Università di Verona

A.A. 2020-21

# Machine Learning & Artificial Intelligence

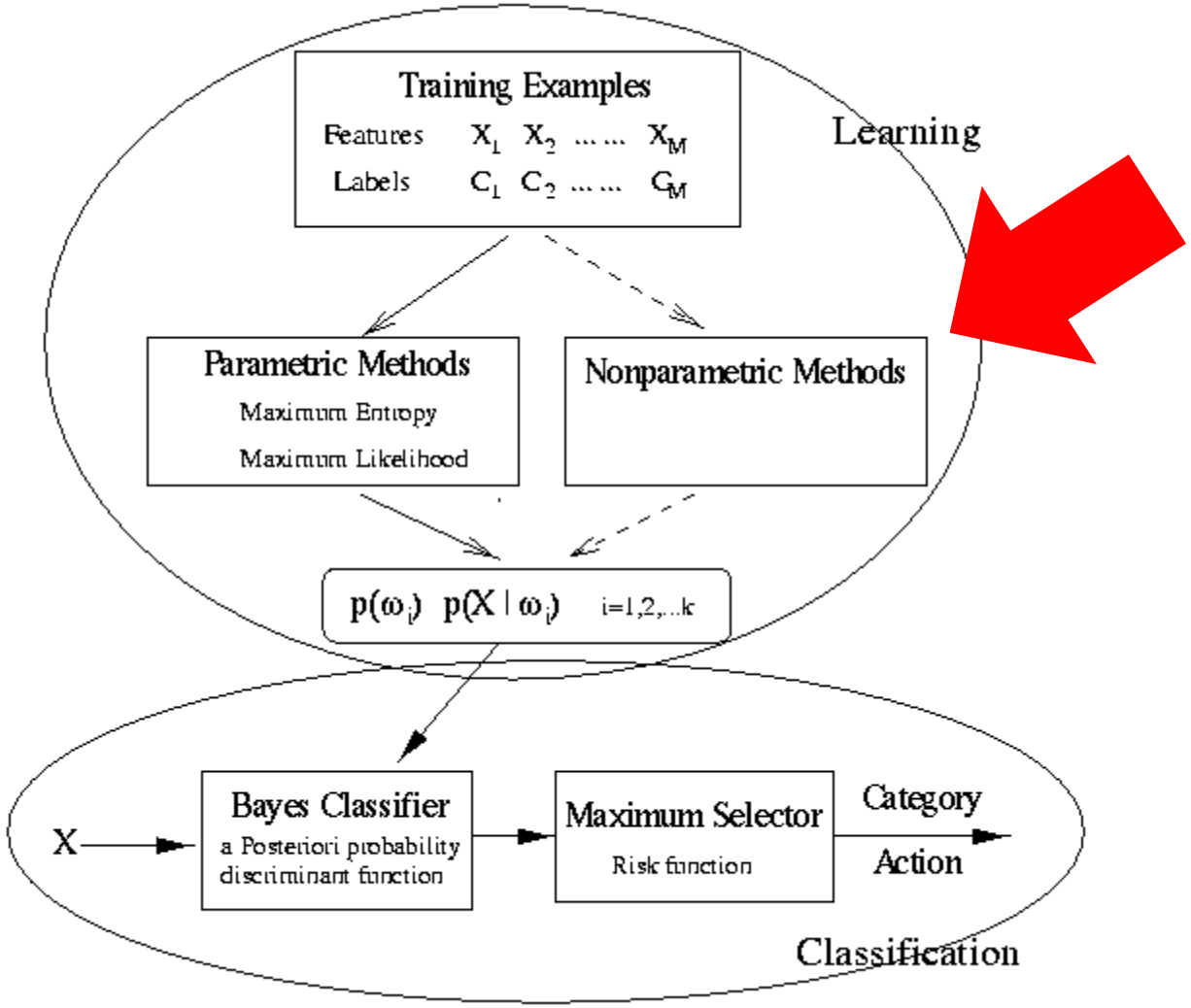
## Metodi non parametrici

Vittorio Murino

# Sommario

- Introduzione
- Le funzioni potenziale
- Stima della densità di probabilità condizionale
- Finestre di Parzen
- Metodo dei prototipi
- Metodo dei *k* punti vicini

# Uno sguardo d'insieme



# Introduzione

- Il classificatore Bayesiano permette di minimizzare la probabilità di errore.
- Il problema di questo classificatore è che si basa sulle probabilità di emissione e condizionale: se queste sono sconosciute devono essere stimate dai dati.
- Esistono tre tipologie di metodi per stimare queste densità:
  - metodi parametrici: si assume nota la forma della densità, se ne stimano i parametri dai dati (esempio gaussiana)
  - metodi non parametrici: nessuna assunzione sulla forma della densità, completamente stimata dai dati (esempio KNN)
  - metodi semiparametrici: tecnica mista, si assume che esista una famiglia piuttosto ampia di funzioni di densità (esempio reti neurali)

- Nei metodi parametrici si assume che la forma delle densità di probabilità sia nota, ma questa assunzione non può essere fatta in molti problemi di riconoscimento.

- In particolare, la maggior parte dei metodi parametrici ipotizza che le densità di probabilità siano unimodali (abbiano cioè un singolo massimo), ma nella realtà molti problemi implicano l'utilizzo di densità multimodali.

- In questa parte verranno illustrati alcuni metodi non parametrici, che permettono di stimare le funzioni di densità di probabilità a partire direttamente dai campioni.

- In particolare, esiste il problema di stimare la quantità

$$p(\mathbf{x}|\omega_i) \equiv \hat{p}_i(\mathbf{x})$$

- Questo problema, se risolto, consente di utilizzare il classificatore Bayesiano (ottimo teorico)

- Tali metodi hanno la caratteristica comune di stimare le funzioni richieste mediante un insieme di funzioni più semplici, in genere associabili ai campioni.

# Le funzioni potenziale

- **Idea di base:** stabilire una analogia tra i campioni, pensati come punti in uno spazio appropriato, e il concetto di carica elettrica.
- Posizionando una carica elettrica in ciascun punto associato ad un campione, si può ipotizzare che, definendo in modo opportuno un potenziale associato alla carica, il risultante potenziale elettrostatico definisca una funzione discriminante per il problema di riconoscimento considerato.
- Questa formulazione del problema implica la capacità di approssimare una funzione globale (cioè riferita all'intero spazio) mediante un insieme di funzioni potenziali.

- Visto sotto quest'ottica, il problema è analogo a trovare un’espressione per la probabilità condizionale a partire dai campioni e da un insieme di funzioni potenziali ad essi associabili.
- D'altronde, i due problemi (definizione della funzione discriminante lineare e probabilità condizionale) sono direttamente correlati.
- In questo caso tuttavia, a differenza, per esempio, del caso Gaussiano, si suppone di non conoscere la forma della funzione di densità di probabilità.
- Pertanto si parla di *metodi non parametrici*, in quanto diventa necessario stimare la densità di probabilità direttamente a partire dai campioni.

- Sia $\gamma(\mathbf{x},\mathbf{y}_j)$ una funzione potenziale per un campione generico $\mathbf{y}_j$ della classe *i*.
- Allora si può scrivere:

$$\hat{p}_i(\mathbf{x}) = \frac{1}{N_i}\sum_{j=1}^{N_i}\gamma(\mathbf{x},\mathbf{y}_j) \qquad \text{dove } N_i = \#\,\text{campioni classe } i$$

- Per costruire una buona approssimazione dobbiamo porre alcuni vincoli sulla forma di $\gamma$.

  1) $\gamma(\mathbf{x},\mathbf{y}) \geq 0$

  2) $\arg\max_{\mathbf{x}} \gamma(\mathbf{x},\mathbf{y}_k) = \mathbf{y}_k$, cioè $\gamma(\mathbf{x},\mathbf{y}_k)$ è massima per $\mathbf{x}=\mathbf{y}_k$;

  3) $\gamma(\mathbf{x},\mathbf{y}_1) \cong \gamma(\mathbf{x},\mathbf{y}_2)$, se $|\mathbf{y}_2-\mathbf{y}_1| < \varepsilon$, cioè se i due vettori dei campioni sono "abbastanza" vicini (questo vincolo serve a garantire che *p* non vari bruscamente o possa avere discontinuità).

  4) $\gamma(\mathbf{x},\mathbf{y})$ continua.

5) $\int_{-\infty}^{+\infty} \gamma(\mathbf{x}, \mathbf{y}_k) d\mathbf{x} = 1$ condizione di normalizzazione.

6) $\gamma(\mathbf{x},\mathbf{y}_k) \cong 0$, se $\mathbf{x}$ è molto lontana da $\mathbf{y}_k$.

- Esistono diverse possibili forme di $\gamma$ che tengono conto di questi vincoli, riferendoci al caso monodimensionale.
- Ci si riferirà in particolare ad una funzione $\gamma(\mathbf{z},\mathbf{y})$ che abbia come argomento una sola variabile $\mathbf{x}$, rappresentata dalla norma euclidea della differenza tra i due vettori $\mathbf{z}$ e $\mathbf{y}$, i.e., $\mathbf{x}=|\mathbf{z}-\mathbf{y}|$.

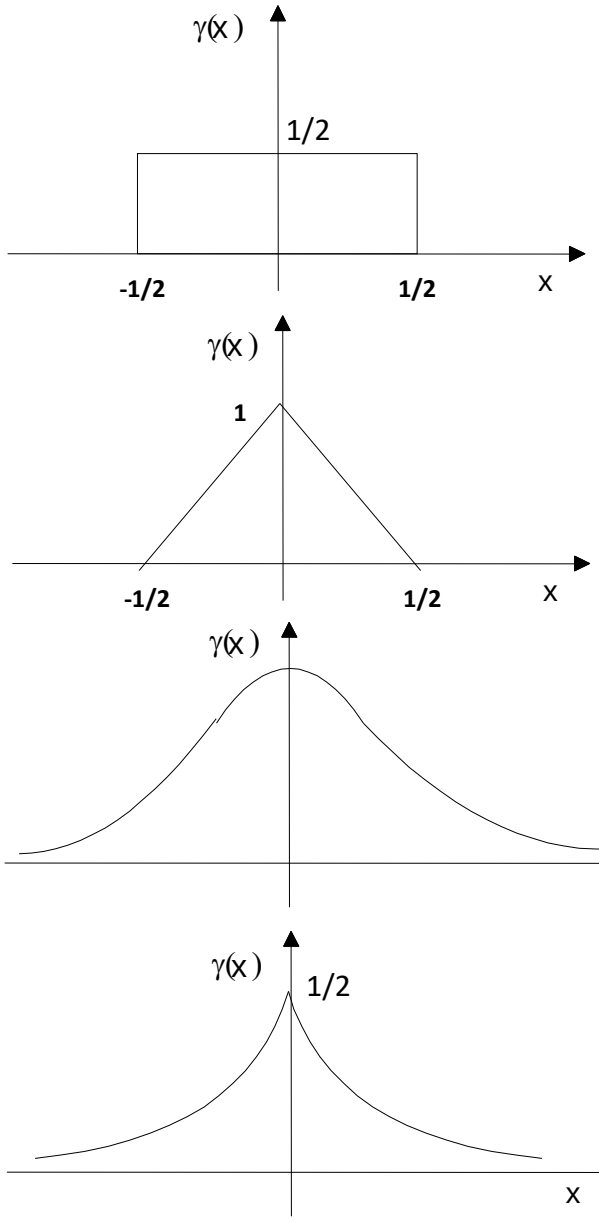
# Esempi di funzioni potenziali

1) $\gamma(\mathbf{x}) = \begin{cases} 0{,}5 & |\mathbf{x}| \le 1 \\ 0 & |\mathbf{x}| > 1 \end{cases}$ Rettangolo

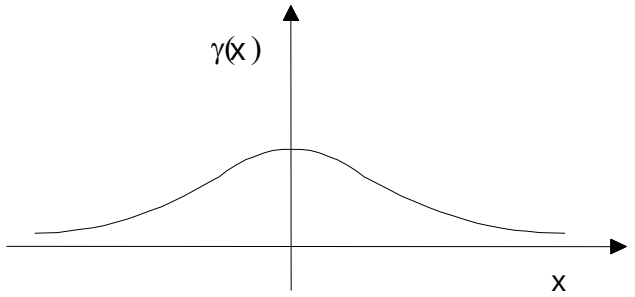
2) $\gamma(\mathbf{x}) = \begin{cases} 1-|\mathbf{x}| & |\mathbf{x}| \le 1 \\ 0 & |\mathbf{x}| > 1 \end{cases}$ Triangolo

3) $\gamma(\mathbf{x}) = (2\pi)^{-\frac{1}{2}} e^{-\left(\frac{\mathbf{x}^2}{2}\right)}$ Gaussiana

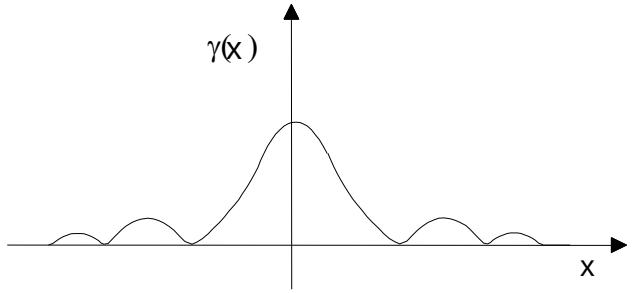
4) $\gamma(\mathbf{x}) = \frac{1}{2} e^{-|\mathbf{x}|}$ Esponenziale decrescente



5) $\gamma(\mathbf{x}) = \left[\pi\left(1+\mathbf{x}^2\right)\right]^{-1}$ Distribuzione di Cauchy




6) $\gamma(\mathbf{x}) = (2\pi)^{-1}\left(\frac{\sin\left(\frac{\mathbf{x}}{2}\right)}{\frac{\mathbf{x}}{2}}\right)^2$ Funzione di tipo (sin x/x)$^2$




- In quest ultimo caso $\gamma$ non è monotona ma si smorza con andamento periodico.
- Nello spazio ad *n* dimensioni $\mathrm{x}$ diventa ovviamente un vettore.
- Problemi:

  1) Scelta delle funzioni potenziali $\gamma$.

  2) Grado di sovrapposizione delle $\gamma$.

# Stima della densità condizionale

## Idea di base:

Problema: stima di $p(\mathbf{x})$

- La probabilità che un vettore $\mathbf{x}$ sia in una regione $\mathcal{R}$ è:

$$P = \int_{\Re} p(\mathbf{x}')d\mathbf{x}' \qquad (1)$$

- $P$ è una versione *smoothed* (o mediata) della densità $p(\mathbf{x})$, e si può stimare il valore *smooth* di $p$ stimando la probabilità $P$.
- Consideriamo un insieme di campioni (i.i.d.) di cardinalità $n$ estratti secondo $p(\mathbf{x})$*:* la probabilità che $k$ punti su $n$ siano in $\mathcal{R}$ è data dalla legge binomiale:

$$P_k = \binom{n}{k} P^k (1-P)^{n-k} \qquad (2)$$

il cui valore aspettato per $k$ è:

$$\mathrm{E}[k] = nP \qquad (3)$$

- La stima ML di $P$ *(= $\theta$)*

  $\max_{\theta}(P_k \mid \theta)$ è data da $\hat{\theta} = \frac{k}{n} \cong P$

- Quindi, il rapporto *k/n* è una buona stima per la probabilità $P$ e così per la densità $p$.

- Se $p(\mathbf{x})$ è continua e la regione $\mathcal{R}$ è così piccola che $p$ non varia significativamente in essa (così da essere approssimabile da una costante), possiamo scrivere:

$$P = \int_{\Re} p(\mathbf{x}')dx' \cong p(\mathbf{x}) \cdot V \qquad (4)$$

  dove $\mathbf{x}$ è in $\mathcal{R}$ e $V$ è il volume incluso in $\mathcal{R}$

Combinando le equazioni (1) , (3) e (4) si ottiene: $$p(x) \cong \frac{k/n}{V} \qquad (5)$$

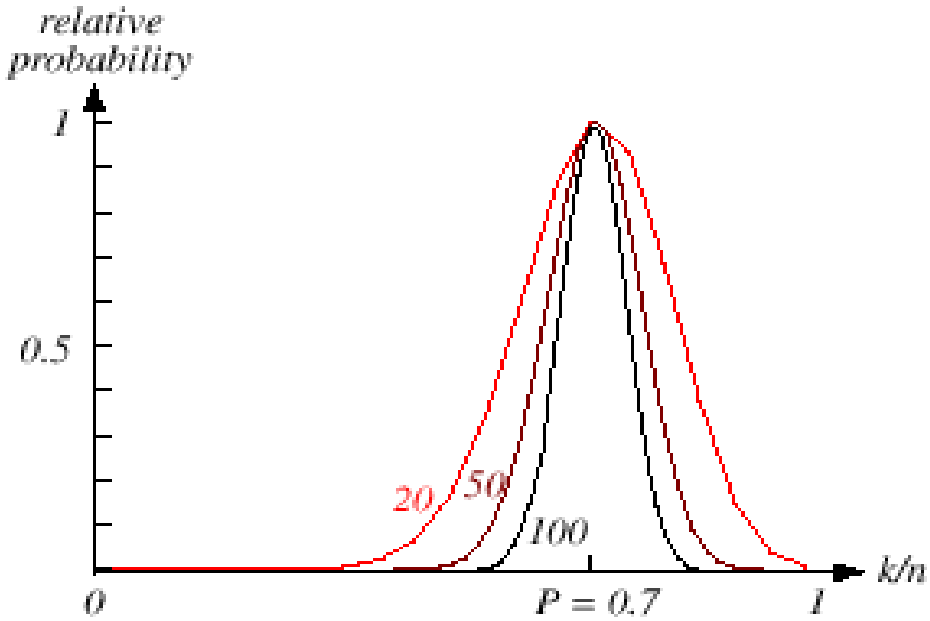



**FIGURE 4.1.** The relative probability an estimate given by Eq. 4 will yield a particular value for the probability density, here where the true probability was chosen to be 0.7. Each curve is labeled by the total number of patterns $n$ sampled, and is scaled to give the same maximum (at the true probability). The form of each curve is binomial, as given by Eq. 2. For large $n$, such binomials peak strongly at the true probability. In the limit $n \to \infty$, the curve approaches a delta function, and we are guaranteed that our estimate will give the true probability. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Condizioni per la convergenza

La frazione $k/(nV)$ è un valore mediato nella regione di $p(\mathbf{x})$.

La $p(\mathbf{x})$ vera si ottiene solo se $V$ diventa piccola arbitrariamente fino a tendere a zero.

$$\lim_{V \to 0,\, \mathrm{k}=0} p(\mathbf{x}) = 0 \;\; (\text{se n fissato})$$

È il caso in cui non ci sono campioni inclusi in $\mathscr{R}$

→ non interessante!

$$\lim_{V \to 0,\, \mathrm{k}\neq 0} p(\mathbf{x}) = \infty$$

In questo caso, la stima diverge → non interessante!

- È necessario che il volume $V$ tenda a 0 in ogni caso se si vuole usare la stima
  - In pratica, $V$ non può diventare piccolo a piacere dato che il numero di campioni è sempre limitato e si deve accettare una certa varianza nel rapporto $k/n$ ed una certa approssimazione (valore mediato) di $p(\mathbf{x})$
  - Teoricamente, se il numero di campioni è illimitato, possiamo superare il problema.
- Per stimare la densità di $\mathbf{x}$, formiamo una sequenza di regioni
  - $\mathcal{R}_1, \mathcal{R}_2,\ldots$ contenenti $\mathbf{x}$: la prima regione con 1 campione, la seconda con 2 campioni e così via.
  - Sia $V_n$ il volume of $\mathcal{R}_n$, $k_n$ il numero di campioni in $\mathcal{R}_n$ e $p_n(\mathbf{x})$ la stima n-esima di $p(\mathbf{x})$, allora:

$$p_n(x) = (k_n/n)/V_n \qquad (7)$$

- Le condizioni per cui $p_n(\mathbf{x})$ converge a $p(\mathbf{x})$ sono*:*

$$1) \lim_{n\to\infty} V_n = 0$$

$$2) \lim_{n\to\infty} k_n = \infty$$

$$3) \lim_{n\to\infty} k_n / n = 0$$

- Ci sono 2 modi diversi di ottenere le sequenze di regioni che soddisfano queste condizioni:

  (a) Fissare una regione iniziale e ridurla specificando il volume $V_n$ secondo una regola che dipende da $n$, tipo $V_n = 1/\sqrt{n}$ e verificando che

$$p_n(\mathbf{x}) \underset{n\to\infty}{\rightarrow} p(\mathbf{x})$$

  Si ottiene il metodo delle **finestre di Parzen**

  (b) Fissare $k_n$ come una qualche funzione di $n$, e.g., $k_n = \sqrt{n}$; e aumentare il volume $V_n$ finchè non includa $k_n$ vicini di $\mathbf{x}$.

  Si ottiene il metodo **k-NN, k nearest neighbor**

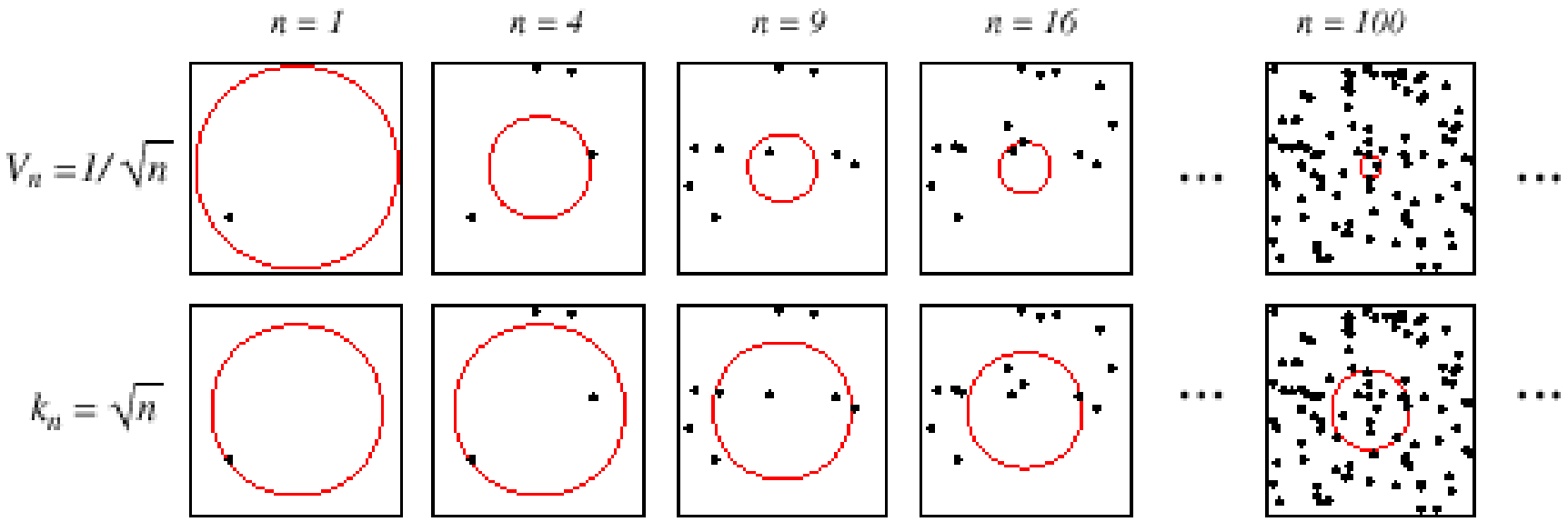
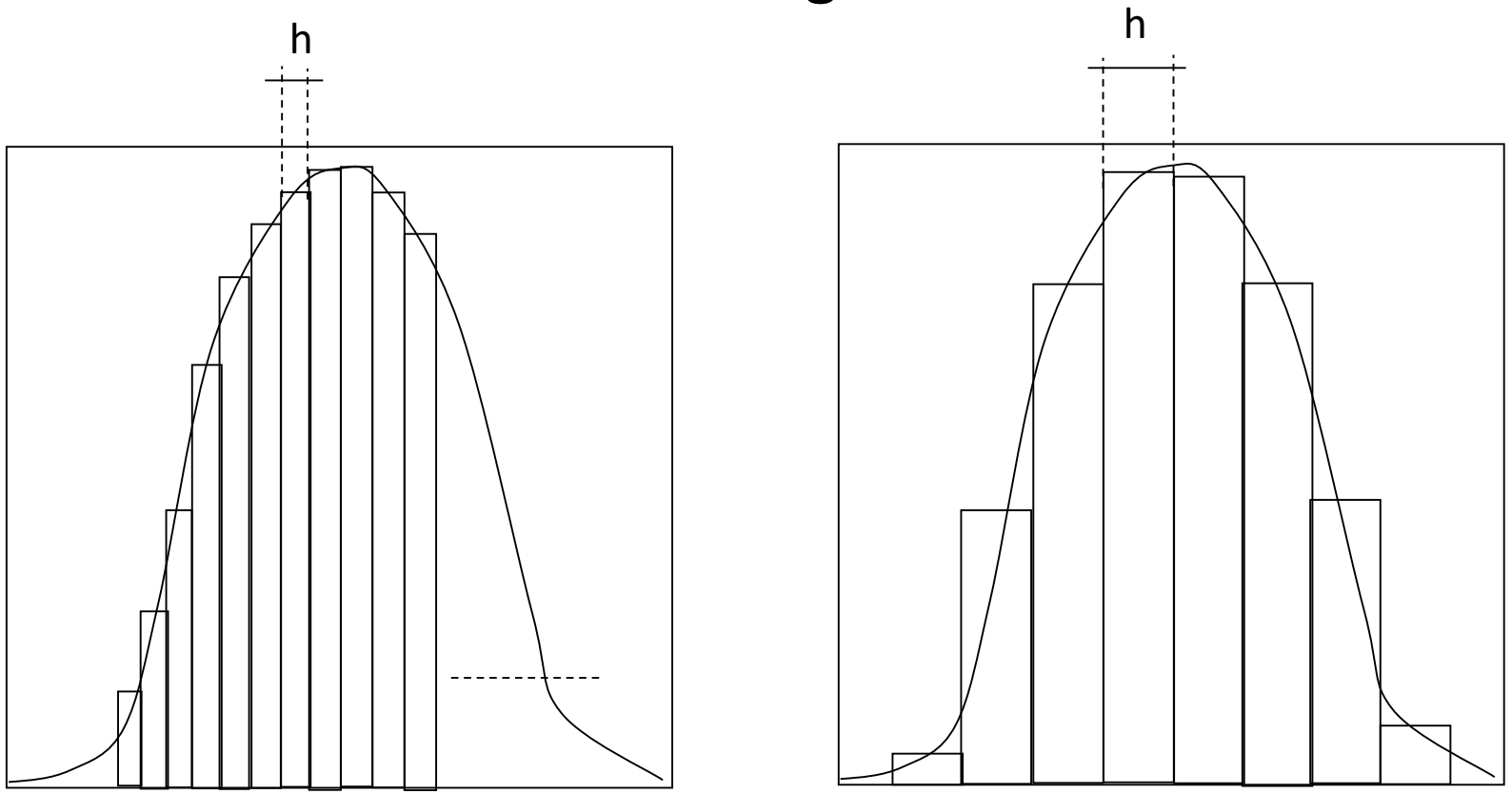



**FIGURE 4.2.** There are two leading methods for estimating the density at a point, here at the center of each square. The one shown in the top row is to start with a large volume centered on the test point and shrink it according to a function such as $V_n = 1/\sqrt{n}$. The other method, shown in the bottom row, is to decrease the volume in a data-dependent way, for instance letting the volume enclose some number $k_n = \sqrt{n}$ of sample points. The sequences in both cases represent random variables that generally converge and allow the true density at the test point to be calculated. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- In sostanza, visto in modo più intuitivo, se consideriamo un caso monodimensionale e abbiamo un insieme di punti $x$ presi da una distribuzione $p(x)$, il modo più semplice di approssimarla è mediante un istogramma:



- La probabilità che un campione $x$ sia in un certo *bin* può essere stimata per ogni bin, quindi dati $n$ campioni e $k$ ($k_n$) di questi si trovano in un bin, la relativa probabilità si può stimare dal rapporto di frequenza $P \cong k/n$

  ... che converge alla vera $P$ se $n \to \infty$ e, assumendo il valore della pdf costante sul bin, abbiamo che questa si può approssimare come

$$\hat{p}(x) \equiv \hat{p}(\hat{x}) \approx \frac{1}{h}\frac{k_n}{n}, \qquad |x - \hat{x}| \leq \frac{h}{2}$$

  dove $\hat{x}$ è il valore medio del bin e nell'ipotesi di $p(x)$ continua e $h$ sufficientemente piccolo.

# Finestre di Parzen

- Se la regione $\mathcal{R}$ è un piccolo ipercubo centrato in $\mathbf{x}$ e si vuole calcolare il numero $k$ di punti al suo interno allora si può definire la seguente funzione

$$\gamma(\mathbf{u}) = \begin{cases} 1, & |u_i| < 1/2 \\ 0, & altrimenti \end{cases} \qquad i = 1,..,D$$

che rappresenta un cubo centrato nell'origine.

- $\gamma(\mathbf{u})$ è un esempio di funzione *kernel*, chiamata in questo contesto finestra di Parzen.
- Quindi:

$$k = \sum_{j=1}^{N} \gamma\left(\frac{\mathbf{x} - \mathbf{y}_j}{h}\right)$$

- E sostituendo nella (5) abbiamo la stima

$$p(x) = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{h^D} \gamma\left(\frac{\mathbf{x} - \mathbf{y}_j}{h}\right), \quad \text{dove } V = h^D$$

- Come per il caso dell'istogramma si rilevano problemi nella presenza di discontinuità tra gli ipercubi, ma che può essere ovviato usando una funzione kernel più dolce
- Il metodo delle funzioni di Parzen può essere visto come un'istanza del metodo più generale delle funzioni potenziale.
- Le finestre di Parzen contribuiscono a stimare la densità di probabilità nel modo seguente (caso monodimensionale, per semplicità):

$$\hat{p}_i(x) = \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{1}{h} \gamma\left(\frac{x - y_j}{h}\right) \qquad \text{dove } N_i = \#\,\text{campioni} \in \omega_i$$

  dove $h$ è la dimensione della finestra di Parzen e $\gamma$ può essere una delle funzioni potenziali viste in precedenza: $h$ regola in pratica la sovrapposizione tra le $\gamma$.
- La scelta di $h$ ha un effetto rilevante sulla stima trovata.
  - se $h$ è troppo grande, la stima sarà caratterizzata da una bassa risoluzione,
  - se $h$ è troppo piccolo il problema sarà quello di una grande variabilità statistica (i campioni interagiscono poco).

- La stima indicata nell'eq. precedente è comunque tanto migliore quanto $N_i$ è più grande.
- Questo implica che si possa scrivere come:

$$\hat{p}_i(x) = \lim_{N_i \to \infty} \left\{ \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{1}{h} \gamma\left( \frac{x - y_j}{h} \right) \right\}$$

dove $y_j$ sono i campioni della classe $\omega_i$.

**Esempio**

- Sia $\gamma$ di tipo Gaussiano, cioè:

$$\gamma(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \qquad \text{e} \qquad \hat{p}(x \mid \omega_k) = \frac{1}{N_k} \sum_{j=1}^{N_k} \frac{1}{h_k} \gamma\left( \frac{x - y_j}{h_k} \right)$$

- Supponiamo i dati usati in ingresso nell'esempio come distribuiti gaussianamente con $N_k$ numero dei campioni del *training set* per la classe $\omega_k$.

- Si cerca di vedere se, usando questi dati, si è in grado con il metodo delle finestre di Parzen e la funzione $\gamma$ introdotta sopra di approssimare la funzione di densità di probabilità Gaussiana che ha originato i dati stessi.
- Il valore $h_{\mathrm{k}}$ usato nella stima può essere considerato in prima battuta dipendente dal tipo di funzione $\gamma$ usato e dal numero di campioni.
- Si introduce una regola pratica per collegare tali misure, per cui: $h_k = \frac{h_1}{N_k}$
- Per vedere l'effetto di una differente scelta dei parametri, si possono provare differenti valori di $h_1$ ed $N_k$, e vedere come varia la stima della funzione di probabilità condizionata al variare del numero di campioni e di $h$.

- In particolare, se si usano i valori di $h_1$ troppo piccoli (o $N_k$ grandi) allora si possono ancora distinguere gli effetti dei singoli campioni nella stima.
- Siamo cioè nel caso di grande variabilità statistica, cioè troppo basso effetto interpolante.
- Al crescere di $h_k$ ho finestre più larghe, cioè sovrapposizione più forte.
- Per scegliere il valore di $h_1$ da utilizzare si valutano i risultati in termini di $p(x)$:
  - una eccessiva presenza di picchi nella densità di probabilità $\hat{p}(\mathbf{x})$ è da evitare, così come un troppo elevato livello di *smoothing* (cioè curva quasi costante o "molto filtrata").
- In definitiva, per applicare questo metodo, occorrono le seguenti condizioni:

  a) un numero elevato di campioni;

  b) verificare la scelta di $\gamma$ ed $h$;

  c) memorizzare tutti i campioni per avere $\hat{p}(\mathbf{x})$.

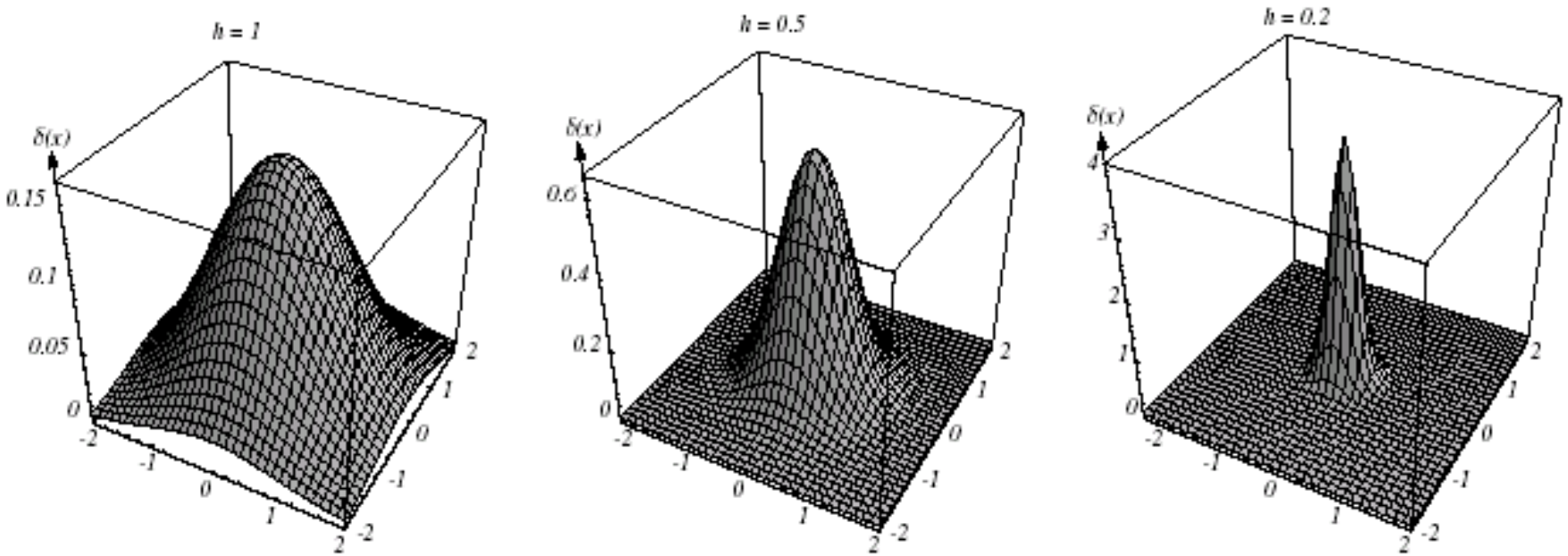



**FIGURE 4.3.** Examples of two-dimensional circularly symmetric normal Parzen windows for three different values of $h$. Note that because the $\delta(\mathbf{x})$ are normalized, different vertical scales must be used to show their structure. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

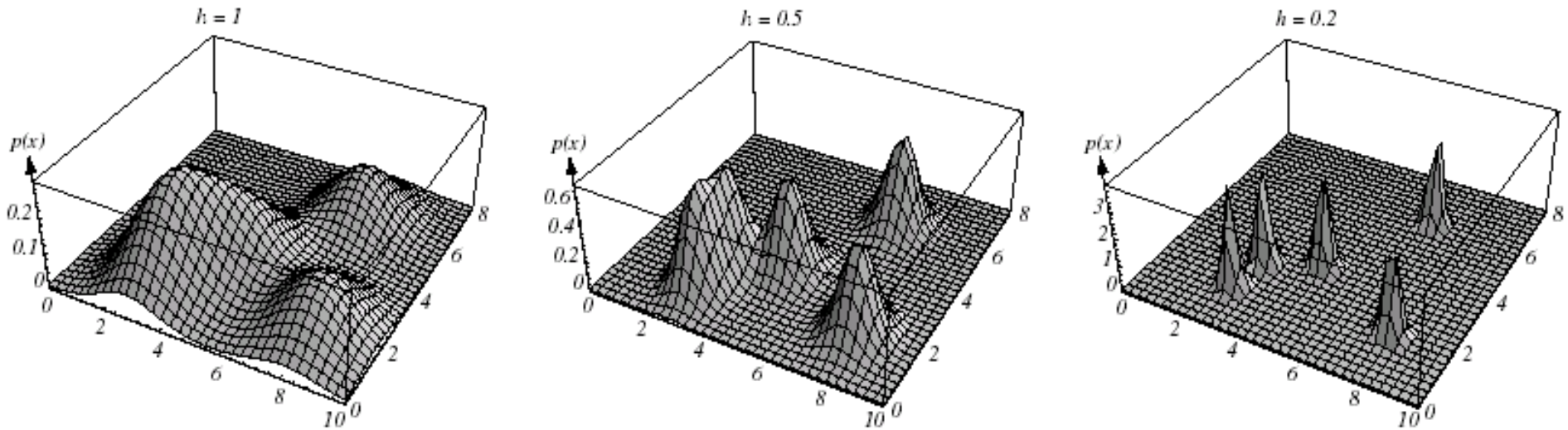



**FIGURE 4.4.** Three Parzen-window density estimates based on the same set of five samples, using the window functions in Fig. 4.3. As before, the vertical axes have been scaled to show the structure of each distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

# Finestre di Parzen con kernel Gaussiano

- Metodo delle finestre di Parzen in caso di funzione potenziale $\gamma$ di tipo Gaussiano (anche chiamate funzioni di Specht).
- In particolare si scrive la probabilità condizionata come:

$$\hat{p}_i(\mathbf{x}) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}} N_i} \sum_{j=1}^{N_i} \exp\left\{ - \frac{\left(\mathbf{x}-\mathbf{y}_j\right)^{\mathrm{t}}\left(\mathbf{x}-\mathbf{y}_j\right)}{\left(2\sigma^2\right)} \right\} \qquad (1)$$

- $\sigma$ è il cosiddetto parametro di *smoothing*.
  - per $\sigma = 0$, la probabilità diventa espressa da una sommatoria di impulsi di Dirac centrati su ogni $\mathbf{y}_j$;
  - per $\sigma = \infty$, $\hat{p}(\mathbf{x})$ diventa costante.

- Dato che il parametro $\sigma$ di smoothing è considerato uniforme su tutte le direzioni è essenziale effettuare una normalizzazione sulle *feature* prima di stimare la probabilità, cioè modificare le *feature* in modo che $\sigma_1 = \sigma_2 = \ldots = \sigma_n = \sigma$.
- Occorre inoltre scegliere $\sigma$ in modo che non si abbia una sovrapposizione eccessiva tra le funzioni potenziali centrate sui diversi campioni.
- A tal fine, si suggerisce di stimare $\sigma$ prendendo gli $L$ punti più vicini al generico campione $\mathbf{y}_i$ e di calcolarne la media delle distanze:

$$\sigma = \frac{1}{L}\sum_{j=1}^{L}\left\|\mathbf{y}_j - \mathbf{y}_i\right\| = \frac{1}{L}\sum_{j=1}^{L}\sqrt{\left(\mathbf{y}_j - \mathbf{y}_i\right)^2}$$

- In genere si sceglie euristicamente $L \cong 0.05N$.
- La scelta di $L$ modifica il grado di sovrapposizione tra le curve.
- Per risalire alla funzione discriminante nel caso di funzioni di Specht, si può anzitutto approssimare l'esponenziale in serie di Taylor intorno al valor medio (pari a zero):

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

- Sostituendo nella (1) si ha:

$$\hat{p}_i(\mathbf{x}) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}} N_i} \mathrm{e}^{-\frac{\mathbf{x}^{\mathrm{t}}\mathbf{x}}{2\sigma^2}} \sum_{j=1}^{N_i} \mathrm{e}^{\frac{\mathbf{x}^{\mathrm{t}}\mathbf{y}_j}{\sigma^2}} \underbrace{\mathrm{e}^{-\frac{\|\mathbf{y}_j\|^2}{2\sigma^2}}}_{\text{noto}} \quad \cong$$

$$\cong \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}} N_i} \mathrm{e}^{-\frac{\mathbf{x}^{\mathrm{T}}\mathbf{x}}{2\sigma^2}} \sum_{j=1}^{N_i} \mathrm{e}^{-c_j} \sum_{h=0}^{r} \left(\mathbf{x}^{\mathrm{t}}\mathbf{y}_j\right)^h \frac{1}{\sigma^{2h} h!}$$

- Come si vede il termine $c_j$ è calcolabile:

$$c_j = \frac{\|y_j\|^2}{2\sigma^2}$$

- L'espressione sopra vale per qualsiasi classe *i,* pertanto, si può scrivere la funzione discriminante della classe $\omega_k$, usando Bayes, come:

$$g_k(\mathbf{x}) = \hat{p}(x|\omega_k)\hat{p}(\omega_k) \cong \frac{\hat{p}(\omega_k)}{N_k}\sum_{j=1}^{N_k} e^{-c_j^k}\sum_{h=0}^{r}\left(\mathbf{x}^t\mathbf{y}_j^k\right)^h \frac{1}{\sigma^{2h}h!}$$

- Questa espressione può essere ottenuta grazie alla semplificazione dei termini comuni del tipo

$$\exp\left\{-\frac{\mathbf{x}^t\mathbf{x}}{\left(2\sigma^2\right)}\right\}$$

- Quando $\sigma$ è abbastanza grande è sufficiente una approssimazione con $r$ piccolo e la funzione discriminante diventa, al limite per $\sigma \to \infty$, lineare (*Metodo dei Prototipi*).

- Nel caso di $\sigma \to 0$, è necessaria una approssimazione con $r$ grande e la funzione discriminante diventa una sommatoria di delta di Dirac (come si può vedere facendo il limite per $\sigma \to 0$ nell'eq. (1)).

- Il metodo diventa così sostanzialmente equivalente al metodo di classificazione dei k-Nearest Neighbors (k-punti vicini, k-NN) (v. in seguito).

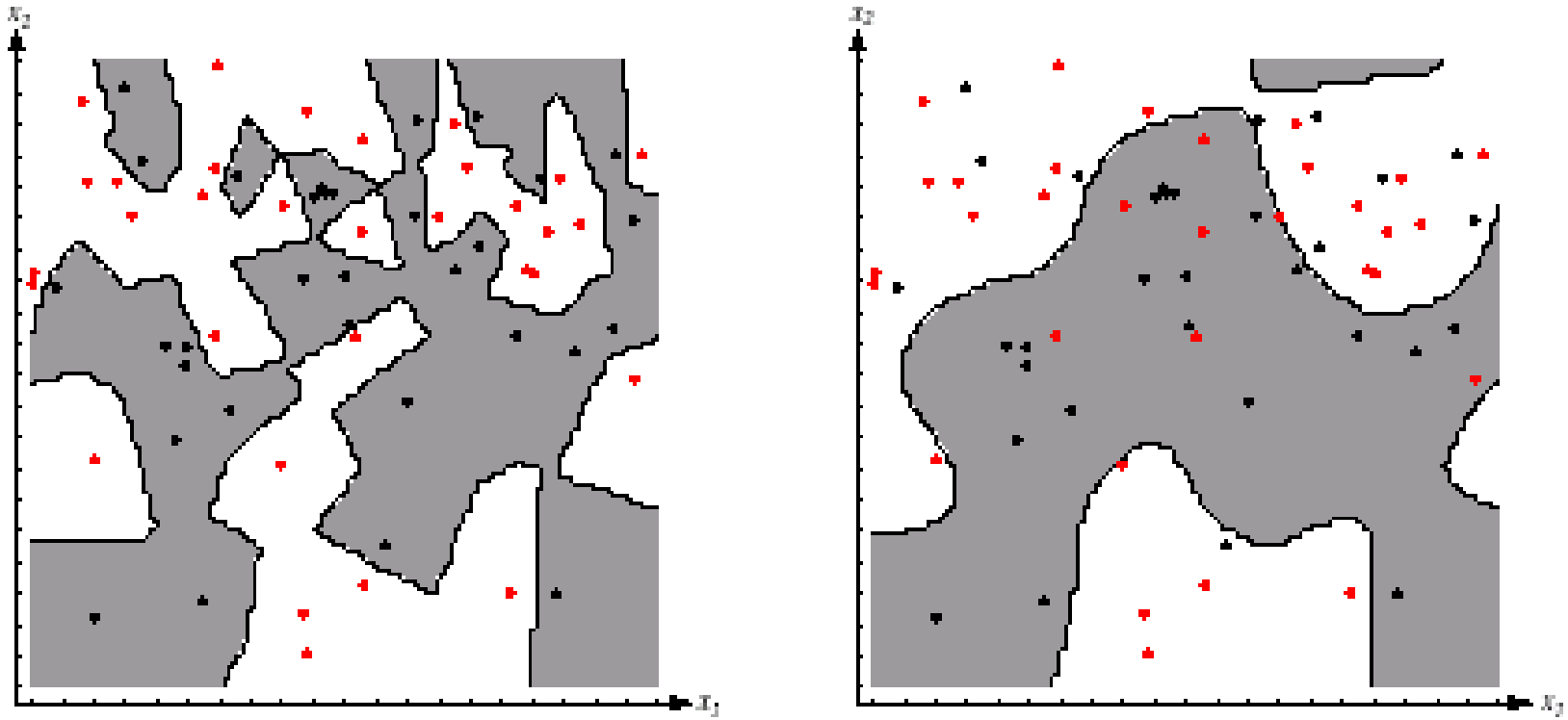



FIGURE 4.8. The decision boundaries in a two-dimensional Parzen-window dichotomizer depend on the window width $h$. At the left a small $h$ leads to boundaries that are more complicated than for large $h$ on same data set, shown at the right. Apparently, for these data a small $h$ would be appropriate for the upper region, while a large $h$ would be appropriate for the lower region; no single window width is ideal overall. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.

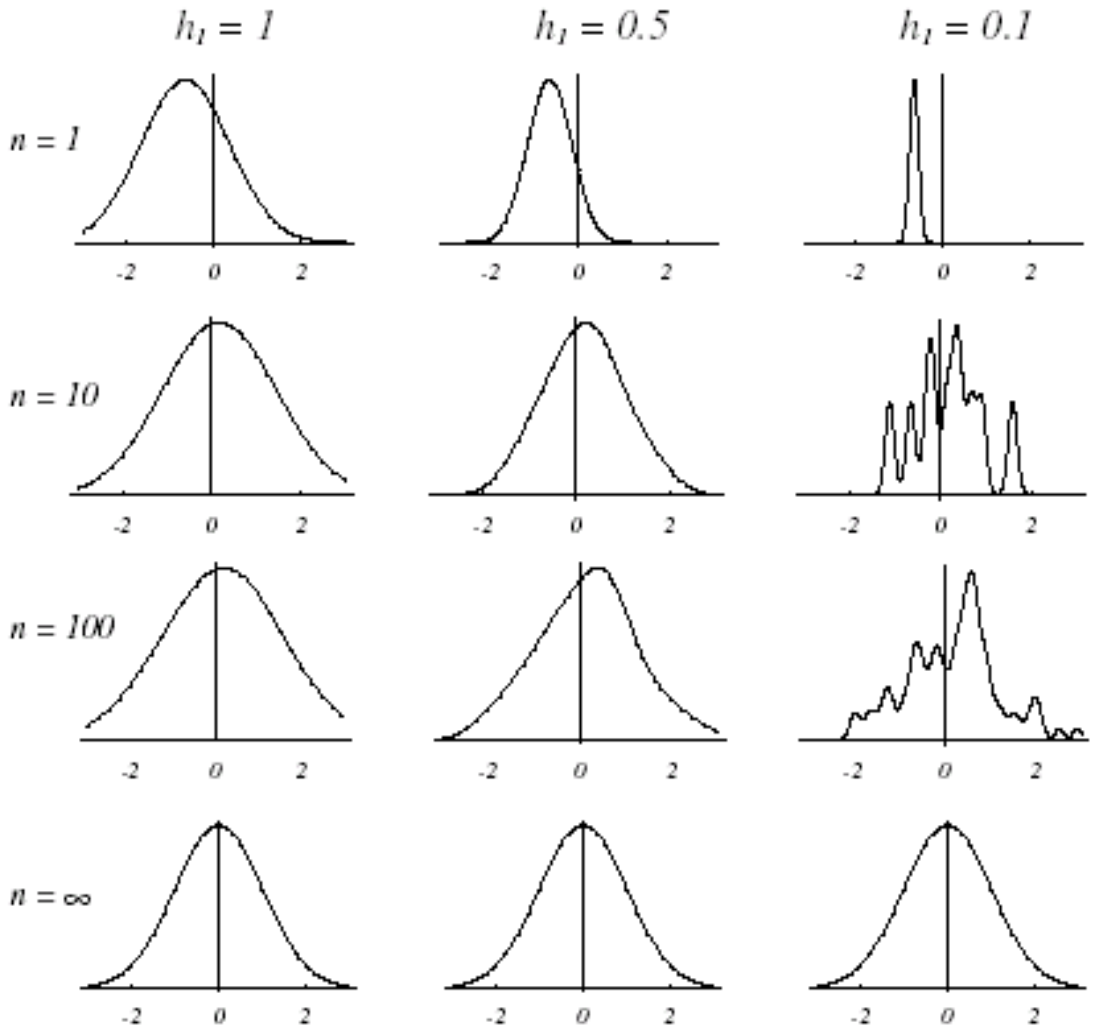
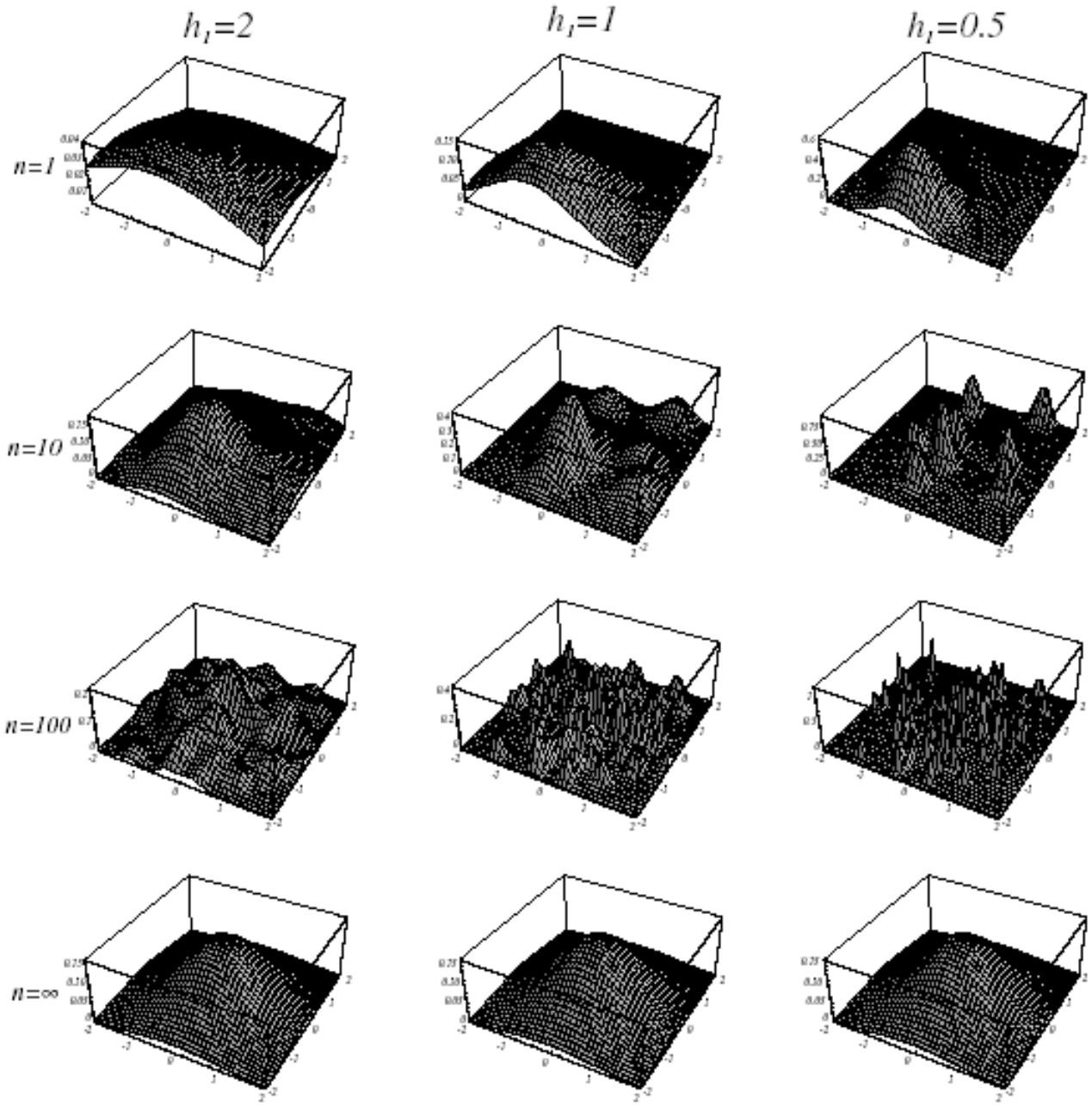
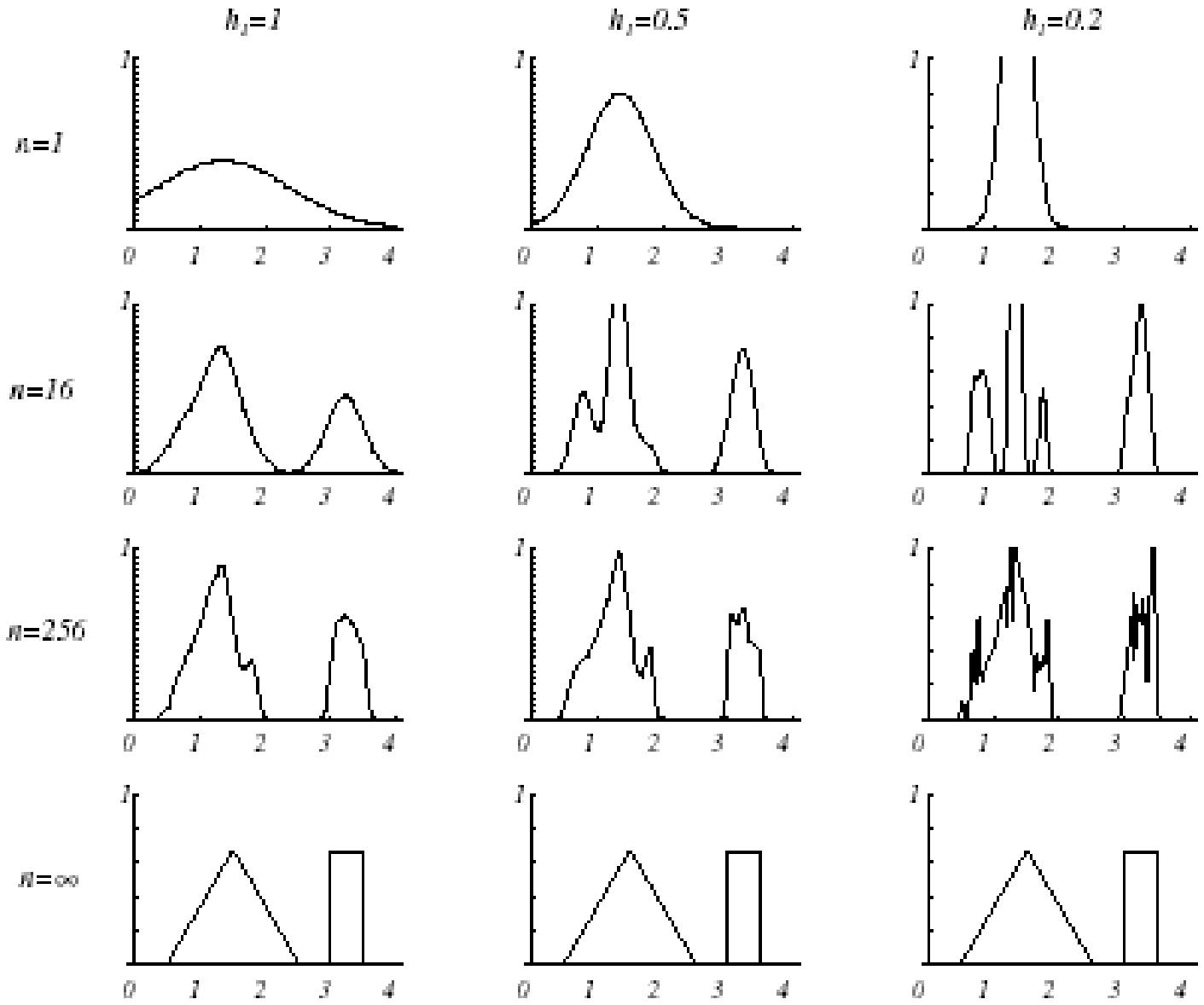



FIGURE 4.5. Parzen-window estimates of a univariate normal density using different window widths and numbers of samples. The vertical axes have been scaled to best show the structure in each graph. Note particularly that the $n = \infty$ estimates are the same (and match the true density function), regardless of window width. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

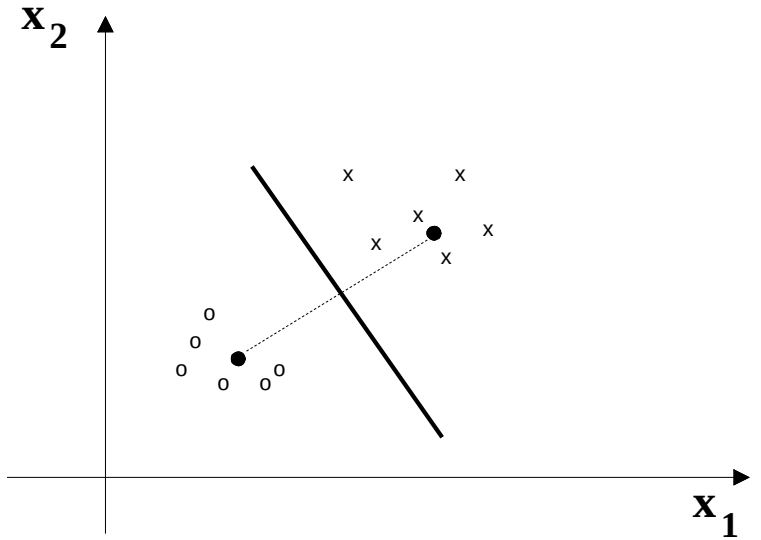

$h_1$=2
$h_1$=1
$h_1$=0.5
n=1
n=10
n=100
n=∞


$h_1=1$
$h_1=0.5$
$h_1=0.2$
$n=1$
$n=16$
$n=256$
$n=\infty$
1
0 1 2 3 4

# Metodo dei prototipi (o degli archetipi)

- Il metodo dei prototipi è in genere applicato per trovare le regioni di decisione di un classificatore basato sulla Minima Distanza, MDC. In generale appartengono alla classe di metodi di Classificazione Semplificata.
- Le distanze utilizzate possono essere molteplici.
- Si suppone che i campioni di una classe tendano a concentrarsi strettamente intorno ad un pattern rappresentativo della classe stessa.
- Questa situazione è propria di casi dove la variabilità del pattern o i disturbi nella fase di osservazione siano regolari e quindi ben modellabili.
- Seguendo questa analogia possiamo dire che il metodo dei prototipi va bene quando i pattern da riconoscere sono una alterazione di una realtà che conosciamo deterministicamente.
- In questo caso i classificatori a minima distanza sono estremamente efficienti.

- Si considerano per esempio tutti i campioni della classe come descritti dal baricentro, $\mathbf{m}$, della classe stessa.




- Sia $D(\mathbf{x},\mathbf{m}_i)$ una metrica nello spazio delle feature.
- La regola di decisione applicata da un MDC è di scegliere la classe $\omega_k$ se $D(\mathbf{x},\mathbf{m}_k) = \min D(\mathbf{x},\mathbf{m}_j)$, per $j = 1,..,M$ e $j \neq k$

- Il metodo per determinare le funzioni discriminanti secondo il metodo dei prototipi è diviso in tre passi.

  1) Si determina il baricentro $\mathbf{m}_i$ della i-esima classe.

  2) Si determina la retta congiungente due baricentri di due classi, caratterizzata dalla equazione $r_{ij}(\mathbf{x})$.

  3) Si calcola l'iperpiano $g_{ij}^*(\mathbf{x})$ perpendicolare a $r_{ij}(\mathbf{x})$.

  4) Si determina il punto di passaggio nello spazio delle feature di tale iperpiano sulla base delle diverse probabilità delle due classi.

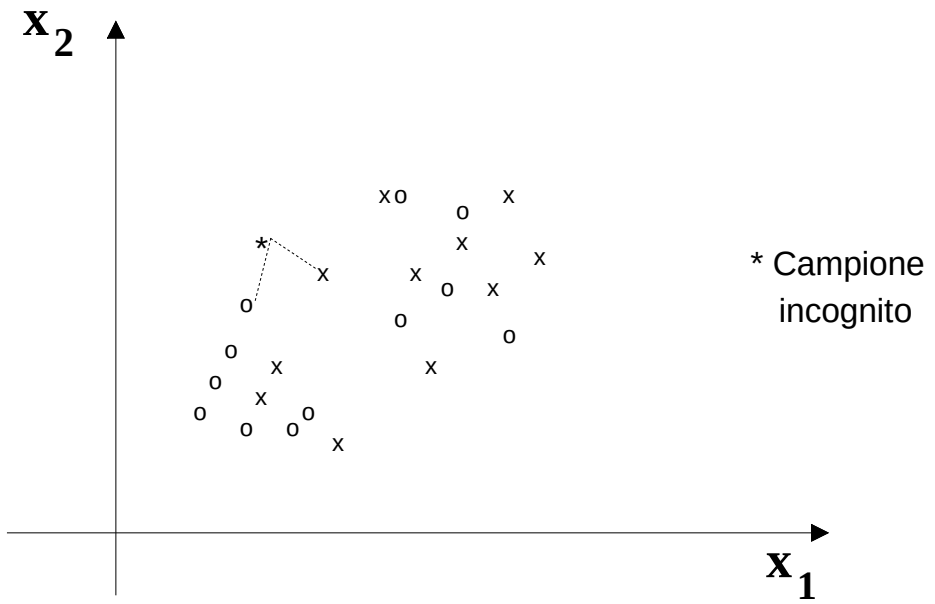
  Più una classe è probabile o ha una importanza maggiore più la funzione discriminante stimata sarà lontana dal baricentro della classe stessa (v. MAP).

  5) Si ripetono i punti 1-4 per tutte le coppie di classi.

- Si determinano così le funzioni discriminanti $f_{ij}(\mathbf{x})$.
- Questo metodo individua funzioni discriminanti lineari. Il metodo dei prototipi è pertanto un caso particolare di un classificatore lineare.
- Vantaggi:
  - semplicità
  - occupazione bassa di memoria.
- Svantaggi:
  - Assunzione che una realtà articolata possa essere rappresentata da un solo campione (prototipo o archetipo della classe), in quanto tale campione potrebbe anche non esistere nella realtà fisica.
- Il classificatore a minima distanza classifica una osservazione sulla base della più vicina distanza (ovvero del "match" più probabile) tra l'osservazione ed i prototipi delle classi.
- Questo approccio è equivalente al metodo di *template matching* (correlazione di un modello con i dati).

# Metodo dei *k* punti vicini (*k* nearest-neighbors, k-NN)

- Il metodo dei *k* punti vicini è derivato da quello dei prototipi, nel senso che applica la stessa regola del classificatore MDC prendendo come riferimento non il baricentro di tutte le classi ma un insieme di punti variabili per un sottinsieme di classi tra quelle trovate nella fase di training.



- Può essere quindi visto come un metodo applicabile al caso in cui *ogni classe sia descritta tramite un insieme allargato di prototipi*.
- Dato un punto incognito *x** si considera $s_i \in \{s_1,\ldots,s_N\} = S$ il punto più vicino (NN) del punto $x^*$ se

$$d(x^*, s_i) = \min_l d(x^*, s_l), \quad l = 1 \ldots N$$

  dove $d(.)$ è una misura di distanza definibile sullo spazio delle feature.
- Se si interpreta ogni punto dell'insieme $S$ come un prototipo di una classe, si può quindi vedere l'eq. sopra come una regola di classificazione 1-NN che associa il campione $x$ alla classe $j$ a cui appartiene il punto $s_i$.

- La regola 1-NN può essere generalizzata per definire una regola k-NN, che consiste nel determinare i $k$ punti appartenenti all'insieme $S$ più vicini alla osservazione $x$.
- La regola di classificazione k-NN associa alla osservazione $x$ la classe $i$ che abbia il maggior numero di elementi tra i $k$ più vicini.
- Chiamiamo $\mathrm{U}(x)$ l'insieme dei $k$ punti più vicini ad $x$.
- Per esempio con $k$ dispari e due classi, $\omega_1$ ed $\omega_2$ la regola di decisione può scegliere la classe con il maggior numero di campioni in $\mathrm{U}(x)$.
- La regola di classificazione alternativa a quella di maggioranza può essere più complicata.

- Decido di caratterizzare in modo univoco i sottoinsiemi di prototipi appartenenti alle diverse classi tra i k-NN, ad esempio calcolandone il baricentro.
- Posso poi misurare la distanza del punto $x$ da tali punti per decidere la classe.
- Per $k = 1$ l'insieme $\mathrm{U}(x)$ è dato dal campione più vicino al campione incognito.
- Per $k$ grande, cioè $k \cong N$, il metodo diventa equivalente al metodo dei prototipi, in quanto il baricentro misurato corrisponde a quello della classe ed uso un classificatore MDC.

- Per utilizzare questo metodo devo tenere in conto le seguenti considerazioni:
  - la metrica deve essere "buona", cioé discriminante;
  - vengono usate solo poche informazioni sulla situazione dello spazio delle feature, cioè solo $k$ punti di tale spazio;
  - è necessario memorizzare tutti i campioni;
  - di conseguenza il metodo è più indicato più il numero di campioni $N$ è basso;
  - bisogna sempre applicare la normalizzazione delle feature in fase di training;
  - le superfici di decisione create sono non-lineari;
  - una scelta tipica è $k \cong \sqrt{N}$ ;
  - l'insieme di test deve essere vasto e possibilmente con pochi errori.

# Come nasce k-NN: stima della pdf

- Il metodo degli archetipi e la sua generalizzazione al k-NN può essere considerato come un metodo di classificazione “semplificato”
- Ma può anche essere usato come un metodo di stima della $p(\mathbf{x})$ e quindi interpretarlo nella sua accezione classica come metodo di classificazione tradizionale
- Invece di fissare $V$ come nel caso dei kernel, si fissa il valore di $k$ e si allarga il raggio della sfera fino a comprendere $k$ campioni
- Vediamo la probabilità a priori: data una classe $\omega_j$, si valuta in genere semplicemente la frequenza di occorrenza dei campioni $N_j$ della classe $j$ rispetto al numero totale di campioni $N$, i.e.,

$$P(\omega_j) \equiv \hat{p}(\omega_j) = \frac{N_j}{N}$$

# Stima della densità e regola dei punti vicini

- 2 classi, $\omega_i$ e $\omega_j$, contenenti $N_i$ ed $N_j$ campioni, $N = N_i + N_j$
- La stima locale di densità per $\omega_i$ si calcola come (e analogamente per $\omega_j$)

$$\hat{p}(x|\omega_i) = \frac{1}{V}\frac{k_i}{N_i}$$

i.e., rapporto tra $k_i$ ($k_j$) punti sugli $N_i$ ($N_j$) totali appartenenti alla classe $\omega_i$ ($\omega_j$) contenuti nel volume $V$

- La regola di Bayes recita $p(x|\omega_i)P(\omega_i) > p(x|\omega_j)P(\omega_j)$, allora

$$\hat{p}(x|\omega_i)\hat{p}(\omega_i) > \hat{p}(x|\omega_j)\hat{p}(\omega_j)$$

$$\Rightarrow \frac{1}{V}\frac{k_i}{N_i}\frac{N_i}{N} > \frac{1}{V}\frac{k_j}{N_j}\frac{N_j}{N} \quad \Rightarrow \quad k_i > k_j$$

- Se ci sono dei casi non determinati (pareggi) ci sono alcune alternative:
  - scelta arbitraria;
  - assegnare $x$ alla classe (tra quelle “in pareggio”) che ha il campione medio più vicino (calcolato tra i $k_i$ campioni);
  - assegnare $x$ alla classe che ha i $k_i$ campioni con distanza minore da esso;
  - etc.

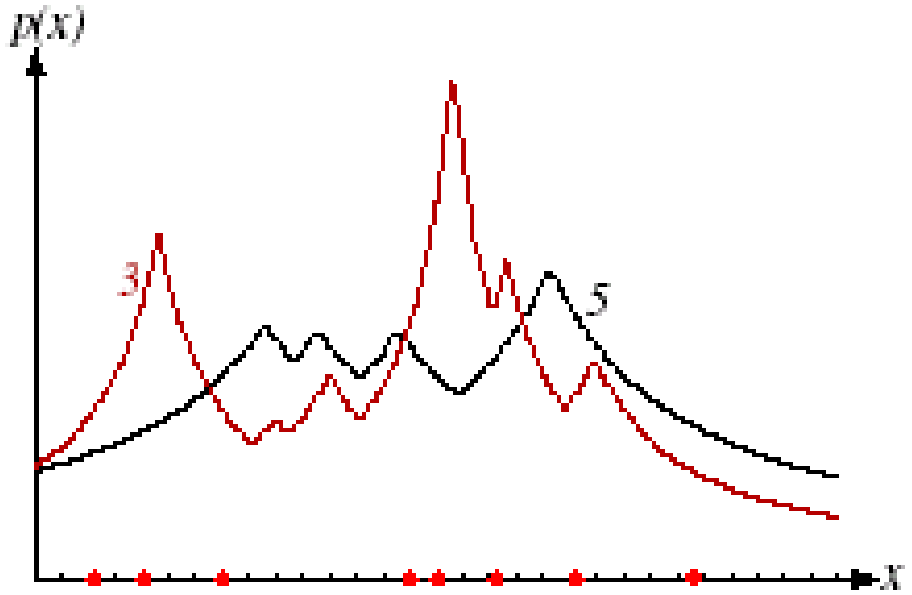



FIGURE 4.10. Eight points in one dimension and the $k$-nearest-neighbor density estimates, for $k = 3$ and 5. Note especially that the discontinuities in the slopes in the estimates generally lie *away* from the positions of the prototype points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

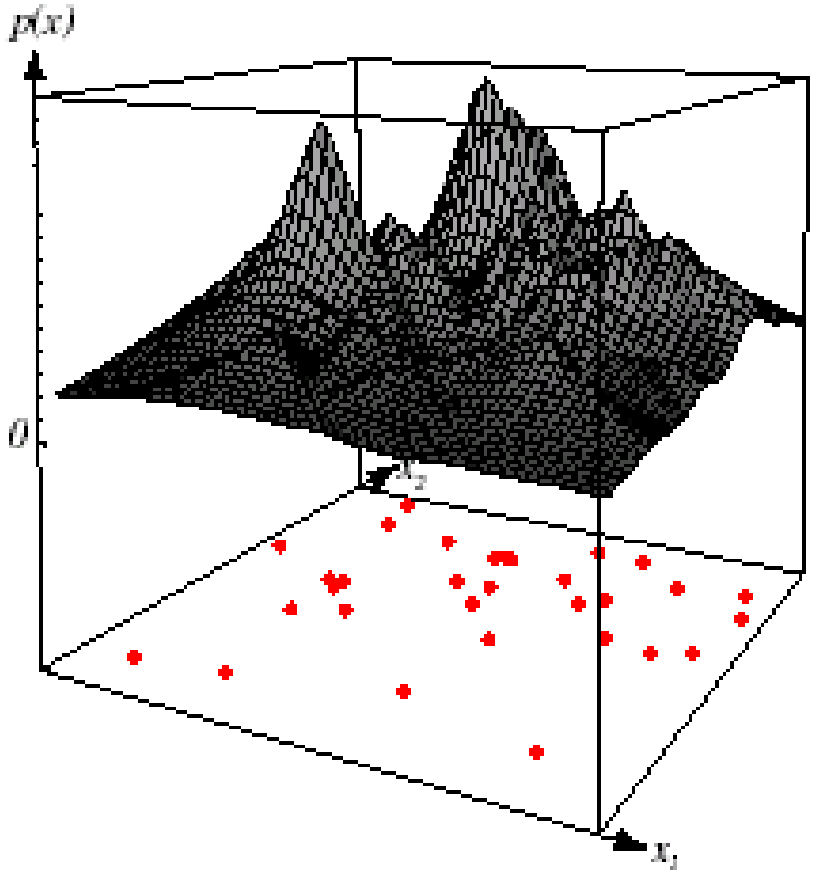



FIGURE 4.11. The $k$-nearest-neighbor estimate of a two-dimensional density for $k = 5$. Notice how such a finite $n$ estimate can be quite "jagged," and notice that discontinuities in the slopes generally occur along lines away from the positions of the points themselves. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.