# Riconoscimento e recupero dell'informazione per bioinformatica

## Teoria della decisione di Bayes

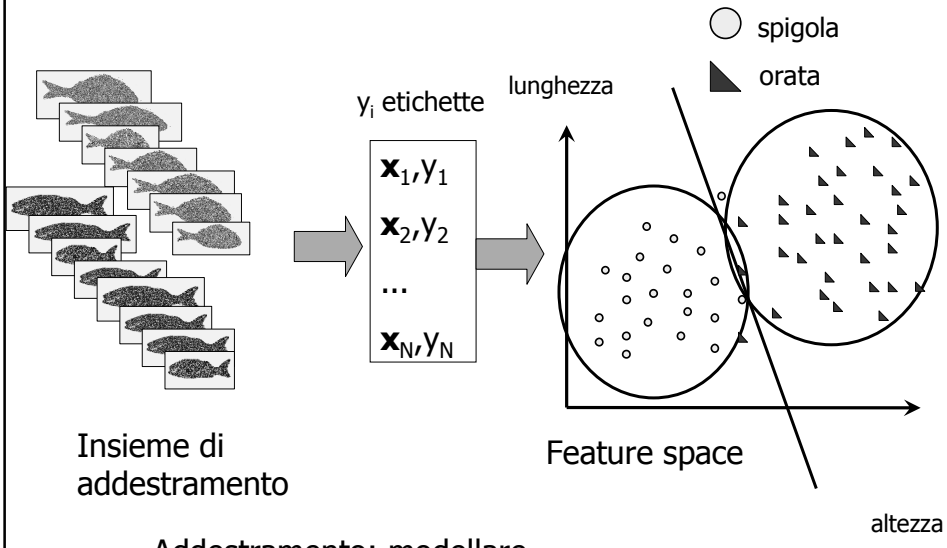
Manuele Bicego

Corso di Laurea in Bioinformatica
Dipartimento di Informatica - Università di Verona

# Sommario

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

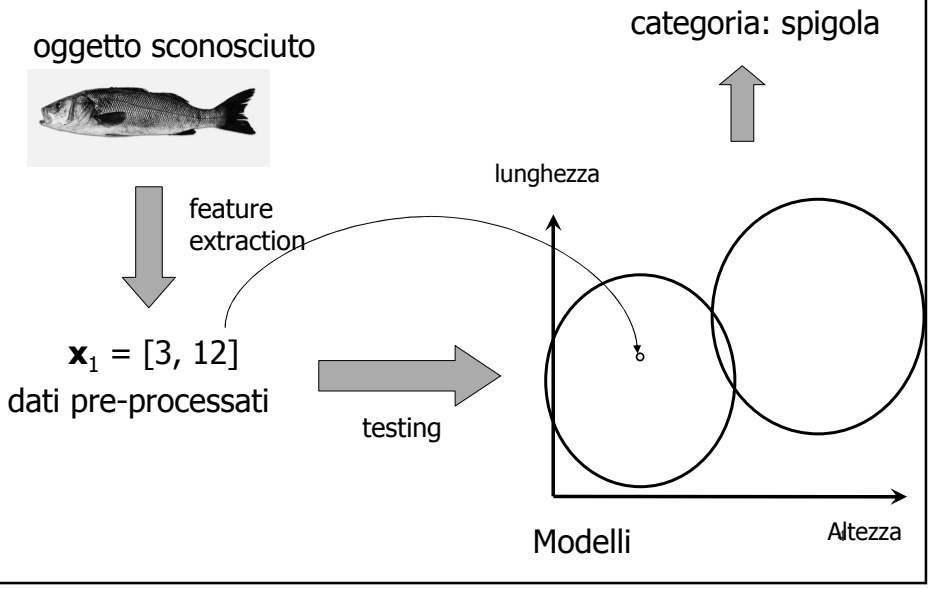
# Esempio: classificazione




Addestramento: modellare (separare) le due classi

3

# Esempio: classificazione/testing



# 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)

5

# 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

6

# Sistema di classificazione

- Concetto di "costo della decisione"
  - quanto costa prendere una decisione sbagliata
- Esempio:
  - oggetto: immagine di un bosco
  - sistema di decisione: deve classificare l'immagine in due classi:
    1. "c'è un incendio"
    2. "non c'è un incendio"
- Costo della decisione: è diverso rilevare un incendio che non c'è o non rilevare un incendio che c'è?

7

# 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 di prima:
  - Azione: nel caso l'oggetto appartenga alla classe 1 ("incendio") viene effettuata la chiamata alle guardie forestali

8

# 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

9

# Teoria della decisione di Bayes

## La teoria della decisione di Bayes

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

11

# Introduzione

- ⇨ Approccio statistico fondamentale di classificazione di pattern
- ⇨ 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 le probabilità ed i costi ad esse associati;

12

# Scenario

- ⇨ Sia **ω** lo stato di natura da descrivere probabilisticamente;
  - ⇨ stato di natura = stato del sistema, classe, categoria,...
- ⇨ **ω** rappresenta le varie classi
  - ⇨ Problema a due classi: **ω** può essere **ω1** o **ω2**
- ⇨ Devo inferire la regola di classificazione (o regola di decisione).
  - ⇨ Dato un oggetto **x**, **ω1** oppure **ω2**?
- ⇨ Quantità che si possono utilizzare:
  - ⇨ Probabilità a priori
  - ⇨ Probabilità condizionale (o likelihood)
  - ⇨ Probabilità a posteriori (regola di Bayes)

13

# Probabilità a priori

- ⇨ Probabilità a priori:
  - ⇨ probabilità P(**ω**): rappresenta la probabilità dello stato nota a priori (senza aver osservato nulla del sistema)
  - ⇨ Esempio: due classi **ω1** and **ω2** per cui sono note
    - ⇨ P(**ω= ω1**) = 0.7
    - ⇨ P(**ω= ω2**) = 0.3
- ⇨ Regola di decisione:
  - ⇨ Decidi **ω1** se P(**ω1**) > P(**ω2**); altrimenti decidi **ω2**
- ⇨ Ovviamente è un sistema limitato:
  - ⇨ Più che decidere, indovino lo stato di natura.

14

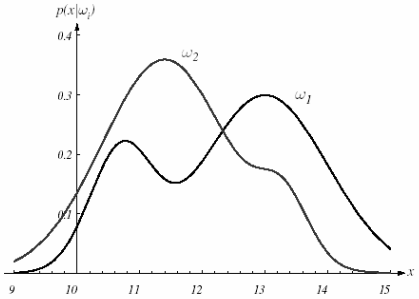
# Probabilità condizionale

⇨ 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 è **ωj**




15

# Probabilità condizionale

⇨ Osservazione: fissata la misurazione x, più è alta p(x|**ωj**) più è probabile che **ωj** sia lo stato giusto

⇨ Regola di decisione (maximum likelihood)

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

⇨ Migliore della regola precedente, ma non tiene conto dell'eventuale conoscenza a priori

16

# Esempio

- ⇨ Discriminare tra calciatore professionista / non calciatore professionista
- ⇨ osservazione x = "stipendio"
- ⇨ Problema: dato uno stipendio x, si deve decidere se è un calciatore o no

- ⇨ Approccio 1: Prob. a priori
  - ⇨ La conoscenza a priori che si ha sul problema dice che la probabilità che una persona sia un calciatore professionista è molto bassa (1%)
  - ⇨ P(**ω= ω1**) = 0.01, P(**ω= ω2**) = 0.99
  - ⇨ Data una persona, e dato il suo stipendio x, si classifica sempre come **ω1**
  - ⇨ Approccio chiaramente limitato

17

# Esempio

- ⇨ Approccio 2: maximum likelihood
  - ⇨ si conosce lo stipendio x. Si sa che un calciatore professionista ha uno stipendio x molto elevato. Si può modellare la probabilità condizionale p(x|**ω1**) e p(x|**ω2**)
  - ⇨ Data una persona, e dato il suo stipendio x, si decide la classe guardando il massimo tra p(x|**ω1**) e p(x|**ω2**)
    - ⇨ probabilmente, se uno ha uno stipendio alto viene classificato come calciatore
  - ⇨ Approccio limitato perché non tiene conto del fatto che pochissime persone sono calciatori professionisti

- ⇨ SOLUZIONE:
  - ⇨ Regola di Bayes: mette assieme probabilità a priori e probabilità condizionale

18

# La regola di Bayes

⇨ (alla lavagna)

19

# Regola di Bayes

Ricapitolando:

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

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!

20

# Regola di Bayes

⇨ 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 $p(x|\omega 1)P(\omega 1) > p(x|\omega 1)P(\omega 1)$, $\omega 1$ altrimenti

Problema principale: le probabilità non sono note, ma occorre stimarle dal training set

21

# Stima delle probabilità

⇨ Stime parametrica: 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

22

# Estensione della teoria di decisione di Bayes

## 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 diventa **x**=$\{x_1,x_2,...,x_d\}$ con $R^d$ spazio delle features

⇨ Più di due stati di natura (classi o categorie)

⇨ **ω1, ω2** diventano {**ω1, ω2,..., ωc**}

⇨ Azioni diverse (associate alla scelta delle classi)

⇨ $\{\alpha_1, ..., \alpha_a\}$ (azioni, ad esempio il rigetto di classificazione)

⇨ Una **funzione di costo** più generale della probabilità di errore, ossia una funzione $\lambda(\alpha_i|\boldsymbol{\omega j})$ che descrive il costo (o la perdita) dell'intraprendere l'azione $\alpha_i$ quando la classe è **ωj**

# Estensione della teoria di decisione di Bayes

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

$$P(\omega_j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_j) P(\omega_j)}{p(\mathbf{x})} \quad , \mathbf{x} = \{x_1, x_2, ..., x_d\}, \mathbf{x} \in \mathrm{R}^d$$

⇨ Supponiamo di osservare un particolare **x**, e decidiamo di effettuare l'azione $\alpha_i$: per definizione, saremo soggetti alla perdita $\lambda(\alpha_i|\boldsymbol{\omega j})$. Poiché **ωj** non si conosce, la perdita attesa (o ***rischio***) *associata a questa decisione* sarà:

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

25

# Estensione della teoria di decisione di Bayes

⇨ 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* $\alpha(\mathbf{x})$:

$$\alpha(\mathbf{x}) \rightarrow \alpha_i \, , \alpha_i \in \{\alpha_1, \alpha_2, ..., \alpha_a\}$$

⇨ tale che $R(\alpha_i|\mathbf{x})$ sia minimo

26

# Estensione della teoria di decisione di Bayes

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

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

⇨ Chiaramente, se $\alpha(\mathbf{x})$ viene scelto in modo che $R(\alpha_i|\mathbf{x})$ sia il minore possibile per ogni $\mathbf{x}$, il rischio complessivo viene minimizzato.

27

# Estensione della teoria di decisione di Bayes

⇨ Quindi, la regola di decisione di Bayes estesa è

1. Calcola $R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathbf{x})$
2. Scegli l' azione $i^* = \min_i R(\alpha_i \mid \mathbf{x})$

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

28

# Caso 2 categorie

⇨ regola che minimizza il rischio nel caso di due classi

⇨ (alla lavagna)

29

# Caso 2 categorie

⇨ Ricapitolando: tre forme

⇨ Posterior

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

⇨ Likelihood e prior

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

⇨ Likelihood ratio

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

30

# Chiusura del cerchio

⇨ Tornando ai problemi di classificazione

⇨ le azioni $\alpha$ i rappresentano la decisione "lo stato giusto è **ωj"**

⇨ In questo caso la funzione di perdita è chiamata "perdita 0-1"

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

⇨ Rischio corrispondente:

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

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

31

# Chiusura del cerchio

⇨ 1-P(**ωj**|x) rappresenta la probabilità di errore

⇨ Minimizzando il Rischio, si minimizza la probabilità di errore

⇨ Minimizzare 1-P(**ωj**|x) per ogni x implica scegliere la j che massimizzi P(**ωj**|x)

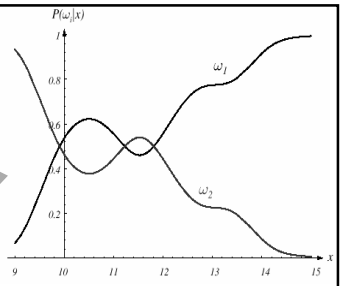
⇨ Si torna alla regola di Bayes (che in questo caso si chiama "Classificazione Minimum Error Rate")

32

# Riassunto

Formula di Bayes

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



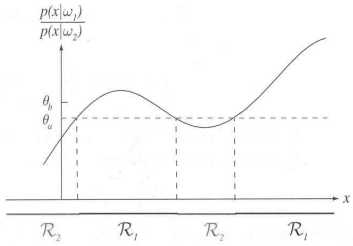

Regola di decisione di Bayes:
Decidi $\omega_1$ se $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$, $\omega_2$ altrimenti

$p(\mathbf{x}|\omega_1)P(\omega_1)$ $p(\mathbf{x}|\omega_2)P(\omega_2)$

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

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

*e permette di minimizzare il rischio!*




*Mettendo a rapporto le likelihood ho*

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

*in cui può essere* (Minimum Error Rate)

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

*da cui mi ricollego alla regola iniziale!*

33

# Classificatori, funzioni discriminanti e superfici di separazione

# Classificatori, funzioni discriminanti e superfici di separazione

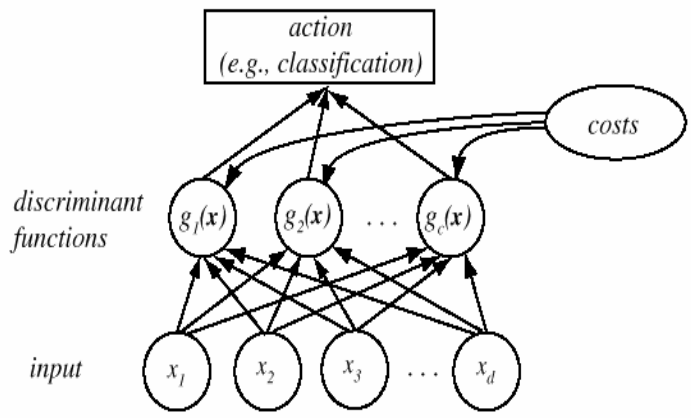
- Uno dei vari metodi per rappresentare classificatori di pattern consiste in un set di ***funzioni discriminanti*** $g_i(\mathbf{x})$, $i=1...c$
- Il classificatore assegna il vettore di feature **x** alla classe $\omega_i$ se

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

- Un classificatore di Bayes si presta facilmente a questa rappresentazione:
  - Rischio generico $g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x})$
  - Minimum Error Rate $g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x})$

35

- Un tale classificatore può essere considerato come una rete che calcola c funzioni discriminanti e sceglie la funzione che discrimina maggiormente




36

- ⇨ 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
  - ⇨ quindi utilizziamo quelle
- ⇨ Esempio: minimum error rate

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

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

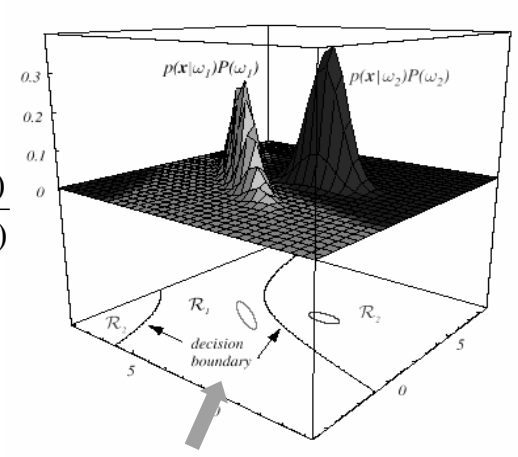
$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i),$$

37

- ⇨ L'effetto di ogni decisione è quello di *dividere lo spazio delle features* in c ***superfici di separazione*** o ***decisione***, $R_1$, ..., $R_c$
- ⇨ Le regioni sono separate con ***confini di decisione***, linee descritte dalle massime funzioni discriminanti.
- ⇨ Nel caso a *due* categorie ho due funzioni discriminanti, g1,g2 per cui assegno x a $\omega$1 se g1 > g2 o g1-g2>0

38

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$$
$$g(\mathbf{x}) = P(\omega_1 \mid \mathbf{x}) - P(\omega_2 \mid \mathbf{x})$$
$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$



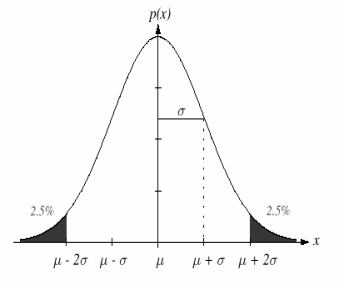

*ho una sola funzione discriminante!*

39

# La distribuzione normale
# Le funzioni discriminanti per la distribuzione 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".

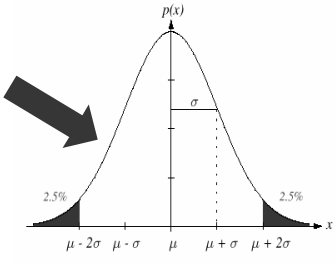



41

# La densità normale univariata

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

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




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

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

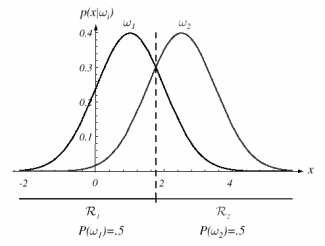
42

# Densità normale multivariata

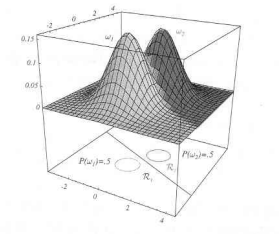
⇨ La generica densità normale multivariata a d dimensioni si presenta nella forma

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

$d$=1




$d$=2



43

# 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

$$\Sigma = \mathrm{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t\right] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t \, p(\mathbf{x}) d\mathbf{x}$$

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

44

# Densità normale multivariata

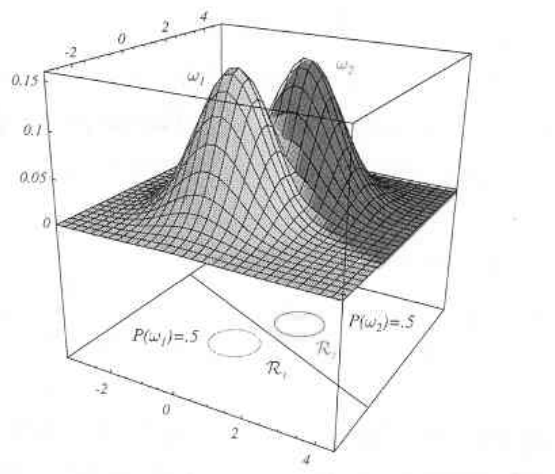
⇨ Caratteristiche della matrice di covarianza
- ⇨Simmetrica
- ⇨Semidefinita positiva ($|\Sigma| \geq 0$)
- ⇨ $\sigma_{ii}$= varianza di $x_i$ (= $\sigma_i^2$ )
- ⇨ $\sigma_{ij}$= covarianza tra $x_i$ e $x_j$

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

45

# Densità normale multivariata

⇨ Forma della matrice di covarianza (alla lavagna)




46

# Funzioni discriminanti per la normale

- Si assuma che le classi siano modellate da distribuzioni normali (stima parametrica delle pdf)
- Vediamo la forma della funzione discriminante nel caso di Minimum Error Rate (formula con la likelihood e prior e il logaritmo)

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

47

# Funzioni discriminanti per la normale

- Alla lavagna:
  - caso generale
  - casi semplificati

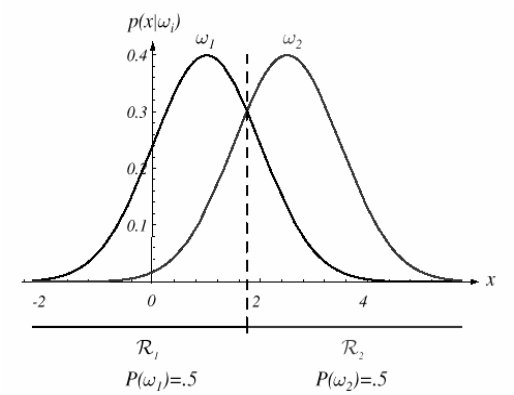
48

# Funzioni discriminanti per la normale

Ricapitolando:

⇨$\Sigma_i=\sigma^2\mathbf{I}$

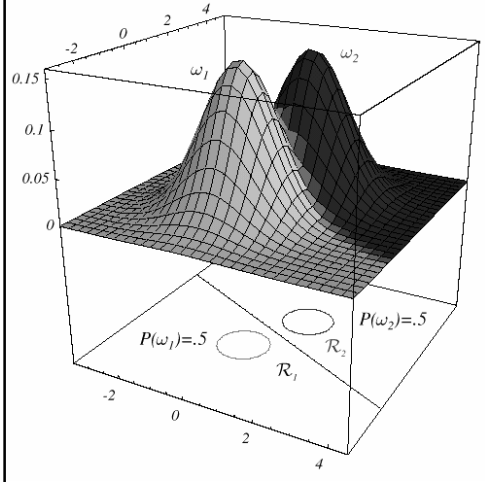
$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x}-\boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i)$$



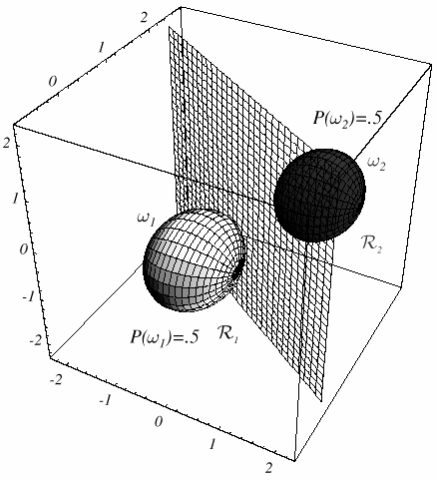

49

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

2-D



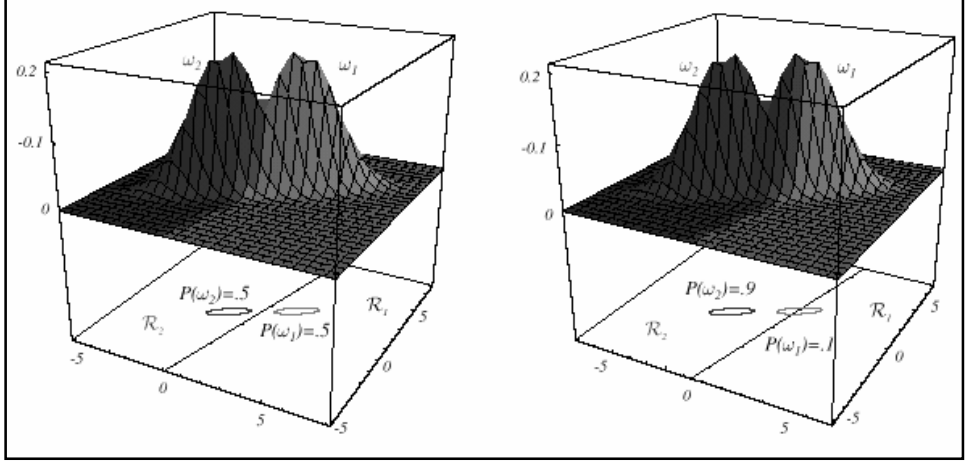

3-D

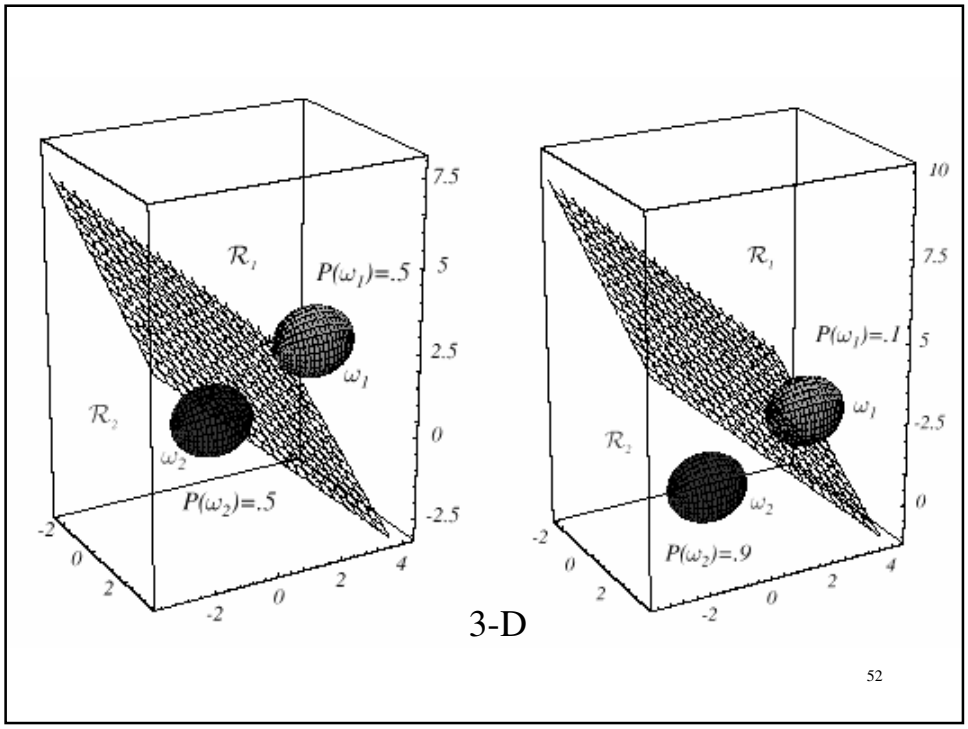



50

⇨$\Sigma_i = \Sigma$

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



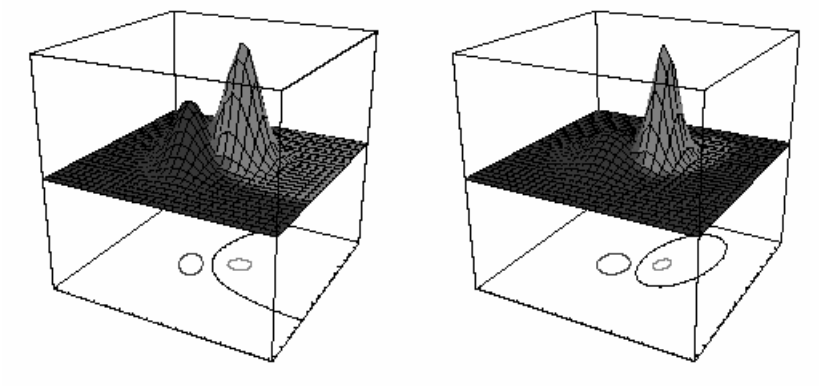




3-D

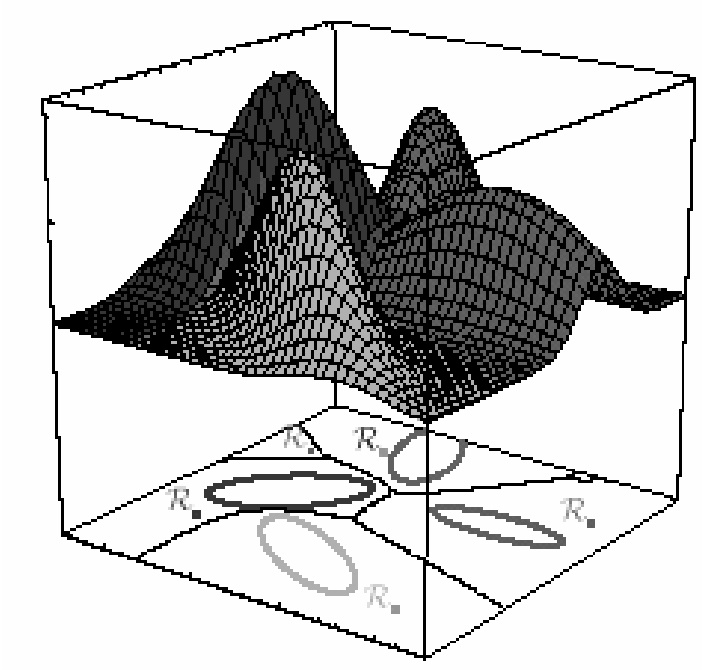
52

⇨ Caso generico: si può semplificare solo il fattore di pi-greco

⇨ il confine di decisione assume molte forme



53




54