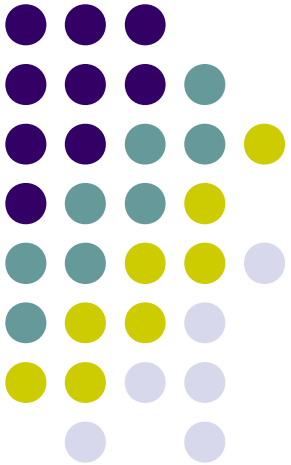
Dalla descrizione alla classificazione

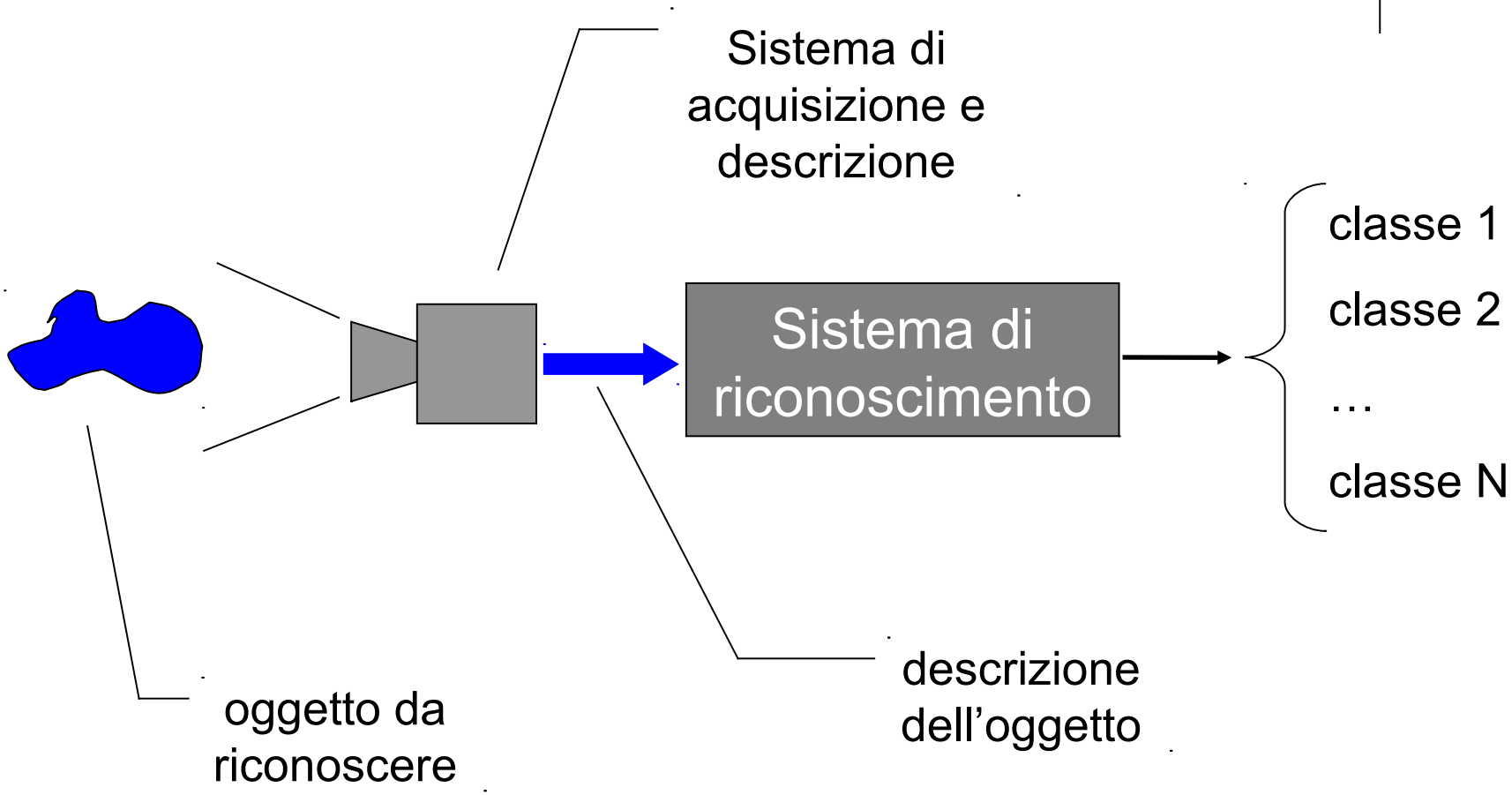
Il problema della classificazione

Teoria bayesiana della decisione

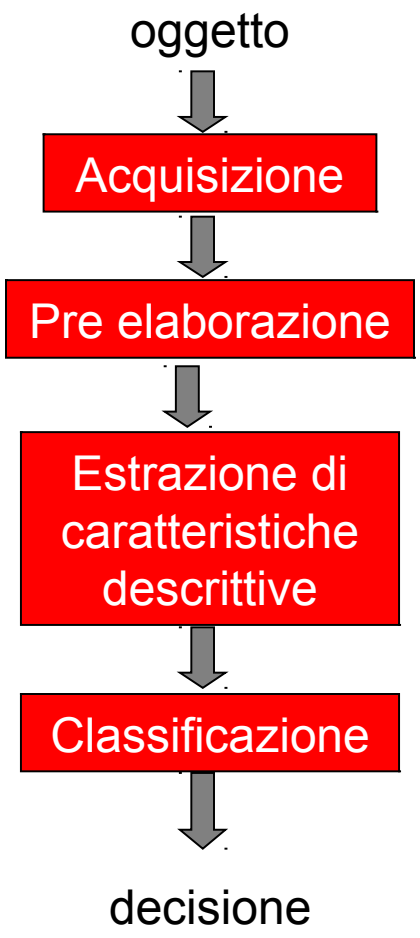
L'approccio Nearest-Neighbor

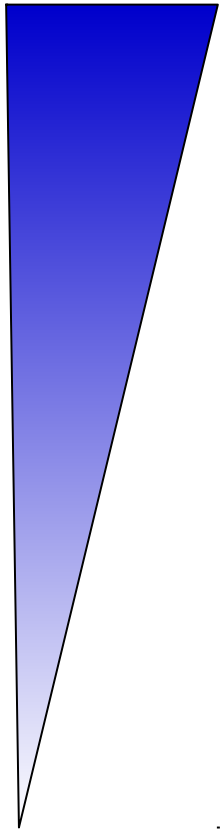


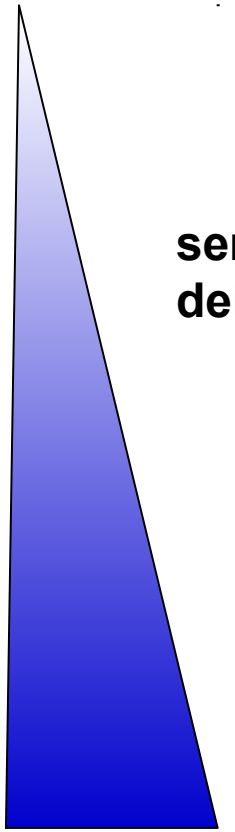
Schema ideale di un sistema di riconoscimento



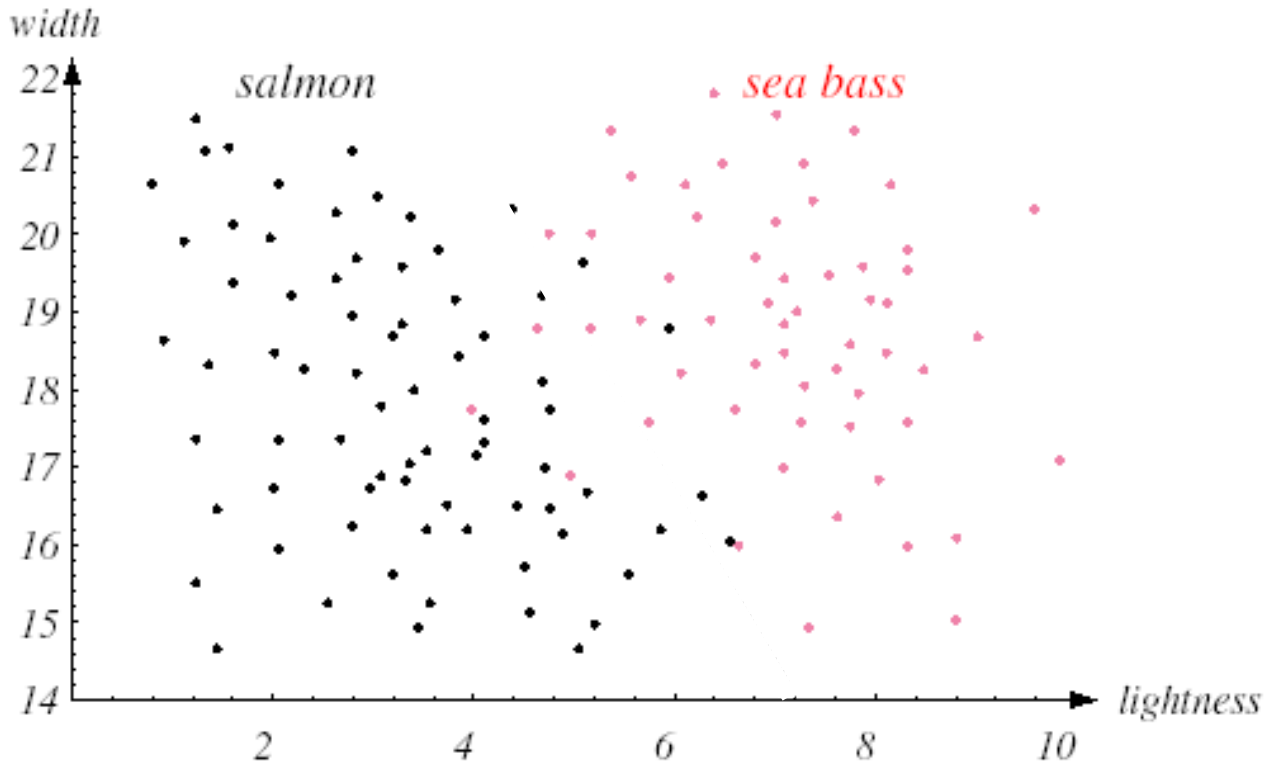
Un processo complesso





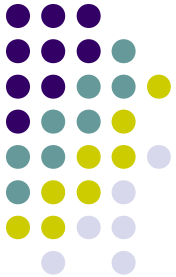


Spazio delle features



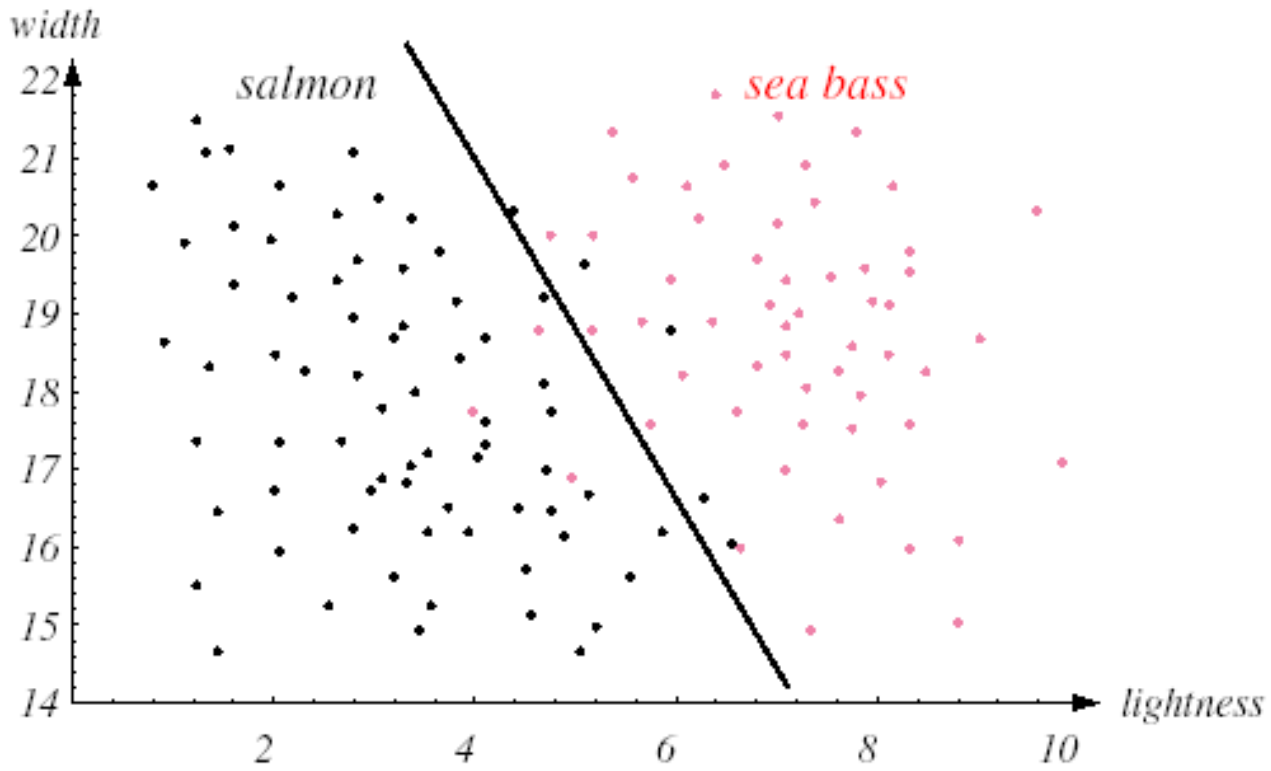
In questo caso si sono scelte due features. Lo spazio delle features è un piano ed ogni campione corrisponde a un feature vector $[x_1\ x_2]^T$.

Regioni di decisione



- Il problema ora è dividere lo spazio delle features in regioni, ognuna delle quali sia ascrivibile ad una delle classi note.
- Si identificano così delle *regioni di decisione* (*decision regions*), separate da una frontiera (*decision boundary*).
- In questo modo è possibile decidere a quale classe assegnare il campione sulla base della posizione del punto nel feature space.

Regioni di decisione

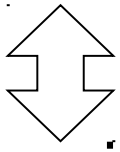


La scelta più immediata è quella di una frontiera semplice, lineare. Gli errori complessivi sono minori rispetto al caso di una sola feature, ma sono comunque presenti.

Si può dare di più ?

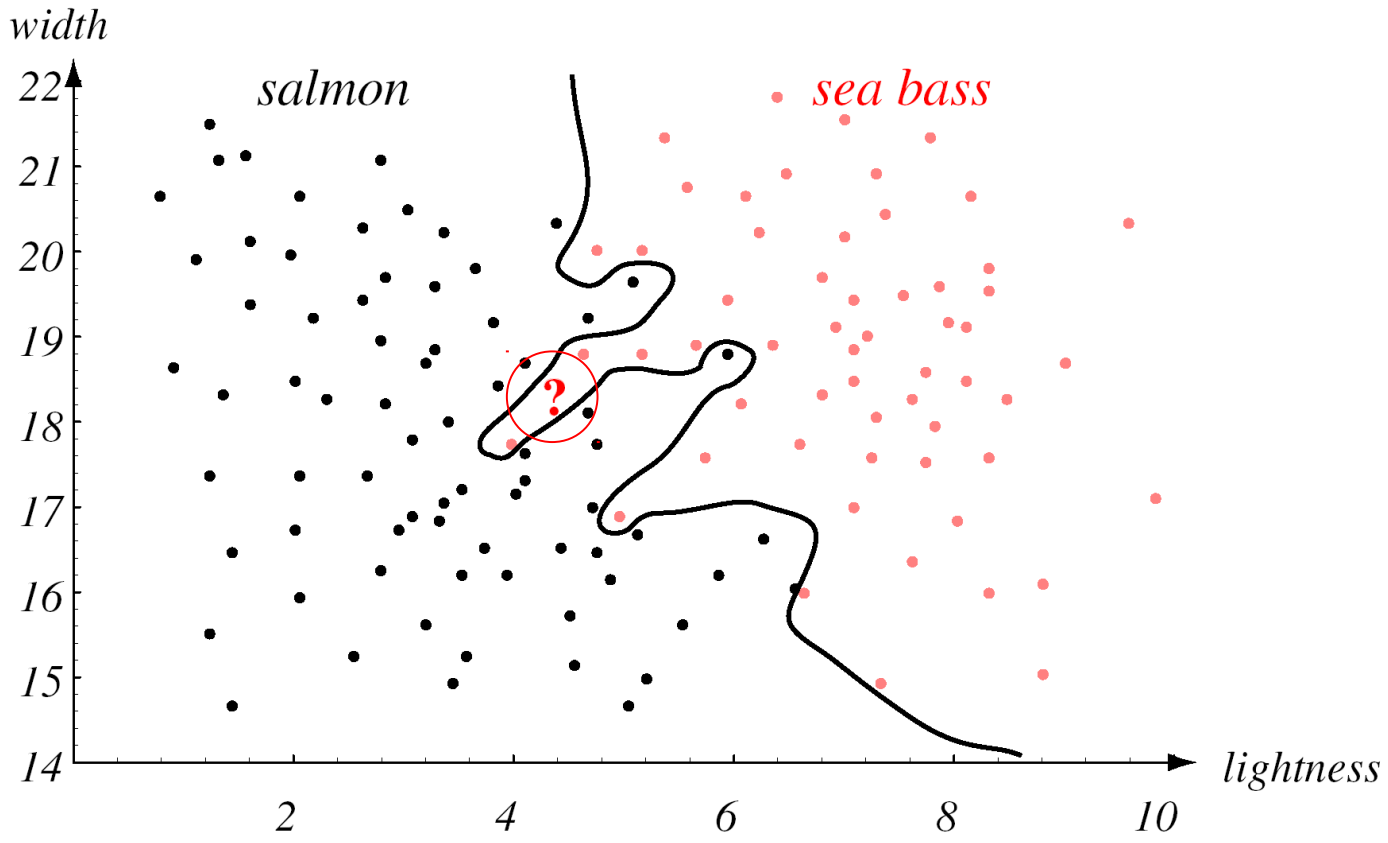
- Le regioni definite in base alla frontiera lineare implicano ancora degli errori.
- Sarebbe possibile eliminare del tutto gli errori con una frontiera meno semplice ?
- Ricordiamo che la frontiera di decisione è generata dal sistema di classificazione; quindi

frontiera meno semplice



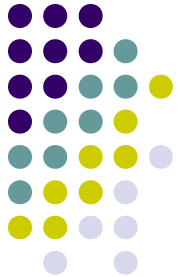
classificatore più complesso

Si può dare di più ?



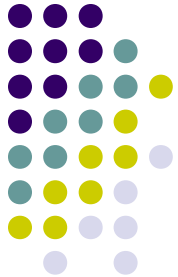
Una frontiera di decisione più complessa che annulla gli errori sul training set. Come sarà classificato il nuovo campione ?

Il problema della generalizzazione



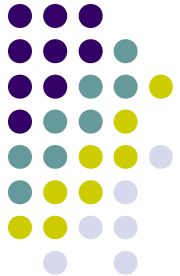
- Con una frontiera complessa è possibile annullare l'errore sul training set.
- Il problema è che in questo modo non si garantisce una buona prestazione del sistema sui campioni che bisognerà classificare in fase operativa.
- Questo aspetto (*generalizzazione*) è fondamentale nella progettazione dell'intero sistema.

Il problema della generalizzazione



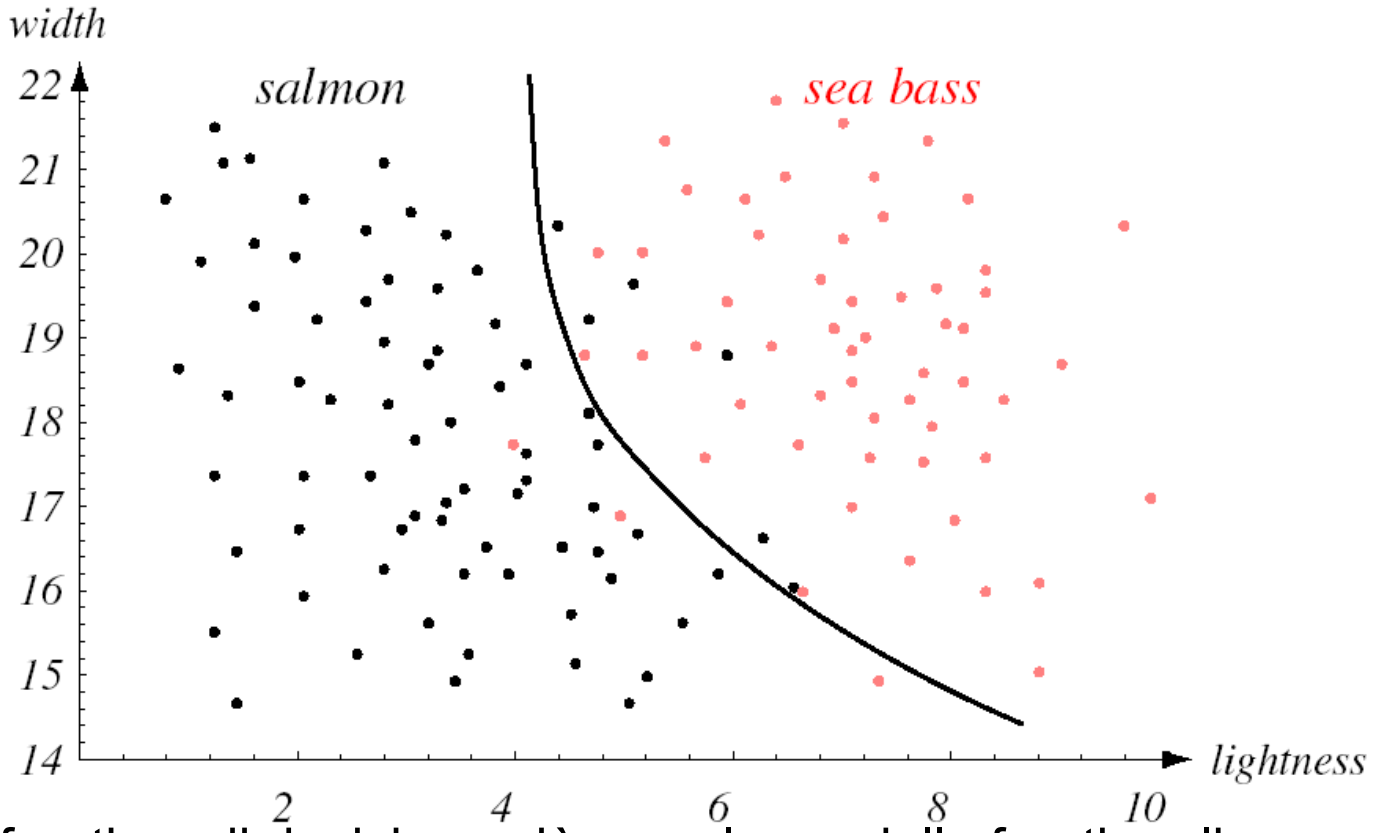
- E' improbabile che un classificatore estremamente complesso garantisca buone capacità di generalizzazione in quanto costruito strettamente sulle caratteristiche dei campioni del particolare training set (e del particolare rumore che si portano dietro).
- Un classificatore efficiente dovrebbe invece essere costruito su caratteristiche più generali che siano valide anche per campioni non appartenenti al training set.

Il problema della generalizzazione



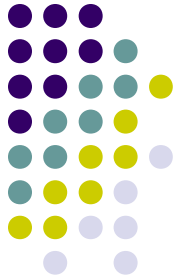
- Si impone quindi di stabilire un compromesso tra:
  - prestazioni del classificatore sul training set
  - capacità di generalizzazione del classificatore (legata alla sua “semplicità”)
- Di conseguenza, è preferibile tollerare qualche errore sul training set se questo porta ad una migliore generalizzazione del classificatore.

Il problema della generalizzazione



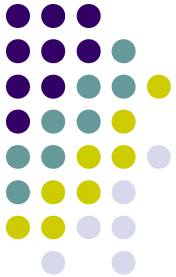
Una frontiera di decisione più complessa della frontiera lineare. Sebbene gli errori sul training set siano ancora presenti, il classificatore sembra garantire una buona capacità di generalizzazione.

Come si costruisce il classificatore ?



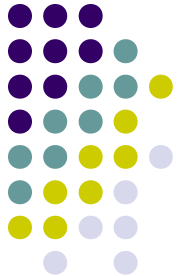
- Appurate le caratteristiche che dovrebbe esibire un classificatore efficiente, si pone il problema della sua costruzione.
- **E' possibile una soluzione algoritmica ?**
  E', cioè possibile, definire un algoritmo per classificare caratteri, parlato, immagini,… ?
- Dopo 40 anni di sforzi in questo senso la risposta è chiaramente negativa.

L'apprendimento



- L'unica alternativa percorribile è quella di apperendere a risolvere i problemi a partire da esempi (*learning by examples*).
- *Apprendimento* (*learning*): ogni metodo che, nella costruzione di un classificatore, combina informazioni empiriche provenienti dall'ambiente e conoscenza a priori del contesto del problema.
- Le informazioni empiriche sono di solito nella forma vista di campioni di esempio (training set).
- Conoscenza a priori: invarianti, correlazioni, …

Paradigmi di apprendimento



Esistono diversi paradigmi di apprendimento:

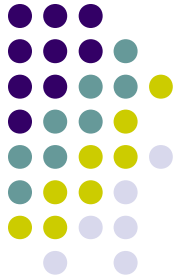
- Apprendimento *supervisionato* (*supervised*):

  per ogni campione del training set è provvista la classe di appartenenza. Obiettivo dell'apprendimento è quello di minimizzare gli errori (o il costo di classificazione).

- Apprendimento *non supervisionato* (*unsupervised*):

  non sono fornite esplicite informazioni sulla classe dei campioni del training set. Obiettivo dell'apprendimento è quello di formare dei *raggruppamenti* (*clusters*) dei campioni generalmente sulla base di una distanza. Spesso è definito dall'esterno il numero dei clusters da produrre. Questo paradigma viene definito anche *clustering*.

Teoria bayesiana della decisione: caratteristiche



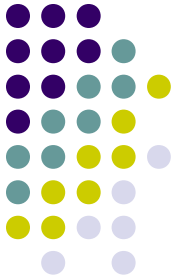
La teoria bayesiana della decisione è un approccio statistico fondamentale all'interno del pattern recognition.

Il suo obiettivo è quello di confrontare quantitativamente diverse decisioni di classificazione utilizzando le probabilità ed i costi cha accompagnano tali decisioni.

Assunzioni fondamentali:

- il problema della decisione è posto in termini probabilistici
- sono noti i valori di tutte le probabilità rilevanti per il problema

Fondamenti



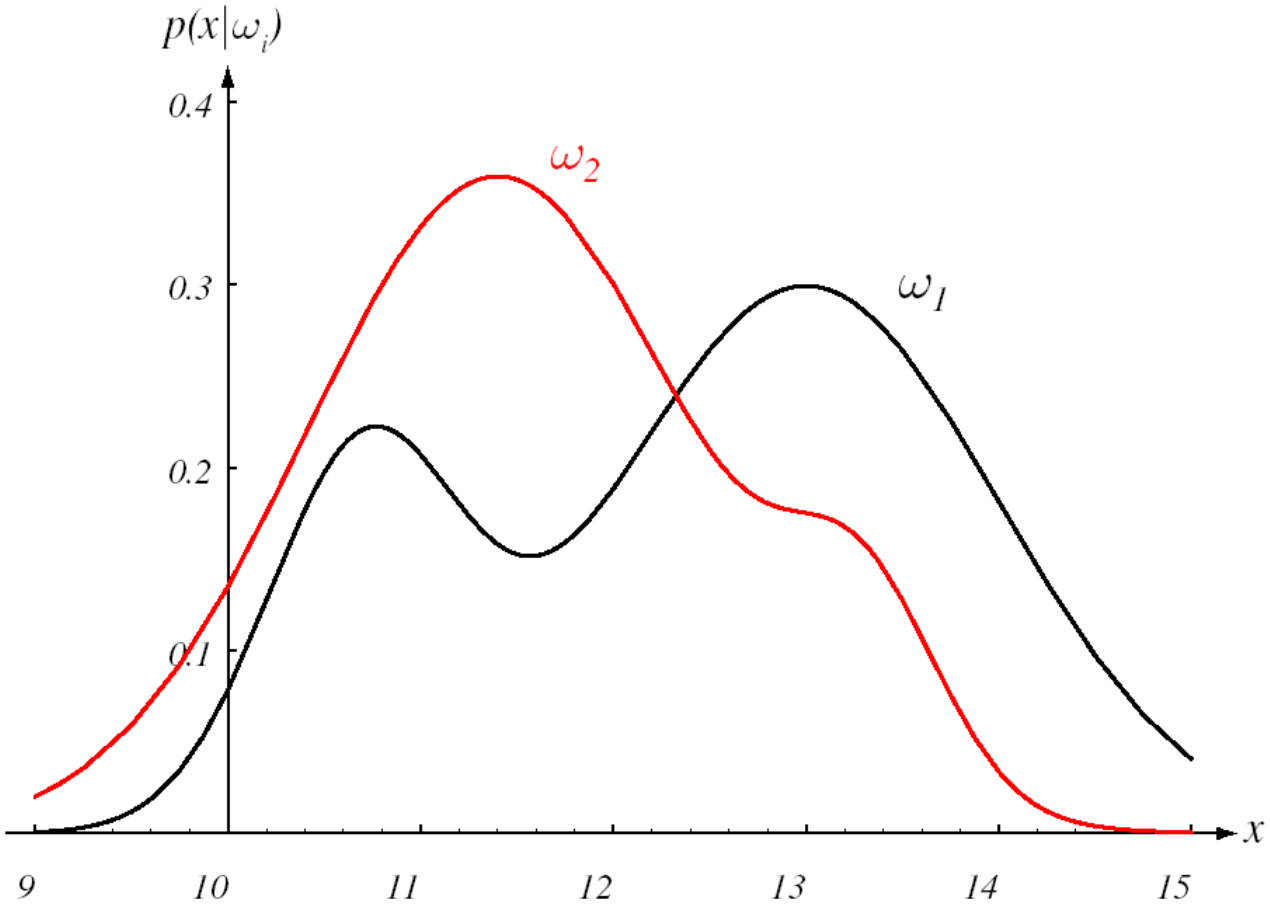
- Consideriamo un problema a C classi, con etichette ω_j con j=1,2,…,C.
- Etichettiamo con α_i i=1,2,…,*a* le decisioni che è possibile prendere.
- Supponiamo di conoscere la probabilità $P(\omega_j)$ che un campione appartenga ad una certa classe (*probabilità a priori*).

Fondamenti

Se non avessimo altre informazioni, la regola di decisione sarebbe basata interamente sulle $P(\omega_j)$.

Supponiamo, invece, di poter utilizzare un feature vector N-dimensionale x che, in questo ambito, è formalizzabile come una variabile aleatoria N-dimensionale.

Conosciamo inoltre la funzione di densità di probabilità condizionata alla classe $p(x|\ \omega_j)$.



Un esempio di densità di probabilità condizionate alle classi con C=2.

Teorema di Bayes

A partire dalle conoscenze descritte, vorremmo stabilire quale sia la probabilità $P(\omega_j|x)$ (*probabilità a posteriori*) che il campione descritto da un feature vector x appartenga alla classe ω_j.

E' possibile ottenere questa informazione grazie al teorema di Bayes per cui:

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

dove

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

Teorema di Bayes

Grazie al teorema di Bayes, è possibile risalire alla probabilità che il feature vector osservato x sia stato prodotto da un campione appartenente alla classe ω_j (prob. a posteriori) a partire dalla probabilità a priori $P(\omega_j)$ e dalle *verosimiglianze* $p(x|\omega_j)$.

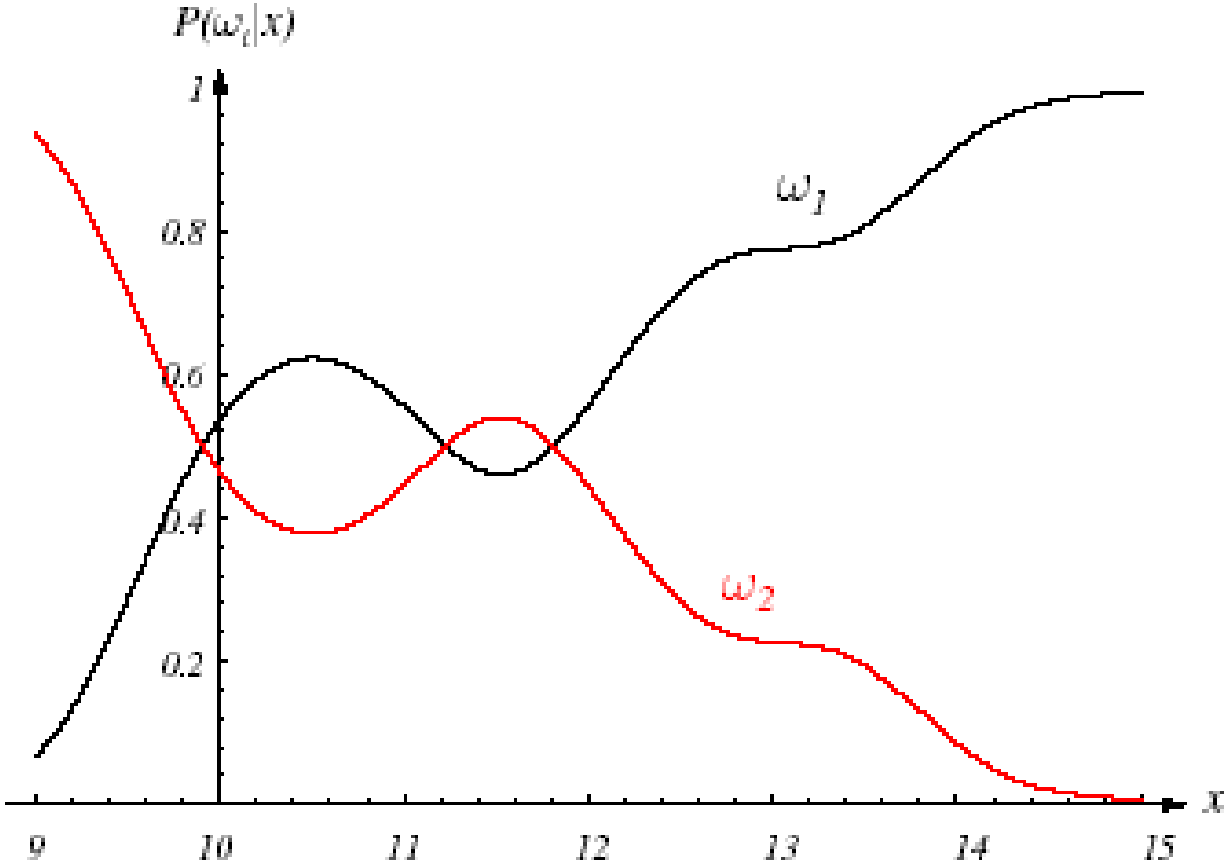
Rev. Thomas Bayes
b. 1702, London
d. 1761, Tunbridge Wells, Kent

Teorema di Bayes

Possiamo esprimere informalmente la formula di Bayes come:

$$\text{Prob. a posteriori} = \frac{(\text{prob. a priori}) \times \text{verosimiglianza}}{\text{evidenza}}$$

In questo modo è chiaro come la conoscenza del valore (misura) x influisce sul nostro giudizio a proposito dello stato di natura



Le probabilità a posteriori relative alle due classi viste prima, assumendo $P(\omega_1)=2/3$ e $P(\omega_2)=1/3$.

Decisione

La decisione tende naturalmente verso la classe cui compete la probabilità a posteriori maggiore:
Decidi ω_1 se $P(\omega_1|x) > P(\omega_2|x)$
altrimenti decidi ω_2

Questa regola di fatto minimizza la probabilità di errore:

$$P(\text{errore}|x)=\min\{P(\omega_1|x) , P(\omega_2|x)\}$$

Decisione

Da un punto di vista operativo, l'evidenza non entra in gioco nella decisione che può quindi ridursi a:

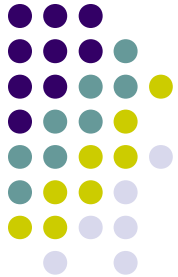
Decidi ω_1 se $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$
altrimenti decidi ω_2

Situazioni particolari:

se $p(x|\omega_1) = p(x|\omega_2)$ l'osservazione del valore x non fornisce informazioni riguardo lo stato di natura ulteriori rispetto alle prob. a priori

se $P(\omega_1) = P(\omega_2)$ la decisione tiene conto solo della verosimiglianza

La regola di decisione



Una *regola di decisione* è una funzione α(x) che indica quale azione intraprendere per ogni possibile valore di x osservato.

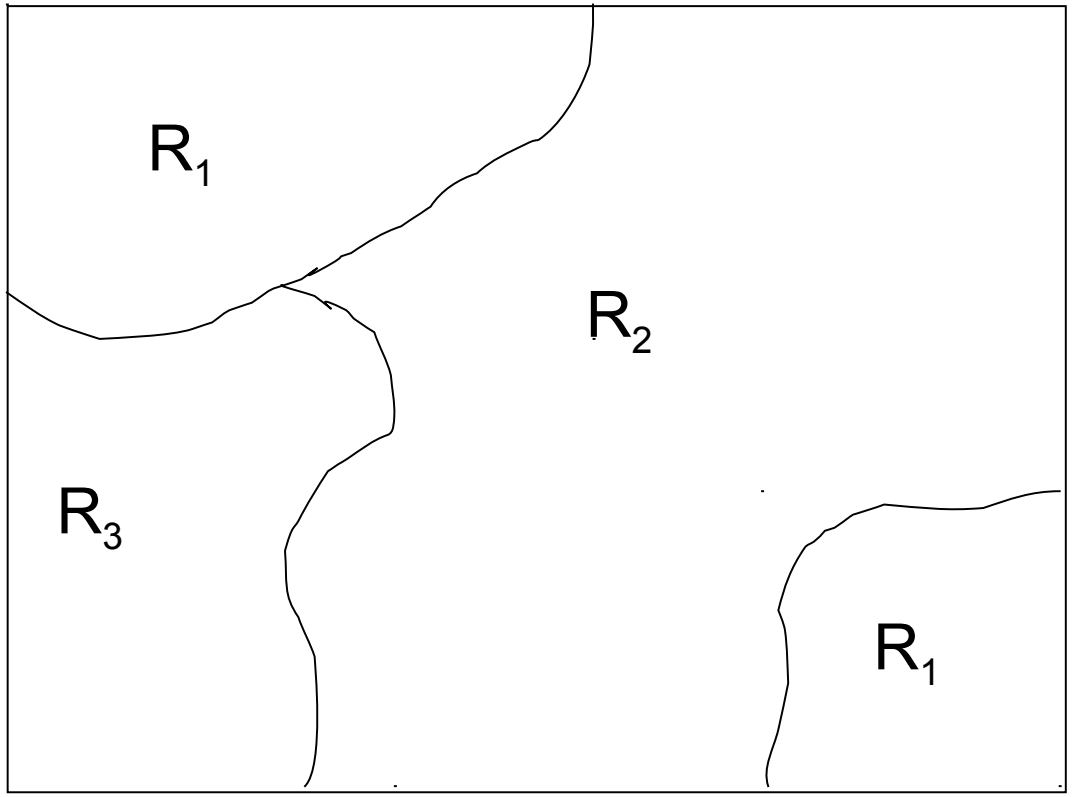
In questo contesto, la *regola di decisione ottima* è quella per cui si ha la massima probabilità a posteriori:

$$\alpha(x) = \underset{j=1,\ldots,C}{\arg\max} P(\omega_j | x)$$

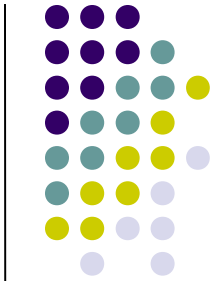
Regioni di decisione

La regola di decisione induce nello spazio delle fatures un insieme di regioni di decisione.

$x \in R_i \Leftrightarrow \alpha(x) = \alpha_i$



Problemi a due classi



Consideriamo il caso particolare di problema a due classi.

La regola di decisione si pone in termini di probabilità a posteriori:

si decide ω_1 se $P(\omega_1|x) > P(\omega_2|x)$

oppure:

$$\frac{P(\omega_1|x)}{P(\omega_2|x)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1$$

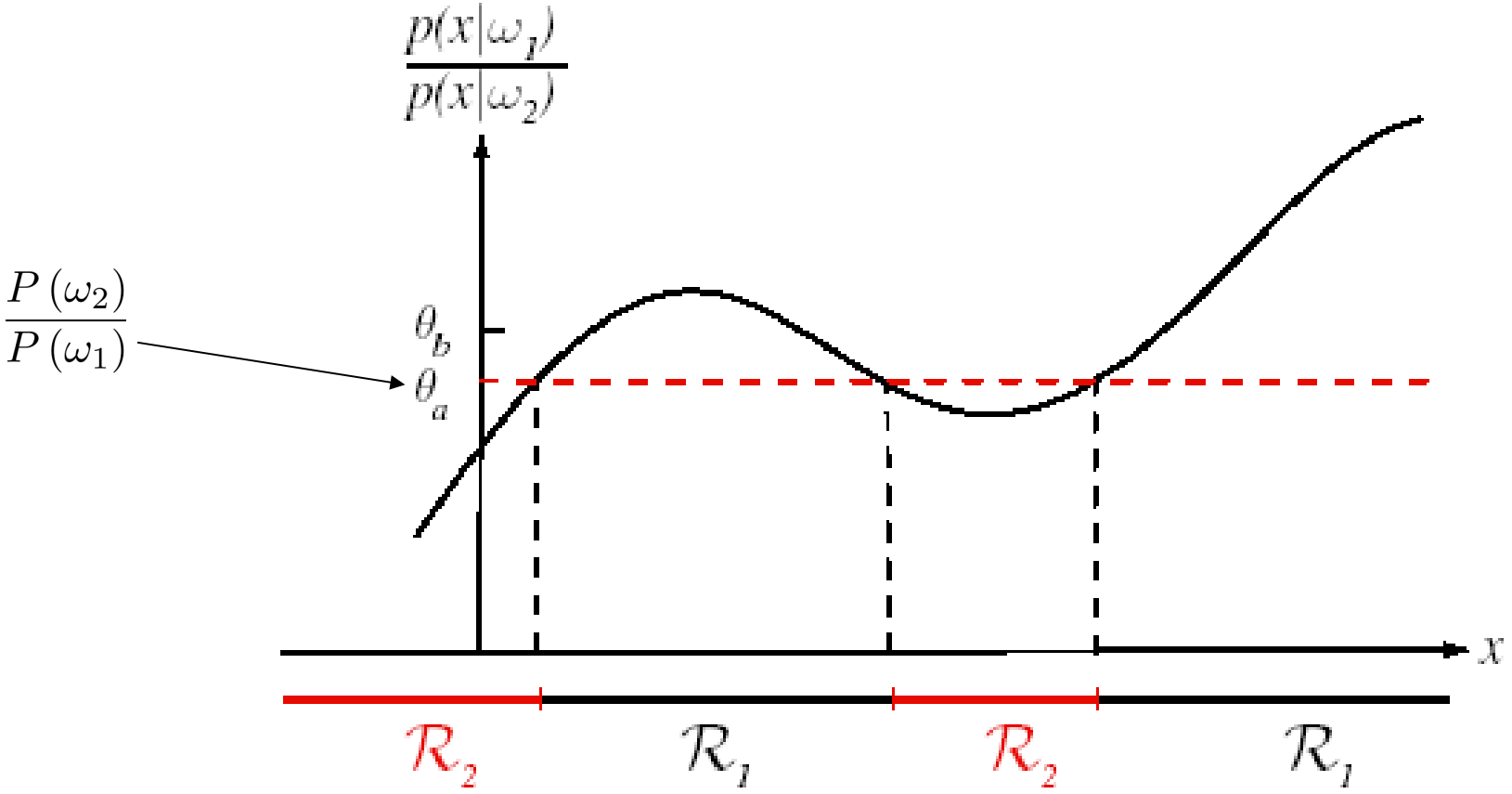
Problemi a due classi

Ricordando il teorema di Bayes, la condizione si può scrivere:

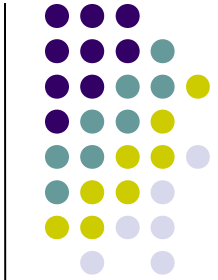
$$\frac{p(x|\omega_1)}{p(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_2)}{P(\omega_1)}$$

dove il membro di sinistra si definisce *rapporto di verosimiglianza* (*likelihood ratio*)

Problemi a due classi



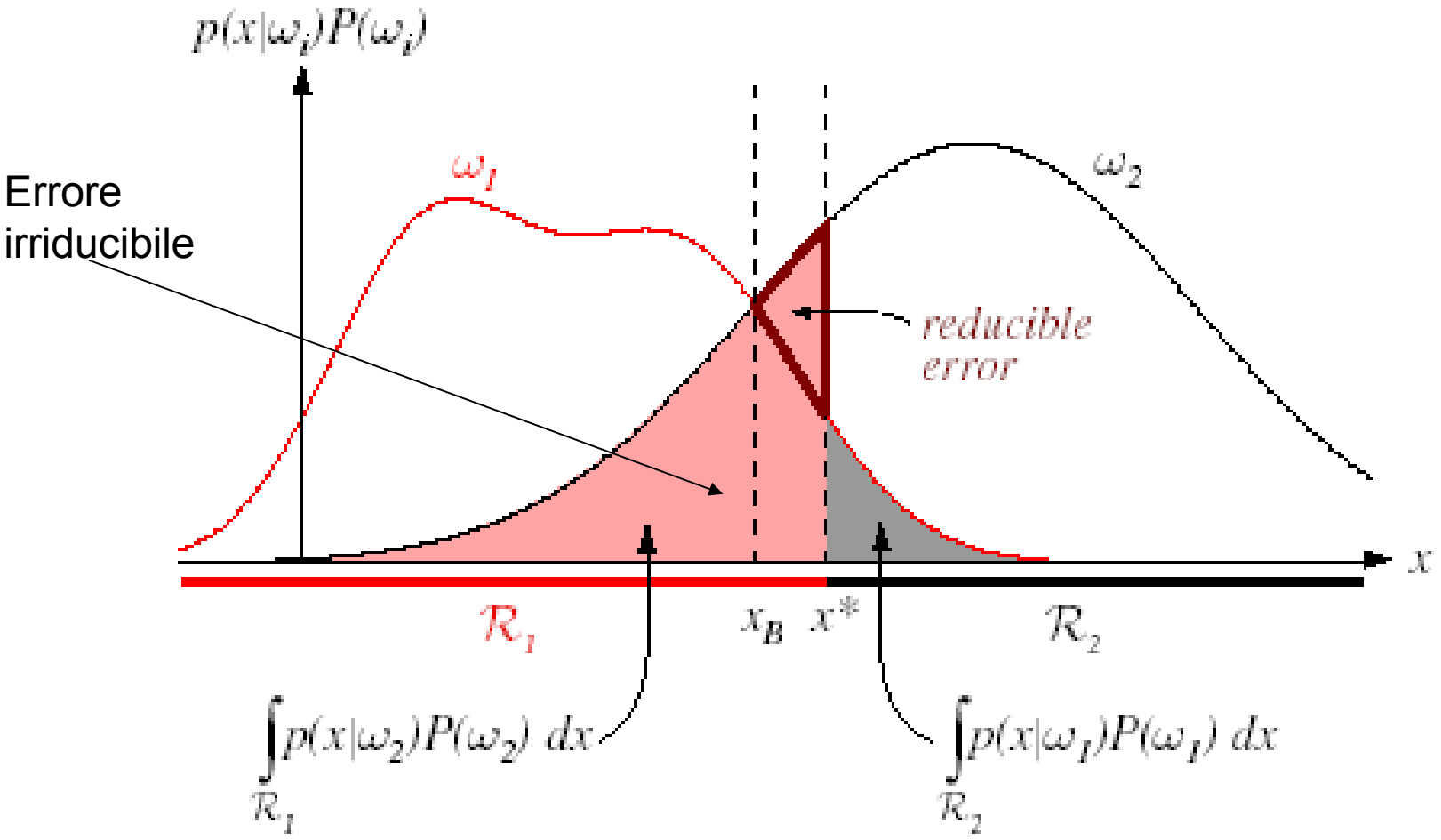
Ottimalità del classificatore bayesiano



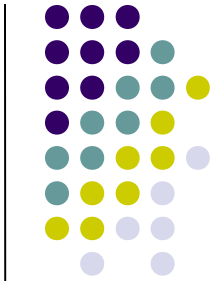
La probabilità di errore minima viene raggiunta con la regola di decisione bayesiana.

Di conseguenza, nei problemi a due classi, il classificatore costruito con questa regola (classificatore bayesiano) è il classificatore ottimo.

Probabilità minima di errore
Problemi a due classi

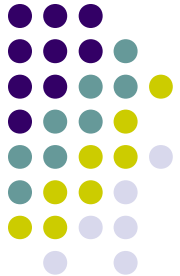


Limitazioni dell'approccio bayesiano



- Con l'approccio bayesiano, sarebbe possibile costruire un classificatore ottimo se si conoscessero:
  - le probabilità a priori $P(\omega_i)$
  - le densità condizionate alla classe $P(x|\omega_i)$
- Informazioni che raramente sono disponibili
- Alternativa: costruire una classificatore da un insieme di esempi (training set)
  - Pro: stima delle $P(\omega_i)$ semplicemente realizzabile
  - Contro: training set troppo limitato per una stima affidabile delle distribuzioni condizionate

Stima a k vicini



- Tra le possibili tecniche per la stima delle funzioni di probabilità richieste dall'approccio bayesiano. Semplice ma piuttosto efficace.
- Consideriamo un insieme Ts di n campioni appartenenti alle varie classi e sia n_i il numero di campioni appartenenti alla classe ω_i.
- Sia x è un campione da classificare non appartenente a Ts.

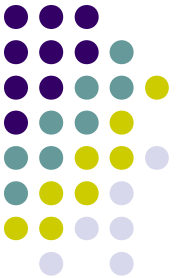
Stima a k vicini

- Si consideri un'ipersfera centrata su x e di raggio tale da includere k campioni di Ts.
- Sia $k_i \leq k$ il numero di campioni interni all'ipersfera appartenenti alla classe ω_i.
- Se V è il volume dell'ipersfera, con il metodo a k vicini si possono stimare
  - La pdf condizionata
  - La pdf incondizionata
  - La probabilità a priori

$$p(x|\omega_i) = \frac{k_i}{n_i \cdot V} \qquad p(x) = \frac{k}{n \cdot V}$$

$$P(\omega_i) = \frac{n_i}{n}$$

Classificatore k-NN



Mettendo tutto insieme, è possibile ottenere una stima della probabilità a posteriori:

$$P\left(\omega_i|x\right)=\frac{p\left(x|\omega_i\right)P\left(\omega_i\right)}{p\left(x\right)}\simeq\frac{k_i}{n_i\cdot V}\frac{n_i}{n}\frac{n\cdot V}{k}=\frac{k_i}{k}$$

In questo modo è possibile definire una regola di classificazione (k *Nearest Neighbor rule* o *k-NN*):

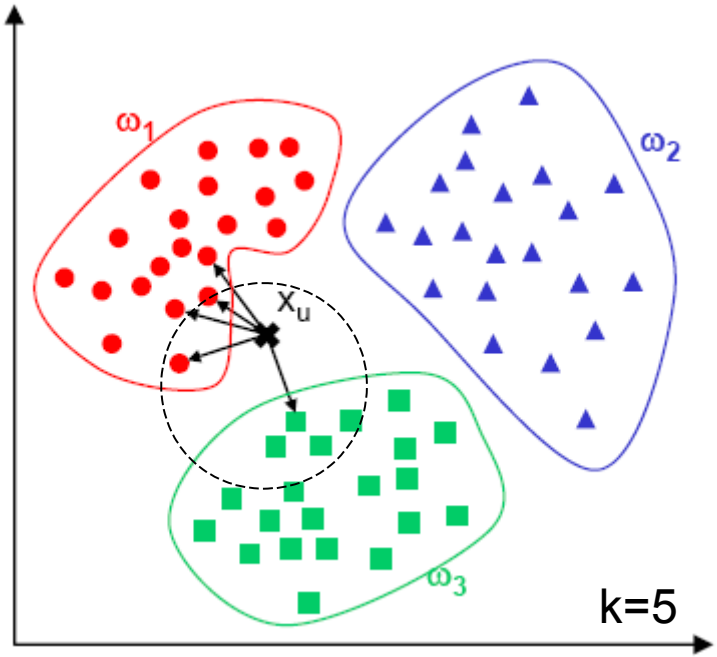
$$\alpha\left(x\right)=\arg\max_{j=1,\ldots,C}\frac{k_j}{k}$$

Classificatore k-NN

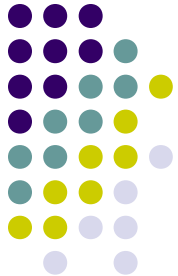
Il classificatore k-NN è un classificatore non parametrico che classifica i campioni sulla base della loro somiglianza con gli esemplari del training set Ts.

Per definire un classificatore k-NN è necessario soltanto

- Scegliere un valore k
- Un insieme di campioni con etichette (training set)
- Una metrica per definire la “vicinanza”



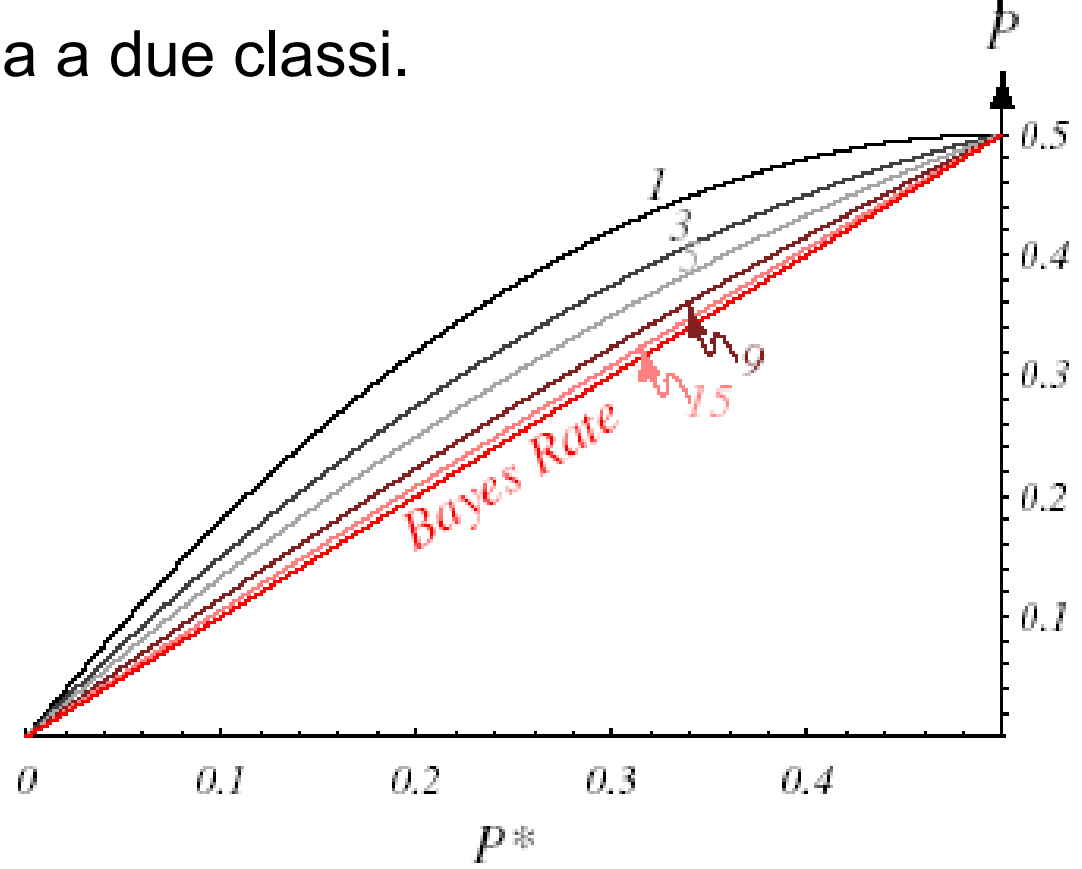
Prestazioni del classificatore k-NN



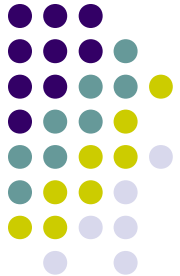
- Il classificatore è sub-ottimo nel senso che non garantisce la probabilità di errore minima esibita dal classificatore bayesiano.
- E' però possibile dimostrare che, con $n \rightarrow \infty$, la probabilità di errore P_e per il classificatore k-NN si avvicina alla probabilità di errore del classificatore bayesiano se $k \rightarrow \infty$.

Prestazioni del classificatore k-NN

Limiti inferiore (errore di Bayes) e superiore alla probabilità di errore del classificatore k-NN per un problema a due classi.



Classificatore Nearest-Neighbor

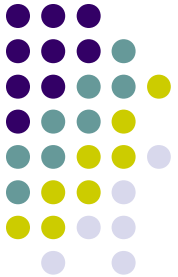


Un caso particolare si ha quando k=1.

Si ottiene un classificatore 1-NN o classificatore "Nearest Neighbor"

La classificazione di un nuovo campione x non appartenente a Ts avviene scegliendo l'etichetta del campione di Ts a minima distanza da x.

Prestazioni del classificatore 1-NN



Anche il classificatore 1-NN è sub-ottimo.

E' però possibile dimostrare che, al crescere di n, la probabilità di errore P_e per il classificatore NN soddisfa la seguente relazione:

$$P_e^* \leq P_e \leq 2P_e^*$$

dove P_{e^*} è la probabilità di errore del classificatore bayesiano.

Classificatore Nearest-Neighbor