Riconoscimento e recupero dell’informazione per bioinformatica

Teoria della decisione di Bayes

Manuele Bicego

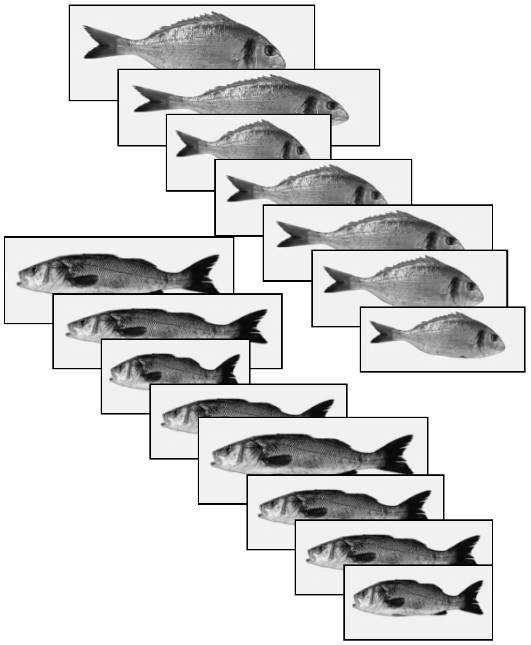
Corso di Laurea in Bioinformatica

Dipartimento di Informatica - Università di Verona

Sommario

- Sistema di classificazione
- La teoria della decisione di Bayes
  - versione base
  - estensioni
- Classificatori, funzioni discriminanti
- Funzioni discriminanti nel caso gaussiano

Esempio: classificazione



Insieme di addestramento

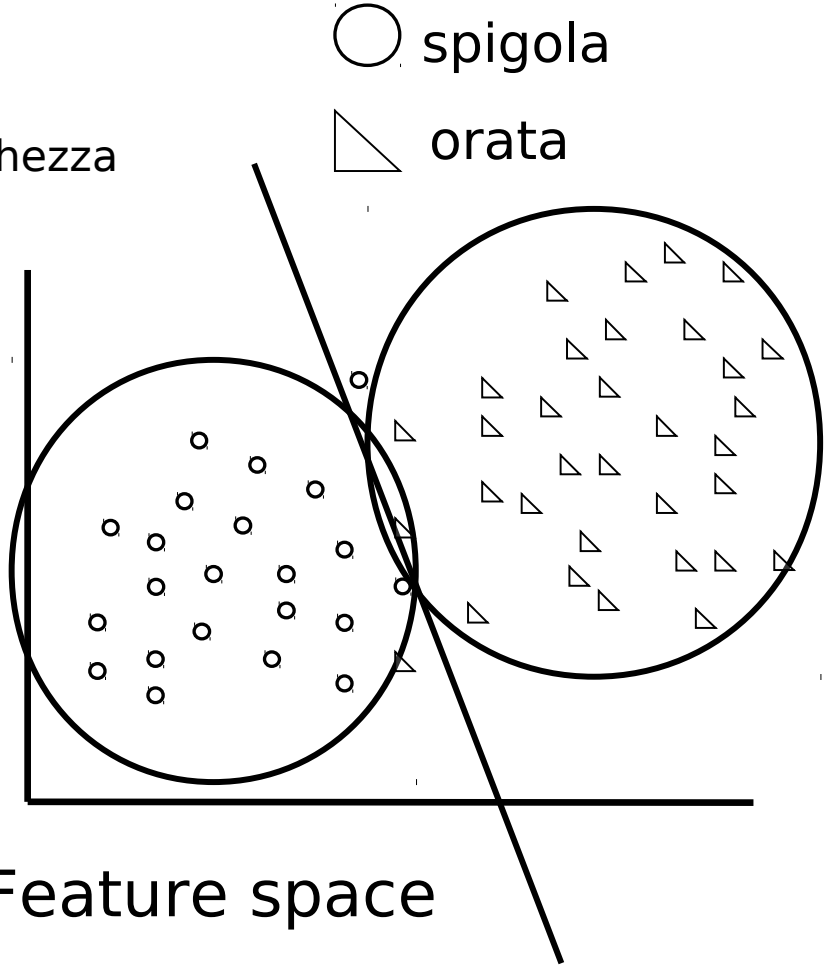
y_i etichette

$\mathbf{x}_1, y_1$

$\mathbf{x}_2, y_2$

...

$\mathbf{x}_N, y_N$



Feature space

Addestramento: modellare (separare) le due classi

Esempio: classificazione/testing

oggetto sconosciuto

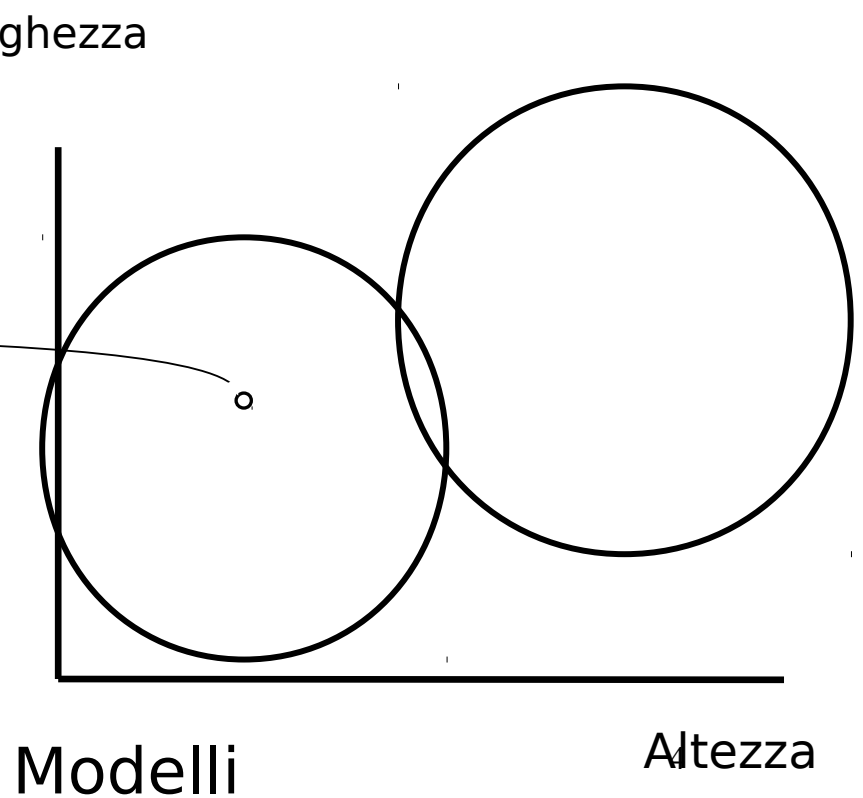
feature extraction

$\mathbf{x}_1 = [3, 12]$

dati pre-processati

testing

categoria: spigola



Sistema di classificazione

- Il fuoco è sul sistema di decisione:
  - un sistema che ci permette di dire, dato un oggetto in ingresso, a quale classe l'oggetto appartiene
  - un sistema che "classifica" l'oggetto: un classificatore

- Dato un oggetto x, un classificatore è una funzione f che ritorna un valore y discreto (una delle possibili categorie/classi)

$$y = f(x)$$

- Differente dalla regressione (y continuo)

Sistema di classificazione

- Goal: stimare la funzione f
- Requisito: si vuole trovare una funzione f che “sbagli” il meno possibile (ridurre gli errori che può fare un sistema di decisione)
  - nel senso dell’errore di generalizzazione
- Errore: un oggetto appartiene alla classe 1 e viene classificato come appartenente alla classe 2

Sistema di classificazione

- Più in generale, il sistema di decisione non solo determina la classe (categoria) dell'oggetto in questione ma permette anche di effettuare un'azione sulla base di tale classe

- Esempio:
  - Se si riconosce il volto di un criminale si chiama la polizia

Sistema di classificazione

- Teorie della decisione: come costruire il classificatore

- Ce ne sono diverse, caratterizzate da:
  - come vengono espresse/caratterizzate le entità in gioco
  - come viene determinata la regola di decisione
  - come possono essere interpretate le soluzioni
- Esempi:
  - Teoria di Bayes: approccio probabilistico
  - Statistical Learning Theory: approccio geometrico

- Non c’è una chiara separazione tra le teorie

Teoria della decisione di Bayes

La teoria della decisione di Bayes

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

Introduzione

- Approccio probabilistico di classificazione di pattern
  - Molto utilizzato
  - Molti i risultati teorici dimostrati
- Ipotesi:
  - Il problema di decisione è posto in termini probabilistici;
  - Tutte le probabilità rilevanti sono conosciute;
- Goal:
  - Discriminare tra le diverse classi (determinare le regole di decisione) usando tali probabilità

Introduzione

- DEFINIZIONE: Sia **ω** la categoria o la classe (“lo stato di natura”) da descrivere probabilisticamente;
- **ω** assume un valore diverso per ogni classe
  - **ω=ω1 -->** prima classe,
  - **ω=ω2 -->** seconda classe,
  - **ω=ω3 -->** terza classe
  - ...
- La regola di classificazione (o regola di decisione) mi permette di rispondere alla seguente domanda:

*“Dato un oggetto **x**, lo assegno a **ω1**, a **ω2** oppure a **ω3**?”* In altre parole: *“A quale classe deve essere assegnato?”*

Introduzione

- Teoria della decisione di Bayes: il problema di decisione è posto in termini probabilistici
- Ci sono diverse probabilità che si possono utilizzare per costruire la regola di decisione
  - Ognuna porta ad una regola di decisione diversa

- In particolare:
  - Probabilità a priori: regola della probabilità a priori
  - Probabilità condizionale: regola Maximum Likelihood
  - Probabilità a posteriori: regola di Bayes

Vediamole con un esempio...

Esempio

- ESEMPIO guida: Discriminare tra un calciatore professionista e il resto del mondo sulla base dello stipendio
  - x = stipendio (pattern da classificare)
  - **ω1 =** calciatore professionista (classe 1)
  - **ω2 =** resto del mondo (classe 2)

Probabilità a priori

- Prima possibilità: utilizzare solo l'eventuale informazione a priori che si ha sul problema:
  - “la probabilità che una persona sia un calciatore professionista è molto bassa (1%)”

Regola di decisione:

Data una persona **x** da classificare: sapendo che il 99% delle persone sono “non calciatori”, la classifico come “non calciatore”

Viene utilizzata la “Probabilità a PRIORI”

Probabilità a priori

- Probabilità a priori:
  - probabilità P(**ω**): rappresenta la probabilità dello stato nota a priori (senza aver osservato nulla del sistema)
  - Probabilità derivanti da conoscenze “esterne”
- Esempio calciatore: P(**ω= ω1**) = 0.01, P(**ω= ω2**) = 0.99

Regola di decisione della Probabilità a PRIORI:

decidi **ω1** se P(**ω1**) > P(**ω2**); altrimenti decidi **ω2**

- Ovviamente è un sistema limitato:
  - Non tiene minimamente conto delle osservazioni (indipendente da x)
- Esempio calciatore:
  - Ogni pattern x (stipendio) è sempre classificato come “resto del mondo”

Probabilità condizionale

- Seconda possibilità: utilizzare dei modelli per gli stipendi delle persone all'interno delle due classi
  - E' presente un training set con cui si costruisce un modello per gli stipendi dei calciatori
  - Si ripete la stessa operazione sugli stipendi dei “non calciatori”

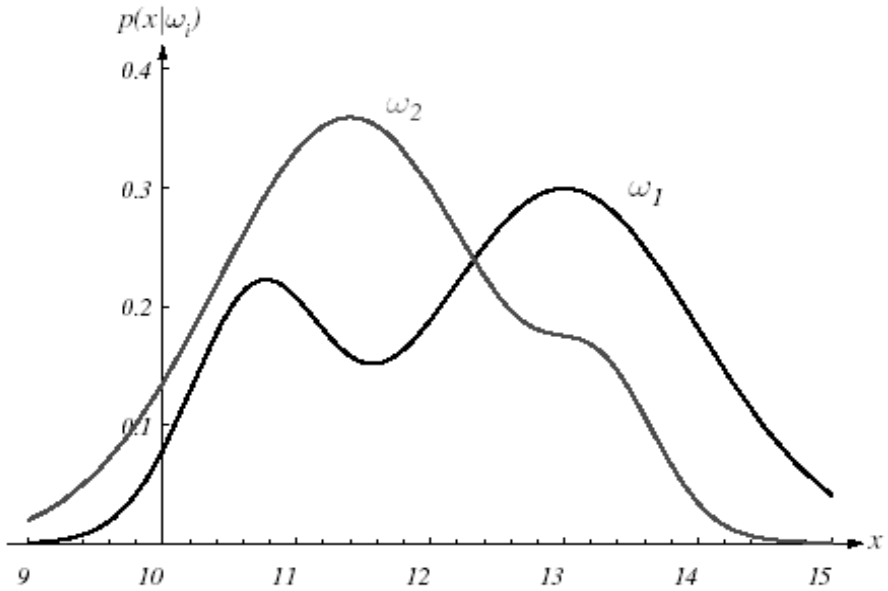
Regola di decisione:

Data una persona da classificare: se il suo stipendio **x** è spiegato meglio dal modello della classe “non calciatore”allora lo assegno alla classe “non calciatore”, altrimenti il contrario

- Modello per x in una classe → PROBABILITA' CONDIZIONALE

Probabilità condizionale

- sia x una misurazione del sistema
  - x è una variabile aleatoria dipendente da **ωj**
- La probabilità condizionale (o likelihood) è definita come P(x|**ωj**)
  - misura la probabilità di avere la misurazione x sapendo che lo stato di natura (la classe è **ωj)**



Regola basata sulla probabilità condizionale

- La regola si basa sulla seguente osservazione ragionevole: fissata la misurazione x, più è alta p(x|**ωj**) più è probabile che **ωj** sia la classe giusta

Regola di decisione (maximum likelihood):

dato x, decidi **ω1** se p(x|**ω1**) > p(x|**ω2**), **ω2** altrimenti

- Migliore della regola basata sulla probabilità a priori perché qui si considera l'osservazione

- E' vero che è sicuramente migliore della regola della probabilità a priori … MA si basa solo sull'osservazione!!

  Esempio calciatore: non si tiene in conto del fatto che pochissime persone sono calciatori professionisti

- SOLUZIONE:
  - Regola di Bayes: utilizza la probabilità a posteriori, che mette assieme probabilità a priori e probabilità condizionale

Regola di Bayes

- IDEA: Prendere la decisione su una persona tenendo conto sia del fatto che ci sono pochi calciatori (proabilità a priori) sia guardando quanto il suo stipendio *x* è spiegato dal modello delle classi (probabilità condizionale)

- Si utilizza la PROBABILITA' A POSTERIORI
  - Probabilità che mette assieme probabilità a priori e probabilità condizionale

(alla lavagna)

Regola di Bayes

Ricapitolando:

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

Regola di decisione di Bayes:

⇨ dato x, decidi **ω1** se p(**ω1**|x) > p(**ω2**|x), **ω2** altrimenti

⇨ La regola di decisione di Bayes minimizza la probabilità di errore (alla lavagna)

Regola di Bayes

Regola di decisione equivalente:

- ⇨ l'evidenza rappresenta un fattore di scala che descrive quanto frequentemente si osserva un pattern x
- ⇨ non dipende da ω1 o da ω2, quindi è ininfluente per la regola di decisione
- ⇨ regola di decisione equivalente:

  Decidi ω1 se $p(x|\omega1)P(\omega1) > p(x|\omega2)P(\omega2)$, ω2 altrimenti

In pratica?

- Problema principale: le probabilità non sono note, come si fa a costruire il classificatore?
- Soluzione: apprendimento da esempi

- Si utilizza un training set per effettuare una “stima” delle probabilità
  - Date le probabilità si può fare “classificazione” con la regola di Bayes

Stima delle probabilità

Diversi approcci:

- Stime parametriche: si conosce la forma della pdf, se ne vogliono stimare i parametri
  - esempio gaussiana, stimo la media
- Stime non parametriche: non si assume nota la forma, la pdf è stimata direttamente dai dati
  - esempio istogramma
- Stime semi-parametriche: ibrido tra le due – i parametri possono cambiare la forma della funzione
  - esempio Neural Networks

Estensione della teoria di decisione di Bayes

Possibili estensioni

Caso 1: il pattern x è composto da più features

Caso 2: ci sono più di due classi

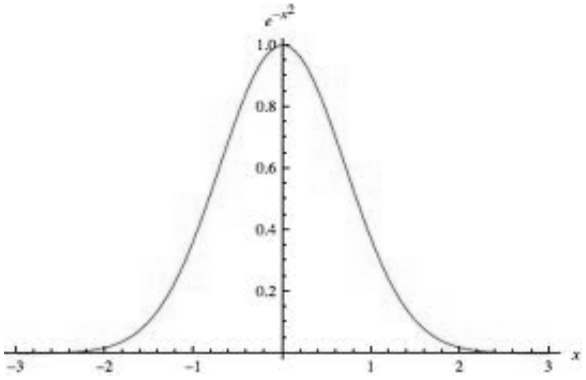
Caso 3: si vogliono fare anche altre azioni (non solo classificare)

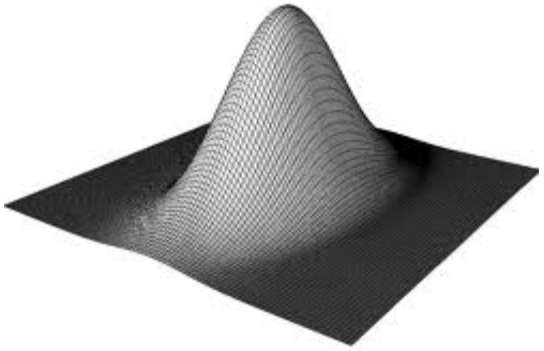
Caso 4: non si vuole solo minimizzare la probabilità di errore

Possibili estensioni: 1

- Il pattern è composto da più features

- Le probabilità sono definite su spazi multidimensionali
  - Esempio: due features, gaussiana con media a 2 dimensioni





Possibili estensioni: 2

- Ci sono più di due classi

- Per ogni classe si può calcolare la probabilità a posteriori

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

- La regola di decisione non cambia e dice di assegnare un oggetto alla classe la cui probabilità a posteriori è massima

$$class(x) = arg\,max_j P(\omega_j|x)$$

Possibili estensioni: 3

- Si possono fare altre operazioni (non solo classificare)

- Un esempio: rigetto della classificazione
  - Un pattern x non viene classificato (non c'è abbastanza confidenza nella risposta)

Chow's rule: rigettare una classificazione se

$$max_j P(\omega_j | x) < T$$

ovviamente la scelta di T è cruciale

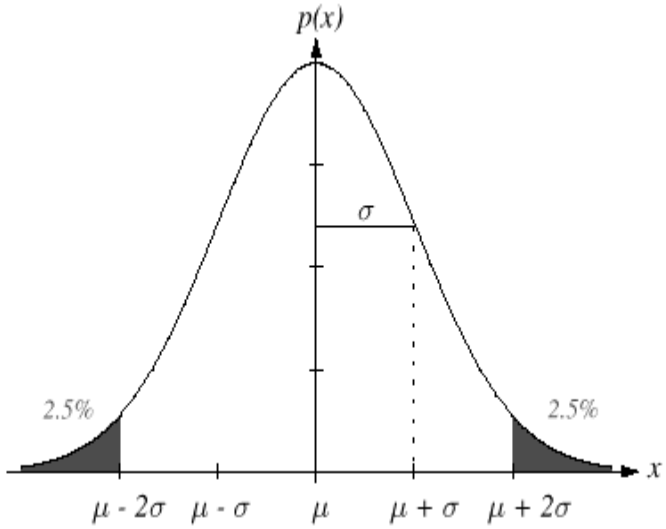
Possibili estensioni: 4

- Minimizzazione di altre funzioni (non solo la probabilità di errore)
- Definizione di “costo della classificazione”
  - quanto costa prendere una decisione sbagliata
  - Sbagliare a predire la classe “c'è un incendio” è più grave che sbagliare a predire “non c'è un incendio”,
- Codificata da una funzione chiamata “Funzione LOSS (perdita)”
- Si può riscrivere la regola di decisione di Bayes per minimizzare il costo (**Regola di Bayes estesa)**

Parentesi: la densità normale

La densità normale

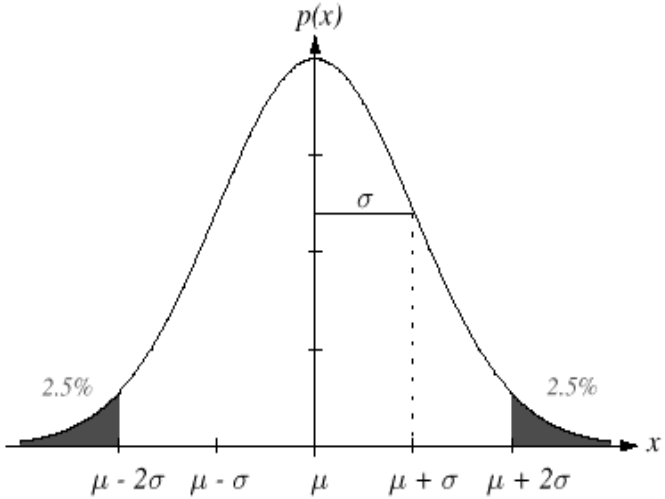
- Una delle più importanti densità è la densità normale o Gaussiana multivariata; infatti:
  - è analiticamente trattabile;
  - più importante, fornisce la migliore modellazione di problemi sia teorici che pratici
    - il teorema del Limite Centrale asserisce che “sotto varie condizioni, la distribuzione della somma di d variabili aleatorie indipendenti tende ad un limite particolare conosciuto come distribuzione normale”.



La densità normale univariata

- Completamente specificata da due parametri, media μ e varianza σ^2,
- si indica con $N(\mu, \sigma^2)$ e si presenta nella forma

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



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

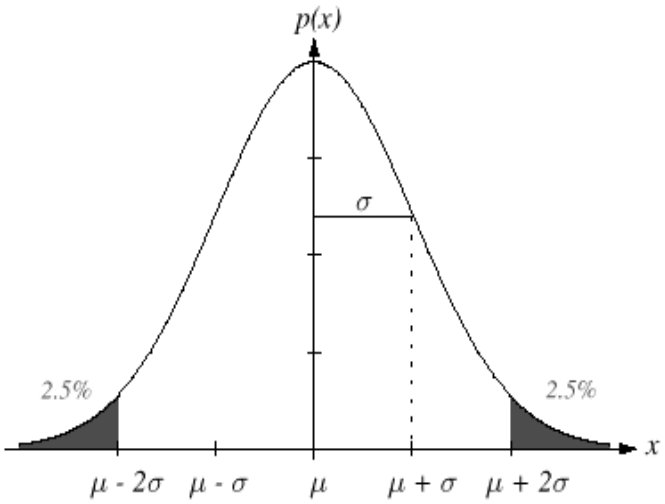
$$\sigma^2 = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, dx$$

Densità normale multivariata

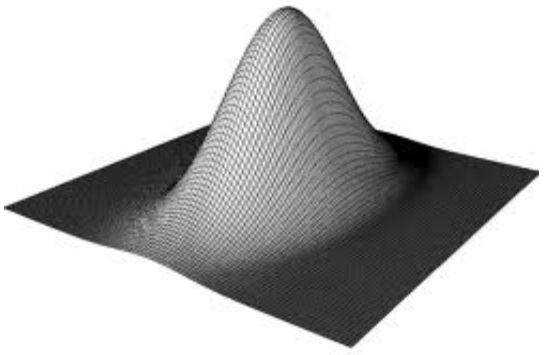
- La generica densità normale multivariata a d dimensioni si presenta nella forma

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

d=1



d=2



Densità normale multivariata

Parametri della densità:

$\boldsymbol{\mu}$ = vettore di ***media*** a *d* componenti

$\boldsymbol{\Sigma}$ = matrice *d* x *d* di ***covarianza***, dove

$|\boldsymbol{\Sigma}|$ = determinante della matrice

$\boldsymbol{\Sigma}^{-1}$ = matrice inversa

$$\Sigma = E\left[(x-\mu)(x-\mu)^t\right] = \int (x-\mu)(x-\mu)^t p(x)\,dx$$

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

Densità normale multivariata

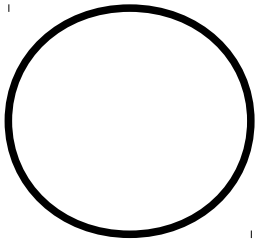
- Caratteristiche della matrice di covarianza
  - Simmetrica
  - Semidefinita positiva ($|\boldsymbol{\Sigma}| \geq 0$)
  - σ_{ii}= varianza di x_i (= σ_i^2)
  - σ_{ij}= covarianza tra x_i e x_j

- se x_i e x_j sono ***statisticamente indipendenti***
  - σ_{ij}= 0
  - $p(\mathbf{x})$ è il prodotto della densità univariata per **x** componente per componente.

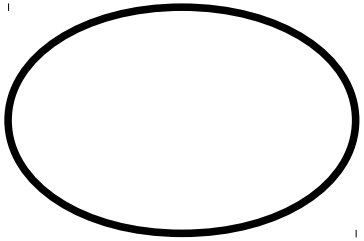
Densità normale multivariata

- La forma della matrice di covarianza porta a diverse distribuzioni normali

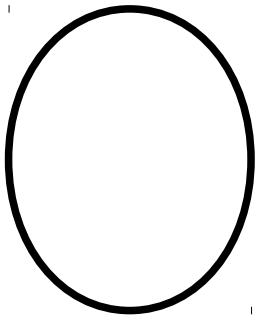
$$\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

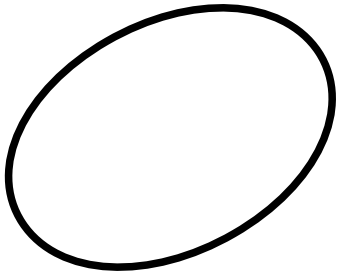


$\sigma_1 > \sigma_2$



$\sigma_2 > \sigma_1$

$\Sigma =$ matrice qualsiasi



Classificatori,
funzioni discriminanti
e superfici di separazione

Classificatori, funzioni discriminanti e superfici di separazione

- Per rappresentare un classificatore spesso si utilizza un insieme di ***funzioni discriminanti*** $g_i(\mathbf{x})$, $i=1\ldots c$

- Il classificatore assegna l'oggetto **x** alla classe ω_i se

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \text{ per ogni } j \neq i$$

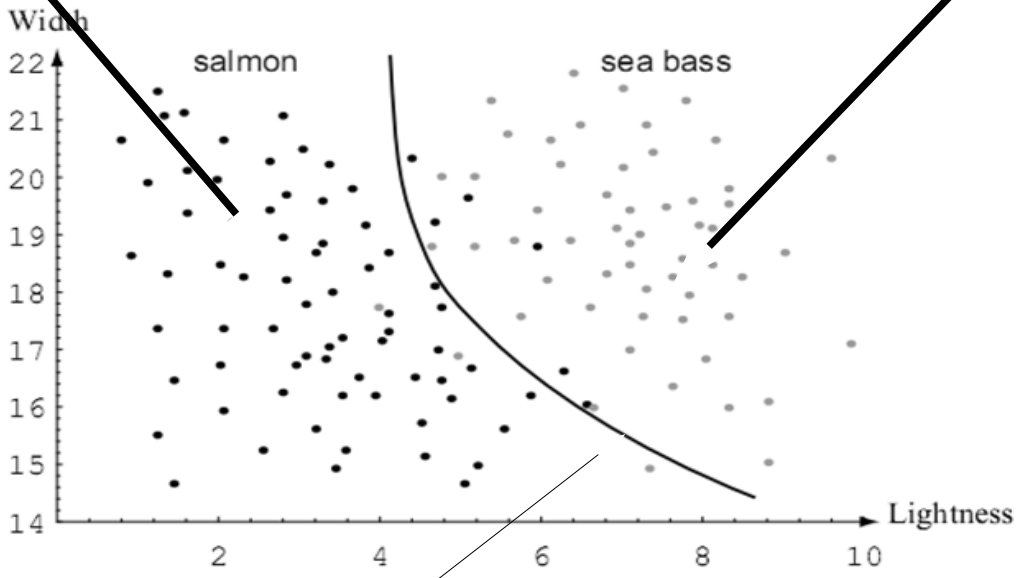
Questo insieme di funzioni suddivide lo spazio delle features in ***Regioni****, separate tra di loro da* ***Confini di decisione*** *(decision boundaries)*

Regione R_1 (tutti i punti che verrebbero classificati come appartenenti alla classe 1)

Punti x per cui $g_1(x) > g_2(x)$

Regione R_2 (tutti i punti che verrebbero classificati come appartenenti alla classe 2)

punti x per cui $g_2(x) > g_1(x)$



Confine di decisione

Punti x per cui $g_1(x) = g_2(x)$

Classificatori, funzioni discriminanti e superfici di separazione

- Un classificatore che utilizza la regola di decisione di Bayes si presta facilmente a questa rappresentazione:

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

Vediamo un esempio con la distribuzione normale...

Funzioni discriminanti per la normale

Occorre calcolare

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

Si assume che ogni classe j sia gaussiana (cioè sia gaussiana la probalità condizionale (likelihood) P(x|**ωj**)

$$p(x|\omega_j) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\right\}$$

(P(x|**ωj**), assieme alla probabilità a priori P(**ωj**), mi permette di calcolare la probabilità a posteriori della classe **ωj**)

Funzioni discriminanti per la normale

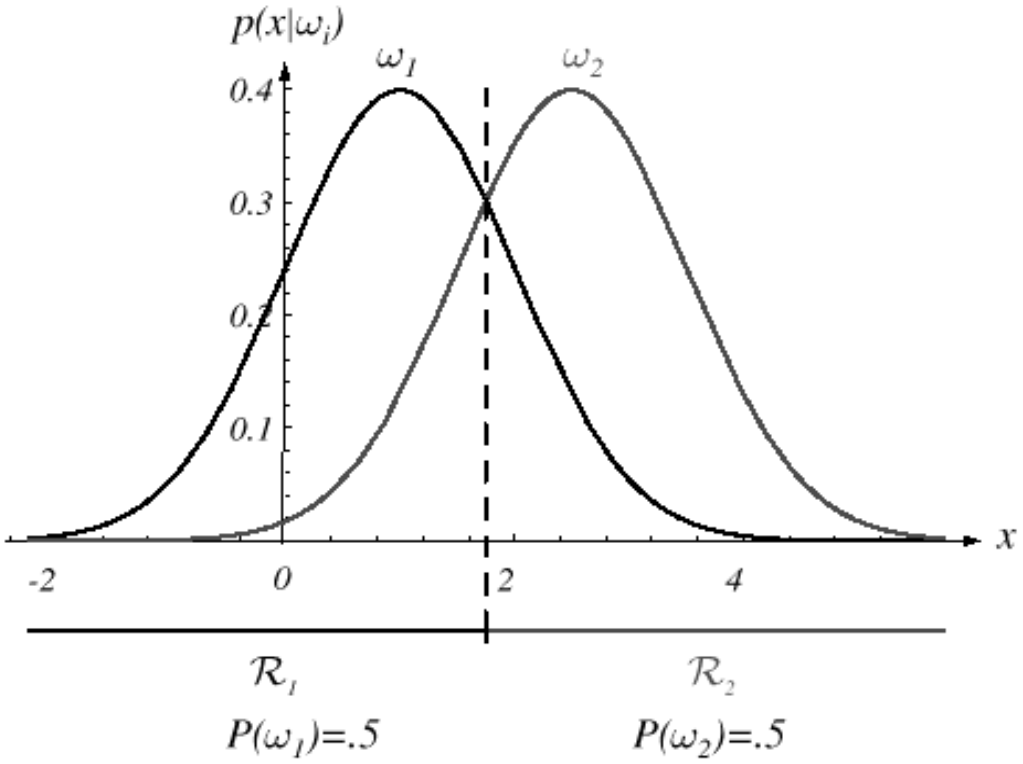
- Alla lavagna:
  - caso generale
  - casi semplificati

Funzioni discriminanti per la normale

Ricapitolando:

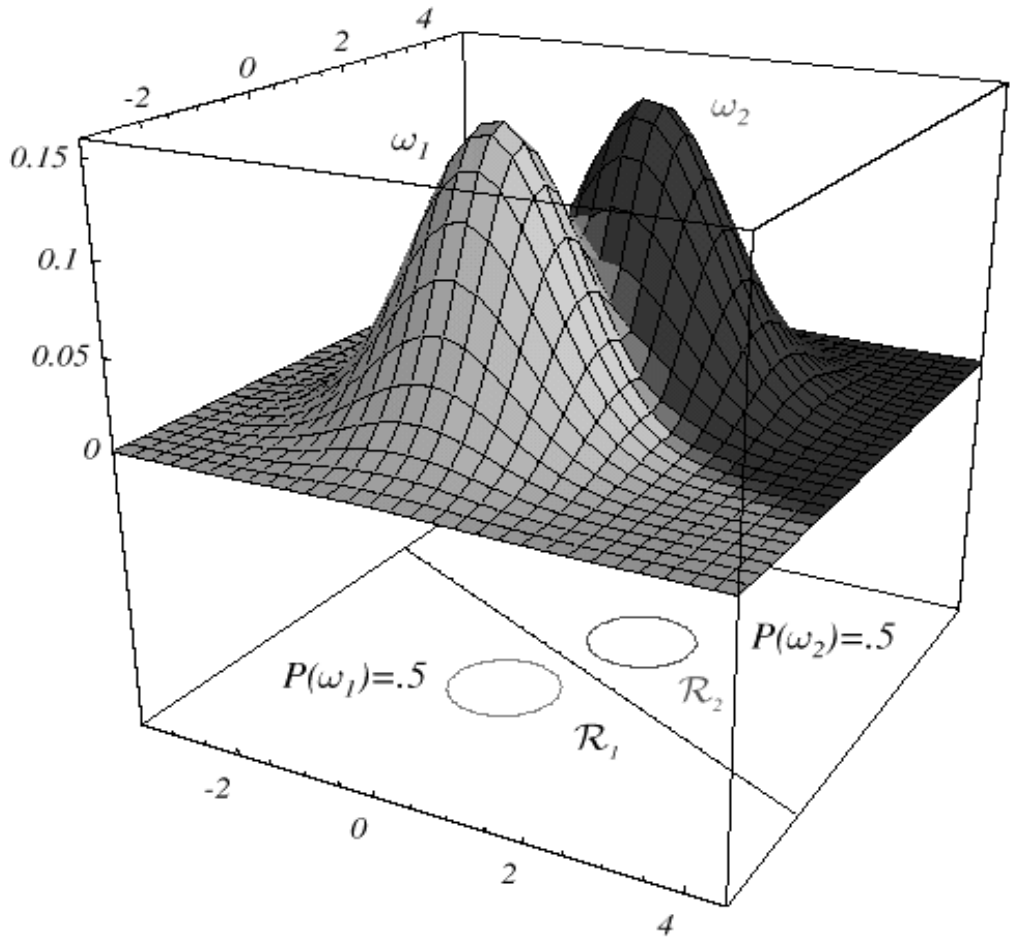
⇨ $\boldsymbol{\Sigma_i = \sigma^2 I}$

$$g_i(x) = -\frac{\|x-\mu_i\|^2}{2\sigma^2} + \ln P(\omega_i)$$

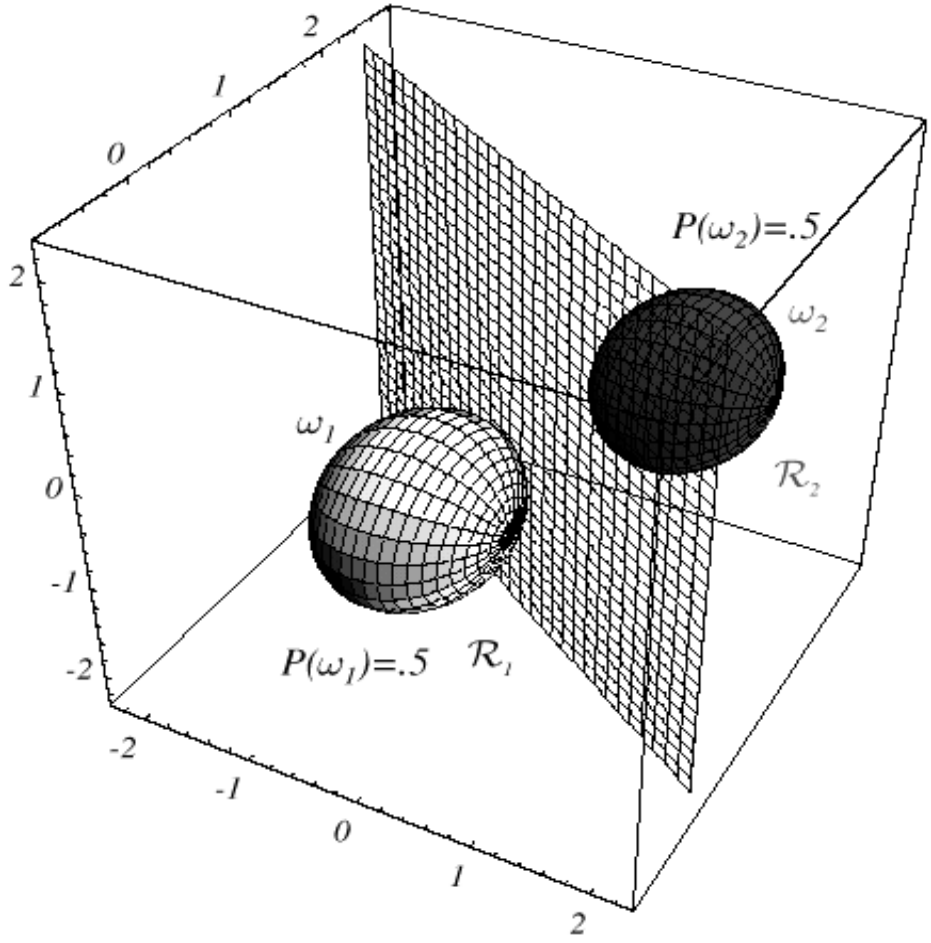


⇨ Il confine di decisione è una retta (un iperpiano)

2-D

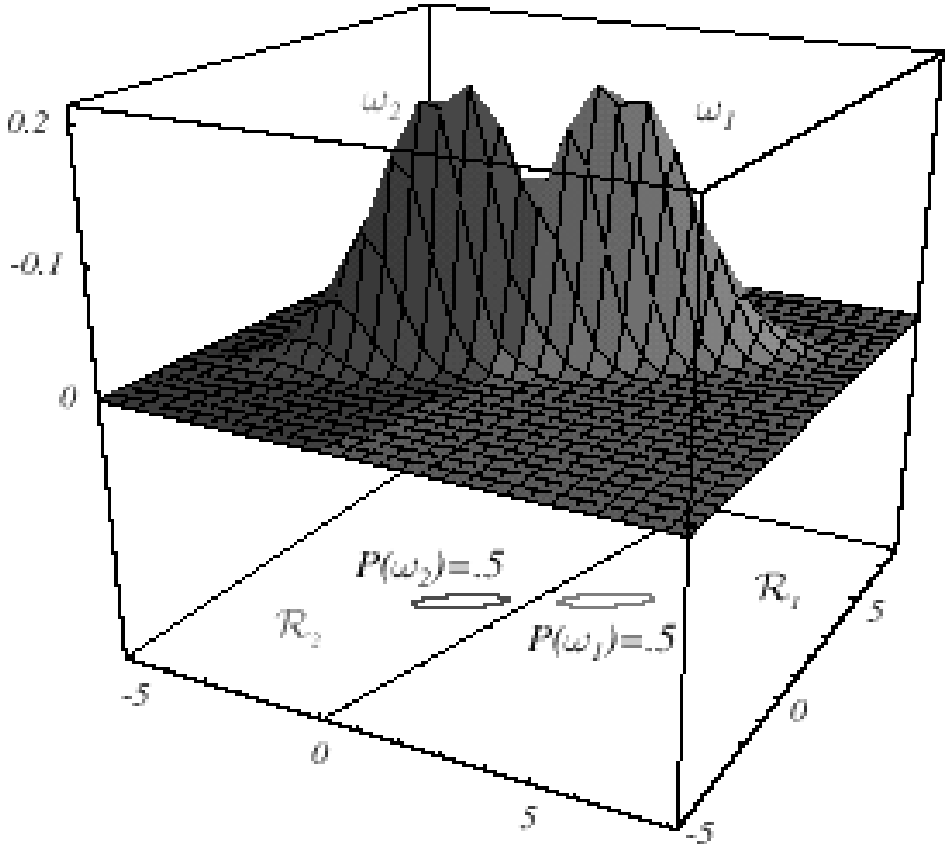


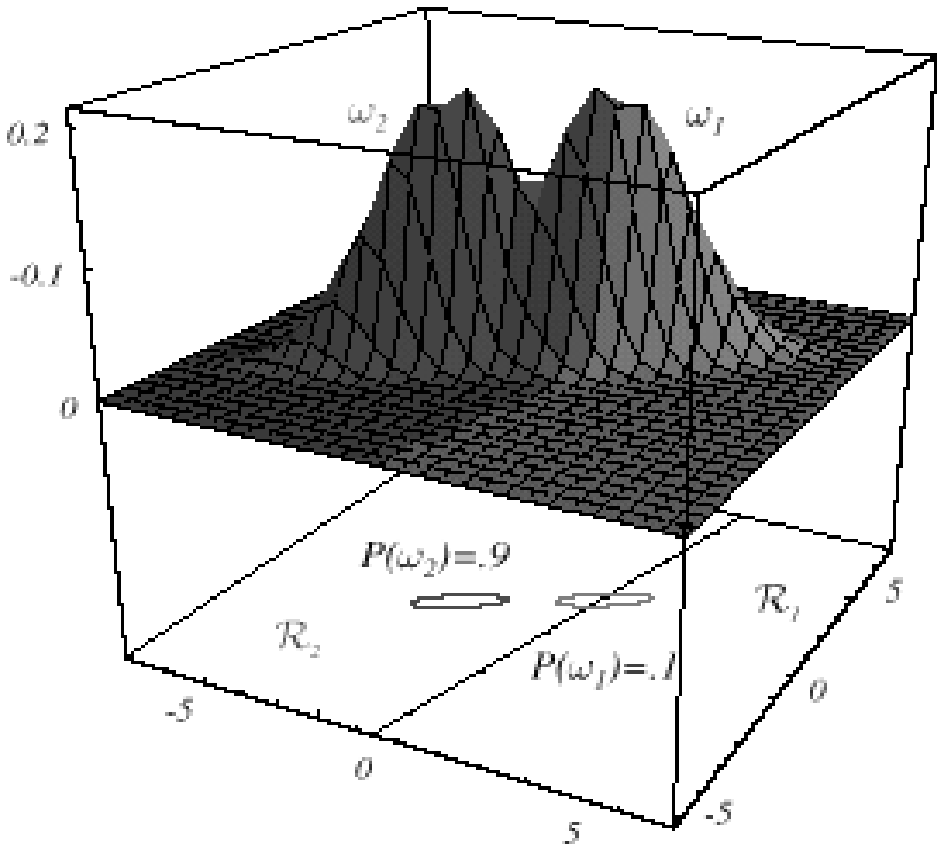
3-D

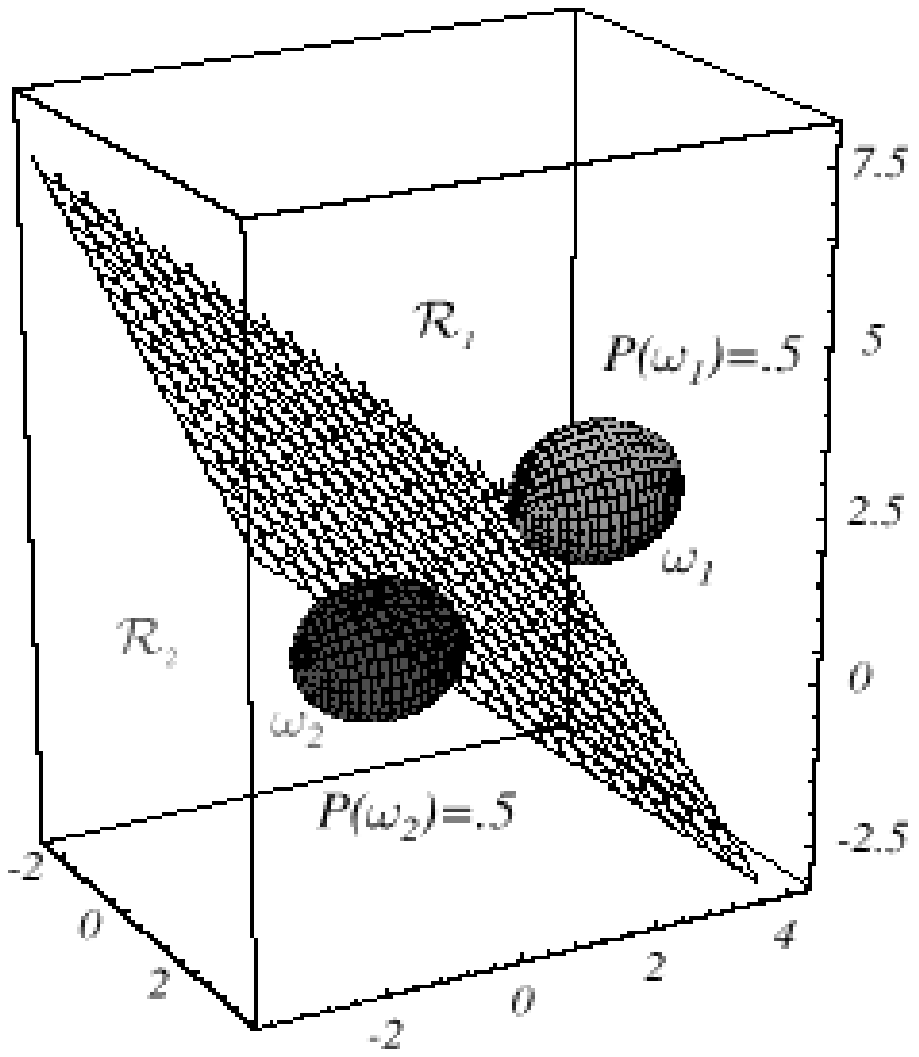


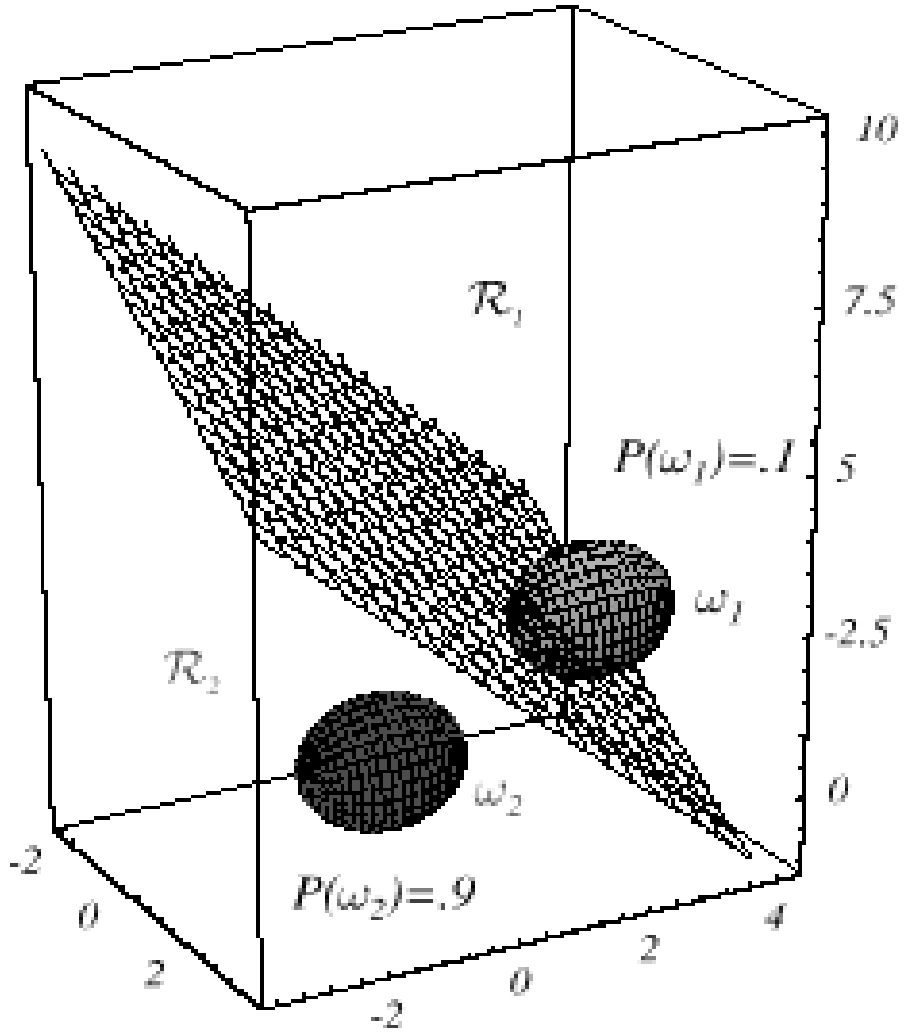
⇨ $\boldsymbol{\Sigma_i = \Sigma}$

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



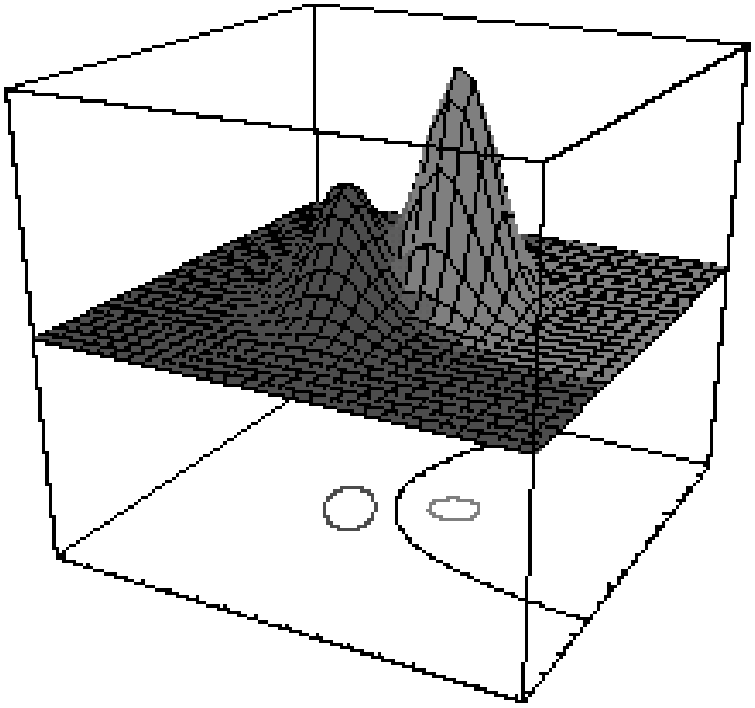


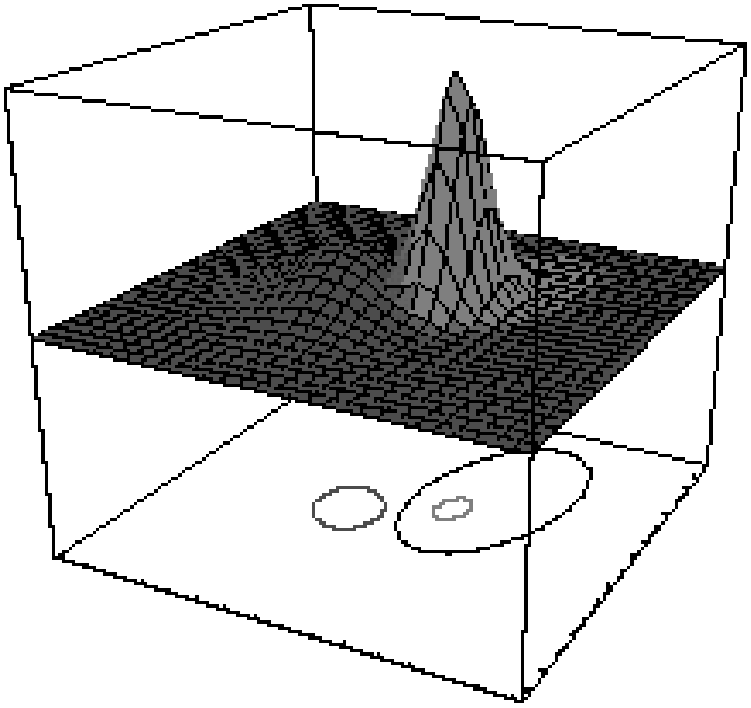

7.5
5
2.5
0
-2.5
$\mathcal{R}_1$
$P(\omega_1)=.5$
ω_1
$\mathcal{R}_2$
ω_2
$P(\omega_2)=.5$
-2
0
2
-2
0
2
4

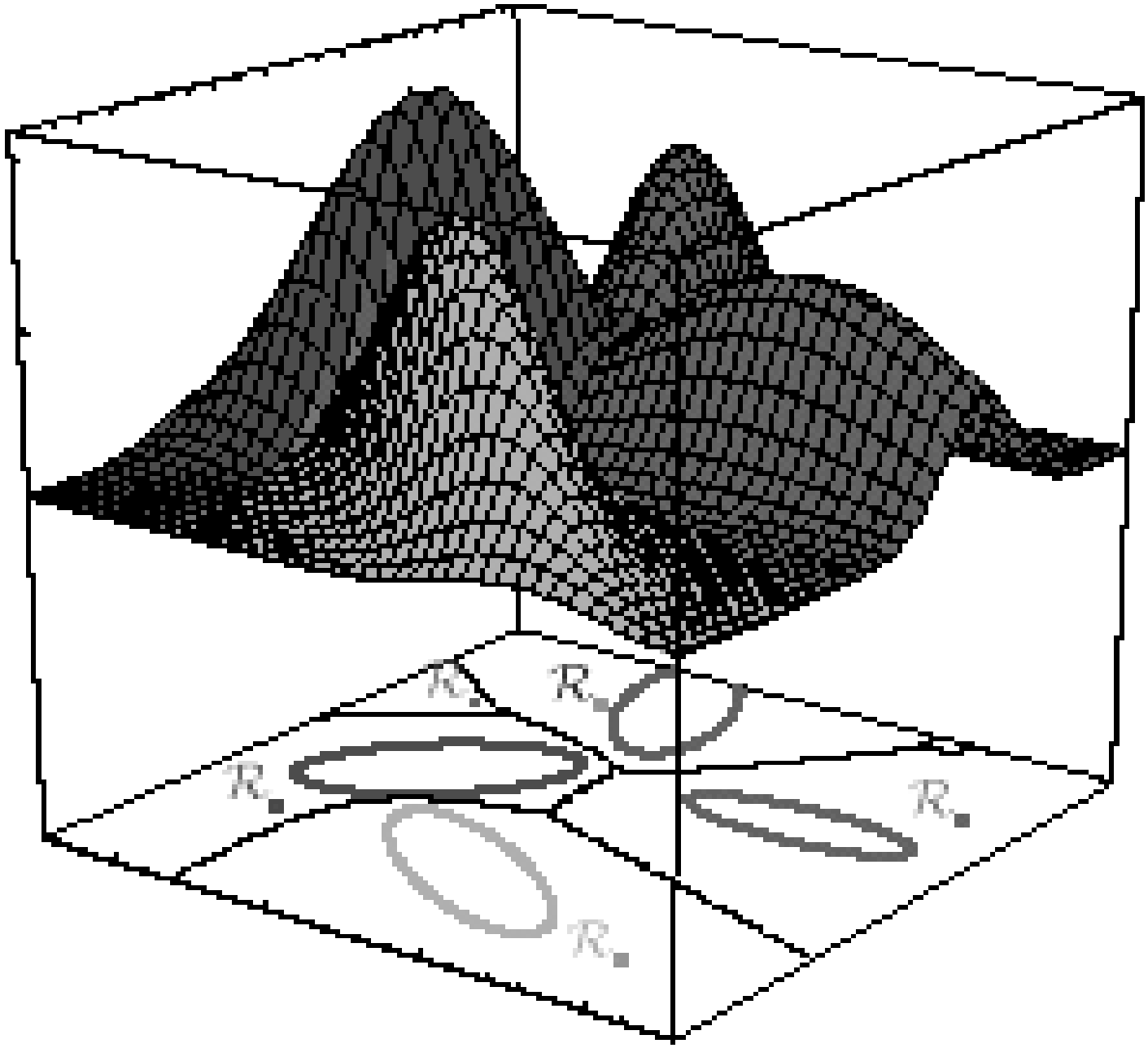

10
7.5
5
-2.5
0
$\mathcal{R}_1$
$P(\omega_1)=.1$
ω_1
$\mathcal{R}_2$
ω_2
$P(\omega_2)=.9$
-2
0
2
-2
0
2
4

3-D

- Caso generico: si può semplificare solo il fattore di pi-greco
  - il confine di decisione assume molte forme





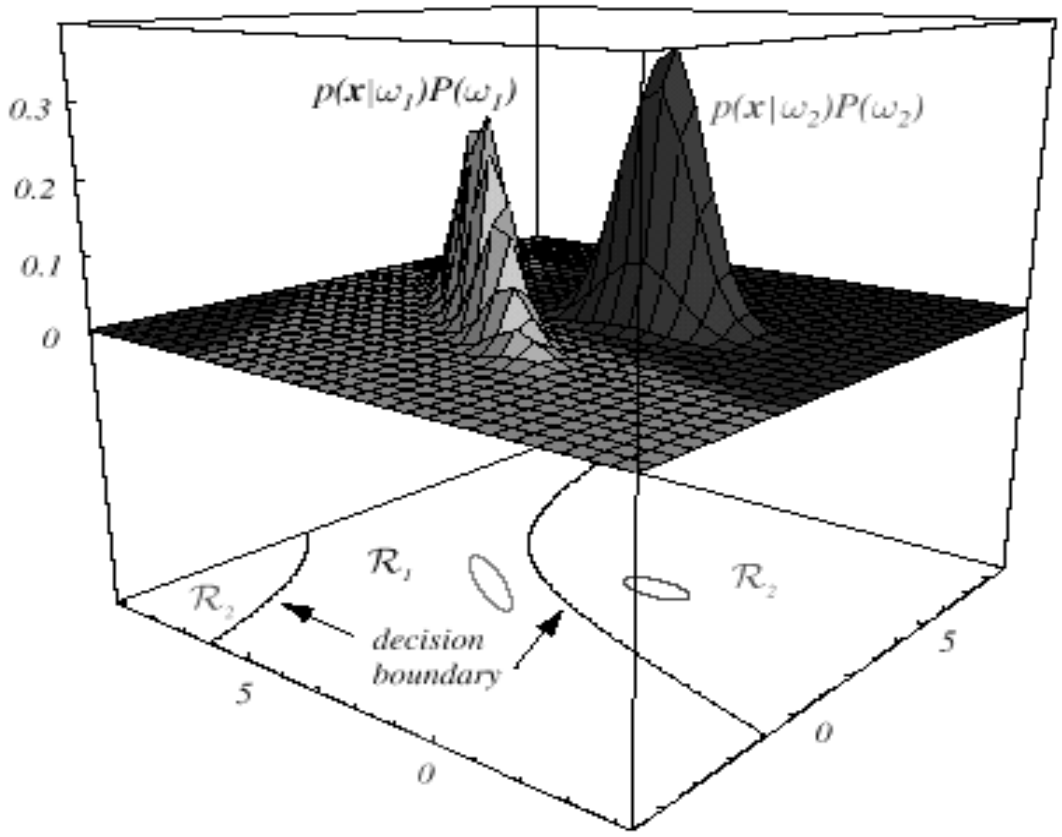

R_1
R_2
R_3
R_4
R_5

Nota finale

- Nel caso a *due* categorie ho due funzioni discriminanti, g1,g2 per cui assegno x a ω1 se g1 > g2

- Posso anche definire una sola funzione discriminante g(x) = g1-g2, e classificare quando g(x) > 0 (g1-g2>0, o g1>g2)

$$g(x) = g_1(x) - g_2(x)$$



ho una sola funzione discriminante!