



**Thomas Bayes**  
1701–1761

Thomas Bayes was born in Tunbridge Wells and was a clergyman as well as an amateur scientist and a mathematician. He studied logic and theology at Edinburgh University and was elected Fellow of the Royal Society in 1742. During the 18<sup>th</sup> century, issues regarding probability arose in connection with

gambling and with the new concept of insurance. One particularly important problem concerned so-called inverse probability. A solution was proposed by Thomas Bayes in his paper 'Essay towards solving a problem in the doctrine of chances', which was published in 1764, some three years after his death, in the *Philosophical Transactions of the Royal Society*. In fact, Bayes only formulated his theory for the case of a uniform prior, and it was Pierre-Simon Laplace who independently rediscovered the theory in general form and who demonstrated its broad applicability.

# Bayesian Decision Theory

chap 2

# Introduction

---

## Bayesian Decision Theory

- ▶ A fundamental **statistical** approach to the problem of pattern classification
  - ▶ based on quantifying the trade offs between various classification decisions using probability and the costs that accompany such decisions
  - ▶ Fundamentals of the theory will be developed in the chapter.
- ▶ State of nature,  $\omega$ 
  - ▶ In the example
    - ▶  $\omega = \omega_1$ , for sea bass and  $\omega = \omega_2$  for salmon
    - ▶ the state of nature is unpredictable
      - $\omega$  to be a variable that must be described probabilistically

# Introduction

---

## ▶ A priori probability

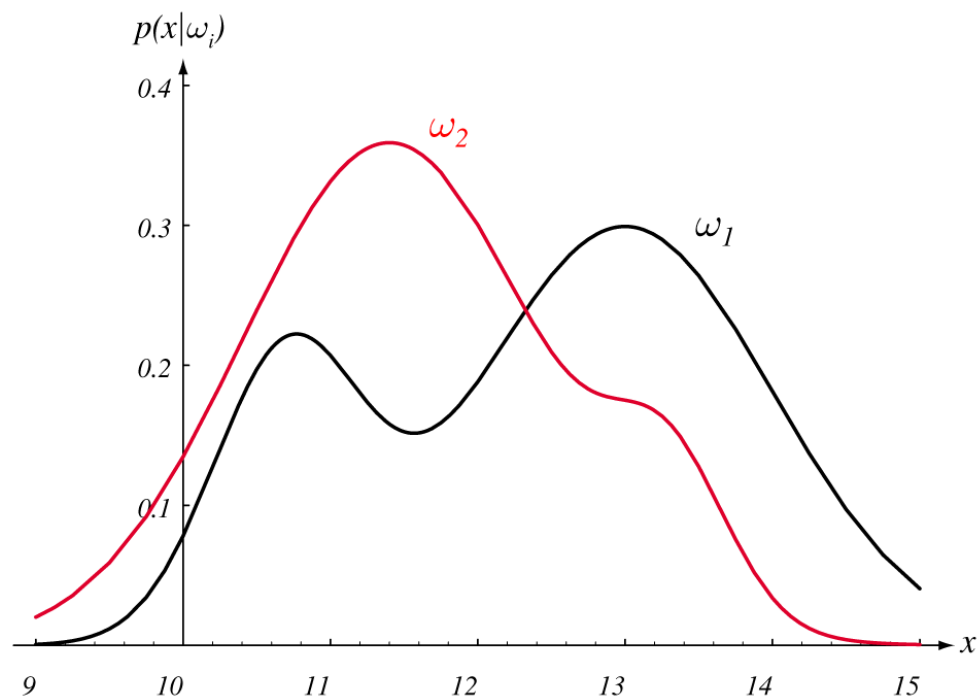
- ▶ prior knowledge of how likely we are to get a sea bass or salmon *before* the fish actually appear
  - ▶  $P(\omega_1)$  : the next fish is sea bass
  - ▶  $P(\omega_2)$  : the next fish is salmon
  - ▶  $P(\omega_1) + P(\omega_2) = 1$  if no other types of fish

## ▶ Decision Rule

- ▶ Without being allowed to see the fish,
  - ▶ Decide  $\omega_1$  if  $P(\omega_1) > P(\omega_2)$ ; otherwise decide  $\omega_2$

# Introduction

- ▶ **Class-conditional probability density function,  $p(x|\omega)$** 
  - ▶ the probability density function for  $x$  given  $\omega$
  - ▶ the difference between  $p(x|\omega_1)$  and  $p(x|\omega_2)$  describes the difference in *lightness* between populations of sea bass and salmon.




# Introduction

---


- ▶ Joint probability density of finding a pattern that is in category  $\omega_j$  and has feature value  $x$  can be written in two ways:  $p(\omega_j, x) = P(\omega_j|x)p(x) = p(x|\omega_j)P(\omega_j)$

- ▶ Bayes formula


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

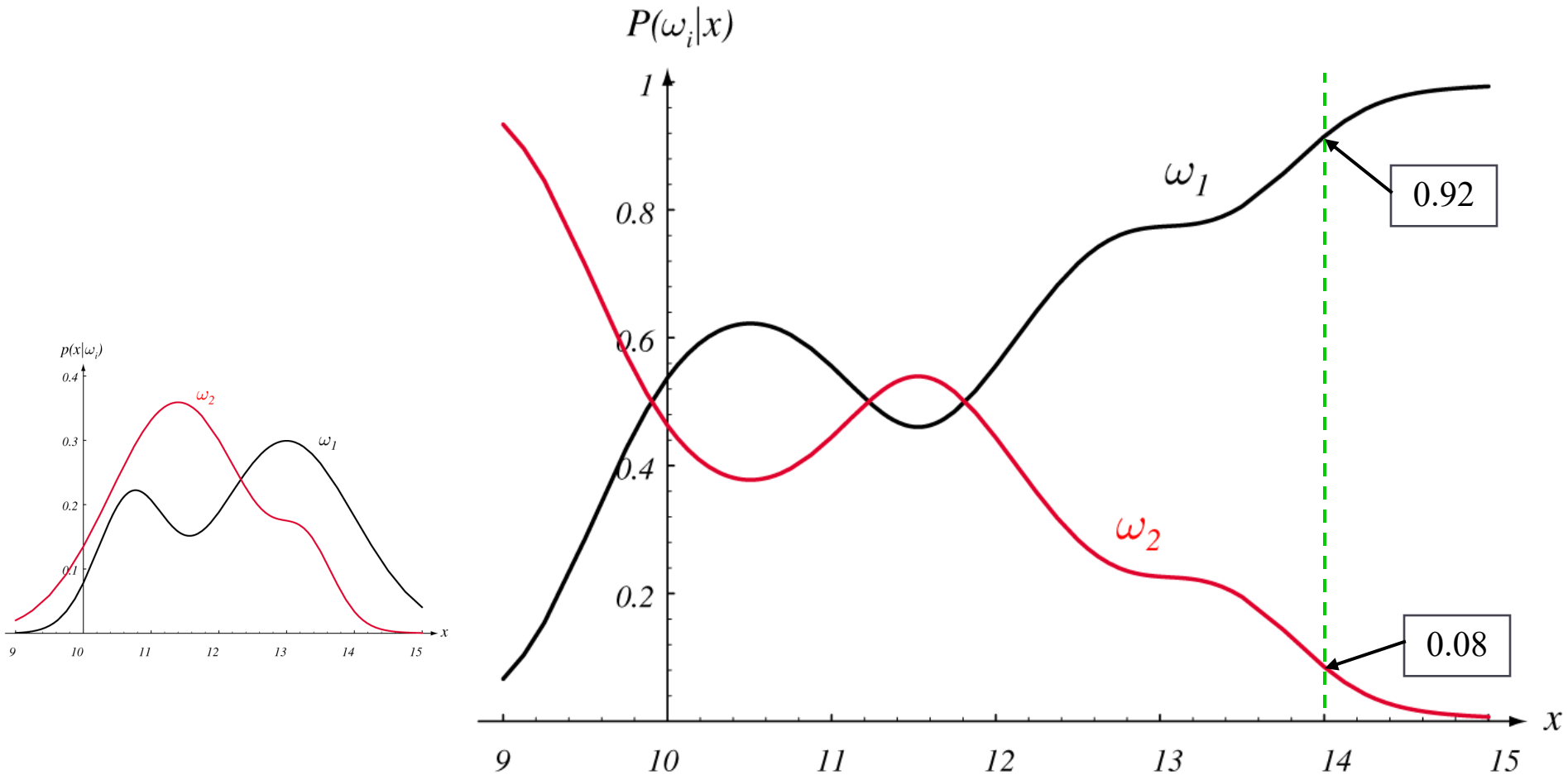
where, in case of two categories

$$p(x) = \sum_{j=1}^2 p(x|\omega_j)P(\omega_j)$$


$$posterior = \frac{likelihood \times prior}{evidence}$$

The formula converts the prior probability  $P(\omega_j)$  to the a posteriori probability  $P(\omega_j|x)$

# Introduction



Posterior probability for  $P(\omega_1)=2/3$  and  $P(\omega_2)=1/3$

# Introduction

---

- ▶ Probability of error when a decision is made

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

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

- ▶ If we ensure that  $P(error|x)$  is as small as possible, then the integral must be as small as possible.

# Introduction

---

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

- ▶ **Bayes Decision Rule** (for minimizing the probability of error)

- ▶ Decide  $\omega_1$  if  $P(\omega_1|x) > P(\omega_2|x)$ ; otherwise decide  $\omega_2$

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

- ▶  $p(x)$ : evidence is unimportant as far as making a decision is concerned.  
( $P(\omega_1|x) + P(\omega_2|x) = 1$ )

- ▶ Decide  $\omega_1$  if  $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$ ; otherwise decide  $\omega_2$

- ▶ Emphasizes the role of the posterior probabilities
  - If  $p(x|\omega_1) = p(x|\omega_2)$ : the decision is based on the prior probabilities
  - If  $P(\omega_1) = P(\omega_2)$ : the decision is based on  $p(x|\omega_j)$



# Bayesian decision theory - continuous features

## ► Generalization of the Bayesian Theory

### ► By allowing the use of more than one feature

- Replacing the scalar  $x$  by the feature vector  $\mathbf{x}$
- $\mathbf{x}$  is in  $\mathbf{R}^d$ , feature space

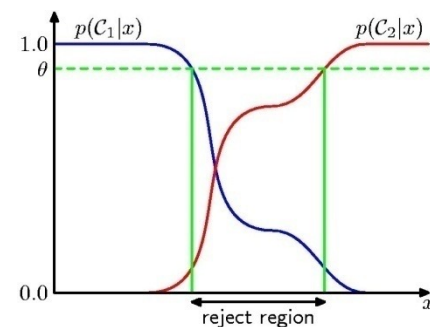
### ► By allowing more than two states of nature

$\{\omega_1, \dots, \omega_c\}$ : the finite set of  $c$  states of nature (categories)

### ► By allowing actions other than merely deciding the state of nature

- To allow the possibility of rejection

$\{\alpha_1, \dots, \alpha_a\}$ : the finite set of  $a$  possible actions



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

- Loss function state exactly how costly each action is, and is used to convert a probability determination into a decision.

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

# Bayesian decision theory - continuous features

$$P(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)P(\omega_j)}{p(\mathbf{x})} \quad p(\mathbf{x}) = \sum_{j=1}^c p(\mathbf{x}|\omega_j)P(\omega_j)$$

- ▶ The expected loss associated with taking action  $\alpha_i$  is merely

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

- ▶ In decision-theoretic terminology, an expected loss is called a *risk* and  $R(\alpha_i|\mathbf{x})$  is called the *conditional risk*.
- ▶ Problem is to find a decision rule against  $P(\omega_j)$  that minimizes the overall risk,

$$R = \int R(\alpha(\mathbf{x})|\mathbf{x})p(\mathbf{x})d\mathbf{x}.$$

Bayes risk:

$R^*$ : the minimum overall risk

- ▶ To minimize the overall risk, compute the conditional risk

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

for  $i = 1, \dots, a$  and then select the action  $\alpha_i$  for which  $R(\alpha_i|\mathbf{x})$  is minimum.

# Bayesian decision theory –

## Two-category Classification

---

- ▶ Action  $\alpha_1$  corresponds to deciding that the true state of nature is  $\omega_1$
- ▶ Action  $\alpha_2$  corresponds to deciding that the true state of nature is  $\omega_2$
- ▶  $\lambda_{ij} = \lambda(\alpha_i | \omega_j)$ : the loss incurred for deciding  $\omega_i$  when the true state of nature is  $\omega_j$ .
- ▶ The conditional risk

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

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

# Bayesian decision theory –

## Two-category Classification

---

► Variety of ways of expressing the Minimum-risk decision rule

1. decide  $\omega_1$  if  $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$ .
2. **decide**  $\omega_1$  if  $(\lambda_{21} - \lambda_{11})P(\omega_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22})P(\omega_2|\mathbf{x})$
3. decide  $\omega_1$  if  $(\lambda_{21} - \lambda_{11})p(\mathbf{x}|\omega_1)P(\omega_1) > (\lambda_{12} - \lambda_{22})p(\mathbf{x}|\omega_2)P(\omega_2)$
4. to decide  $\omega_1$  if

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

under the reasonable assumption that  $\lambda_{21} > \lambda_{11}$ .

$$\begin{aligned} R(\alpha_1|\mathbf{x}) &= \lambda_{11}P(\omega_1|\mathbf{x}) + \lambda_{12}P(\omega_2|\mathbf{x}) \\ R(\alpha_2|\mathbf{x}) &= \lambda_{21}P(\omega_1|\mathbf{x}) + \lambda_{22}P(\omega_2|\mathbf{x}) \end{aligned}$$

Likelihood ratio: the Bayes decision rule can be interpreted as calling for deciding  $\omega_1$  **if the likelihood ratio exceeds a threshold value that is independent of the observation  $\mathbf{x}$ .**

# Minimum-error-rate Classification

---

- ▶ If errors are to be avoided, it is natural to seek a decision rule that minimizes the probability of error, the *error rate*.
- ▶ Zero-one loss function

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

- ▶ No loss to a correct decision
- ▶ A unit loss to any error: equally costly
- ▶ Risk

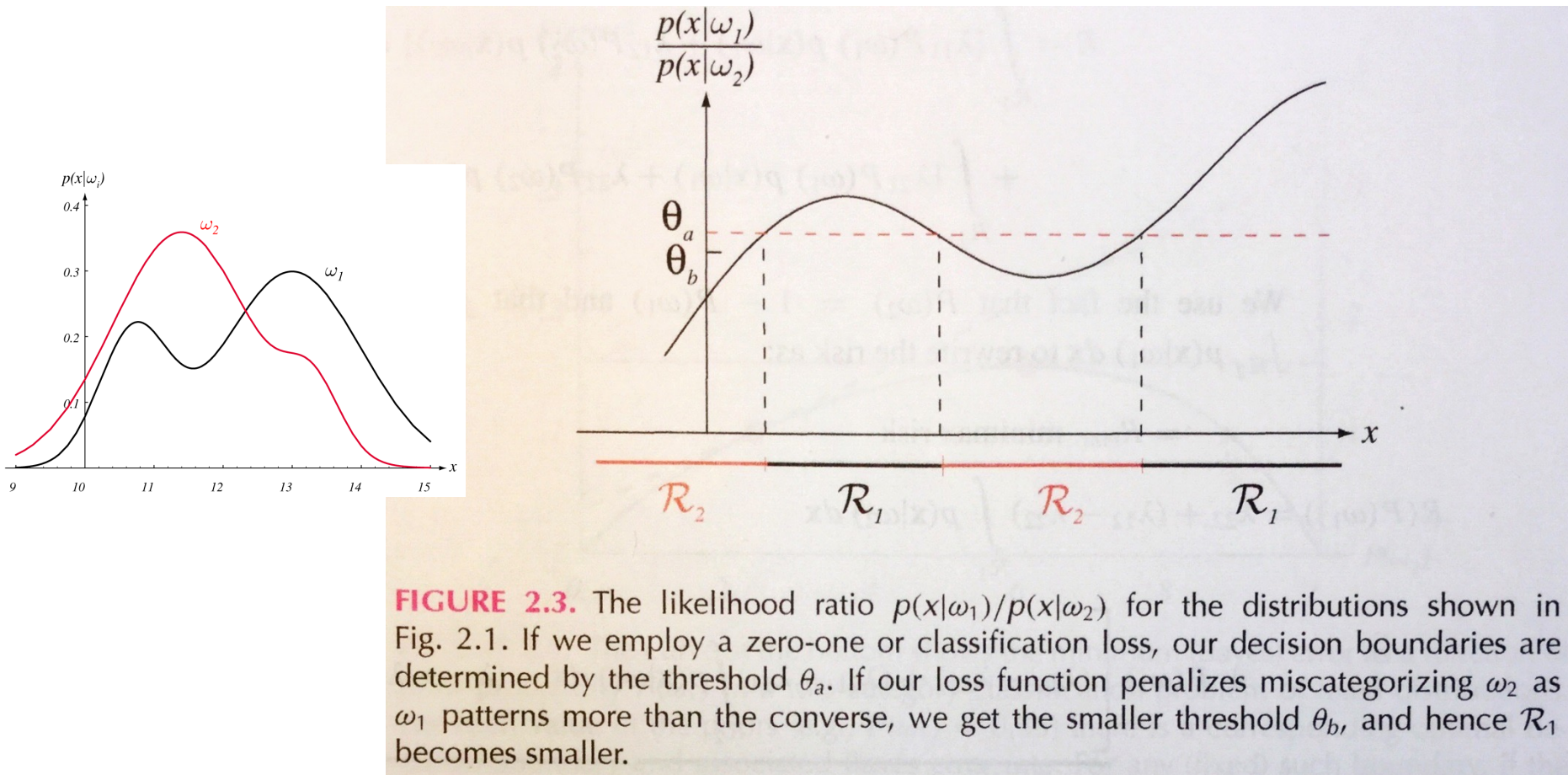
$$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{j=1}^c \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}) \\ &= \sum_{j \neq i} P(\omega_j|\mathbf{x}) \\ &= 1 - P(\omega_i|\mathbf{x}) \end{aligned}$$

For minimum error rate decide  $\omega_i$   
if  $P(\omega_i|\mathbf{x}) > P(\omega_j|\mathbf{x})$  for all  $j \neq i$ .

# Minimum-error-rate Classification

- The likelihood ratio  $p(\mathbf{x}|\omega_1)/p(\mathbf{x}|\omega_2)$

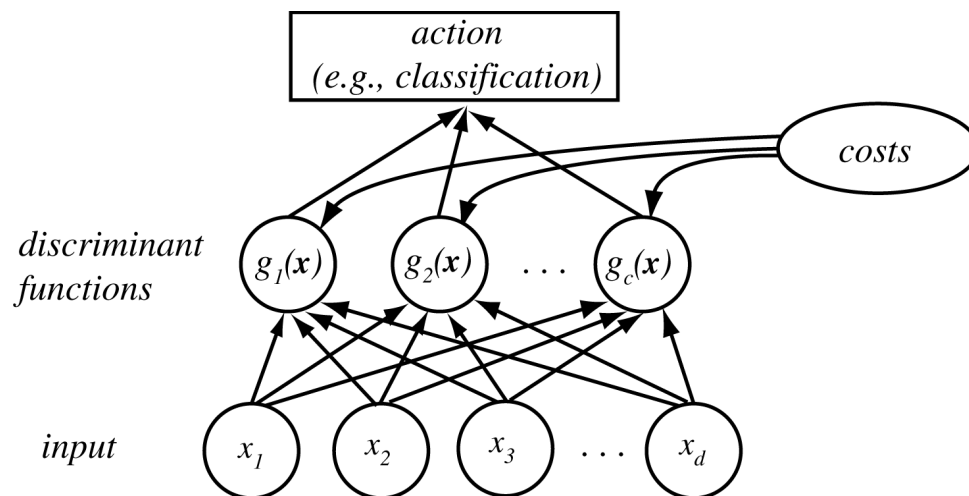
$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{\lambda_{12} - \lambda_{22} P(\omega_2)}{\lambda_{21} - \lambda_{11} P(\omega_1)}$$



# Classifiers, Discriminant functions, and Decision surfaces

## ► The multicategory case

- One of the most useful way of representing pattern classifiers is in terms of *discriminant functions*  $g_i(\mathbf{x}), i = 1, \dots, c$ .
  - A feature vector  $\mathbf{x}$  is assigned to class  $\omega_i$  if  $g_i(\mathbf{x}) > g_j(\mathbf{x})$  for all  $j \neq i$ .
- A classifier:
  - a network or machine that computes  $c$  discriminant functions and selects the category corresponding to the largest discriminant.



# Classifiers, Discriminant functions, and Decision surfaces

---

- ▶ The choice of discriminant functions is not unique.

$$g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x}) \text{ (for risk)}$$

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) \text{ (for minimum-error-rate)}$$

- ▶ Modification of the discriminant function is possible.
  - ▶ If we replace  $g_i(\mathbf{x})$  by  $f(g_i(\mathbf{x}))$ , where  $f(\cdot)$  is a monotonically increasing function, the resulting classification is unchanged.

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^c p(\mathbf{x}|\omega_j)P(\omega_j)}$$

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

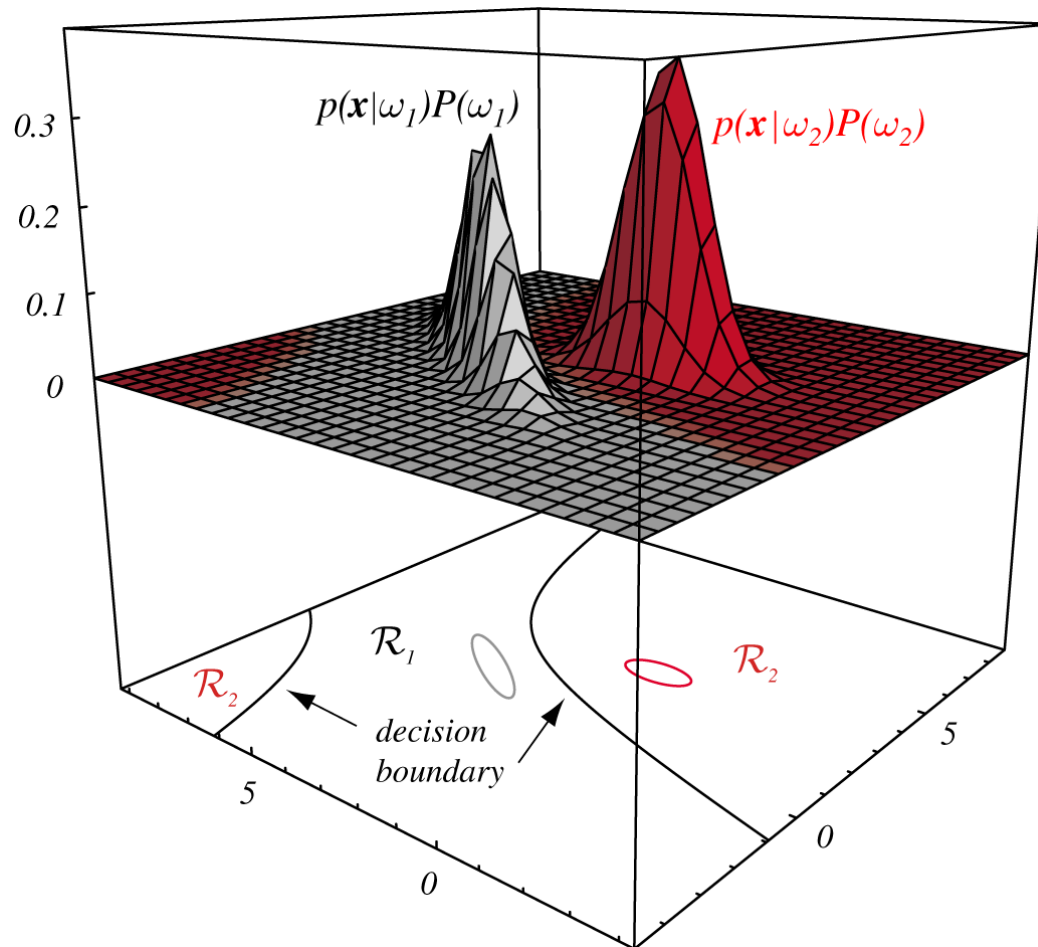
$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$$

- ▶ The effect of any decision rule is to divide the feature space into  $c$  decision regions,  $\mathcal{R}_1, \dots, \mathcal{R}_c$ . The regions are separated by decision boundaries.



# Classifiers, Discriminant functions, and Decision surfaces

---



# Classifiers, Discriminant functions, and Decision surfaces

---

## ▶ The two-category case

### ▶ Dichotomizer

- ▶ A classifier that places a pattern in one of only two categories.
- ▶ Instead of using two discriminant functions, it is more common to define a single discriminant function

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

- ▶ Decide  $\omega_1$  if  $g(\mathbf{x}) > 0$ ; otherwise  $\omega_2$ .

$$g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x})$$
$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

# The normal density

---

## ► Why Normal density?

- Due to analytical tractability, the multivariate normal or Gaussian density has received a large extent of attention.
- An appropriate model for an important situation, the case where the feature vector  $\mathbf{x}$  for a given class  $\omega_i$  are continuous-valued, randomly corrupted versions of a single prototype vector  $\boldsymbol{\mu}_i$ .

## ***Expectation (expected value)***

$$E[f(x)] \equiv \int_{-\infty}^{\infty} f(x)p(x)dx$$

If the values of the feature  $x$  are restricted to points in a discrete set  $D$

$$E[f(x)] \equiv \sum_{x \in D} f(x)P(x)$$

# The normal density - Univariate Density

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

- ▶ **mean**

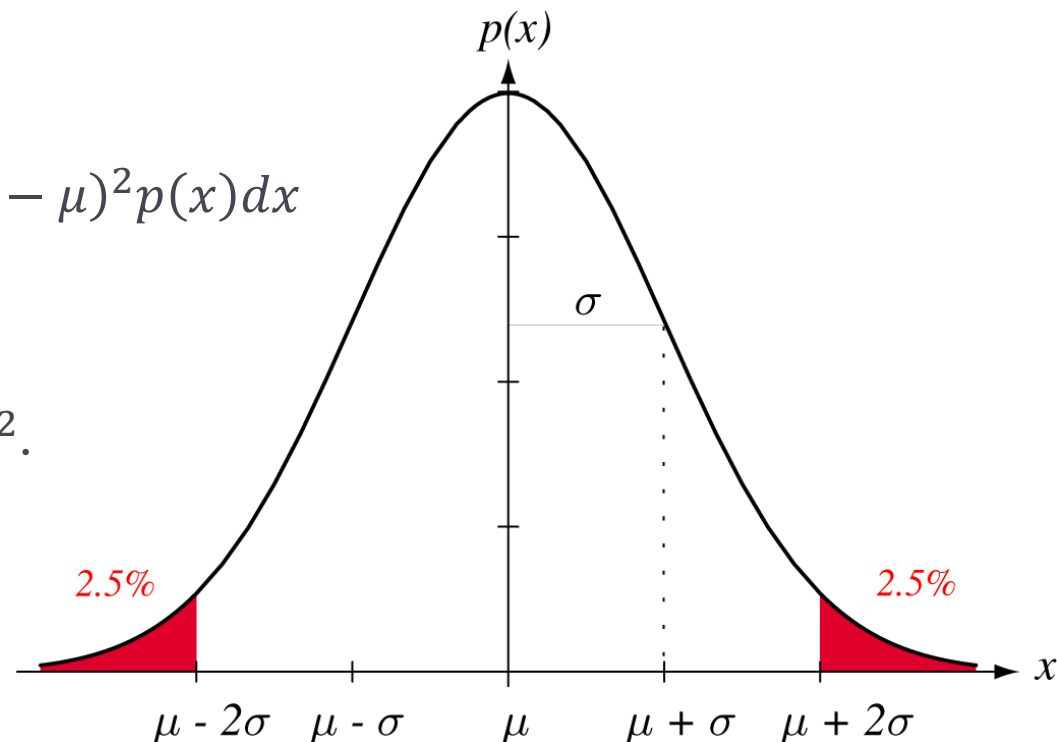
- ▶  $\mu = E[x] = \int_{-\infty}^{\infty} xp(x)dx$

- ▶ **variance**

- ▶  $\sigma^2 = E[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx$

- ▶  $p(x) \sim N(\mu, \sigma^2)$

- ▶  $x$  is distributed *normally* with mean  $\mu$  and variance  $\sigma^2$ .



# The normal density - multivariate Density

---

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

- ▶ **mean vector**

- ▶  $\boldsymbol{\mu} = E[\mathbf{x}] = \int_{-\infty}^{\infty} \mathbf{x} p(\mathbf{x}) d\mathbf{x}$

- ▶ **covariance matrix**

- ▶  $\boldsymbol{\Sigma} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t] = \int_{-\infty}^{\infty} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t p(\mathbf{x}) d\mathbf{x}$

- ▶ **Statistical independence**

- ▶ If  $x_i$  and  $x_j$  are statistically independent, then  $\sigma_{ij} = 0$ , and  $p(\mathbf{x})$  reduces to the product of the univariate normal densities for the components of  $\mathbf{x}$ .

# The normal density – multivariate Density

- ▶ Linear combination of jointly normally distributed random variables, independent or not, are normally distributed.

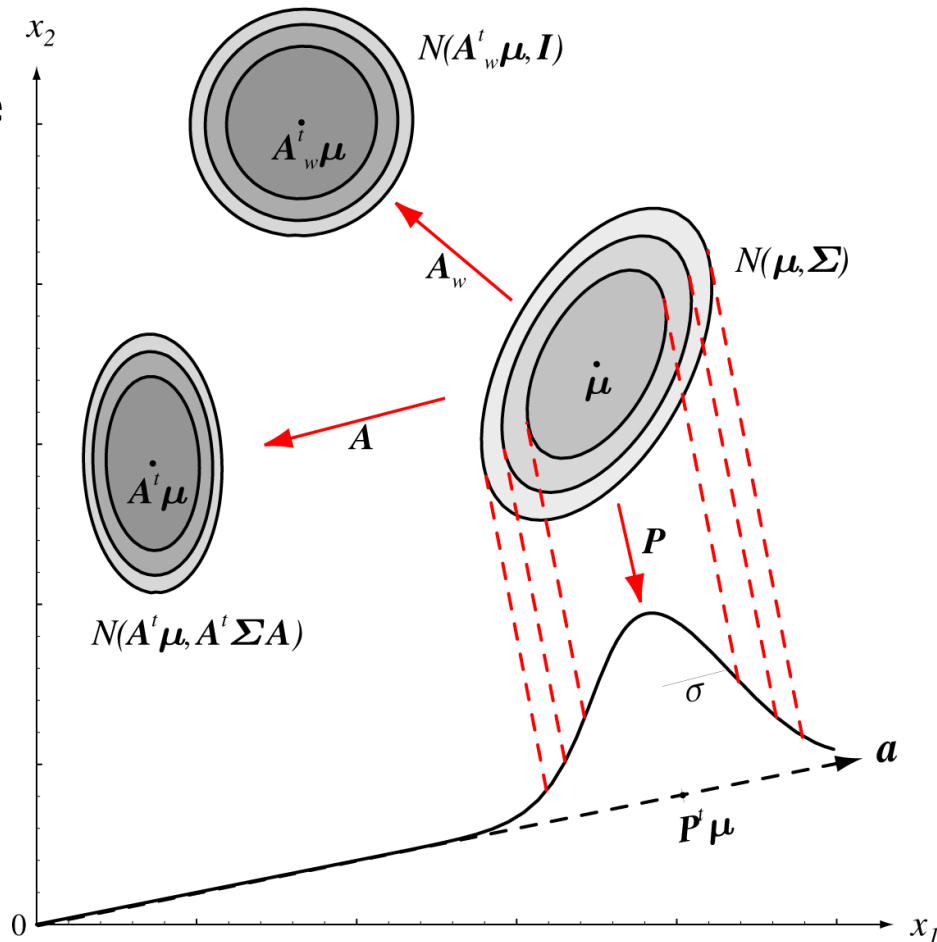
$$p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\mathbf{y} = \mathbf{A}^t \mathbf{x} \Rightarrow p(\mathbf{y}) \sim N(\mathbf{A}^t \boldsymbol{\mu}, \mathbf{A}^t \boldsymbol{\Sigma} \mathbf{A})$$

- ▶ Performing a coordinate transformation that converts an arbitrary multivariate normal distribution into a spherical one.

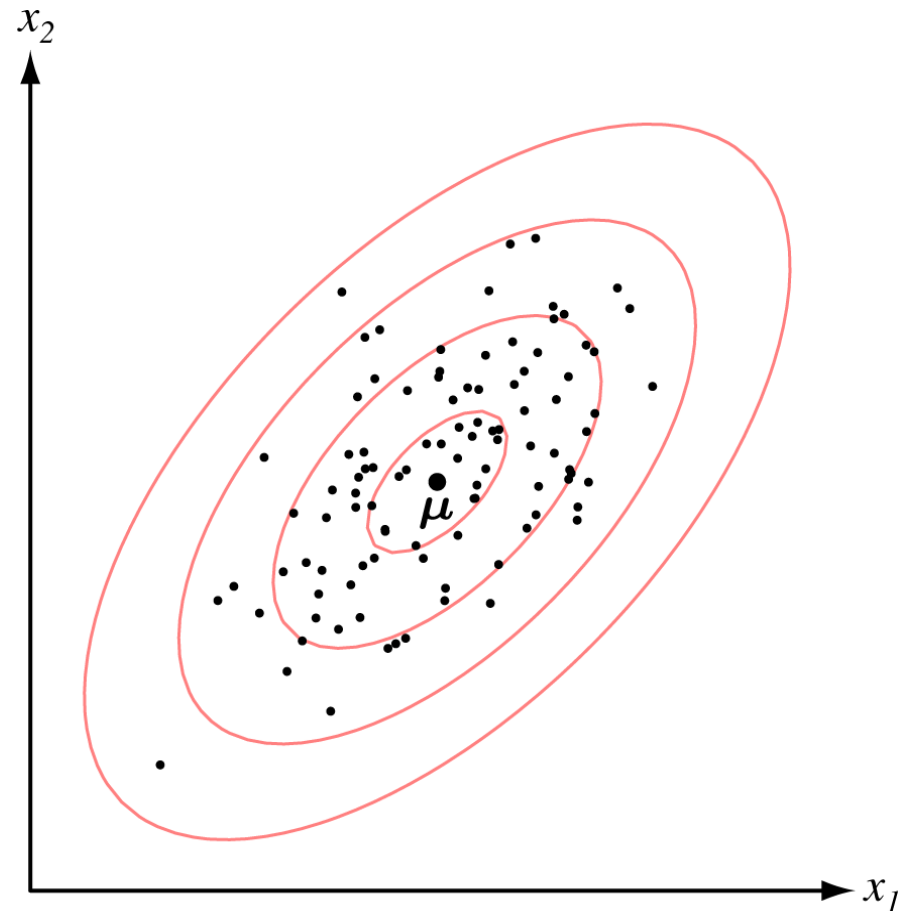
$$\mathbf{A}_\omega = \boldsymbol{\Phi} \boldsymbol{\Lambda}^{-1/2}$$

- ▶  $\boldsymbol{\Phi}$ : the matrix whose columns are the orthogonal eigenvectors of  $\boldsymbol{\Sigma}$ .
- ▶  $\boldsymbol{\Lambda}$ : the diagonal matrix of the corresponding eigenvalues.



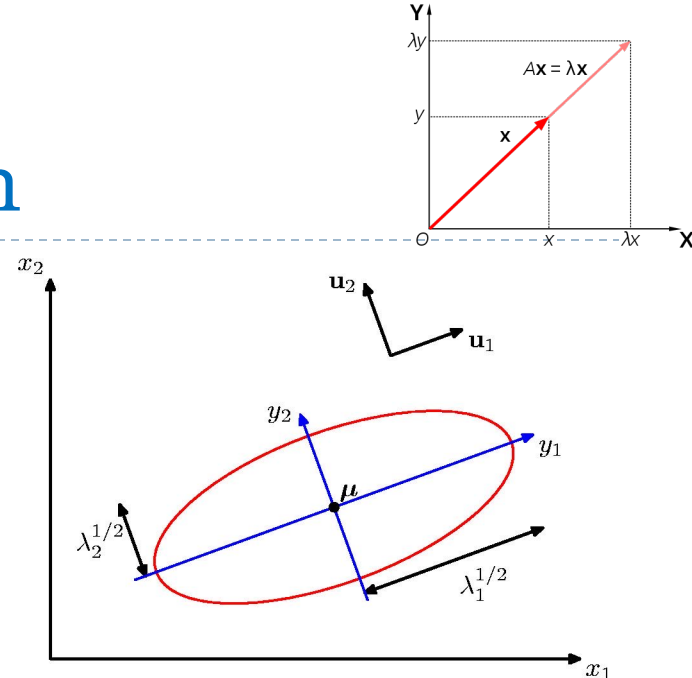
# The normal density - multivariate Density

- ▶ The multivariate normal density is completely specified by  $d + d(d + 1)/2$  parameters.
- ▶ In the figure right:
  - ▶ The center of the cluster is determined by the mean vector.
  - ▶ The shape of the cluster is determined by the covariance matrix.
  - ▶ The loci of points of constant density are hyperellipsoids for which the quadratic form  $(\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$  is constant.
  - ▶ The principal axes of these hyperellipsoids are given by the eigen vectors of  $\boldsymbol{\Sigma}$ ; the eigenvalues determine the lengths of these axes.
- ▶ Mahalanobis distance (from  $\mathbf{x}$  to  $\boldsymbol{\mu}$ )
$$r^2 = (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

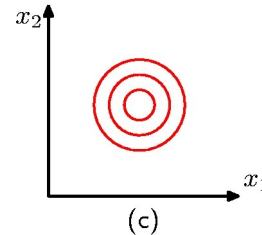
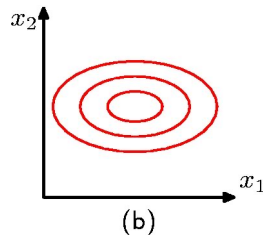
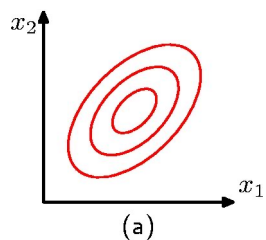


# The Gaussian Distribution

The red curve shows the elliptical surface of constant probability density for a Gaussian in a 2-dim space  $\mathbf{x} = (x_1, x_2)$ . The major axis of the ellipse are defined by the eigenvectors  $\mathbf{u}_i$  of the covariance matrix, with corresponding eigenvalues  $\lambda_i$ .



Contours of constant probability density for a Gaussian distribution in 2-dims in which the covariance matrix is (a) of general form, (b) diagonal, in which the elliptical contours are aligned with the coordinate axes, and (c) proportional to identity matrix, in which the contours are concentric circles.





# Discriminant functions for the normal density

---

- ▶ Discriminant function for the minimum-error-rate classification

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

with normal distribution

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

- ▶ Three cases of the discriminant functions:

1.  $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}$
2.  $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$
3.  $\boldsymbol{\Sigma}_i = \text{arbitrary}$

# Discriminant functions for the normal density

---

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

### ▶ The simplest case

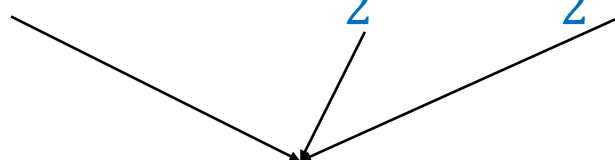
- ▶ The features are statistically independent and each feature has the same variance,  $\sigma^2$ .
- ▶ The samples fall in equal-size hyperspherical clusters.
- ▶ The clusters for the  $i$ th class is centered about the mean vector  $\mu_i$ .
- ▶ The computation of the determinant and the inverse of  $\Sigma_i$  is easy.

$$|\Sigma_i| = \sigma^{2d} \quad \Sigma_i^{-1} = \frac{1}{\sigma^2} \mathbf{I}$$

# Discriminant functions for the normal density

---

► Case 1 :  $\Sigma_i = \sigma^2 \mathbf{I}$  (cont.)

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


The diagram consists of three lines originating from the terms  $-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \Sigma_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)$ ,  $-\frac{d}{2} \ln 2\pi$ , and  $-\frac{1}{2} \ln |\Sigma_i|$  and converging to a single point below them.

independent of  $i$

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

where,  $\|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = (\mathbf{x} - \boldsymbol{\mu}_i)^t (\mathbf{x} - \boldsymbol{\mu}_i)$  is called **Euclidean norm**.

# Discriminant functions for the normal density

## ► Case 1 : $\Sigma_i = \sigma^2 \mathbf{I}$ (cont.)

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

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + \omega_{i0} \quad \text{where, } \mathbf{w}_i = \frac{1}{\sigma^2} \boldsymbol{\mu}_i \quad \text{and} \quad \omega_{i0} = \frac{-1}{2\sigma^2} \boldsymbol{\mu}_i^t \boldsymbol{\mu}_i + \ln P(\omega_i)$$

Linear discriminant  
function

threshold

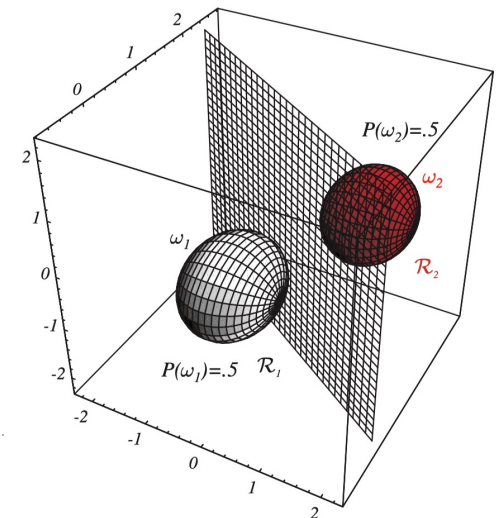
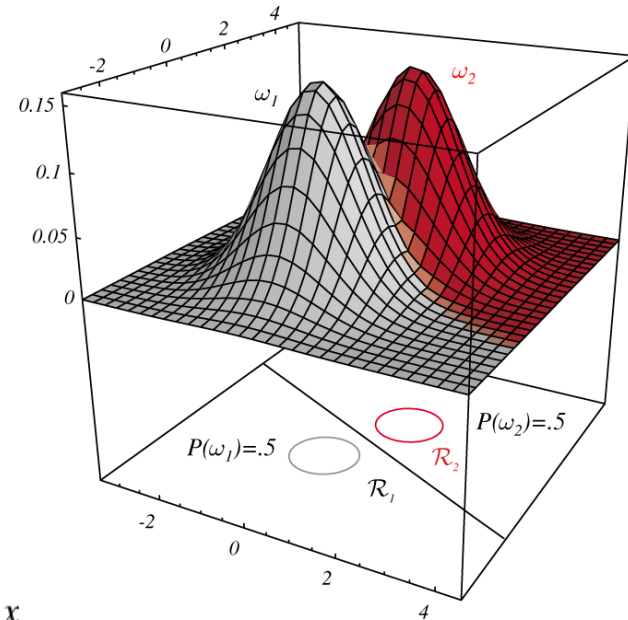
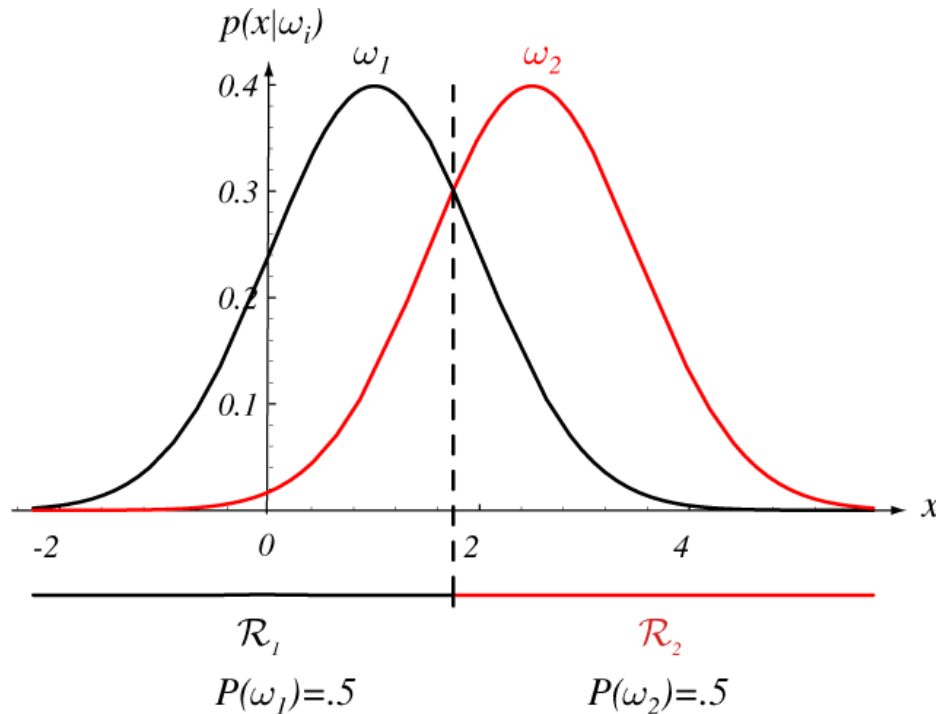
Hyperplanes defined by the linear equations  $g_i(\mathbf{x}) = g_j(\mathbf{x})$

$$\mathbf{w}^t (\mathbf{x} - \mathbf{x}_0) = 0 \quad \text{where, } \mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \quad \text{and}$$

$$\mathbf{x}_0 = \frac{1}{2} (\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

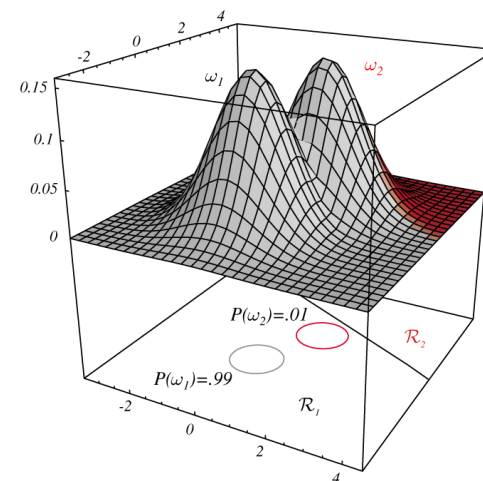
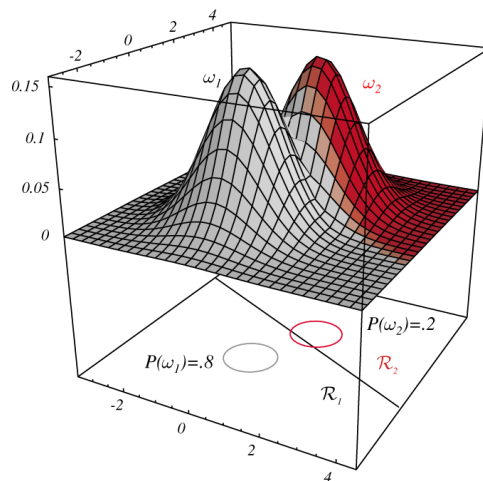
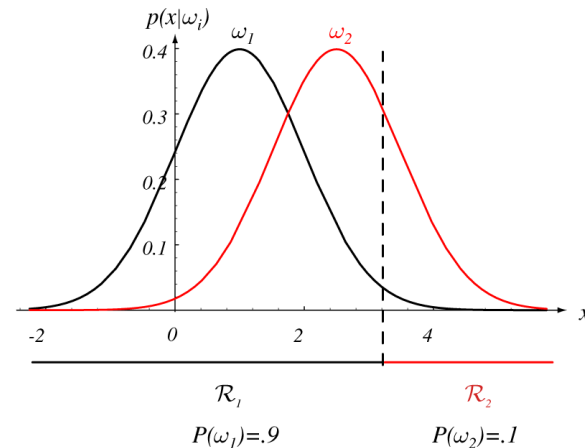
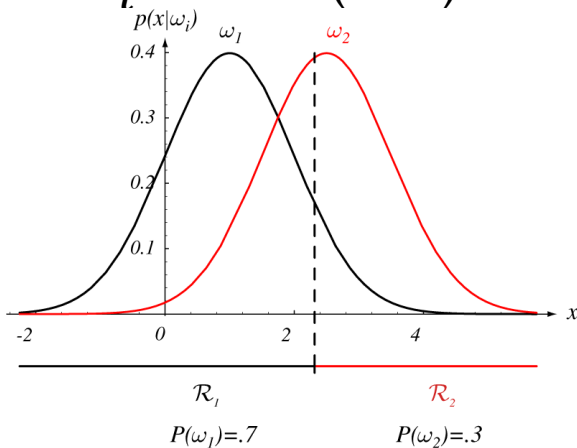
# Discriminant functions for the normal density

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



# Discriminant functions for the normal density

## ► Case 1 : $\Sigma_i = \sigma^2 \mathbf{I}$ (cont.)



# Discriminant functions for the normal density

## ▶ Case 2 : $\Sigma_i = \Sigma$

- ▶ The covariance matrices for all of the classes are identical.
  - ▶ This corresponds to the situation in which the samples fall in hyperellipsoidal clusters of **equal size and shape**.
  - ▶ The cluster for the  $i$ th class is centered about the mean vector  $\mu_i$ .

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_i)^t \Sigma_i^{-1}(\mathbf{x} - \mu_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)$$

independent of  $i$



$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_i)^t \Sigma_i^{-1}(\mathbf{x} - \mu_i) + \ln P(\omega_i)$$

# Discriminant functions for the normal density

---

- ▶ Case 2 :  $\Sigma_i = \Sigma$  (cont.)

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

- ▶ After dropping a quadratic term  $\mathbf{x}^t \Sigma^{-1} \mathbf{x}$  which is independent of  $i$ ,

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + \omega_{i0} \text{ where, } \mathbf{w}_i = \Sigma^{-1} \boldsymbol{\mu}_i \text{ and } \omega_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^t \Sigma^{-1} \boldsymbol{\mu}_i + \ln P(\omega_i)$$

The resulting decision boundaries are hyperplanes

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0 \quad \text{where,} \quad \mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \quad \text{and}$$

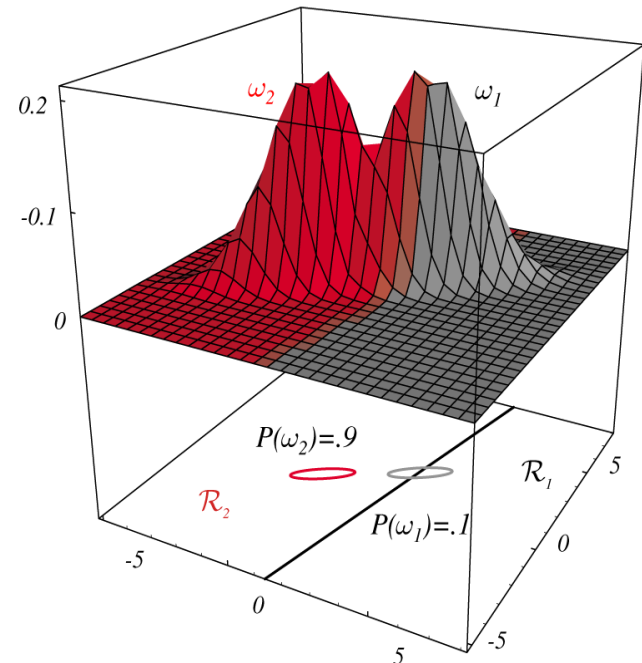
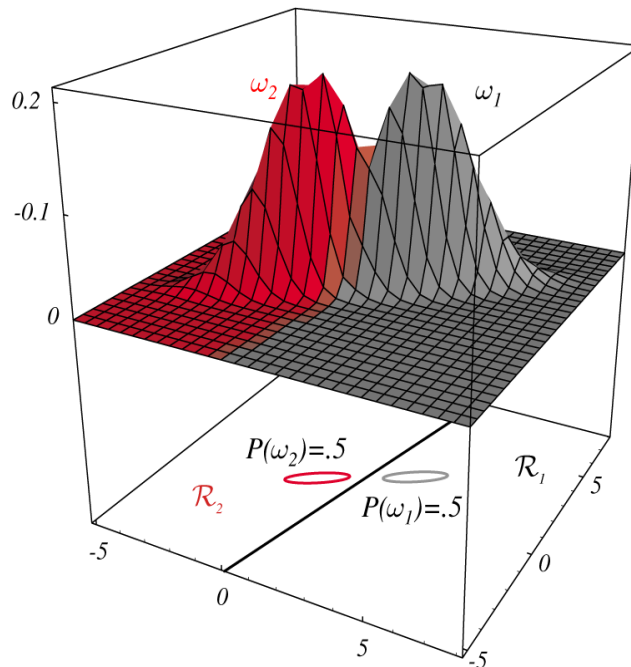
$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln[P(\omega_i)/P(\omega_j)]}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$



# Discriminant functions for the normal density

## ► Case 2 : $\Sigma_i = \Sigma$ (cont.)

Because  $\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$  is generally not in the direction of  $\boldsymbol{\mu}_i - \boldsymbol{\mu}_j$  the hyperplane separating the regions is generally not orthogonal to the line between the means.



# Discriminant functions for the normal density

## Case 3 : $\Sigma_i = \text{arbitrary}$

- The covariance matrices are different for each category.

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

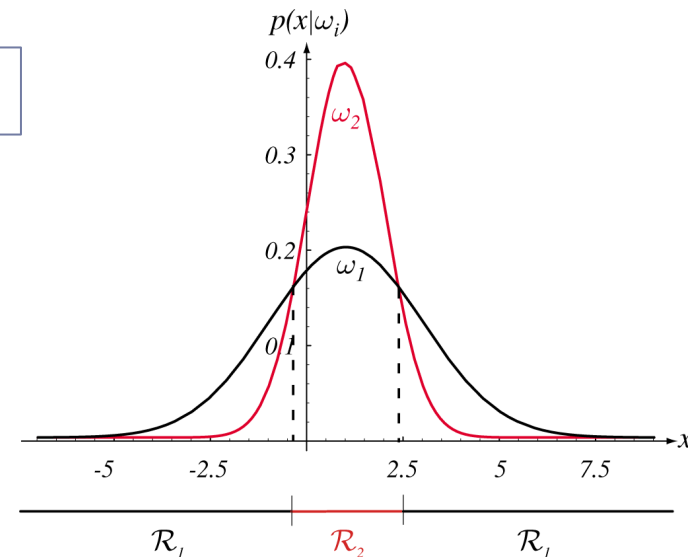
independent of  $i$



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

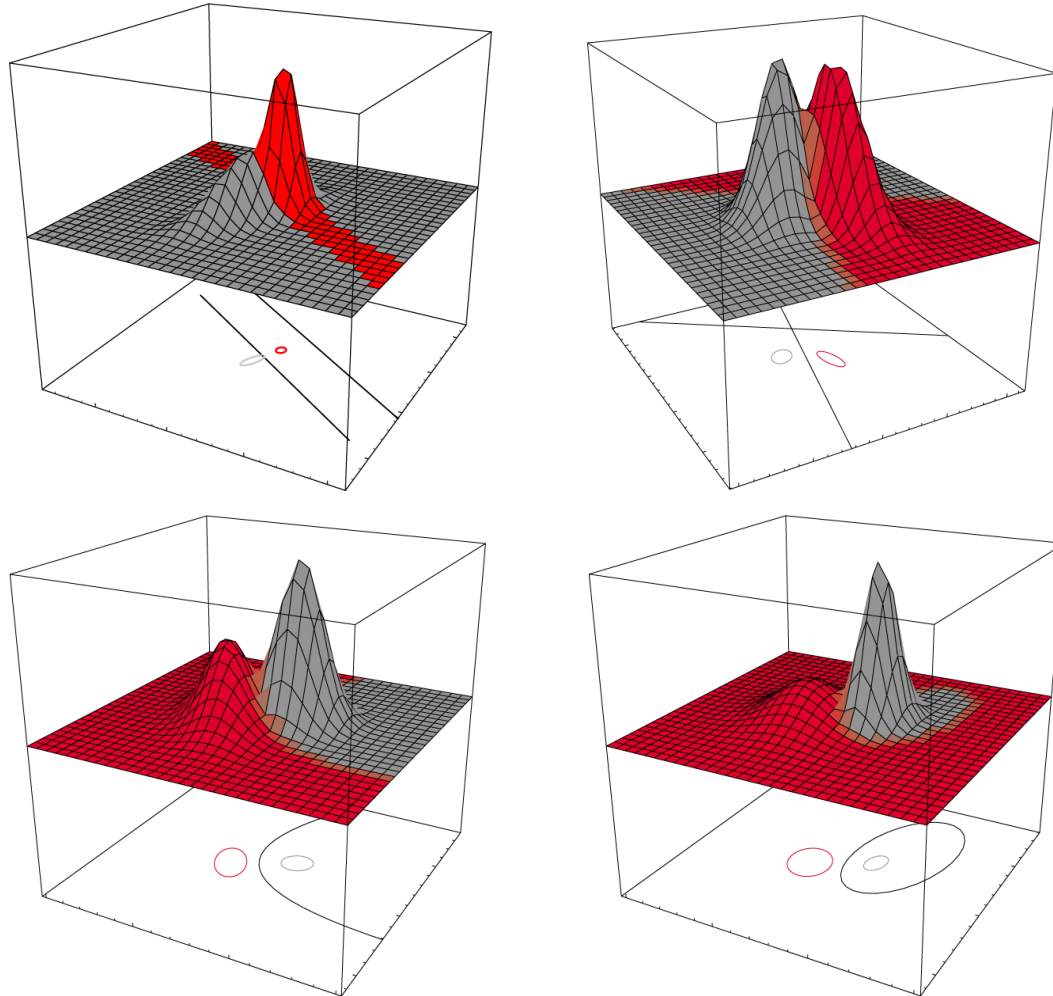
where,  $\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}$ ,  $\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$  and

$$\omega_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$



# Discriminant functions for the normal density

## ► Case 3 : $\Sigma_i = \text{arbitrary}$ (cont.)

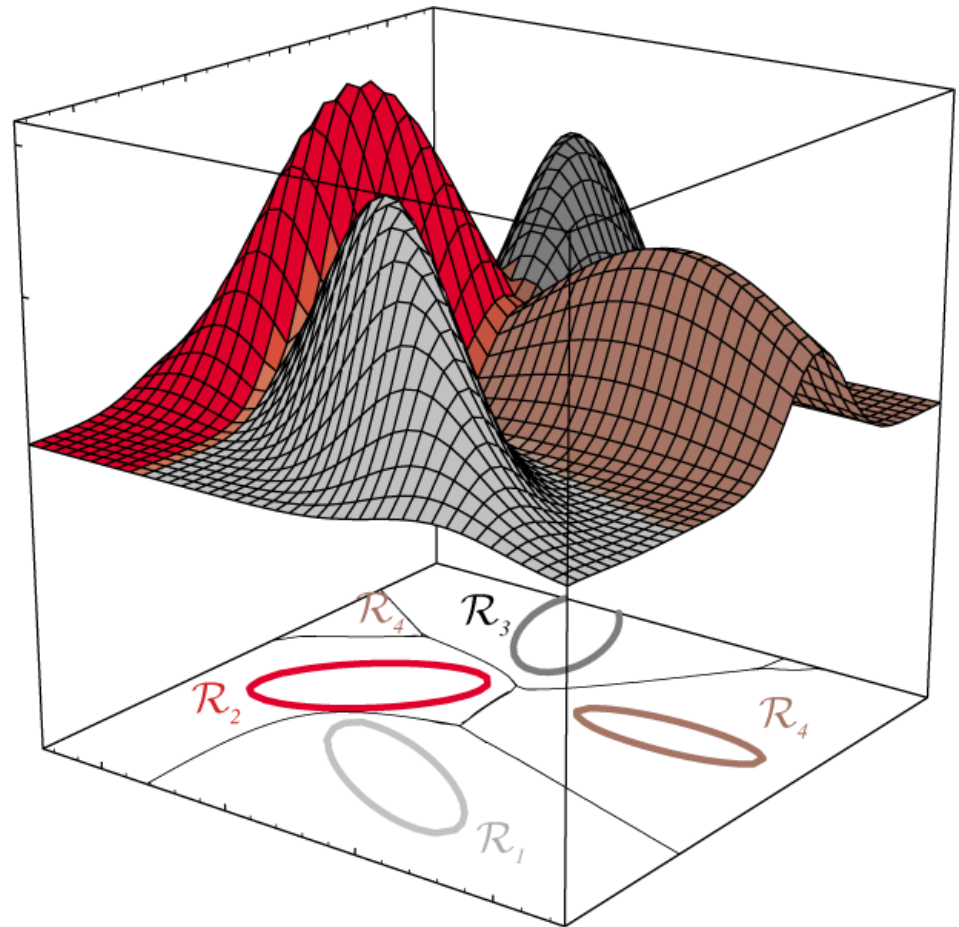


# Discriminant functions for the normal density

## ► Case 3 : $\Sigma_i = \text{arbitrary}$ (cont.)

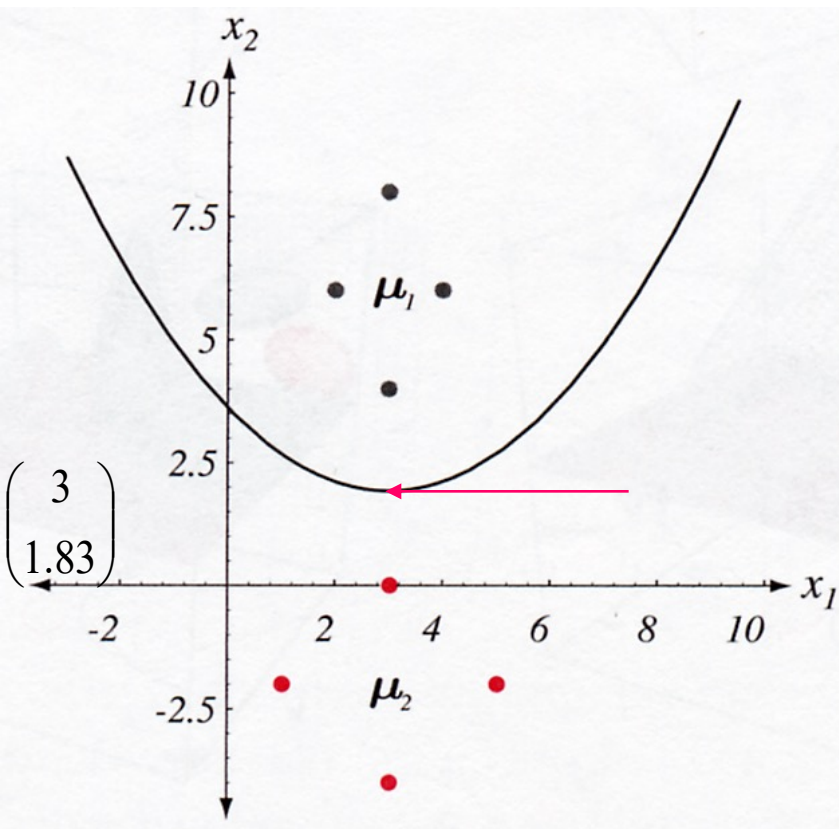
The decision regions for four normal distributions.

See example 1 at page 44.



# Discriminant functions for the normal density

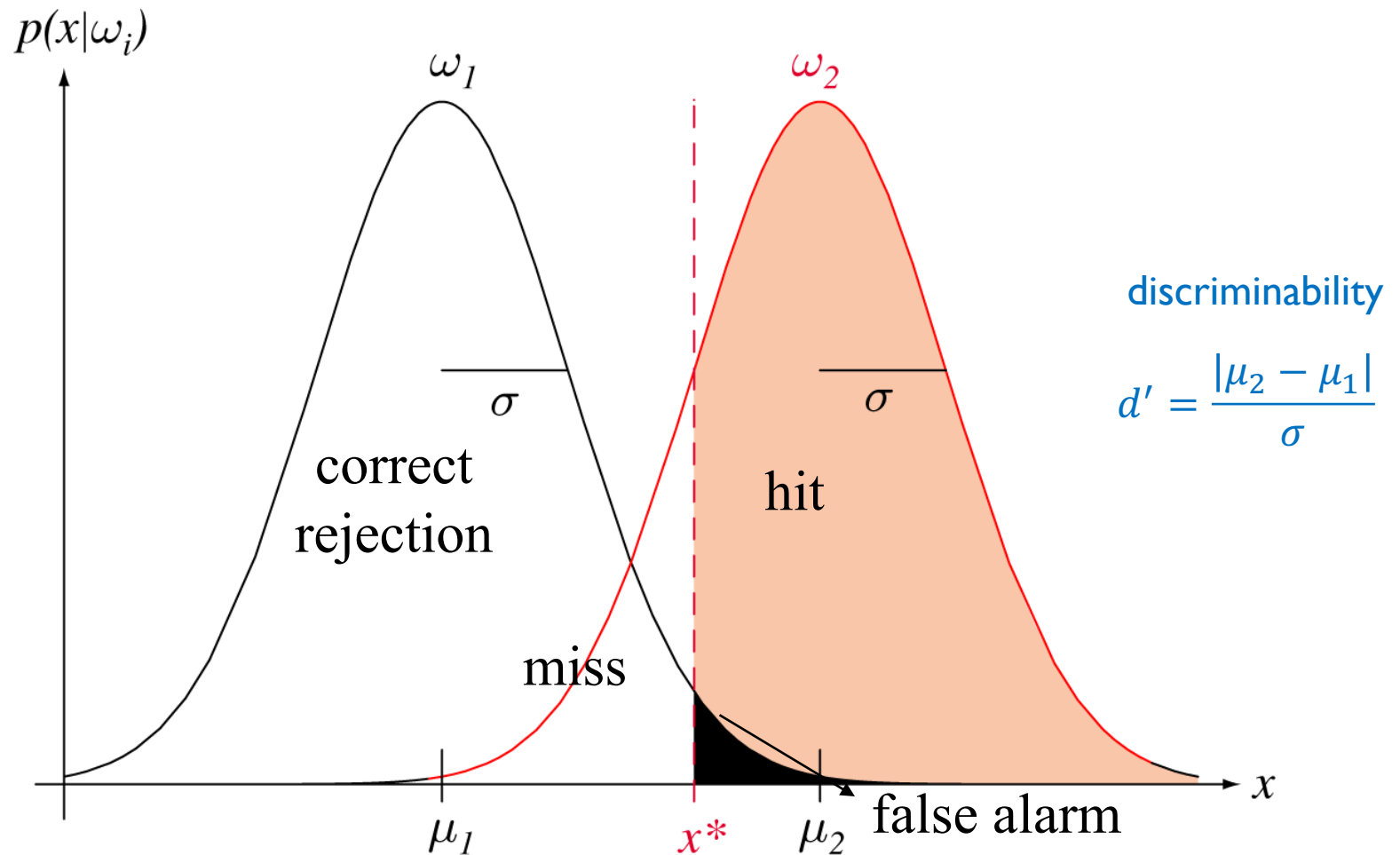
## EXAMPLE I: Decision regions for two-dimensional Gaussian data



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

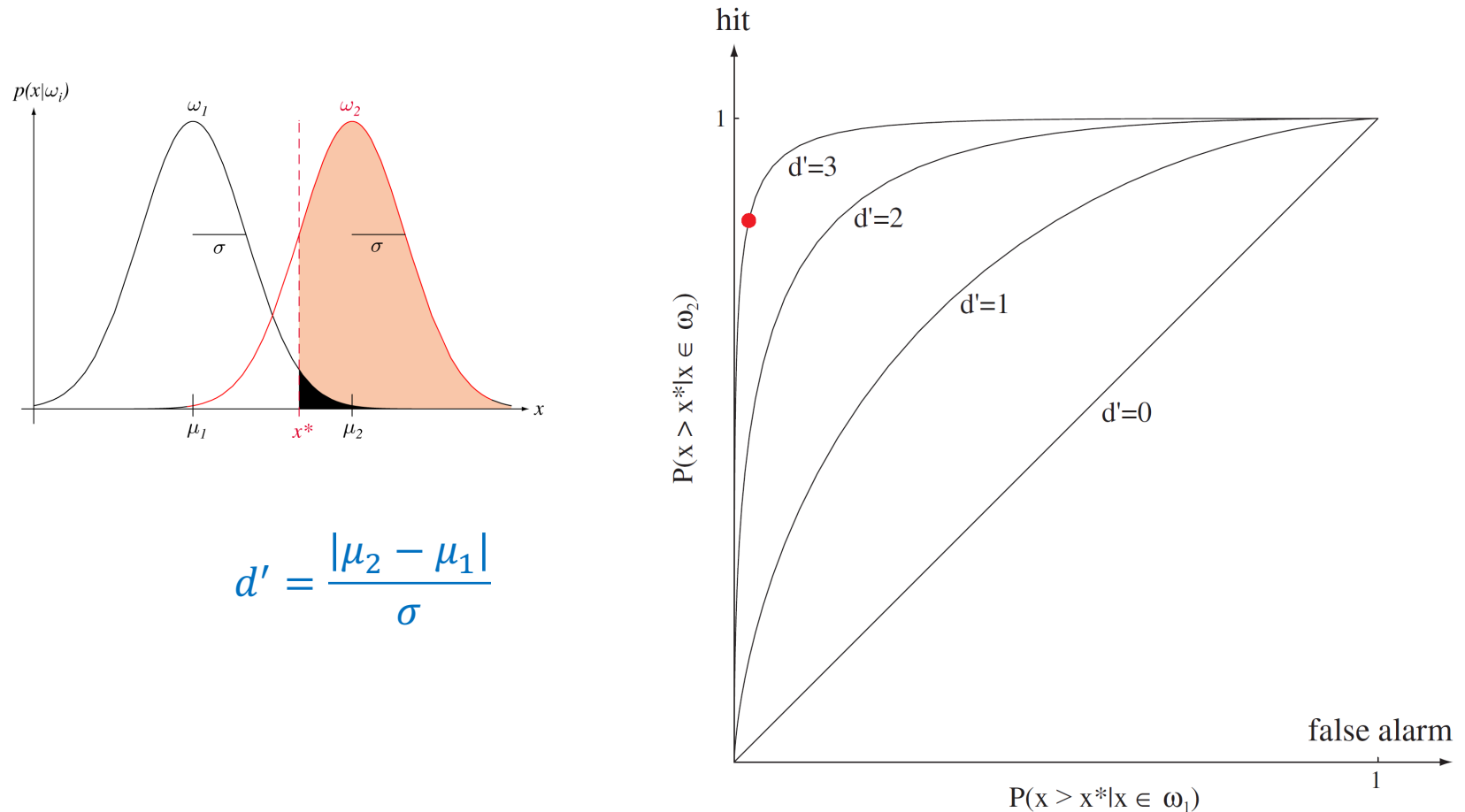
$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

# Error bounds for normal distribution



# Error bounds for normal distribution

## ► ROC(receiver operating characteristic)



# Bayes decision theory – discrete features

---

- ▶ In many applications the component of  $\mathbf{x}$  are binary-, ternary-, or higher-integer-valued, so that  $\mathbf{x}$  can assume only one of  $m$  discrete values,  $\mathbf{v}_1, \dots, \mathbf{v}_m$ .

$$\int p(\mathbf{x}|\omega_j) d\mathbf{x} \quad \longrightarrow \quad \sum_{\mathbf{x}} P(\mathbf{x}|\omega_j)$$

- ▶ Then, Bayes formula involves probability, rather than probability densities:

$$P(\omega_j|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_j)P(\omega_j)}{P(\mathbf{x})} \qquad P(\mathbf{x}) = \sum_{j=1}^c P(\mathbf{x}|\omega_j)P(\omega_j)$$

- ▶ The definition of the conditional risk is unchanged, and the fundamental Bayes decision rule remains the same.
- ▶ The basic rule to minimize error rate by maximizing the posterior probability is also unchanged.



# Bayes decision theory – discrete features

---

## Independent binary features

- ▶ Consider the two-category problem in which the components of the feature vector are binary-valued and conditionally independent.

Let  $\mathbf{x} = (x_1, \dots, x_d)^t$ , where  $x_i$  are either 0 or 1, with probabilities

$$p_i = \Pr[x_i = 1|\omega_1] \quad \text{and} \quad q_i = \Pr[x_i = 1|\omega_2]$$

➡  $P(\mathbf{x}|\omega_1) = \prod_{i=1}^d p_i^{x_i} (1 - p_i)^{1-x_i} \quad \text{and} \quad P(\mathbf{x}|\omega_2) = \prod_{i=1}^d q_i^{x_i} (1 - q_i)^{1-x_i}$

Then the likelihood ratio

$$\frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} = \prod_{i=1}^d \left(\frac{p_i}{q_i}\right)^{x_i} \left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}$$

# Bayes decision theory – discrete features

## Independent binary features

$$g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x}) \quad (30)$$

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)} \quad (31)$$

The discriminant function

$$g(x) = \sum_{i=1}^d \left[ x_i \ln \frac{p_i}{q_i} + (1 - x_i) \ln \frac{1 - p_i}{1 - q_i} \right] + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

The function is linear in the  $x_i$

$$g(\mathbf{x}) = \sum_{i=1}^d \omega_i x_i + \omega_0$$

$$\text{where, } \omega_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \text{ and } \omega_0 = \sum_{i=1}^d \ln \frac{1-p_i}{1-q_i} + \ln \left( \frac{P(\omega_1)}{P(\omega_2)} \right)$$

See example 3 at page 53.