

Facoltà di Scienze MM. FF. NN.

Università di Verona

A.A. 2010-11

**Teoria e Tecniche del  
Riconoscimento**

**Teoria della decisione di Bayes**

# Rev. Thomas Bayes, F.R.S (1702-1761)



# Introduzione

- Approccio statistico fondamentale di classificazione di pattern
- Ipotesi:
  1. Il problema di decisione è posto in termini probabilistici;
  2. Tutte le probabilità rilevanti sono conosciute;
- Goal:

Discriminare le differenti *regole di decisione* usando le *probabilità* ed i *costi* ad esse associati;

# Un esempio semplice

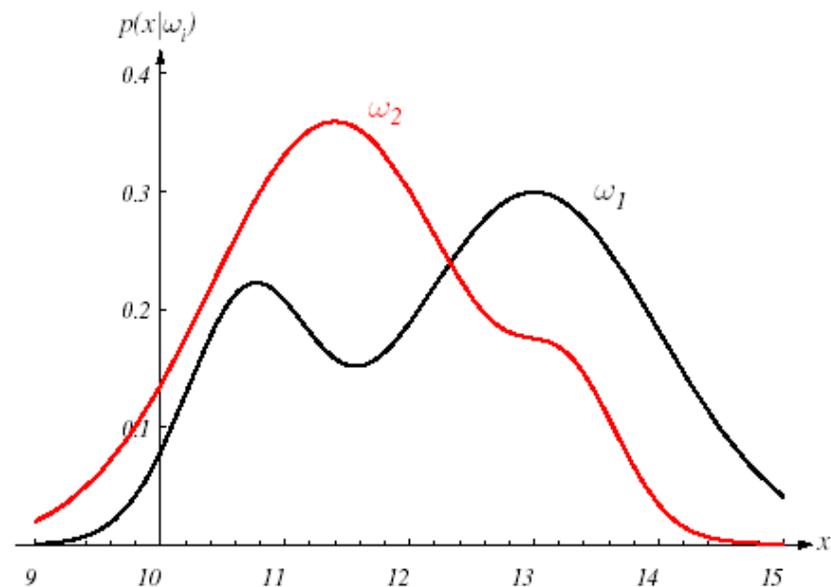
- Sia  $\omega$  lo *stato di natura* da descrivere probabilisticamente;
- Siano date:
  1. Due classi  $\omega_1$  and  $\omega_2$  per cui sono note
    - a)  $P(\omega = \omega_1) = 0.7$
    - b)  $P(\omega = \omega_2) = 0.3$**= Probabilità a priori o Prior**
  2. Nessuna misurazione.
- Regola di decisione:
  - Decidi  $\omega_1$  se  $P(\omega_1) > P(\omega_2)$ ; altrimenti decidi  $\omega_2$
- Più che decidere, *indovino* lo stato di natura.

# Altro esempio – Formula di Bayes

- Nell'ipotesi precedente, con in più la singola misurazione  $x$ , v.a. dipendente da  $\omega_j$ , posso ottenere

$$p(x | \omega_j)_{j=1,2} = \text{Likelihood, o Probabilità stato-condizionale}$$

ossia *la probabilità di avere la misurazione  $x$  sapendo che lo stato di natura è  $\omega_j$*   
Fissata la misurazione  $x$  più è alta  $p(x | \omega_j)$  più è probabile che  $\omega_j$  sia lo stato “giusto”.



# Altro esempio – Formula di Bayes (2)

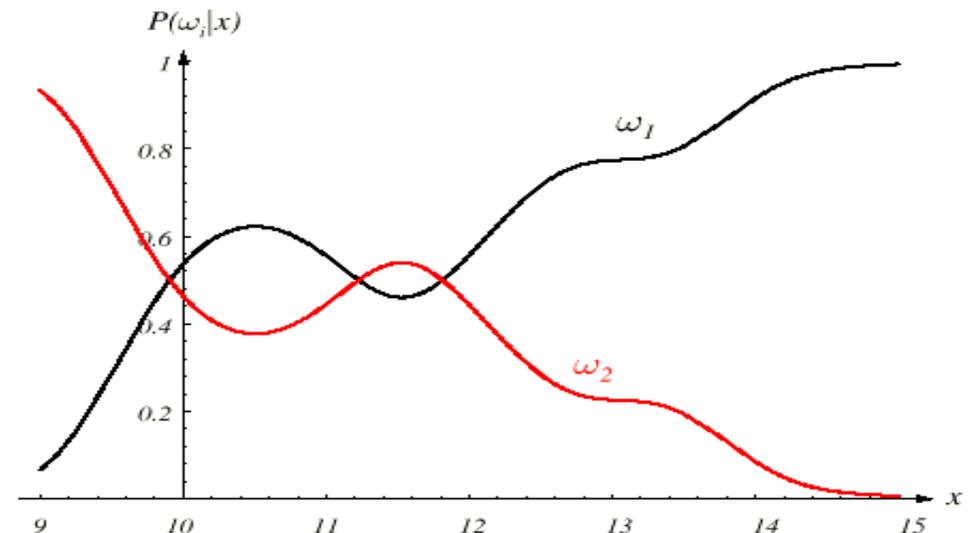
- Note  $P(\omega_j)$  e  $p(x | \omega_j)$ , la decisione dello stato di natura diventa, per Bayes

$$p(\omega_j, x) = P(\omega_j | x) p(x) = p(x | \omega_j) P(\omega_j)$$

ossia

$$P(\omega_j | x) = \frac{p(x | \omega_j) P(\omega_j)}{p(x)} \propto p(x | \omega_j) P(\omega_j) \quad , \text{dove:}$$

- $P(\omega_j)$  = Prior
- $P(x | \omega_j)$  = Likelihood
- $P(\omega_j | x)$  = **Posterior**
- $p(x) = \sum_{j=1}^J p(x | \omega_j) P(\omega_j)$   
= **Evidenza**



# Regola di decisione di Bayes

$$P(\omega_j | x) = \frac{p(x | \omega_j)P(\omega_j)}{p(x)} \iff \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

- Ossia il *Posterior* o **probabilità a posteriori** è la probabilità che lo stato di natura sia  $\omega_j$  data l'osservazione  $x$ .
- Il fattore più importante è il prodotto *likelihood*  $\times$  *prior* ;  
l'evidenza  $p(x)$  è semplicemente un fattore di scala, che assicura che

$$\sum_j P(\omega_j | x) = 1$$

- Dalla formula di Bayes deriva **la regola di decisione di Bayes:**

Decidi  $\omega_1$  se  $P(\omega_1/x) > P(\omega_2/x)$  ,  $\omega_2$  altrimenti

## Regola di decisione di Bayes (2)

- Per dimostrare l'efficacia della regola di decisione di Bayes:

1) Definisco la *probabilità d'errore* annessa a tale decisione:

$$P(\text{error} | x) = \begin{cases} P(\omega_1 | x) & \text{se decido } \omega_2 \\ P(\omega_2 | x) & \text{se decido } \omega_1 \end{cases}$$

2) Dimostro che *la regola di decisione di Bayes minimizza la probabilità d'errore*.

Decido  $\omega_1$  se  $P(\omega_1 | x) > P(\omega_2 | x)$  e viceversa.

3) Quindi se voglio *minimizzare la probabilità media di errore* su tutte le osservazioni possibili,

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

se per ogni  $x$  prendo  $P(\text{error}/x)$  più piccola possibile mi assicuro la probabilità d'errore minore (come detto il fattore  $p(x)$  è influente).

# Regola di decisione di Bayes (3)

In questo caso tale probabilità d'errore diventa

$$P(\text{error}/x) = \min[P(\omega_1/x), P(\omega_2/x)];$$

*Questo mi assicura che la regola di decisione di Bayes*

*Decidi  $\omega_1$  se  $P(\omega_1/x) > P(\omega_2/x)$ ,  $\omega_2$  altrimenti  
minimizza l'errore!*

- ***Regola di decisione equivalente:***

- La forma della regola di decisione evidenzia *l'importanza della probabilità a posteriori*, e sottolinea *l'ininfluenza dell'evidenza*, un fattore di scala che mostra quanto frequentemente si osserva un pattern  $x$ ; eliminandola, si ottiene la equivalente regola di decisione:

*Decidi  $\omega_1$  se  $p(x/\omega_1)P(\omega_1) > p(x/\omega_2)P(\omega_2)$ ,  $\omega_2$  altrimenti*

# Estensione della teoria di decisione di Bayes

- È possibile estendere l'approccio Bayesiano utilizzando:

- Più di un tipo di osservazioni o **feature**  $x$ , p.e., *peso, altezza, ...*

$$x \rightarrow \mathbf{x} = \{x_1, x_2, \dots, x_d\} \in \mathbb{R}^d \text{ con } \mathbb{R}^d \text{ spazio delle feature}$$

- Più di due stati di natura o **categorie**

$$\omega_1, \omega_2 \rightarrow \{\omega_1, \omega_2, \dots, \omega_c\}$$

- **Azioni diverse**, oltre alla scelta degli stati di natura

$$\{\alpha_1, \alpha_2, \dots, \alpha_a\}$$

- Una **funzione di costo** più generale della probabilità di errore, ossia  $\lambda(\alpha_i | \omega_j)$  che descrive il costo (o la perdita) dell'azione  $\alpha_i$  quando lo stato è  $\omega_j$ ;

## Estensione della teoria di decisione di Bayes (2)

- Le estensioni mostrate non cambiano la forma della probabilità a posteriori, che rimane:

$$P(\omega_j | \mathbf{x}) = \frac{p(\mathbf{x} | \omega_j)P(\omega_j)}{p(\mathbf{x})}, \mathbf{x} = \{x_1, x_2, \dots, x_d\}, \mathbf{x} \in \mathbb{R}^d$$

- Supponiamo di osservare un particolare  $\mathbf{x}$ , e decidiamo di effettuare l'azione  $\alpha_i$ : per definizione, saremo soggetti alla perdita  $\lambda(\alpha_i | \omega_j)$ . Data l'indeterminazione di  $\omega_j$ , la perdita attesa (o *rischio*) associata a questa decisione sarà:

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

- In questo caso la teoria di decisione di Bayes indica di effettuare l'azione che minimizza il rischio condizionale ossia, formalmente, una *funzione di decisione*  $\alpha(\mathbf{x})$  tale che:

$$\alpha(\mathbf{x}) \rightarrow \alpha_i, \alpha_i \in \{\alpha_1, \alpha_2, \dots, \alpha_a\}, \text{ tale che } R(\alpha_i | \mathbf{x}) \text{ sia minimo.}$$

## Estensione della teoria di decisione di Bayes (3)

- Per valutare una simile funzione si introduce il **Rischio complessivo**, ossia *la perdita attesa data una regola di decisione*; dato che  $R(\alpha_i | x)$  è il rischio condizionale associato all'azione e visto che la regola di decisione specifica l'azione, il rischio complessivo risulta

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

- Chiaramente, se  $\alpha(\mathbf{x})$  viene scelto in modo che  $R(\alpha_i | x)$  sia il minore possibile per ogni  $\mathbf{x}$ , il rischio complessivo viene minimizzato. Quindi la regola di decisione di Bayes estesa è:

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

2) *Scegli l'azione*  $i^* = \min_i R(\alpha_i | \mathbf{x})$

Il risultante rischio complessivo minimo prende il nome di **Rischio di Bayes**  $R^*$  ed è *la migliore performance che può essere raggiunta*.

# Problemi di classificazione a due categorie

- Consideriamo la regola di decisione di Bayes applicata ai problemi di classificazione con due stati di natura possibili  $\omega_1, \omega_2$ , con  $\alpha_i \rightarrow$  lo stato giusto è  $\omega_i$ . Per definizione,  $\lambda_{ij} = \lambda(\alpha_i | \omega_j)$

- Il rischio condizionale diventa

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

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

- Vi sono molti modi equivalenti di esprimere la regola di decisione di minimo rischio, ognuno con i propri vantaggi:

- *Forma fondamentale*: scegli  $\omega_1$  se  $R(\alpha_1 | \mathbf{x}) < R(\alpha_2 | \mathbf{x})$
- In termini di *probabilità a posteriori* scegli  $\omega_1$  se

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

## Problemi di classificazione a due categorie (2)

- Ordinariamente, la perdita per una decisione sbagliata è maggiore della perdita per una decisione giusta, pertanto

$$(\lambda_{21} - \lambda_{11}), (\lambda_{12} - \lambda_{22}) > 0$$

- Quindi, in pratica, la nostra decisione è determinata dallo stato di natura più probabile (indicato dalla probabilità a posteriori), sebbene scalato dal fattore differenza (comunque positivo) dato dalle perdite.
- Utilizzando Bayes, sostituiamo la probabilità a posteriori con

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

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

ottenendo la forma equivalente *dipendente da prior e densità condizionali*

## Problemi di classificazione a due categorie (3)

- Un'altra forma alternativa, valida per l'assunzione ragionevole che  $\lambda_{21} > \lambda_{11}$  è di decidere  $\omega_1$  se

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

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

Questa forma di regola di decisione si focalizza sulla dipendenza da  $\mathbf{x}$  delle densità di probabilità. Consideriamo  $p(\mathbf{x} | \omega_j)$  una funzione di  $\omega_j$ , cioè la funzione di likelihood, e formiamo il *likelihood ratio*, che traduce la regola di Bayes come *la scelta di  $\omega_1$  se il rapporto di likelihood supera una certa soglia*

# Classificazione *Minimum Error Rate*

- Nei problemi di classificazione ogni stato è associato ad una delle  $c$  classi  $\omega_j$  e le azioni  $\alpha_i$  significano “lo stato giusto è  $\omega_i$ ”.
- La funzione perdita associata a questo caso viene definita *di perdita 0-1* o *simmetrica*

$$\lambda(\alpha_i | \omega_j) = \begin{cases} 0 & \text{se } i = j \\ 1 & \text{se } i \neq j \end{cases}$$

- Il rischio corrispondente a questa funzione di perdita è *la probabilità media di errore*, dato che il rischio condizionale è

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

e  $P(\omega_i | \mathbf{x})$  è la probabilità che l'azione  $\alpha_i$  sia corretta.

## Classificazione Minimum Error Rate (2)

- Per minimizzare il rischio totale ossia in questo caso minimizzare la probabilità media di errore, dobbiamo scegliere  $i$  che massimizzi la probabilità a posteriori  $P(\omega_i | \mathbf{x})$ , ossia, per il *Minimum Error Rate*:

Decidi  $\omega_i$  se  $P(\omega_i/\mathbf{x}) > P(\omega_j/\mathbf{x})$  per ogni  $j \neq i$

# Riassunto

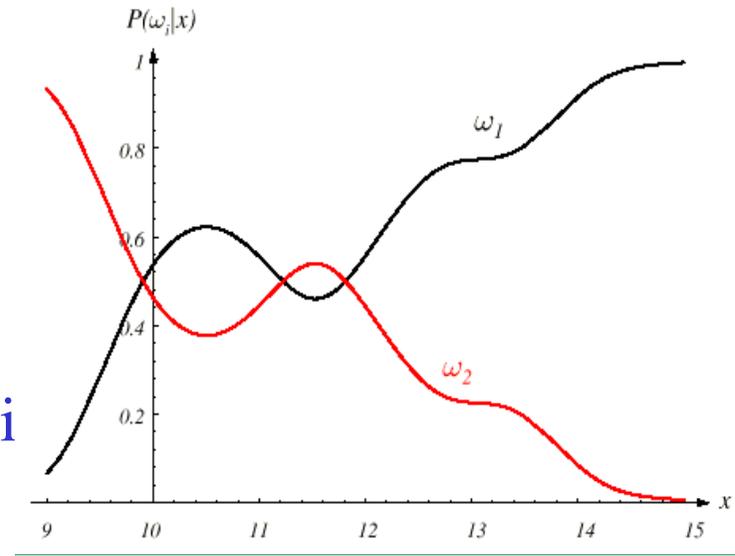
Formula di Bayes



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

Regola di decisione di Bayes:

Decidi  $\omega_1$  se  $P(\omega_1/\mathbf{x}) > P(\omega_2/\mathbf{x})$ ,  $\omega_2$  altrimenti  
 $p(\mathbf{x}/\omega_1)P(\omega_1)$      $p(\mathbf{x}/\omega_2)P(\omega_2)$



Con la **funzione di perdita**, la regola non cambia

Decidi  $\omega_1$  se  $(\lambda_{21} - \lambda_{11})p(\mathbf{x} | \omega_1)P(\omega_1) > (\lambda_{12} - \lambda_{22})p(\mathbf{x} | \omega_2)P(\omega_2)$ ,  $\omega_2$  altrimenti  
e permette di minimizzare il rischio!

Mettendo a rapporto le likelihood ho

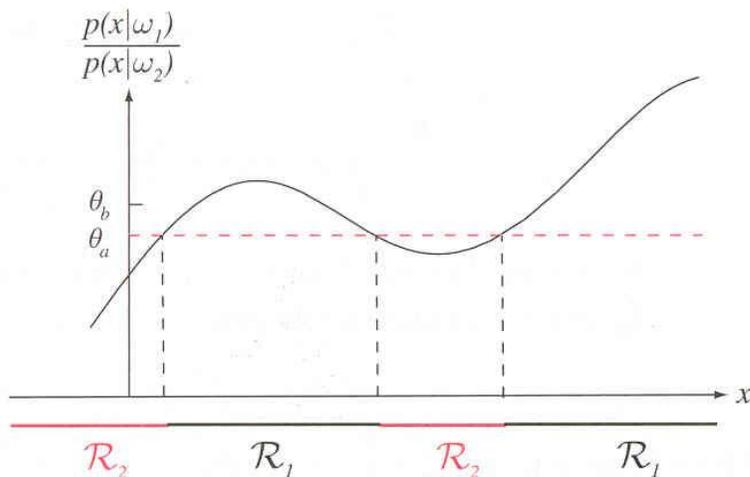
$$\frac{p(\mathbf{x} | \omega_1)}{p(\mathbf{x} | \omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

in cui può essere (**Minimum Error Rate**)

$$\lambda_{ij} = \lambda(\alpha_i | \omega_j) = \begin{cases} 0 & \text{se } i = j \\ 1 & \text{se } i \neq j \end{cases}$$



da cui mi ricollego alla regola iniziale!



# Teoria della decisione

- Quindi il problema può essere scisso in una fase di *inferenza* in cui si usano i dati per addestrare un modello  $p(\omega_k|\mathbf{x})$  e una seguente fase di *decisione*, in cui si usa la posterior per fare la scelta della classe
- Un'alternativa è quella di risolvere i 2 problemi contemporaneamente e addestrare una funzione che mappi l'input  $\mathbf{x}$  direttamente nello spazio delle decisioni, cioè delle classi  $\rightarrow$  *funzioni discriminanti*
- Esistono 3 approcci per risolvere il problema della decisione (in ordine decrescente di complessità):
  1. Risolvere prima il problema di inferenza per determinare le densità class-conditional per ogni singola classe, inferire anche i prior e quindi usare Bayes per trovare la posterior e quindi determinare la classe (sulla base della teoria della decisione)

**$\rightarrow$  Modelli generativi**

2. Risolvere prima il problema di inferenza per determinare *direttamente* la posterior e quindi usare la teoria della decisione per decidere la classe

➔ **Modelli discriminativi**

3. Trovare una funzione  $f(\mathbf{x})$ , chiamata *funzione discriminante*, che mappi l'input  $\mathbf{x}$  direttamente nell'etichetta di una classe

- Tuttavia, stimare la posterior è molte volte utile in quanto:
  - si combinano i modelli nel caso in cui un problema complesso debba essere suddiviso in problemi più semplici e quindi “fondere” i risultati (naive Bayes sotto l’ipotesi di indipendenza condizionale)

$$\begin{aligned}
 p(\omega_j | \mathbf{x}_A, \mathbf{x}_B) &\propto p(\mathbf{x}_A, \mathbf{x}_B | \omega_j) p(\omega_j) \\
 &\propto p(\mathbf{x}_A | \omega_j) p(\mathbf{x}_B | \omega_j) p(\omega_j) \quad \textit{naive Bayes} \\
 &\propto \frac{p(\omega_j | \mathbf{x}_A) p(\omega_j | \mathbf{x}_B)}{p(\omega_j)}
 \end{aligned}$$

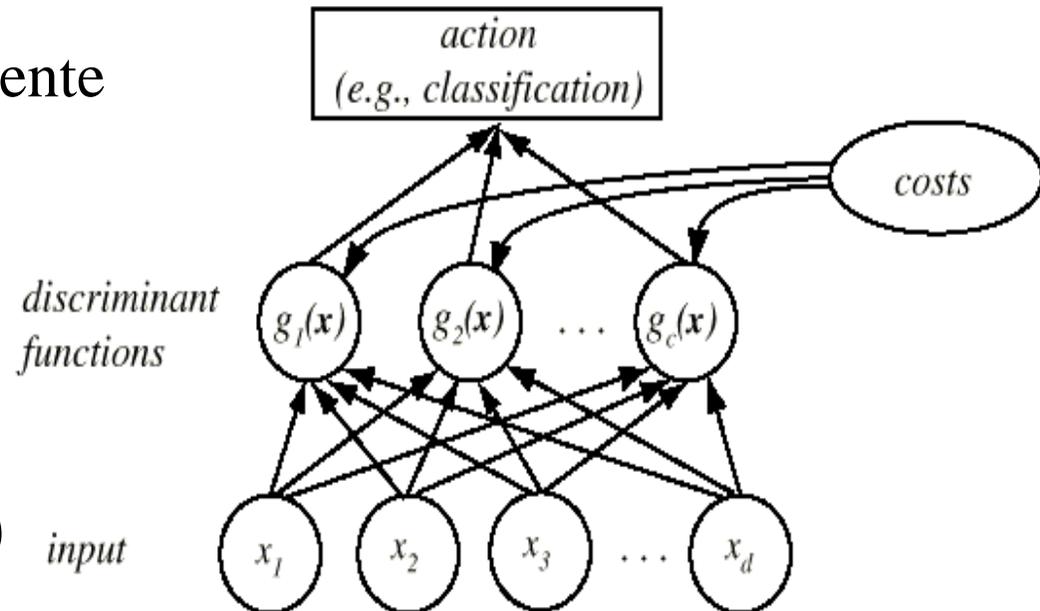
# Classificatori, funzioni discriminanti e superfici di separazione

- Uno dei vari metodi per rappresentare classificatori di pattern consiste in un set di *funzioni discriminanti*  $g_i(\mathbf{x})$ ,  $i=1\dots c$
- Il classificatore assegna il vettore di feature  $\mathbf{x}$  alla classe  $\omega_i$  se
$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \text{ per ogni } j \neq i$$
- Un tale classificatore può essere considerato come una rete che calcola  $c$  funzioni discriminanti e sceglie la funzione che discrimina maggiormente

- Un classificatore di Bayes si presta facilmente a questa rappresentazione:

*Rischio generico*  $g_i(\mathbf{x}) = -R(\alpha_i | \mathbf{x})$

*Minimum Error Rate*  $g_i(\mathbf{x}) = P(\omega_i | \mathbf{x})$  input



# Classificatori, funzioni discriminanti e superfici di separazione (2)

- Esistono molte funzioni discriminanti equivalenti. Per esempio, tutte quelle per cui i risultati di classificazione sono gli stessi
  - Per esempio, se  $f$  è una funzione monotona crescente, allora

$$g_i(\mathbf{x}) \Leftrightarrow f(g_i(\mathbf{x}))$$

- Alcune forme di funzioni discriminanti sono più semplici da capire o da calcolare

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

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

Minimum  
Error Rate



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

# Classificatori, funzioni discriminanti e superfici di separazione (3)

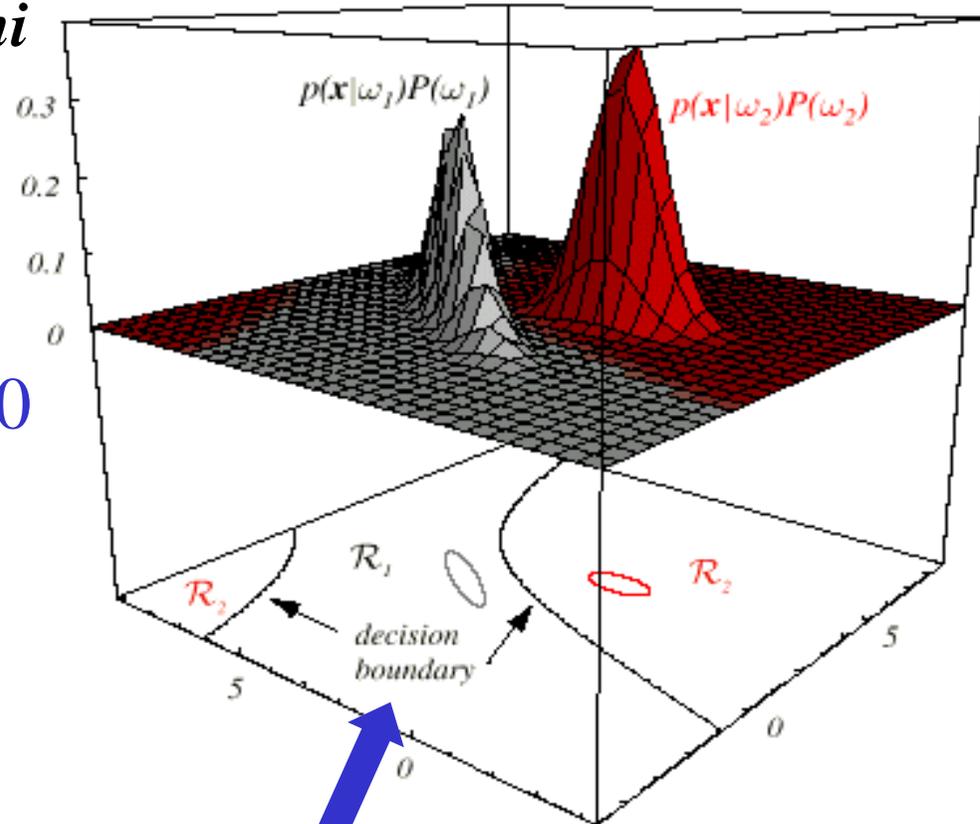
- L'effetto di ogni decisione è quello di *dividere lo spazio delle features* in *c superfici di separazione o decisione*,  $R_1, \dots, R_c$ 
  - Le regioni sono separate con *confini di decisione*, linee descritte dalle massime funzioni discriminanti.
  - Nel caso a *due* categorie ho due funzioni discriminanti,  $g_1, g_2$  per cui assegno  $\mathbf{x}$  a  $\omega_1$  se  $g_1 > g_2$  o  $g_1 - g_2 > 0$

– Usando

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

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

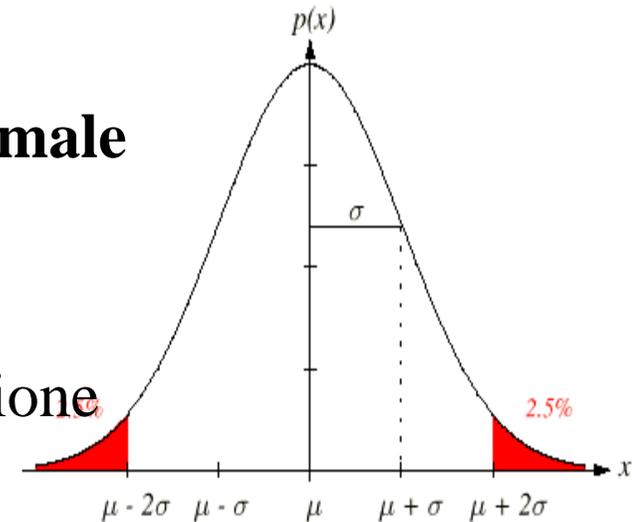
$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x} | \omega_1)}{p(\mathbf{x} | \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$



*ho una sola funzione discriminante!*

# La densità normale

- La struttura di un classificatore di Bayes è determinata da:
  - Le densità condizionali  $p(\mathbf{x} | \omega_i)$
  - Le probabilità a priori  $P(\omega_i)$
- Una delle più importanti densità è **la densità normale** o **Gaussiana multivariata**; infatti:
  - è analiticamente trattabile;
  - più importante, fornisce la migliore modellazione di problemi sia teorici che pratici
    - il teorema del Limite Centrale asserisce che “*sotto varie condizioni, la distribuzione della somma di  $d$  variabili aleatorie indipendenti tende ad un limite particolare conosciuto come distribuzione normale*”.



## La densità normale (2)

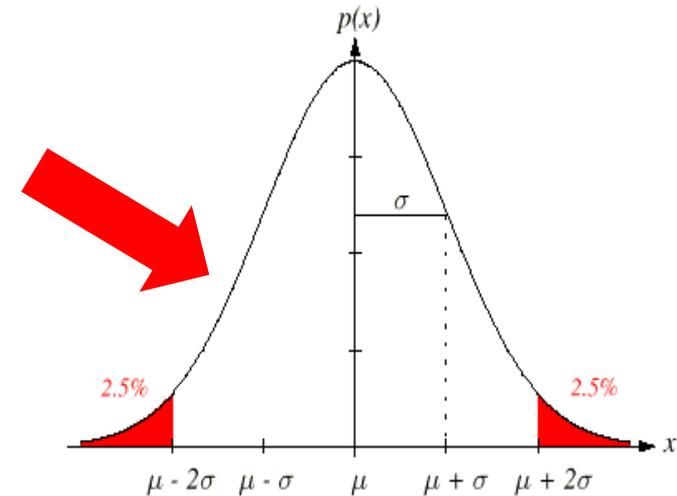
- La funzione Gaussiana ha altre proprietà
  - La trasformata di Fourier di una funzione Gaussiana è una funzione Gaussiana;
  - È ottimale per la localizzazione nel tempo o in frequenza
    - Il principio di indeterminazione stabilisce che la localizzazione non può avvenire simultaneamente in tempo e frequenza

# Densità normale univariata

- Iniziamo con la densità normale univariata. Essa è completamente specificata da due parametri, *media*  $\mu$  e *varianza*  $\sigma^2$ , si indica con  $N(\mu, \sigma^2)$  e si presenta nella forma

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

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



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

- Fissata media e varianza la densità Normale è quella dotata di massima entropia;
  - L'entropia misura l'incertezza di una distribuzione o la quantità d'informazione necessaria in media per descrivere la variabile aleatoria associata, ed è data da

$$H(p(x)) = - \int p(x) \ln p(x) dx$$

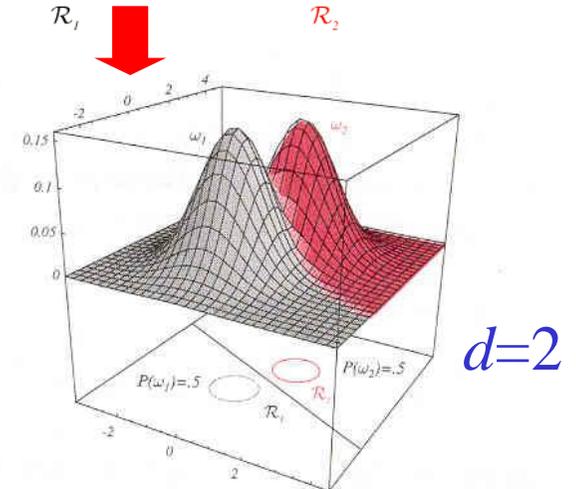
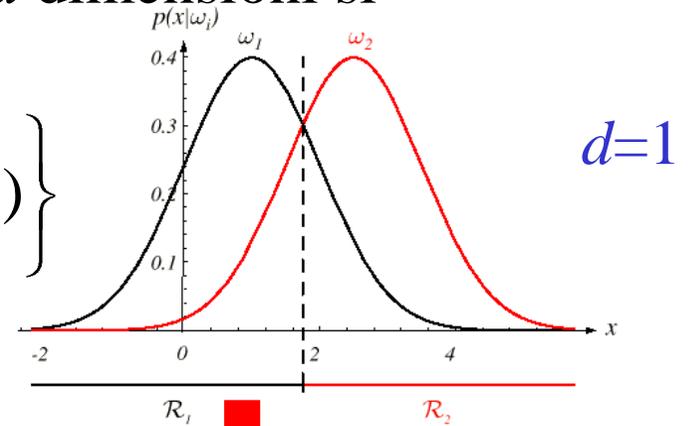
# Densità normale multivariata

- La generica densità normale multivariata a  $d$  dimensioni si presenta nella forma

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

in cui

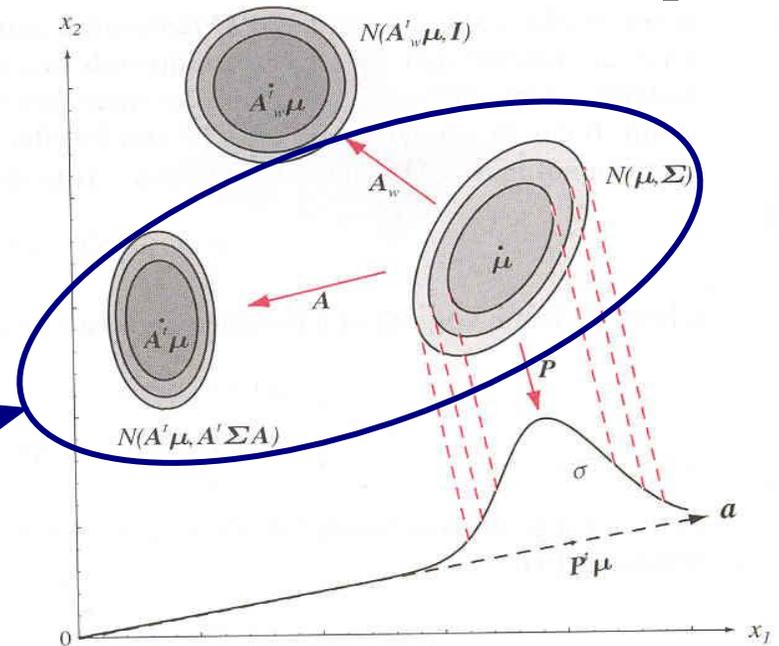
- $\boldsymbol{\mu}$  = vettore di *media* a  $d$  componenti
- $\Sigma$  = matrice  $d \times d$  di *covarianza*, dove
  - $|\Sigma|$  = determinante della matrice
  - $\Sigma^{-1}$  = matrice inversa



- Analiticamente  $\Sigma = E \left[ (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t \right] = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t p(\mathbf{x}) d\mathbf{x}$
- Elemento per elemento  $\sigma_{ij} = E \left[ (x_i - \mu_i)(x_j - \mu_j) \right]$

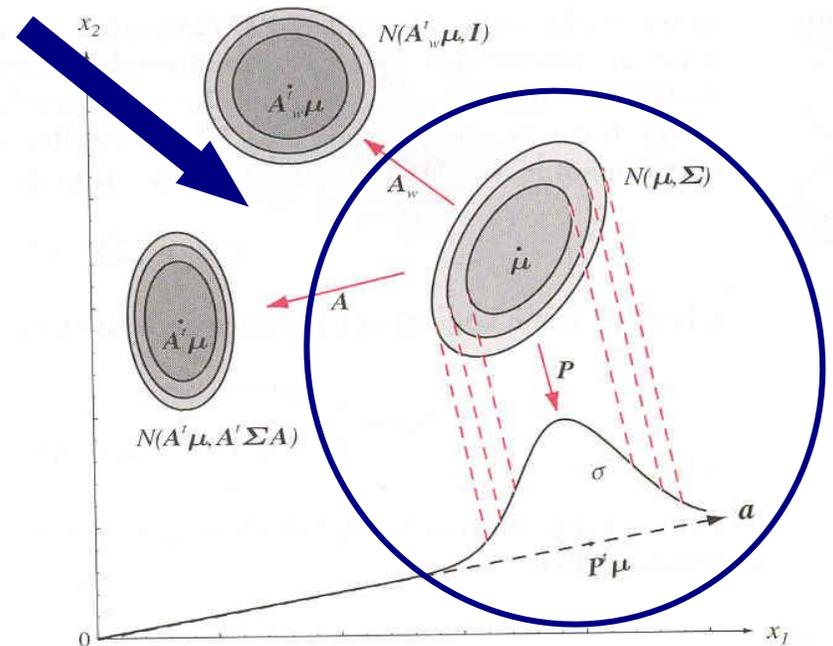
# Densità normale multivariata (2)

- Caratteristiche della matrice di covarianza
    - Simmetrica
    - Semidefinita positiva ( $|\Sigma| \geq 0$ )
    - $\sigma_{ii}$  = varianza di  $x_i$  ( $= \sigma_i^2$ )
    - $\sigma_{ij}$  = covarianza tra  $x_i$  e  $x_j$  (se  $x_i$  e  $x_j$  sono *statisticamente indipendenti*  $\sigma_{ij} = 0$ )
    - Se  $\sigma_{ij} = 0 \quad \forall i \neq j$   $p(\mathbf{x})$  è il prodotto della densità univariata per  $\mathbf{x}$  componente per componente.
    - Se
      - $p(\mathbf{x}) \approx N(\boldsymbol{\mu}, \Sigma)$
      - A matrice  $d \times k$
      - $\mathbf{y} = \mathbf{A}^t \mathbf{x}$
- $p(\mathbf{y}) \approx N(\mathbf{A}^t \boldsymbol{\mu}, \mathbf{A}^t \Sigma \mathbf{A})$



## Densità normale multivariata (3)

- CASO PARTICOLARE:  $k = 1$ 
  - $p(\mathbf{x}) \approx N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - $\mathbf{a}$  vettore  $d \times 1$  di lunghezza unitaria
  - $y = \mathbf{a}^t \mathbf{x}$
  - $y$  è uno scalare che rappresenta la proiezione di  $\mathbf{x}$  su una linea in direzione definita da  $\mathbf{a}$
  - $\mathbf{a}^t \boldsymbol{\Sigma} \mathbf{a}$  è la *varianza* di  $\mathbf{x}$  su  $\mathbf{a}$
- In generale  $\boldsymbol{\Sigma}$  ci permette di calcolare la *dispersione* dei dati in ogni superficie, o sottospazio.

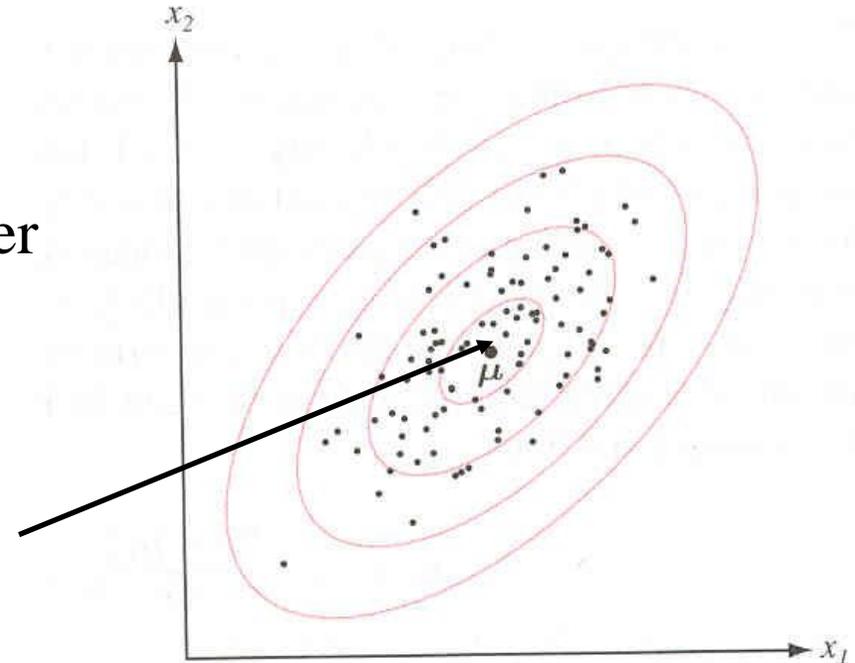


## Densità normale multivariata (4)

- Siano (trasf. sbiancante, *whitening transform*)
  - $\Phi$  la matrice degli autovettori di  $\Sigma$  in colonna;
  - $\Lambda$  la matrice diagonale dei corrispondenti autovalori;
- La trasformazione  $A_w = \Phi\Lambda^{-1/2}$ , applicata alle coordinate dello spazio delle feature, assicura una distribuzione con matrice di covarianza  $= \mathbf{I}$  (matrice identica)
- La densità  $N(\boldsymbol{\mu}, \Sigma)$  d-dimensionale necessita di  $d + d(d+1)*2$  parametri per essere definita
- Ma cosa rappresentano graficamente

$\Phi$  e  $\Lambda$  ?

**Media**  
**individuata dalle**  
**coordinate di  $\boldsymbol{\mu}$**



# Densità normale multivariata (5)

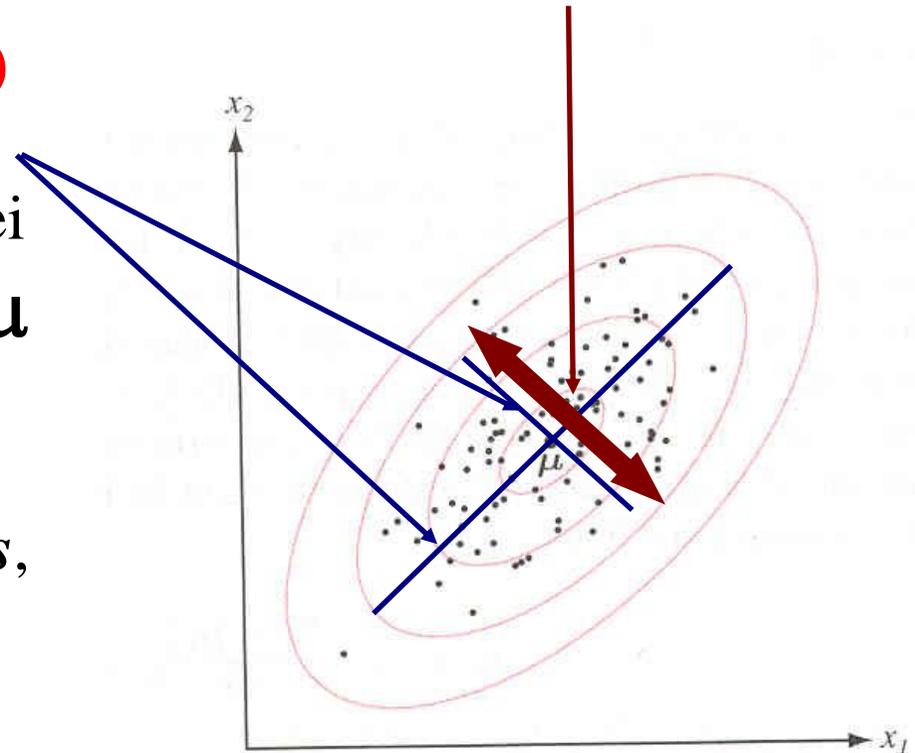
Gli assi principali degli iperellissoidi sono dati dagli autovettori di  $\Sigma$  (descritti da  $\Phi$ )

Gli iperellissoidi sono quei luoghi dei punti per i quali la distanza di  $\mathbf{x}$  da  $\boldsymbol{\mu}$

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

detta anche *distanza di Mahalanobis*, è costante

Le lunghezze degli assi principali degli iperellissoidi sono dati dagli autovalori di  $\Sigma$  (descritti da  $\Lambda$ )



# Funzioni discriminanti - Densità Normale

- Tornando ai classificatori Bayesiani, ed in particolare alle funzioni discriminanti, analizziamo la funzione discriminante come si traduce nel caso di densità Normale e *minimum error rate*

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



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

- A seconda della natura di  $\Sigma$ , la formula soprascritta può essere semplificata. Vediamo alcuni esempi.

# Funzioni discriminanti - Densità Normale $\Sigma_i = \sigma^2 \mathbf{I}$

- È il caso più semplice in cui le feature sono statisticamente indipendenti ( $\sigma_{ij} = 0, i \neq j$ ), ed ogni classe ha la stessa varianza (caso 1-D):

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

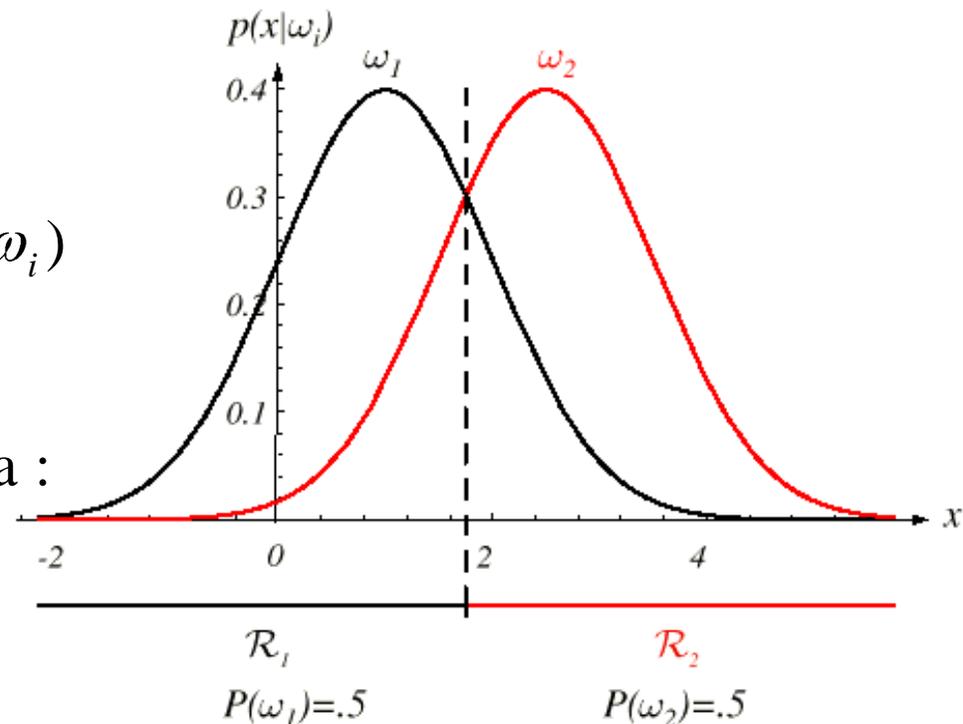
$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2} \left[ \mathbf{x}^t \mathbf{x} - 2\boldsymbol{\mu}_i^t \mathbf{x} + \boldsymbol{\mu}_i^t \boldsymbol{\mu}_i \right] + \ln P(\omega_i)$$

dove il termine  $\mathbf{x}^t \mathbf{x}$ , uguale per ogni  $\mathbf{x}$ , può essere ignorato giungendo alla forma :

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

dove

$$\mathbf{w}_i = \frac{1}{\sigma^2} \boldsymbol{\mu}_i \quad \text{e} \quad w_{i0} = -\frac{1}{2\sigma^2} \boldsymbol{\mu}_i^t \boldsymbol{\mu}_i + \ln P(\omega_i)$$



## Funzioni discriminanti - Densità Normale $\Sigma_i = \sigma^2 \mathbf{I}$ (2)

- Le funzioni precedenti vengono chiamate *funzioni discriminanti lineari* (o *linear machine*)
  - I **confini di decisione** sono dati da  $g_i(\mathbf{x}) = g_j(\mathbf{x})$  per le due classi con più alta probabilità a posteriori
  - In questo caso particolare abbiamo:

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

dove

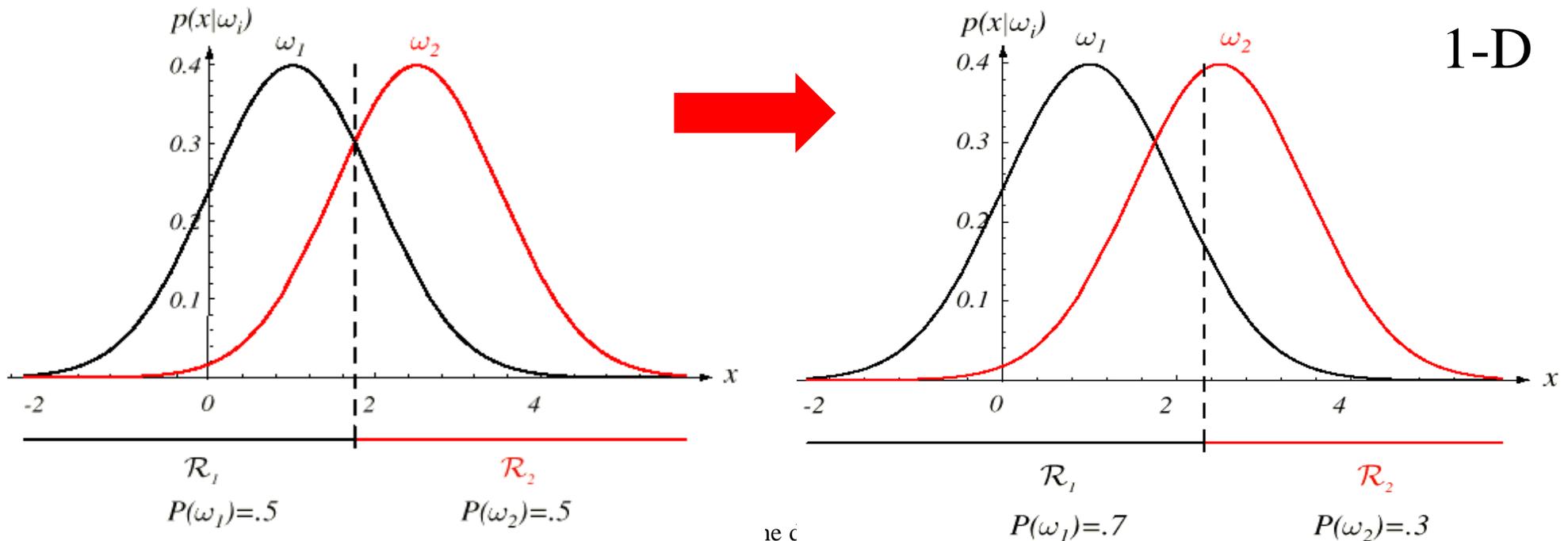
$$\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$$

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

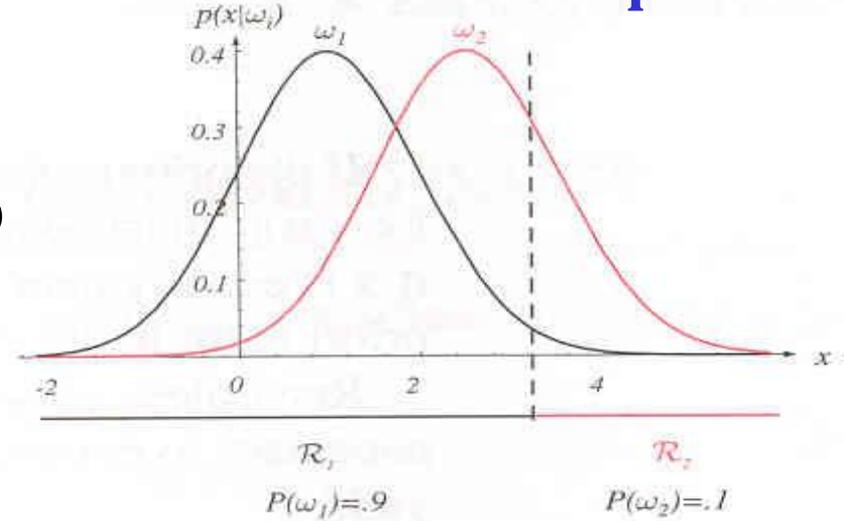
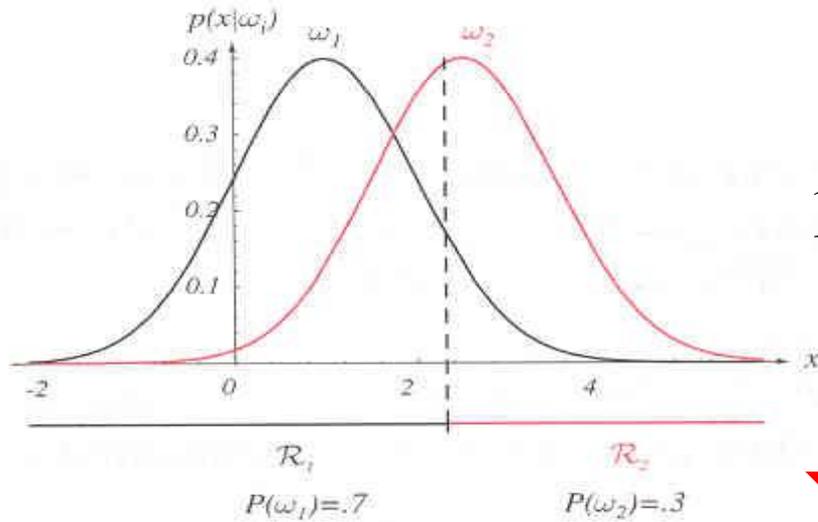
NB: se  $\sigma^2 \ll \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2$   
la posizione del confine di  
decisione è insensibile ai prior!

# Funzioni discriminanti - Densità Normale $\Sigma_i = \sigma^2 \mathbf{I}$ (3)

- Le funzioni discriminanti lineari definiscono un iperpiano passante per  $\mathbf{x}_0$  ed ortogonale a  $\mathbf{w}$ :  
dato che  $\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ , l'iperpiano che separa  $R_i$  da  $R_j$  è *ortogonale* alla linea che unisce le medie.
- Dalla formula precedente si nota che, a parità di varianza, il prior maggiore determina la classificazione.

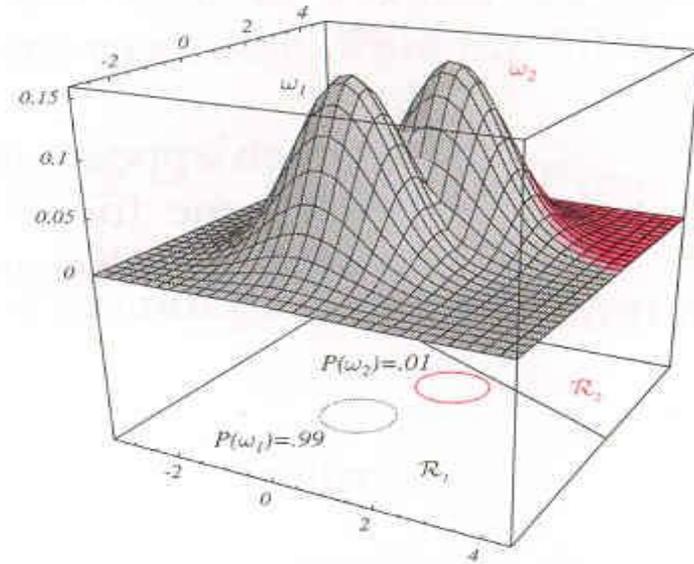
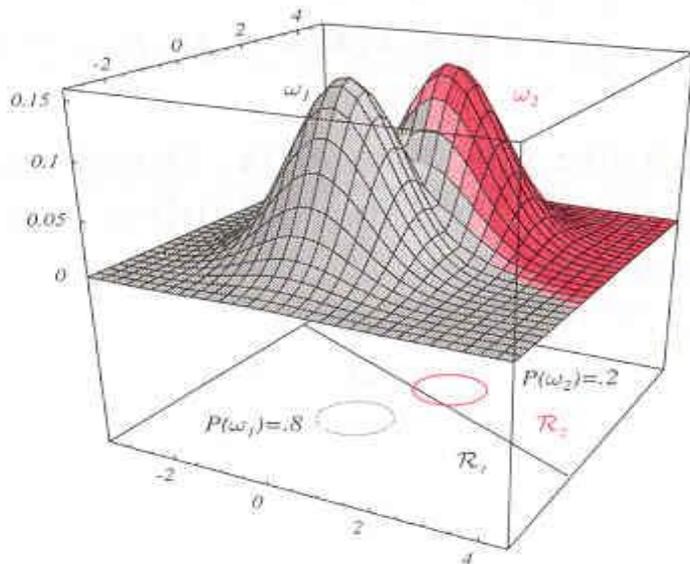


# Funzioni discriminanti - Densità Normale $\Sigma_i = \sigma^2 I$ (4)

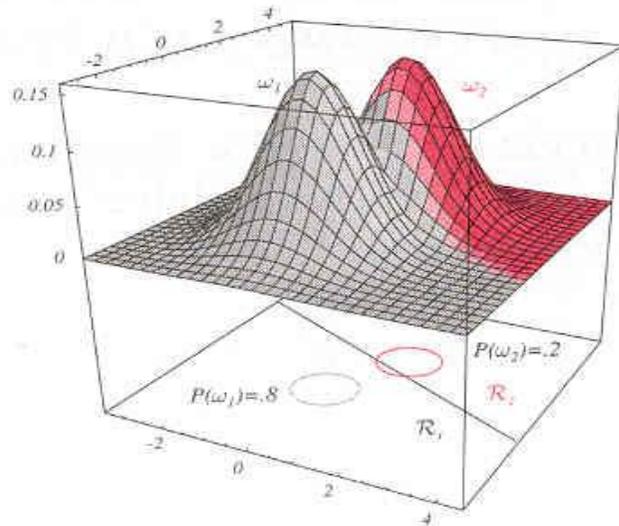


1-D

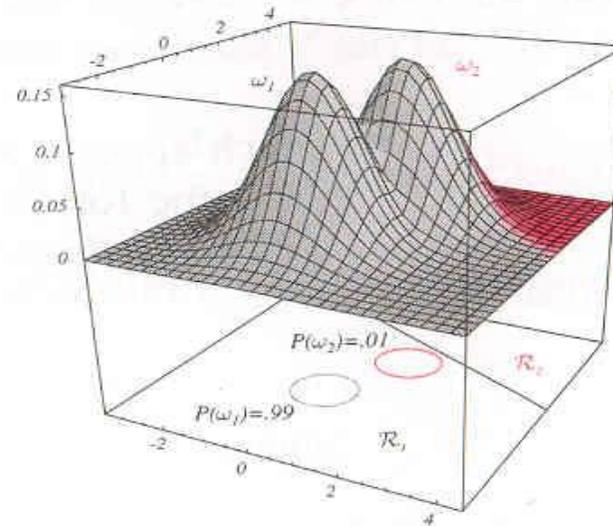
2-D



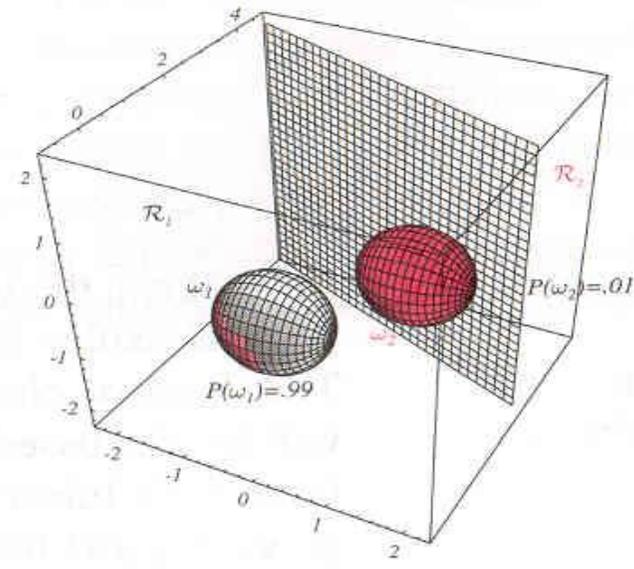
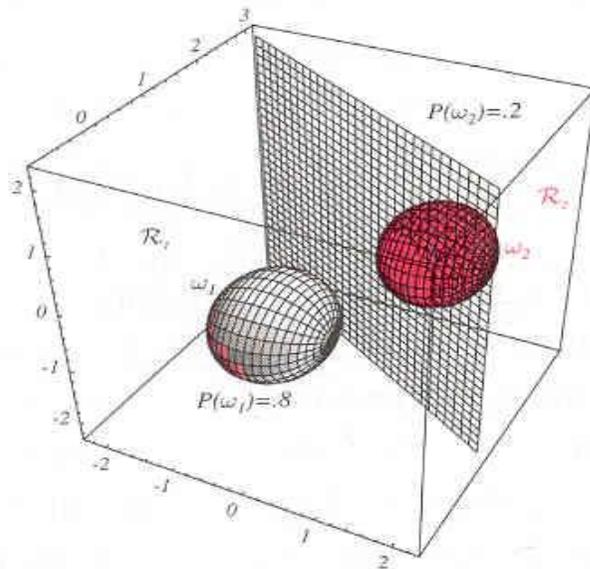
# Funzioni discriminanti - Densità Normale $\Sigma_i = \sigma^2 \mathbf{I}$ (5)



2-D



3-D



## Funzioni discriminanti - Densità Normale $\Sigma_i = \sigma^2 \mathbf{I}$ (6)

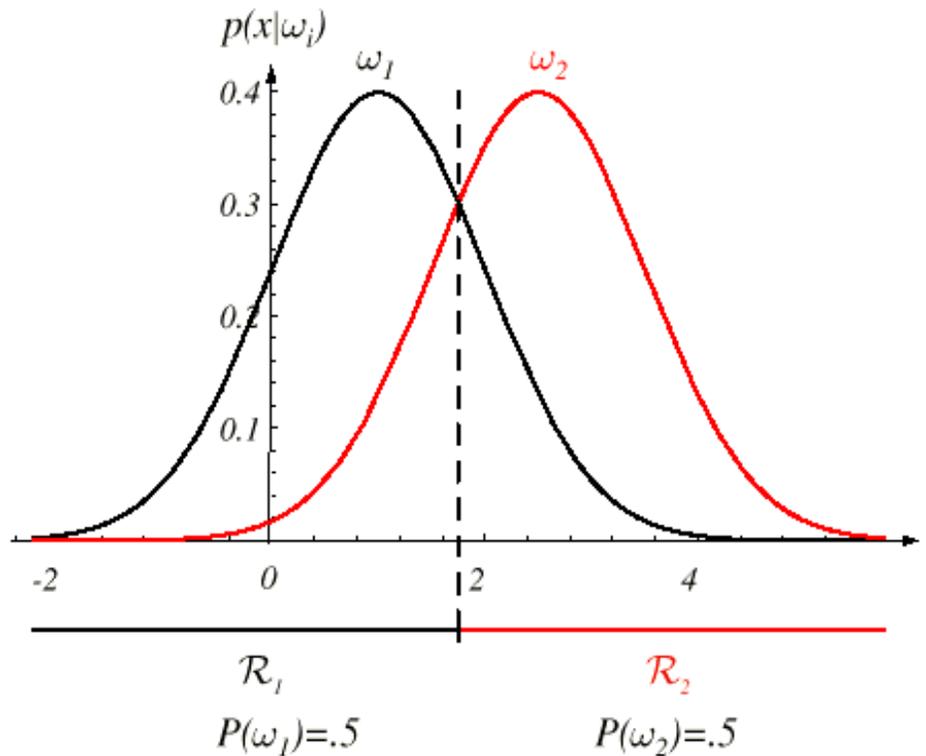
$$\mathbf{x}_0 = \frac{1}{2} (\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$



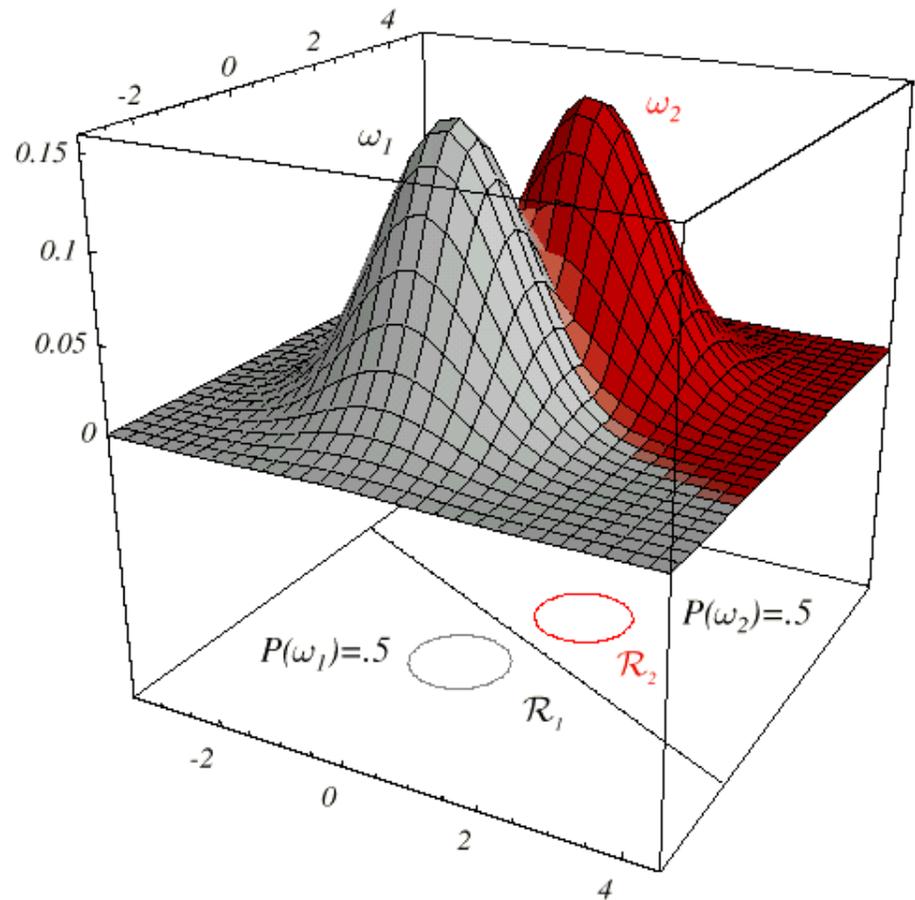
- NB.: Se le probabilità prior  $P(\omega_i)$ ,  $i=1, \dots, c$  sono *uguali*, allora il termine logaritmico può essere ignorato, riducendo il classificatore ad un *classificatore di minima distanza*.
- In pratica, la regola di decisione ottima ha una semplice interpretazione geometrica
  - Assegna  $\mathbf{x}$  alla classe la cui media  $\boldsymbol{\mu}$  è più vicina

# Funzioni discriminanti - Densità Normale $\Sigma_i = \sigma^2 \mathbf{I}$ (7)

1-D

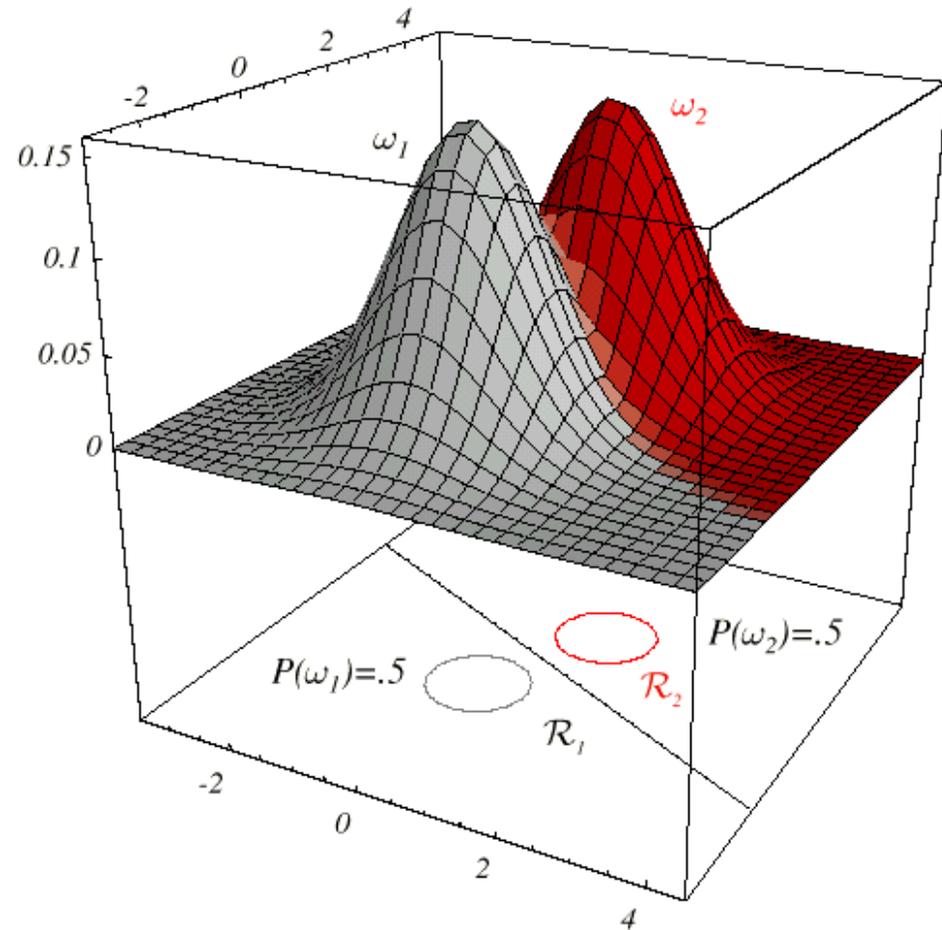


2-D

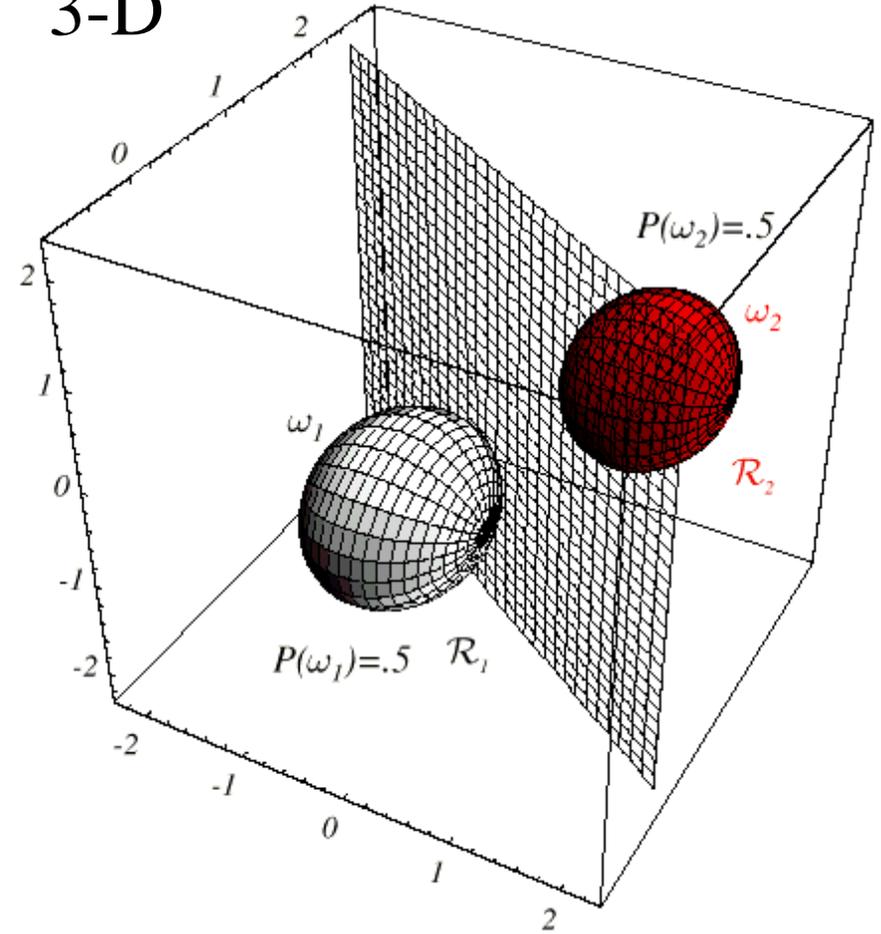


# Funzioni discriminanti - Densità Normale $\Sigma_i = \sigma^2 \mathbf{I}$ (8)

2-D



3-D



## Funzioni discriminanti - Densità Normale $\Sigma_i = \Sigma$

- Un altro semplice caso occorre quando le matrici di covarianza per tutte le classi sono uguali, ma arbitrarie.
- In questo caso l'ordinaria formula

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

può essere semplificata con

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

che è ulteriormente trattabile, con un procedimento analogo al caso precedente (sviluppando il prodotto ed eliminando il termine  $\mathbf{x}^t \Sigma^{-1} \mathbf{x}$ )

## Funzioni discriminanti - Densità Normale $\Sigma_i = \Sigma$ (2)

- Otteniamo così funzioni discriminanti ancora lineari, nella forma:

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

dove

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

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

- Poiché i discriminanti sono lineari, i *confini di decisione* sono ancora iperpiani

## Funzioni discriminanti - Densità Normale $\Sigma_i = \Sigma$ (3)

- Se le regioni di decisione  $R_i$  ed  $R_j$  sono contigue, il confine tra esse diventa:

$$\mathbf{w}'(\mathbf{x} - \mathbf{x}_0) = 0,$$

where

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

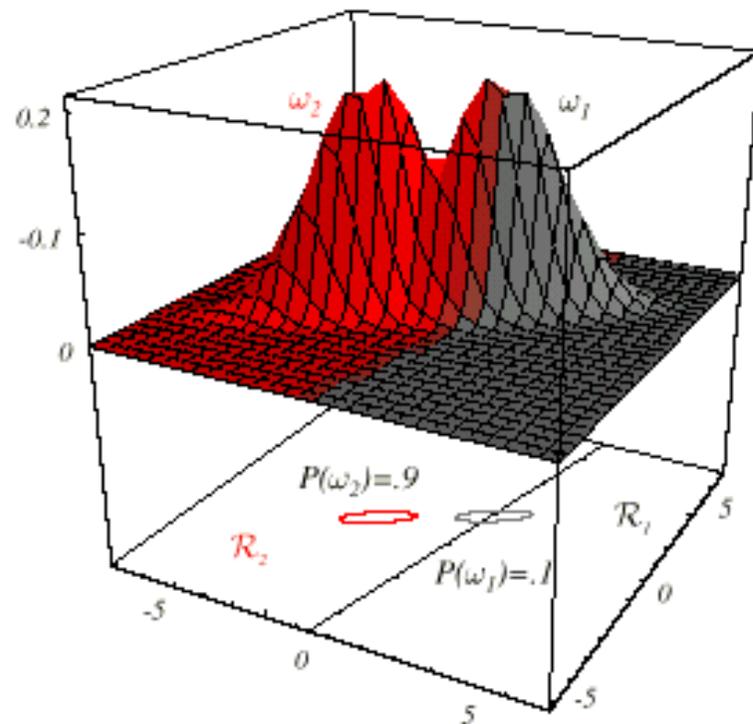
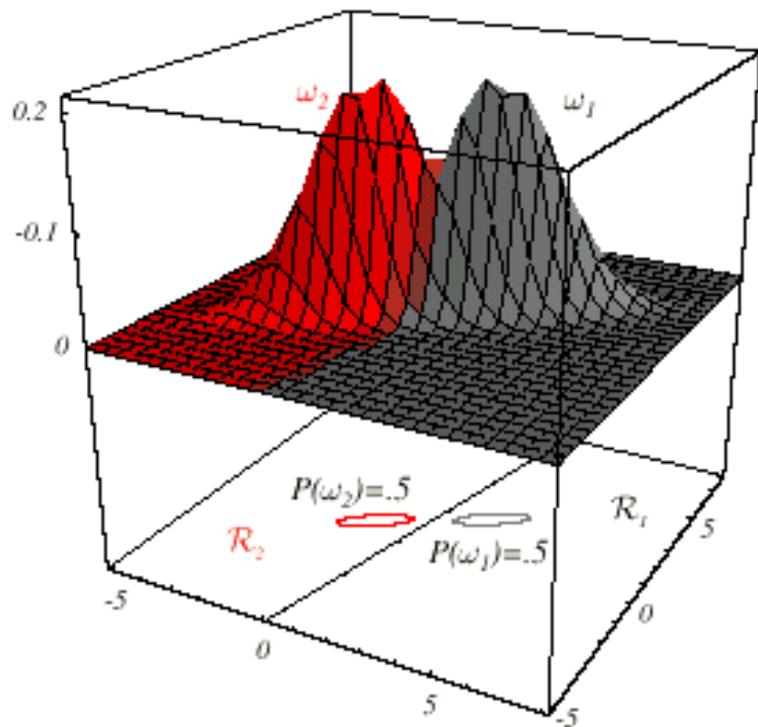
and

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

# Funzioni discriminanti - Densità Normale $\Sigma_i = \Sigma$ (4)

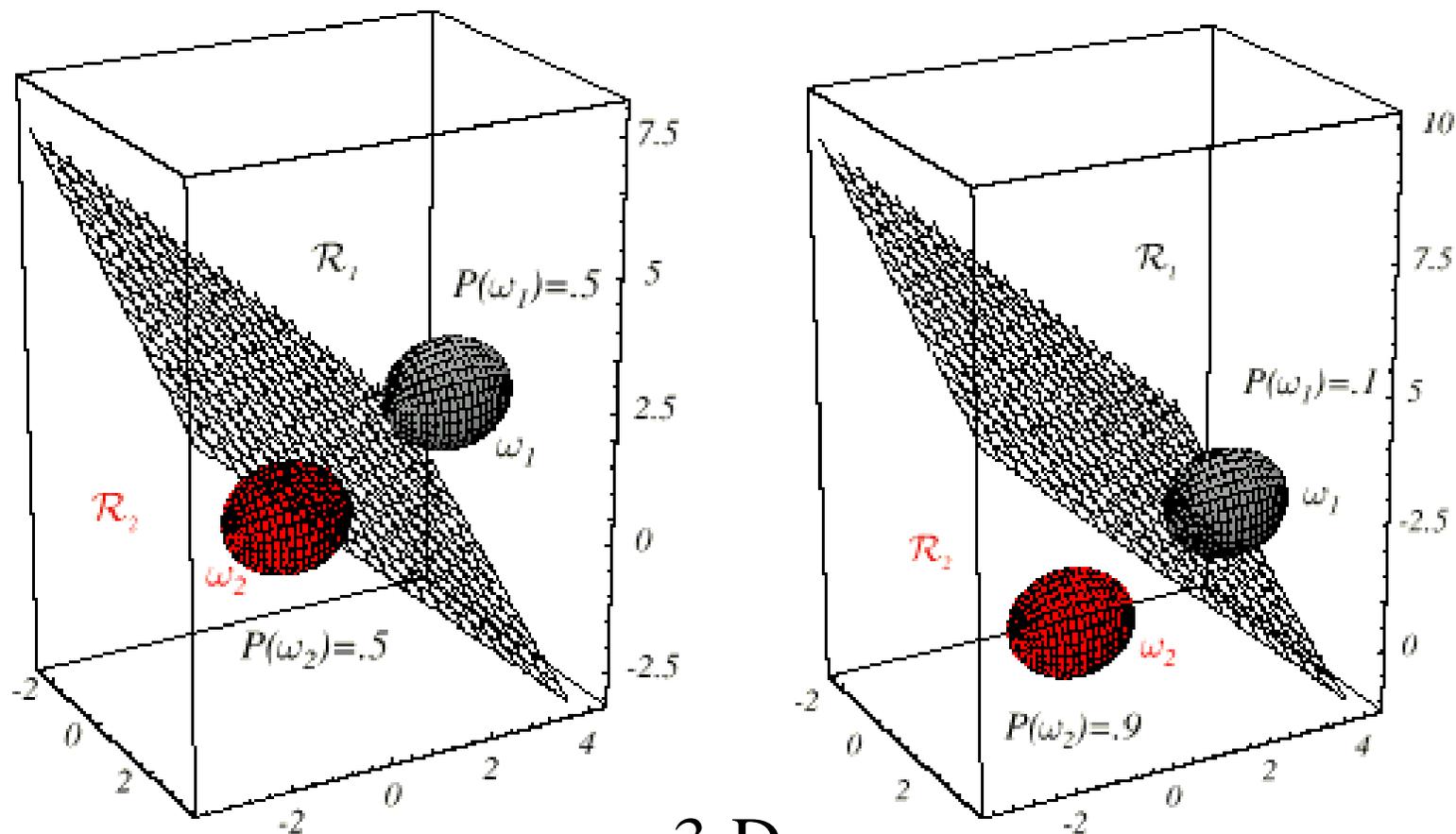
- Poiché  $\mathbf{w}$  in generale (differentemente da prima) non è il vettore che unisce le 2 medie ( $\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ ), l'iperpiano che divide  $R_i$  da  $R_j$  non è quindi ortogonale alla linea tra le medie; comunque, esso interseca questa linea in  $\mathbf{x}_0$
- Se i *prior* sono uguali, allora  $\mathbf{x}_0$  si trova in mezzo alle medie, altrimenti l'iperpiano ottimale di separazione si troverà spostato verso la media meno probabile.

# Funzioni discriminanti - Densità Normale $\Sigma_i = \Sigma$ (5)



2-D

# Funzioni discriminanti - Densità Normale $\Sigma_i = \Sigma$ (6)



3-D

## Funzioni discriminanti - Densità Normale $\Sigma_i$ arbitraria

- Le matrici di covarianza sono differenti per ogni categoria;
- Le funzioni discriminanti sono inerentemente quadratiche;

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

where

$$\mathbf{W}_i = -\frac{1}{2} \Sigma_i^{-1},$$

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

and

$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i' \Sigma_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i).$$

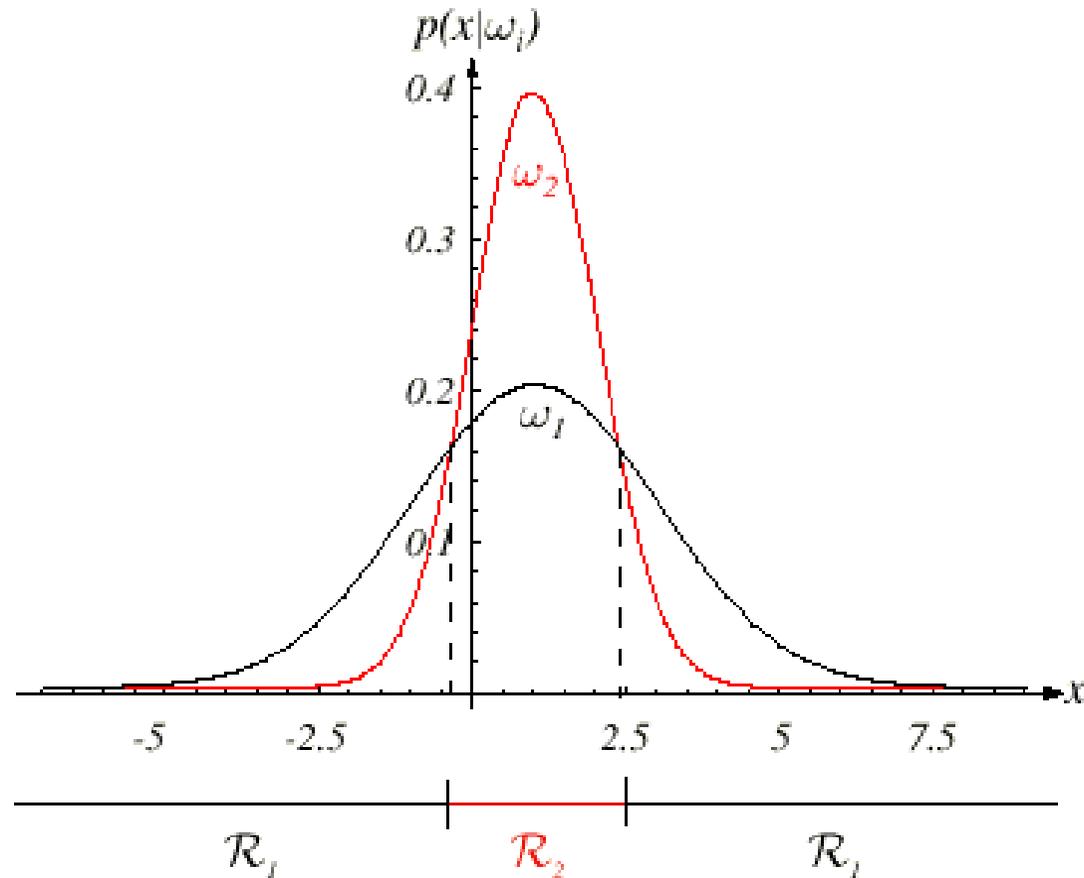
# Funzioni discriminanti

## Densità Normale $\Sigma_i$ arbitraria (2)

- Nel caso 2-D le superfici di decisione sono *iperquadriche*:
  - Iperpiani
  - Coppia di iperpiani
  - Ipersfere
  - Iperparaboloidi
  - Iperiperboloidi di vario tipo
- Anche nel caso 1-D, per la varianza arbitraria, le regioni di decisione di solito sono non connesse.

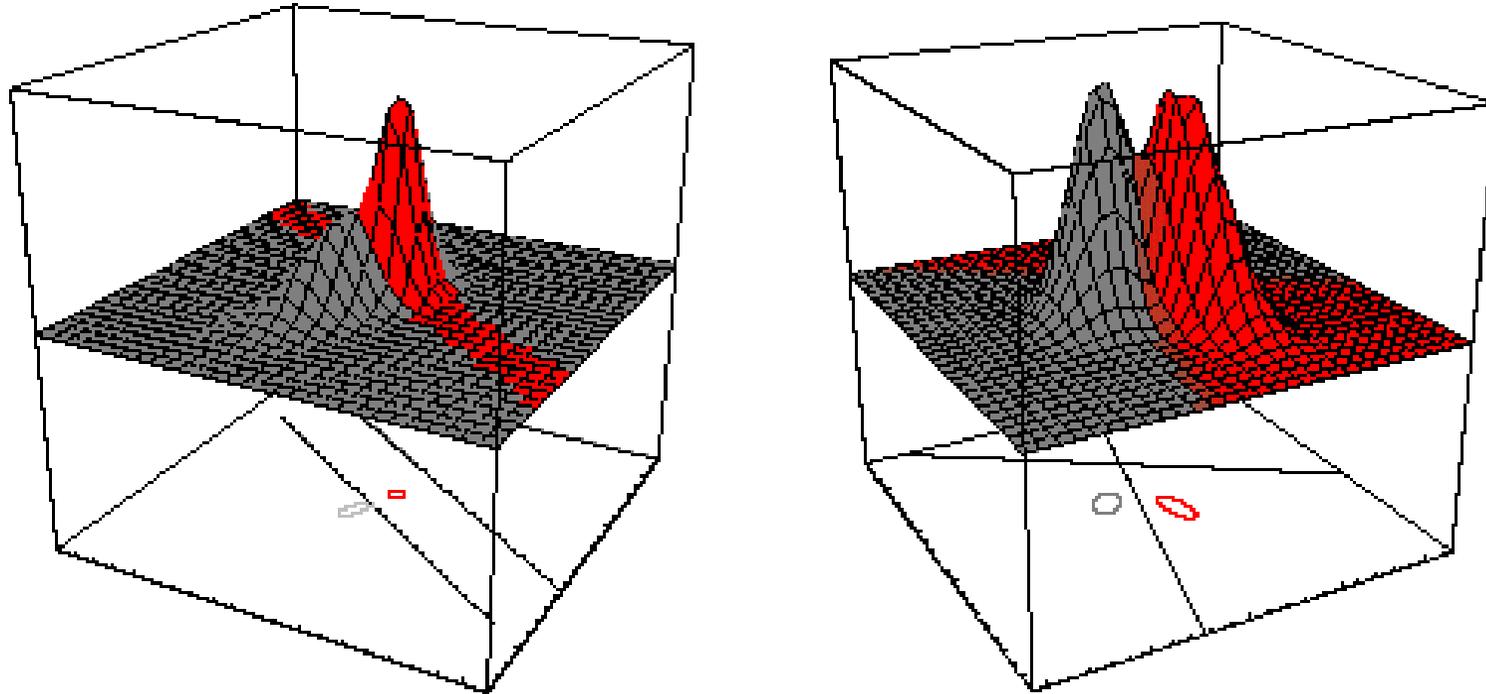
# Funzioni discriminanti

## Densità Normale $\Sigma_i$ arbitraria (3)



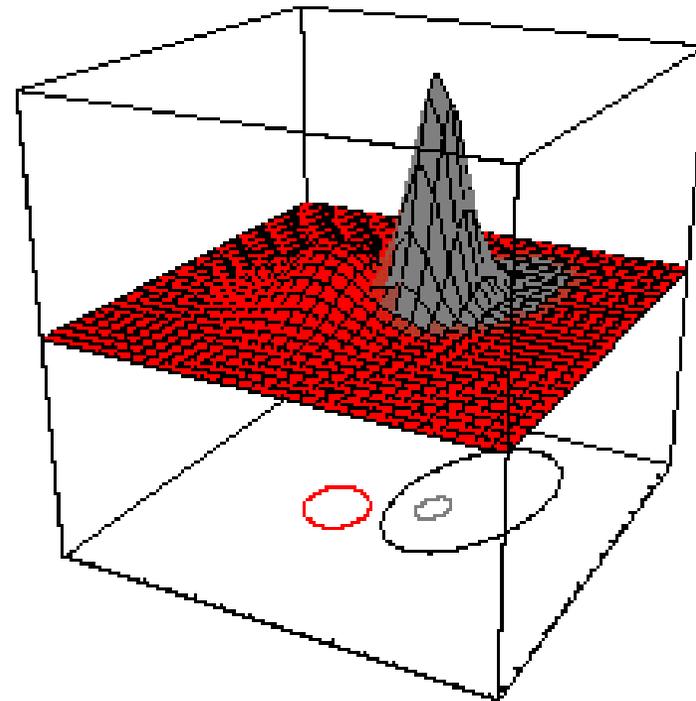
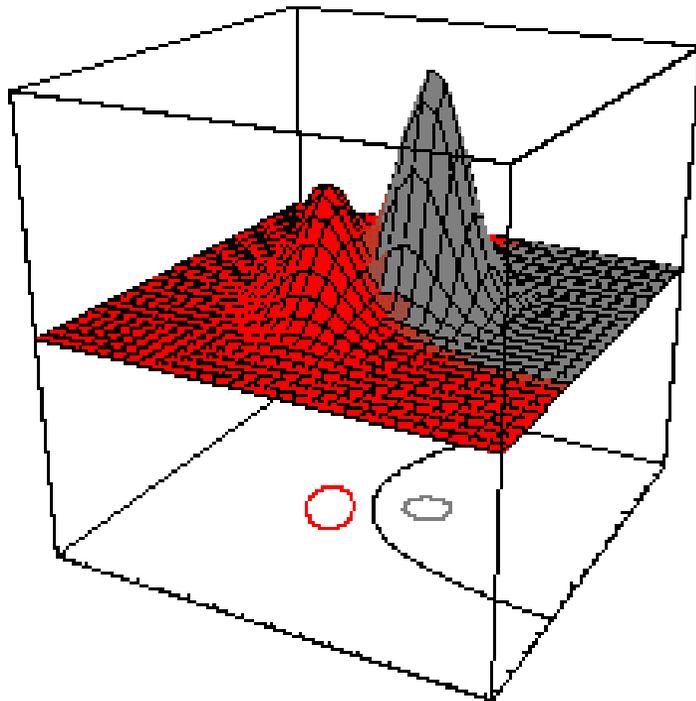
# Funzioni discriminanti

## Densità Normale $\Sigma_i$ arbitraria (4)



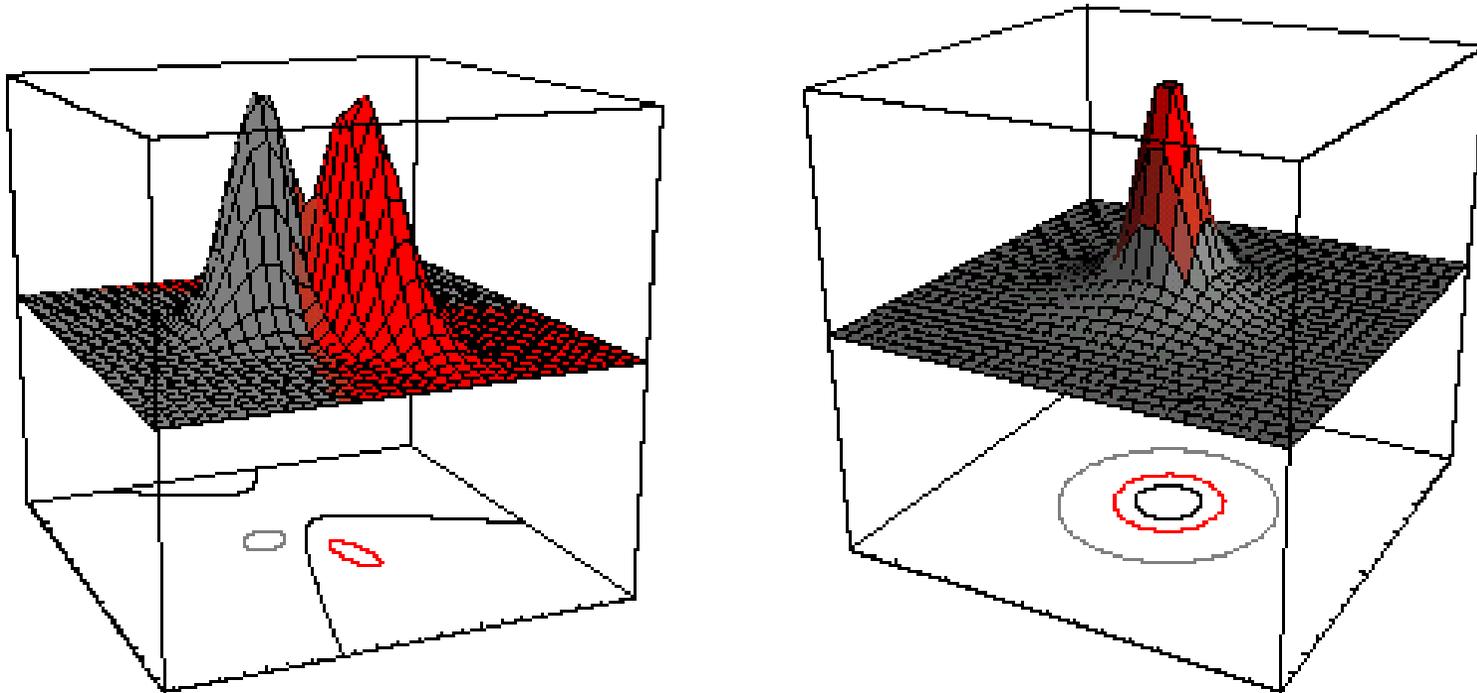
# Funzioni discriminanti

## Densità Normale $\Sigma_i$ arbitraria (5)



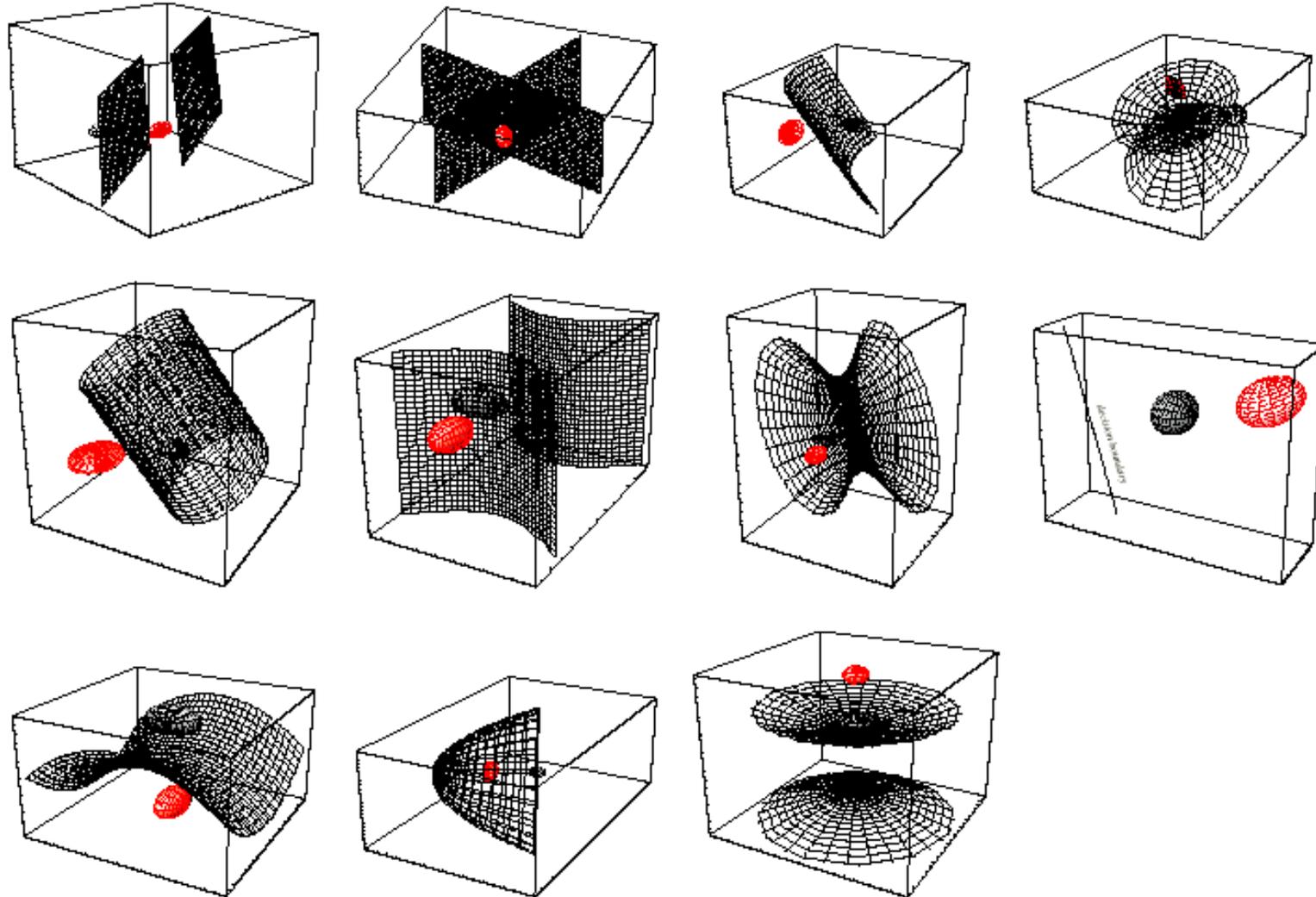
# Funzioni discriminanti

## Densità Normale $\Sigma_i$ arbitraria (6)



# Funzioni discriminanti

## Densità Normale $\Sigma_i$ arbitraria (7)



# Funzioni discriminanti

## Densità Normale $\Sigma_i$ arbitraria (8)

