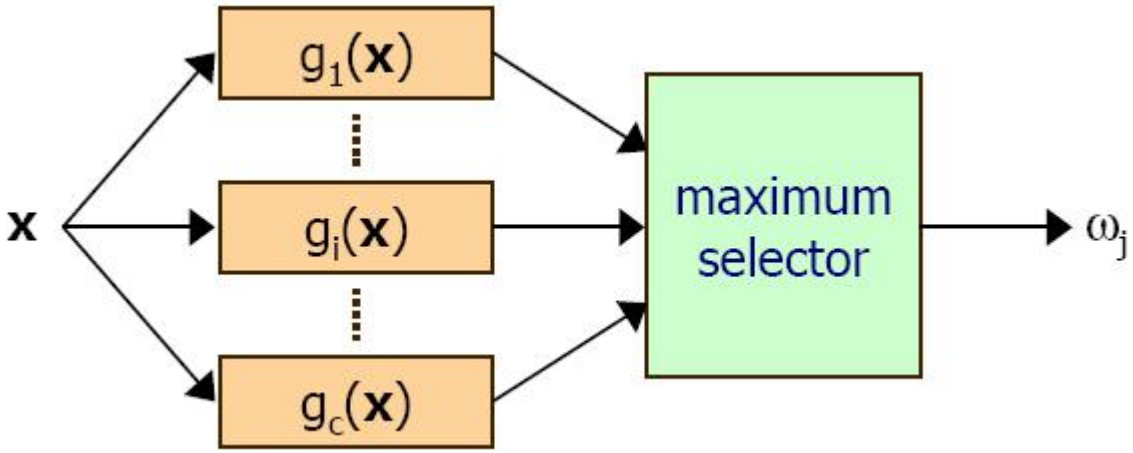


Minimax Rule (Independent of prior probability)

Discriminant Function (判别函数)

- The most commonly used expression of classifier is discriminant function $g_i(\mathbf{x}), i=1,\ldots,c$, Each class corresponds to a discriminant function
- Decision Rule Based on Discriminant Function

If $g_j(\mathbf{x}) \geq g_i(\mathbf{x}), \forall i \neq j$, decide ω_j



Discriminant Function

- Bayesian classifier based on minimum total risk

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

- Bayesian classifier based on minimum error probability

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

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

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

- Different discriminant functions may be used to express the same decision rule, provided that the following conditions are satisfied:

- Replace $g_i(\mathbf{x})$ with $f(g_i(\mathbf{x}))$, where $f(\cdot)$ is a *monotonically increasing function*
  - $g_i(\mathbf{x}) \longrightarrow k\, g_i(\mathbf{x})$, where $k > 0$
  - $g_i(\mathbf{x}) \longrightarrow g_i(\mathbf{x}) + k$

Discriminant Function

- Two-class Classification
  - Only a discriminant function is needed

    $g(\mathbf{x}) \equiv g_1(\mathbf{x}) - g_2(\mathbf{x})$

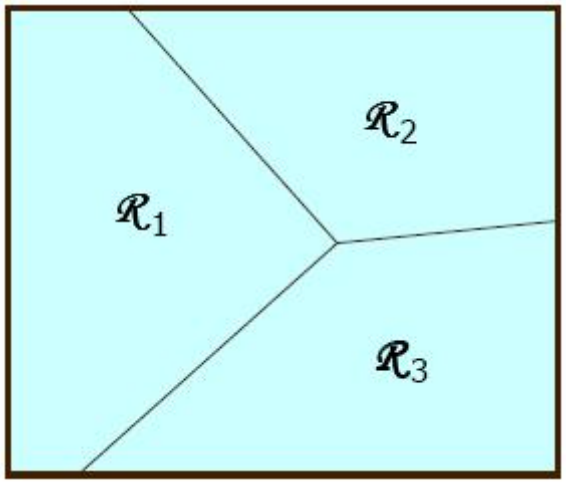
- Disicion Rule

  If $g(\mathbf{x}) \geq 0$, decide ω_1 , otherwise decide ω_2

- Examples
  - $g(\mathbf{x}) = P(\omega_1 \mid \mathbf{x}) - P(\omega_2 \mid \mathbf{x})$
  - $g(\mathbf{x}) = \log \frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} + \log \frac{P(\omega_1)}{P(\omega_2)}$

Discriminant Function

- **Decision region (判决区域)**
  - Decision region $\mathcal{R}_i$ is a subspace in the feature space，the decision rules classify all sample x falling into $\mathcal{R}_i$ into class ω_i
- **Decision boundary (判决边界)**
  - Surface in feature space where ties occur among several largest discriminant functions



Multivariate Gaussian Density Function

- d-dimensional Gaussian density function

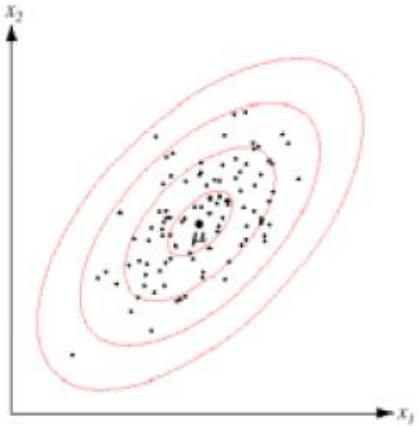
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\,|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}-\boldsymbol{\mu})\right]$$

  - $\boldsymbol{\mu}$ is d-dimensional mean vector
  - $\boldsymbol{\Sigma}$ is d×d covariance matrix，usually a symmetric semi-positive definite matrix

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

  - Take log

$$\log p(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}-\boldsymbol{\mu}) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}|$$



Summary (Cont.)

- Discriminant functions
  - General way to represent classifiers
  - One function per class
  - 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**

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

Discriminant Function

- Class conditional probability density function

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

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

- Discriminant Function Based on Minimum Error Probability Classification

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

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

Constant

Special Cases

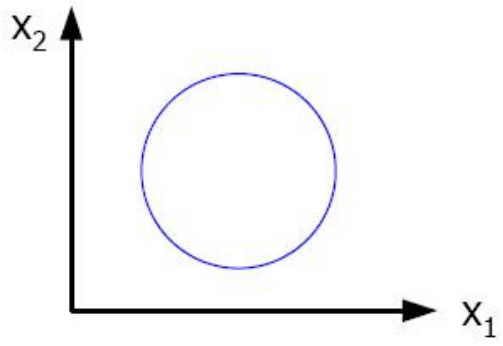
- Case 1：Uniform prior probability

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

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| \qquad \boldsymbol{\Sigma}_i = \sigma^2\mathbf{I}$$

- Case1a：Uniform prior probability，and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$



  - Statistical independence of features
  - All features have the same variance

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i) = -\frac{1}{2\sigma^2}\left\|\mathbf{x}-\boldsymbol{\mu}_i\right\|^2$$

Square Euclidean distance

Special Cases

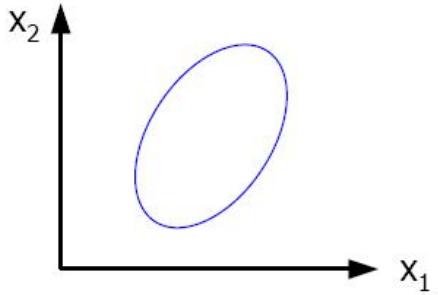
- Case 1：Uniform prior probability

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

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| \qquad \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$$

  - Case1b：Uniform prior probabilty，and
    - All classes of data have

      the same covariance matrix



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

Square Mahalanobis distance（马氏距离）

Case 1a and1b can be considered as a minimum distance classifier，that is, classify **x** to be the class which the nearest mean vector $\boldsymbol{\mu}_i$ belong to.

Special Cases

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

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

Same for all i

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

Let $\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i$, $w_{i0} = \frac{-1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \log P(\omega_i)$

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

Threshold in i direction (bias)

A classifier using a linear discriminant function is called a linear classifier (linear machine)

Special Cases

- Case 2: $\Sigma_i = \sigma^2 \mathbf{I}$
  - Decision plane of a linear machine is a number of hyperplanes, each of which is determined by the equality of the discriminant functions of two classes with the maximum posterior probability: :

$$g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$$

$$\Leftrightarrow \left[2\boldsymbol{\mu}_i^t\mathbf{x} - 2\boldsymbol{\mu}_j^t\mathbf{x} - \boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \boldsymbol{\mu}_j^t\boldsymbol{\mu}_j\right] + 2\sigma^2 \log\frac{P(\omega_i)}{P(\omega_j)} = 0$$

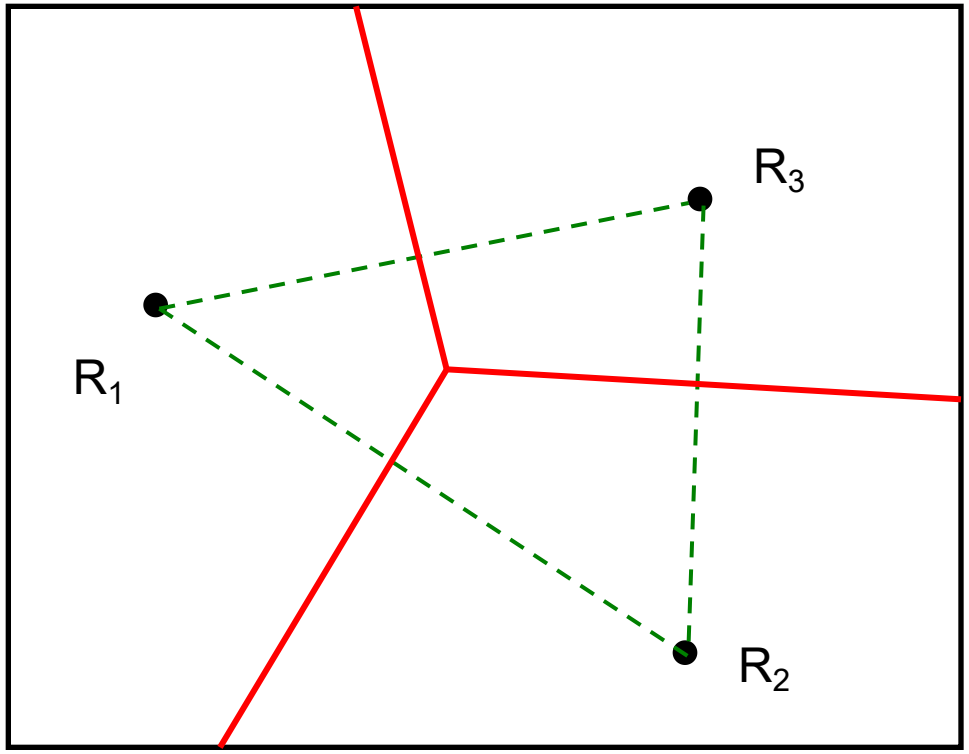
$$\Leftrightarrow (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t(2\mathbf{x} - \boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + 2\sigma^2 \log\frac{P(\omega_i)}{P(\omega_j)}\frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} = 0$$

$$\Leftrightarrow (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\left[\mathbf{x} - \left(\frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_j}{2} - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}\log\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right)\right] = 0$$

$\mathbf{x_0}$

The hyperplane passes through $\mathbf{x_0}$, and is perpendicular to $(\mu_i - \mu_j)$, that is, the connection between two types of mean points

Special Cases



Special Cases

- Return to Case 1a： Uniform prior probability,

  and $\Sigma_i = \sigma^2 \mathbf{I}$

  - The discriminant function is only related to the mean values in every classes

    $$g_i(x) = 2\mu_i^t x - \mu_i^t \mu_i$$

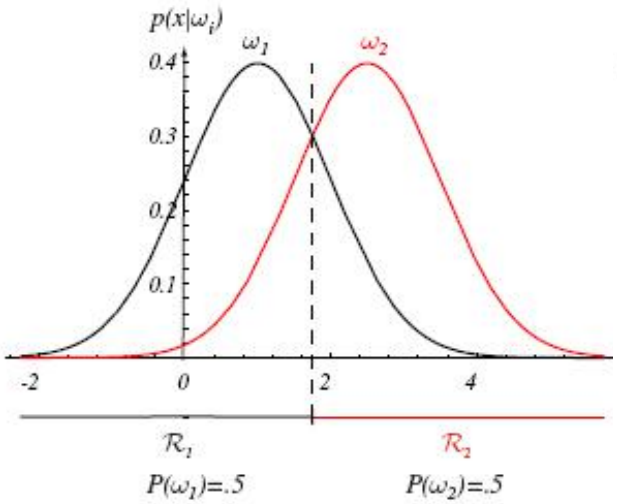
  - Decision plane

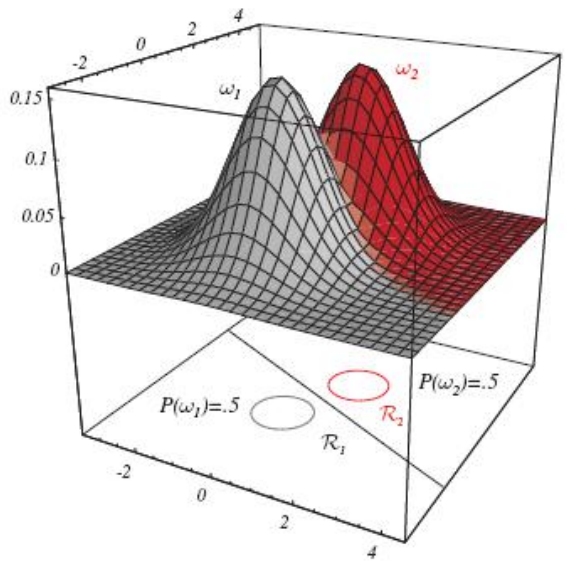
    $$(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \left(\mathbf{x} - \frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_j}{2} \right) = 0$$

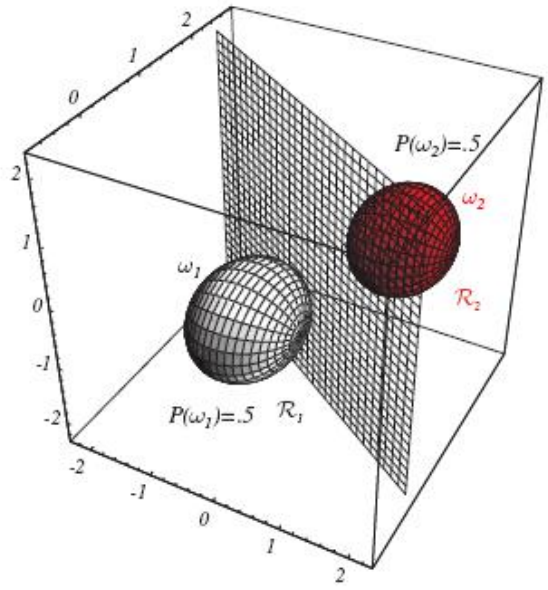
Decision surface is a vertical median dividing line connecting two mean vectors in two classes

Special Cases

- Return to Case 1a： Uniform prior probability, and $\Sigma_i = \sigma^2 \mathbf{I}$

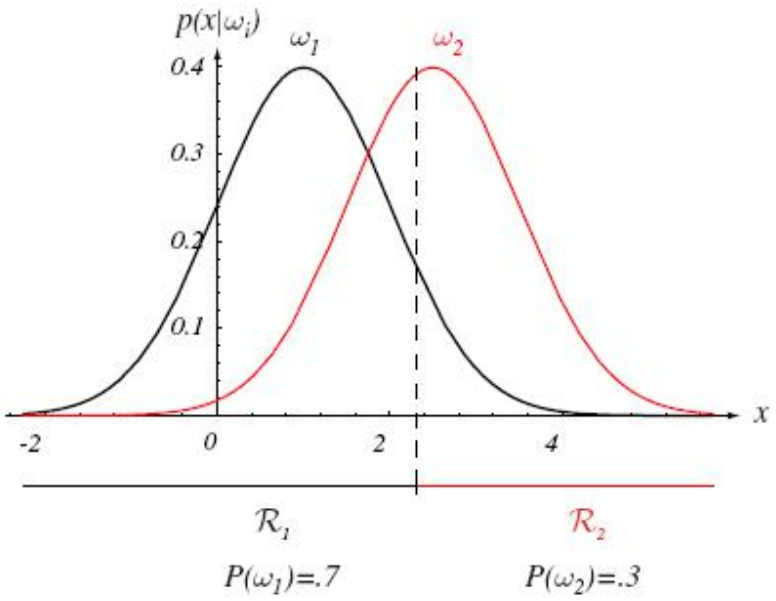




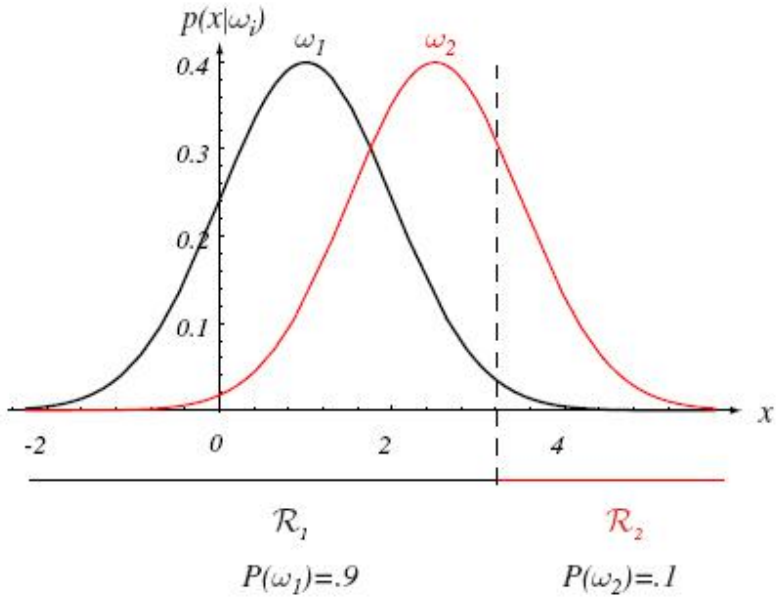


Special Cases

- How about the prior probability is different? —— one-dimensional



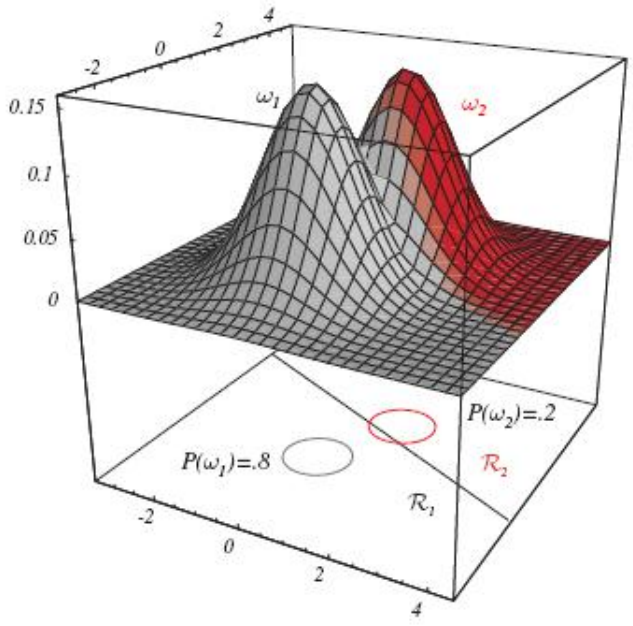
$P(\omega_1) = 0.7 \quad P(\omega_2) = 0.3$



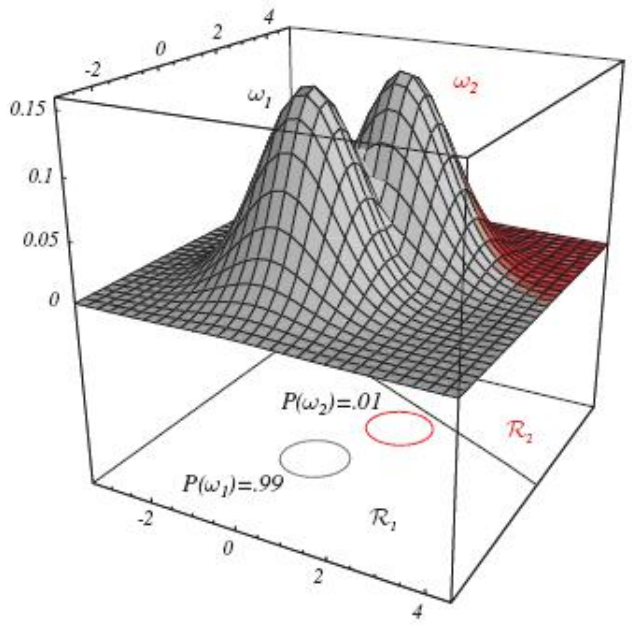
$P(\omega_1) = 0.9 \quad P(\omega_2) = 0.1$

Special Cases

- How about the prior probability is different? —— two-dimensional



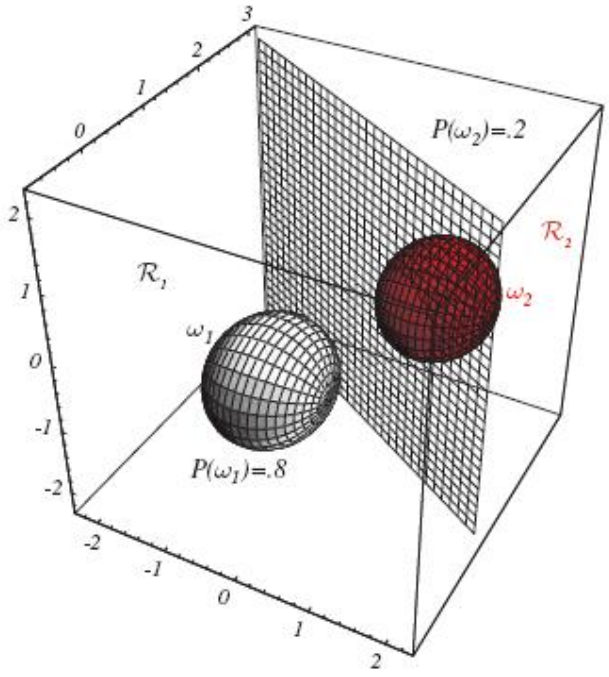
$P(\omega_1) = 0.8 \quad P(\omega_2) = 0.2$



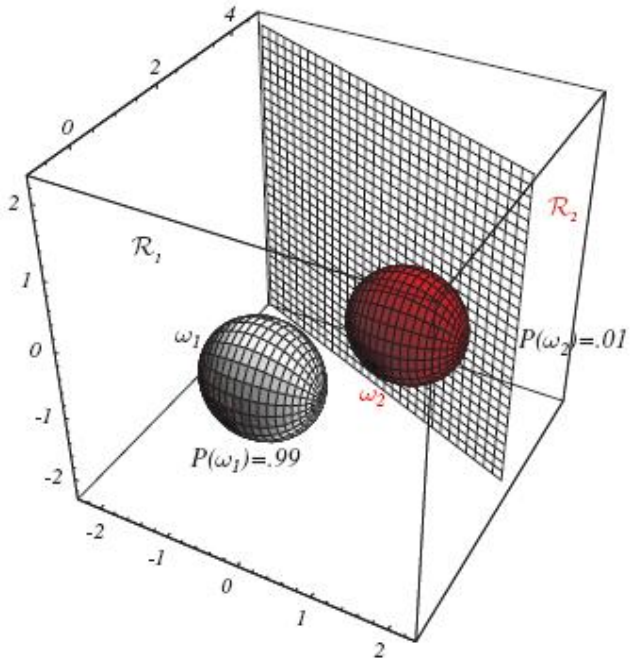
$P(\omega_1) = 0.99 \quad P(\omega_2) = 0.01$

Special Cases

- How about the prior probability is different? —— three-dimensional



$P(\omega_1) = 0.8 \quad P(\omega_2) = 0.2$



$P(\omega_1) = 0.99 \quad P(\omega_2) = 0.01$

Special Cases

- Case 3: $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$

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

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

Same for all i

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

Special Cases

- Case 3: $\Sigma_i = \Sigma$
  - Decision plane consists of a hyperplane in the following form

$$g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$$

$$\Leftrightarrow \left[2\,\boldsymbol{\mu}_i^t\,\boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\,\boldsymbol{\mu}_j^t\,\boldsymbol{\Sigma}^{-1}\mathbf{x} - \boldsymbol{\mu}_i^t\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \boldsymbol{\mu}_j^t\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j\right] + 2\log\frac{P(\omega_i)}{P(\omega_j)} = 0$$

$$\Leftrightarrow (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(2\mathbf{x} - \boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + 2\log\frac{P(\omega_i)}{P(\omega_j)}\frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} = 0$$

$$\Leftrightarrow \left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right]^t\left[\mathbf{x} - \left(\frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_j}{2} - \frac{1}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}\log\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right)\right] = 0$$

$\mathbf{x}_0$

The hyperplane passes through $\mathbf{x}_0$, and usually is not perpendicular to $(\mu_i - \mu_j)$

Special Cases

- Return to Case 1b： Uniform prior probability,

  and $\Sigma_i = \Sigma$

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

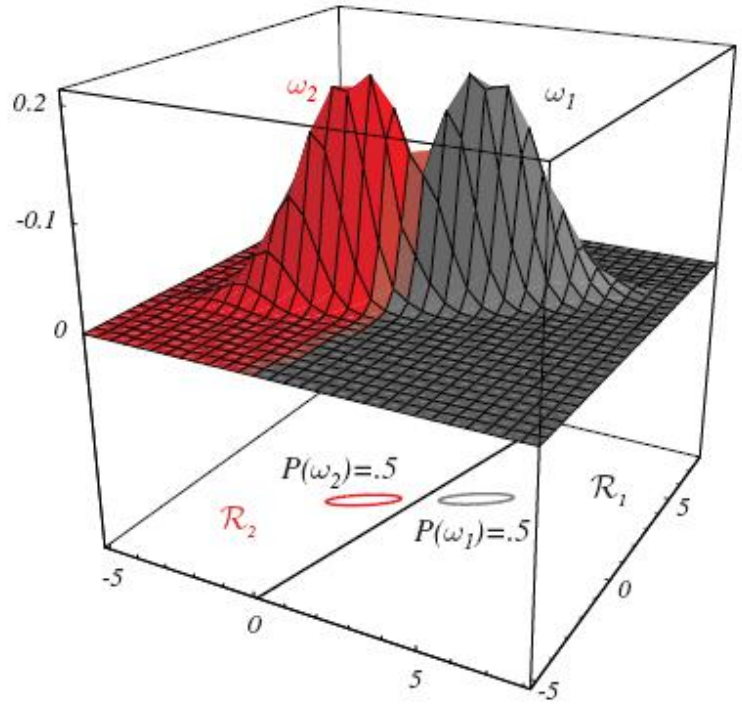
  - Decision plane

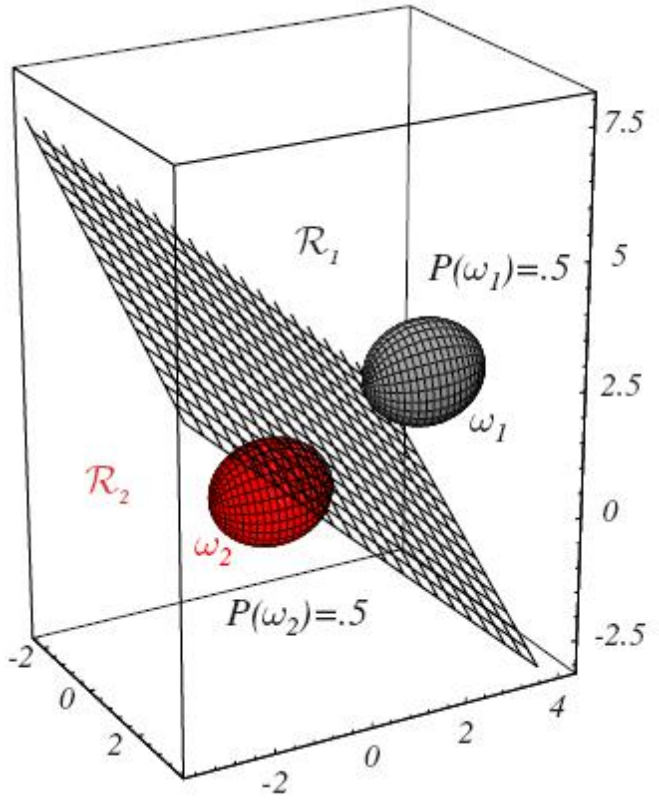
$$\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right]^t\left(\mathbf{x} - \frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_j}{2}\right) = 0$$

Decision plane pass through the mid-point of the line which connects the two mean vectors

Special Cases

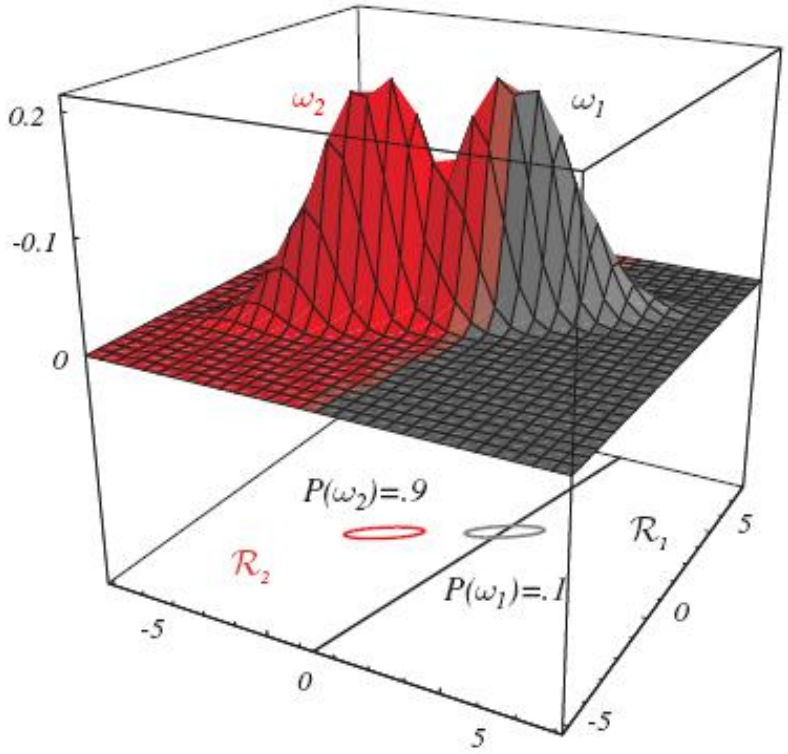
- Return to Case 1b： Uniform prior probability, and $\Sigma_i = \Sigma$

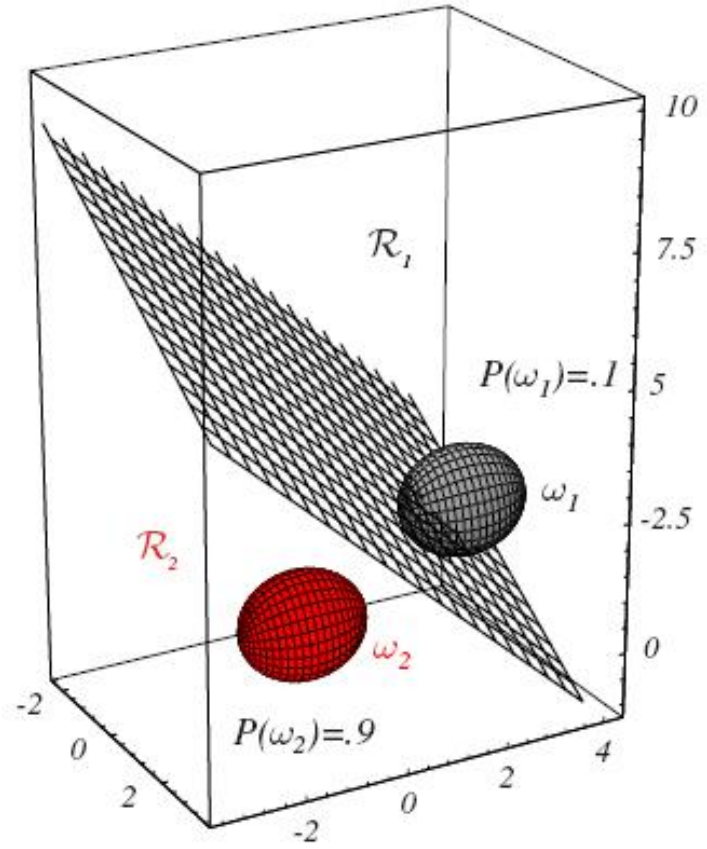




Special Cases

- How about the prior probability is different?





General Cases

- Random Gaussian density function

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

Let $\mathbf{W}_i = -\frac{1}{2}\Sigma_i^{-1}$, $\mathbf{w}_i = \Sigma_i^{-1}\boldsymbol{\mu}_i$, $w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t\Sigma_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\log|\Sigma_i| + \log P(\omega_i)$

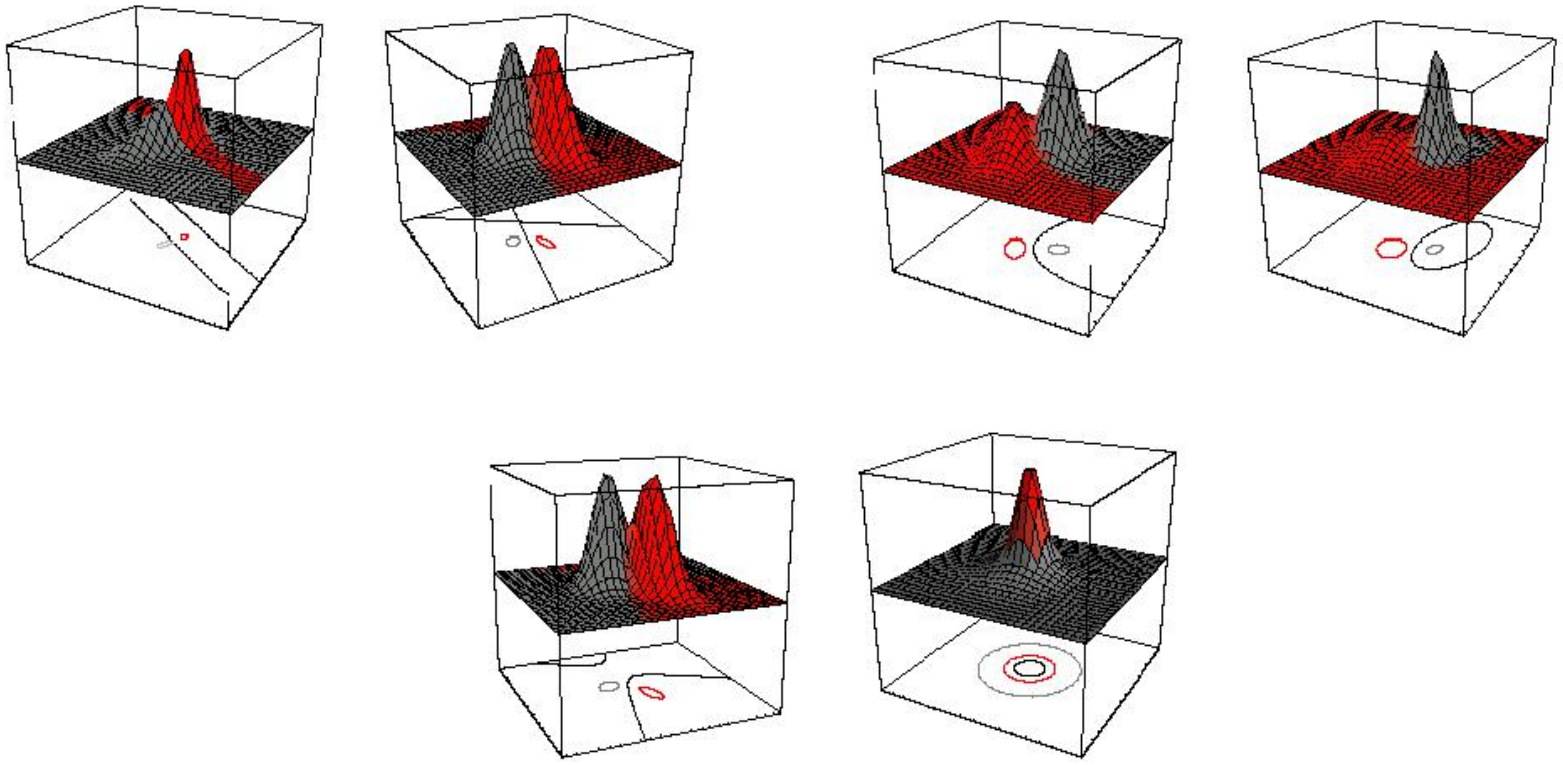
Quadratic discriminant function

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

General Cases

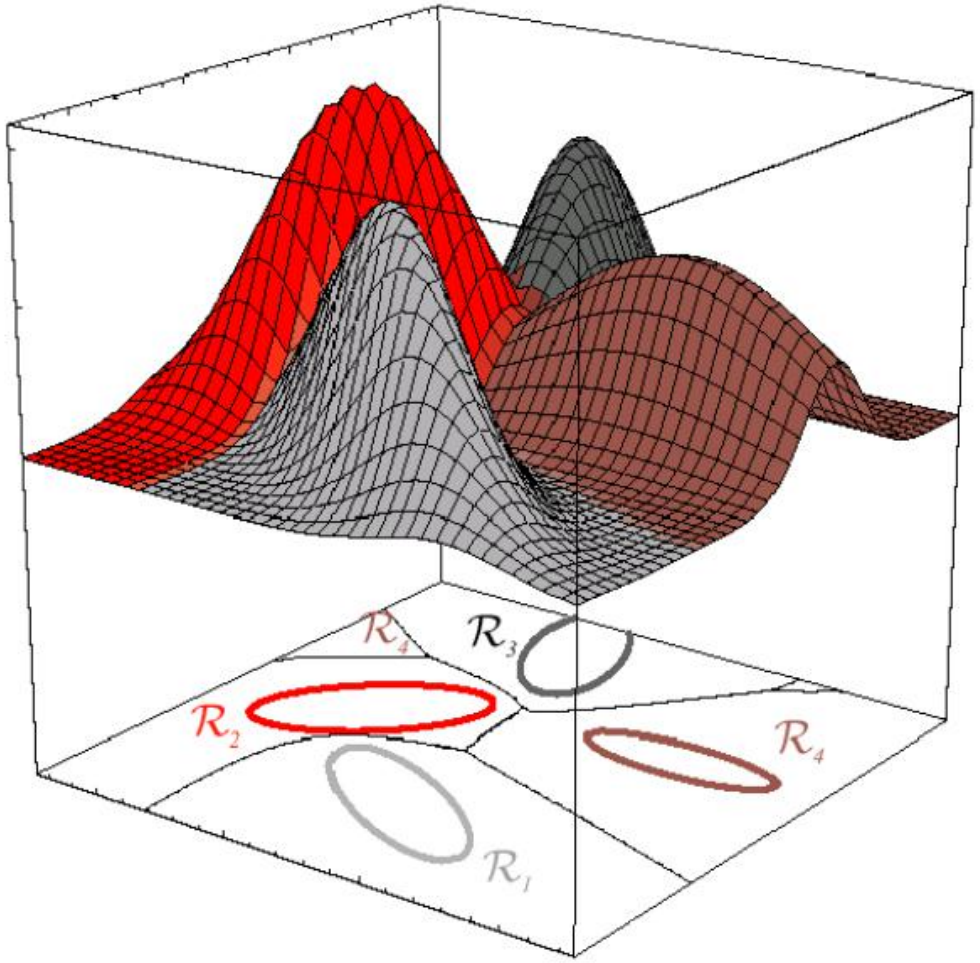
- Random Gaussian density function

  In two-class classification，desicion plane is a superquadratic surface



General Cases

- Random Gaussian density function （multi-class classification）



Example

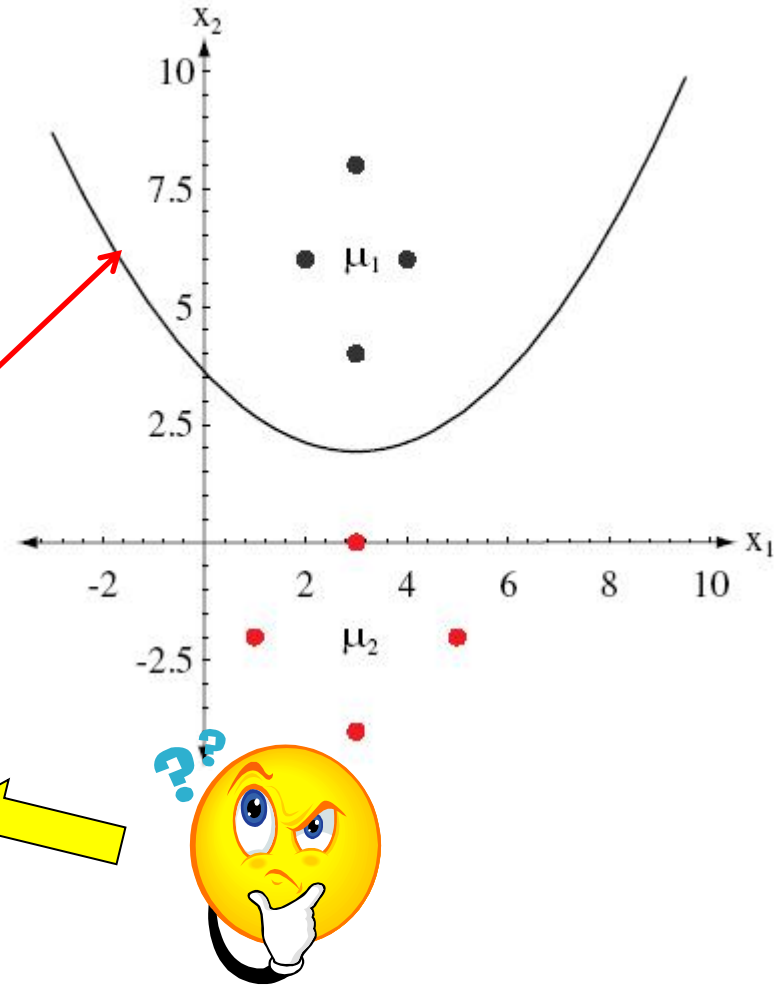
$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

$$P(\omega_1) = P(\omega_2) = 0.5$$

$$\Sigma_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad \text{and} \quad \Sigma_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

Decision boundary

$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2.$$



Decision boundary does not pass through the mid-point $\begin{pmatrix} 3 \\ 2 \end{pmatrix}$ of μ1 and μ2, but is a lower point $\begin{pmatrix} 3 \\ 1.83 \end{pmatrix}$