Riconoscimento e recupero dell’informazione per bioinformatica

Classificatori generativi

Manuele Bicego

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

Sommario

- Approcci alla classificazione:
  - classificazione generativa vs classificazione discriminativa
- Classificazione generativa: stima parametrica della pdf
  - Stima maximum Likelihood
  - Stima Bayesiana
- Classificazione generativa: stima non parametrica
  - Parzen Windows
  - K-Nearest Neighbor

Approcci alla classificazione

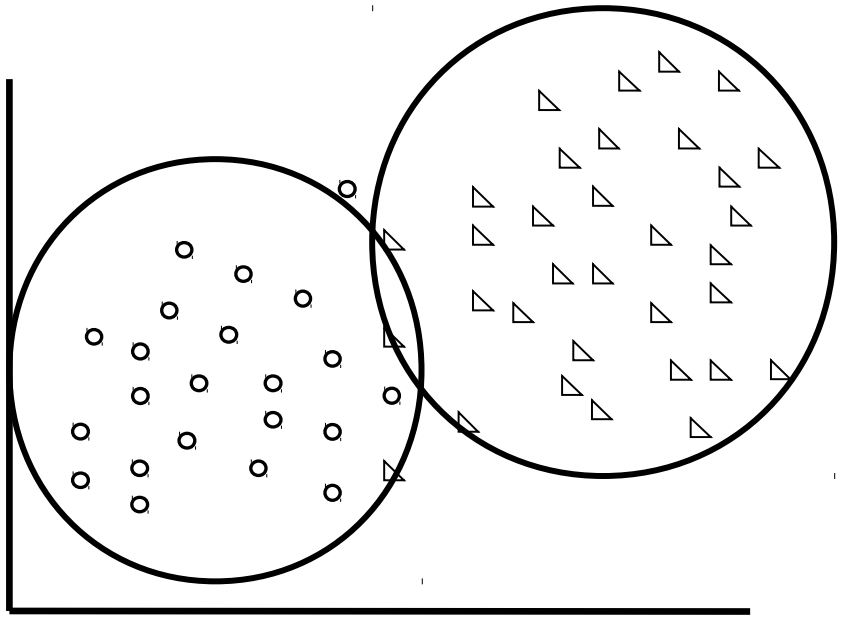
Approcci alla classificazione

RIASSUNTO:

- Per la classificazione, la regola che minimizza la probabilità di errore è quella di Bayes (assegnare un oggetto alla classe la cui posterior è maggiore)
- La regola utilizza le posterior (che non sono note)
- Per stimare le posterior si possono o non si possono utilizzare la likelihood e il prior
- Approcci:
  - Approcci generativi: si calcolano likelihood e prior
  - Approcci discriminativi: si calcola direttamente le posterior e il corrispondente confine di decisione
  - (si può anche stimare direttamente la funzione f(x) che mappa gli oggetti nelle classi)

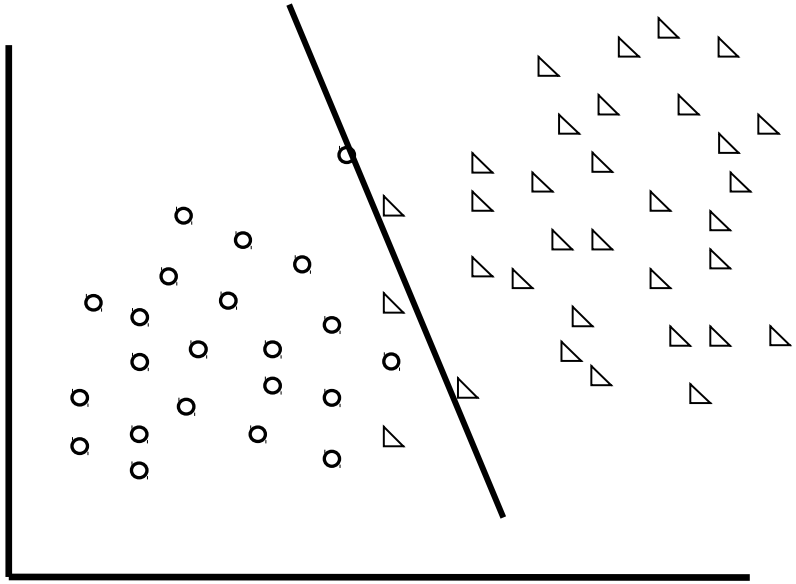
Approcci alla classificazione

Generativi: un modello per ogni classe



$$\tilde{y} = \text{argmax}_y P(y|x)$$

Discriminativi: modellano direttamente il confine



$$\tilde{y} = sign(w \cdot x + b)$$

Approcci alla classificazione

Aspetto	**Meglio generativo o discriminativo?**	**Motivazione**
Capacità descrittiva per le singole classi	Generativo	Un modello permette di generare nuovi dati (descrizione completa delle classi)
Efficacia della descrizione	Discriminativo	Un modello generativo descrive anche parti dello spazio del problema non utili ai fini della discriminazione
Capacità di gestire dati non vettoriali	Generativo	Modelli probabilistici specifici per dati non vettoriali (esempio Modelli di Markov per sequenze)

Approcci alla classificazione

Aspetto	**Meglio generativo o discriminativo?**	**Motivazione**
Facilità di addestramento	Generativo	Il discriminativo deve considerare più di una classe al colpo
Estendibilità a nuove classi	Generativo	Basta addestrare un nuovo modello, per il discriminativo occorre riaddestrare tutto
Velocità di testing	Discriminativo	Tipicamente il generativo richiede una soluzione iterativa per prendere una decisione

Approcci alla classificazione

Aspetto	**Meglio generativo o discriminativo?**	**Motivazione**
Efficacia di classificazione	Discriminativo	Sistema costruito specificatamente per risolvere il problema di classificazione
Eleganza matematica	Generativo	Teoria e interpretabilità probabilistica dei modelli generativi
Flessibilità	Discriminativo	Problemi nell'approccio generativo se si assume un modello sbagliato

Approccio generativo alla classificazione

Approccio generativo

RIASSUNTO

- Si stimano le probabilità a priori e le probabilità condizionali

- Si combinano a formare le probabilità a posteriori

- Si classifica con la regola di Bayes

Stima delle probabilità

Problema: le probabilità sono sconosciute, occorre stimarle dal training set!

- ⇨ 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 nessuna forma per la pdf, ma viene stimata direttamente dai dati
  - ⇨ esempio istogramma

Stima parametrica

Introduzione

Stima parametrica

Per costruire un classificatore bayesiano **con stima parametrica** si procede in questo modo:

- Si stima dal training set la probabilità a priori per ogni classe
- Per la probabilità condizionale:
  - Si decide (*o si stima*) la forma per ogni classe (ad esempio gaussiana)
  - Si stimano i parametri a partire dai dati di training (un insieme di parametri per ogni classe)
- Si usano le stime risultanti come se fossero i valori veri e si utilizza la teoria di decisione Bayesiana per costruire il classificatore

Stima dei parametri – Probabilità a priori

Stima della probabilità a priori: più facile

⇨ Si ha a disposizione il training set: un insieme di n dati di training in cui ad ogni pattern x_j è assegnata un'etichetta ω

⇨ Allora si può stimare $p(\omega_i)$ come

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

dove n_i è il numero di campioni con etichetta ω_i

Stima della probabilità condizionale

La stima della prob. condizionale rappresenta il vero problema!!

⇨ L'obiettivo è stimare i **parametri sconosciuti** della **funzione conosciuta** p(x|ωj)

Per es., stimare il vettore $\theta_j = (\mu_j, \Sigma_j)$ sapendo che

$$p(x|\omega_j) \approx N(\mu_j, \Sigma_j)$$

Esiste anche un caso più difficile*:* si vuole stimare sia la forma che i parametri della funzione sconosciuta p(x|ωj)

Stima dei parametri – Probabilità condizionale

- Il training set **T** può essere diviso in sottoinsiemi $\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_c$ (uno per ogni classe)

- Assunzioni che si fanno
  - $\mathbf{T}_j$ contiene campioni generati dalla probabilità $p(x|\omega_j)$
  - I campioni appartenenti al set $\mathbf{T}_i$ non danno informazioni relative ai parametri di $p(x|\omega_j)$ se $i \neq j$.

Stima dei parametri – Due approcci

Essendo i diversi $\mathbf{T}_j$ indipendenti, basta trovare la soluzione al seguente problema (da applicare poi a tutti i $\mathbf{T}_j$)

Dato un set di dati di training $D=\{x_1, x_2, ..., x_n\}$ (estratti indipendentemente), si vuole stimare il parametro $\boldsymbol{\theta}$ (che determina in maniera univoca $p(x|\omega)$) a partire da D

$\boldsymbol{\theta}$ è un vettore che rappresenta i parametri necessari a descrivere in modo univoco $p(x|\omega)$

p.e., $\theta=(\mu,\Sigma)$ se $p(x|\omega)\approx N(\mu,\Sigma)$

⇨ Esistono due approcci
- ⇨ Stima Maximum-likelihood (ML)
- ⇨ Stima di Bayes

Stima parametrica

Stima Maximum Likelihood

Approccio Maximum Likelihood

Punto di partenza:

- Definizione: la likelihood del training set **D,** cioè la "probabilità di tutti i punti del training set"
  - Si indica con P(**D**|θ)
  - Probabilità che dipende dal vettore di parametri θ

- Sapendo che i pattern del set D sono i.i.d., si ha che

$$p(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$$

Approccio Maximum Likelihood

⇨ NOTA: Dato un $\boldsymbol{\theta}$, $P(\mathbf{D}|\theta)$ può essere vista come una misura di "quanto bene il dataset è spiegato dal modello di parametro $\boldsymbol{\theta}$"

La stima di Maximum Likelihood di $\boldsymbol{\theta}$ è, per definizione, il valore $\hat{\boldsymbol{\theta}}$ che massimizza $p(\mathbf{D}|\boldsymbol{\theta})$;

"Si cerca il parametro $\boldsymbol{\theta}$ che meglio spiega il dataset"

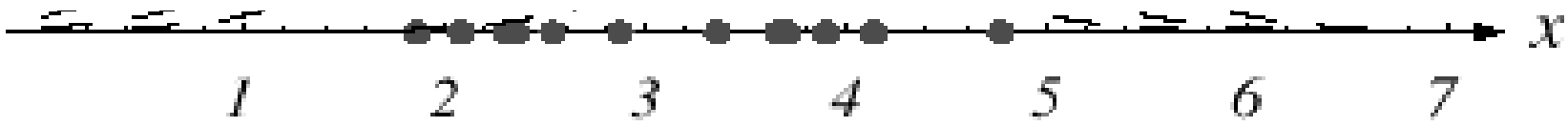
⇨ Vista come funzione di $\boldsymbol{\theta}$, $P(\mathbf{D}|\theta)$ viene chiamata ***likelihood*** *di* θ *rispetto al set di campioni* ***D.***

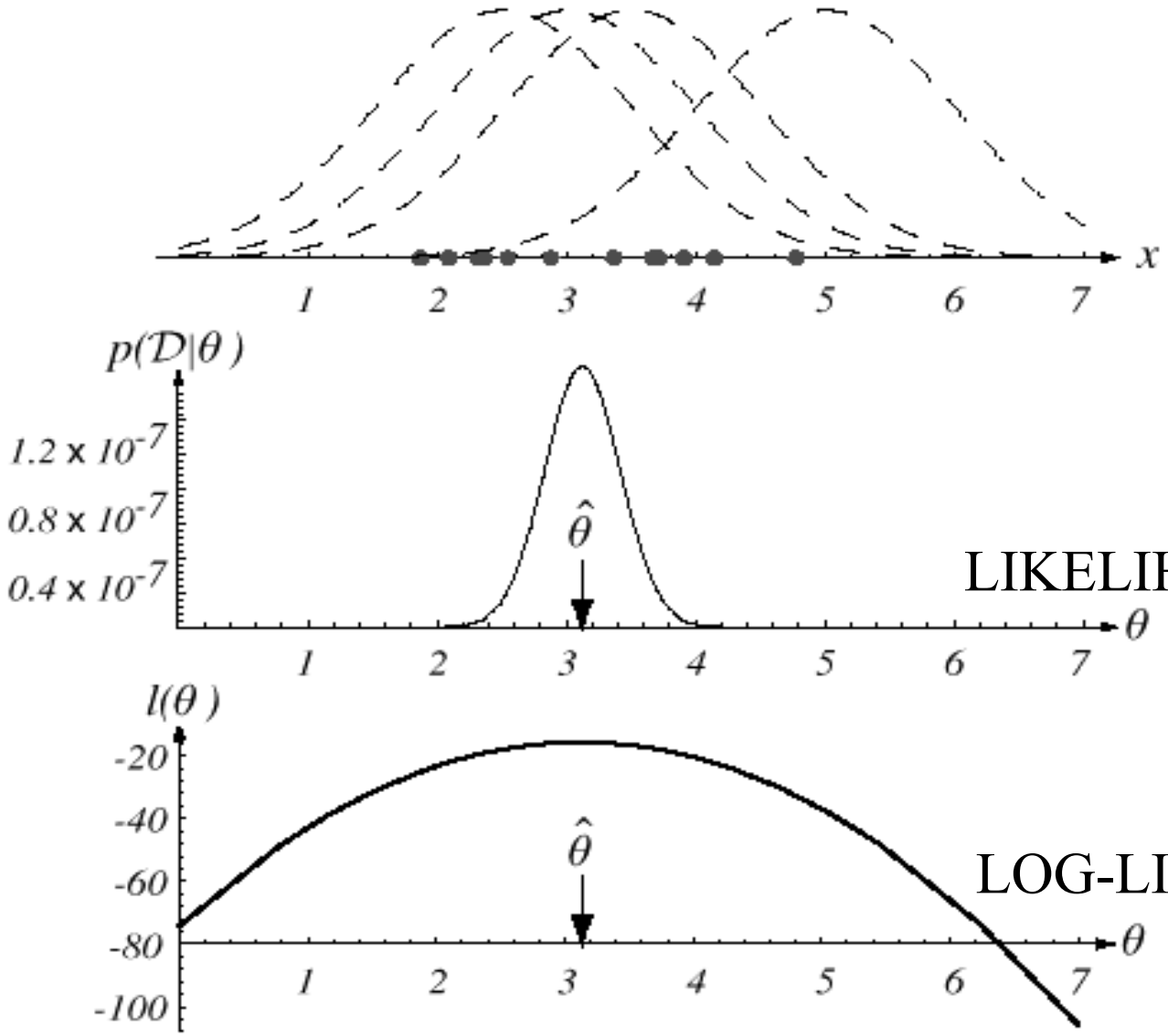
Approccio Maximum Likelihood

- Assunzione: $\boldsymbol{\theta}$ è fissato ma sconosciuto

Esempio:

- Punti di training 1-d generati da una densità gaussiana di varianza nota ma media sconosciuta
- Goal: stimare i parametri (cioè la media)





4 delle infinite possibili gaussiane

LIKELIHOOD

LOG-LIKELIHOOD

la stima ML corrisponde al valore che “spiega meglio” (è in accordo) le osservazioni (il training set D)

Approccio Maximum Likelihood

NOTE IMPORTANTI:

- La likelihood p(**D**|θ) è funzione di θ, mentre la densità condizionale p(x|θ) è funzione di x
- Più dati ci sono nel training set D, più è stretto il picco attorno al valore massimo

Goal: ottimizzare la likelihood

- Per scopi analitici risulta più semplice lavorare con il logaritmo della likelihood.
- Definiamo quindi l(**θ**) come ***funzione di log-likelihood***

$$l(\theta) \equiv \ln p(D|\theta) = \sum_{k=1}^{n} \ln p(x_k|\theta)$$

Approccio Maximum Likelihood

- Lo scopo è di ottenere quindi il vettore

$$\hat{\theta} = \arg\max_{\theta} l(\theta)$$

- dove $\boldsymbol{\theta}=(\theta_1 \ldots \theta_p)^t$ è il vettore di p parameteri da stimare
  - (la dipendenza sul data set **D** è implicita)

Defininiamo l'operatore gradiente

$$\nabla\theta \equiv \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix}$$

Approccio Maximum Likelihood

⇨ Per ricavare il max:

$$l(\theta) \equiv \ln p(D|\theta) = \sum_{k=1}^{n} \ln p(x_k|\theta)$$

$$\nabla_\theta l(\theta) = \sum_{k=1}^{n} \nabla_\theta \ln p(x_k|\theta)$$

NOTA: è evidente il vantaggio dell'utilizzare il logaritmo (derivata della somma e non del prodotto)

⇨ vogliamo ottenere $\nabla_\theta l(\theta) = 0$

Maximum Likelihood: caso gaussiano

Applichiamo ora l'approccio ML ad alcuni casi specifici

- Consideriamo che i campioni siano generati da una popolazione normale multivariata di media $\boldsymbol{\mu}$ e covarianza $\boldsymbol{\Sigma}$.

- Due semplici casi:
  - monodimensionale, media sconosciuta varianza nota (alla lavagna)
  - monodimensionale, media e varianza sconosciuta

ML – modello d’errore

- In generale, se i modelli parametrici sono validi, il classificatore *maximum-likelihood* fornisce risultati eccellenti.
- Invece, se si usano famiglie parametriche scorrette, il classificatore produce forti errori
  - Questo accade anche se è nota la famiglia parametrica da usare ma si sbaglia qualcosa (Esempio di prima (media sconosciuta, varianza nota), stima scorretta della varianza)
- Di fatto *manca un modello d’errore che dia un voto alla parametrizzazione ottenuta*.
- Inoltre, per applicare la stima di Maximum-Likelihood, tutti i dati di training devono essere disponibili
  - Se vogliamo utilizzare nuovi dati di training, è necessario ricalcolare la procedura di stima Maximum-Likelihood

Stima parametrica

Stima di Bayes

Riassunto

- Riassunto: approccio generativo alla classificazione:
  - Dati C training set $\mathbf{T}_1$,$\mathbf{T}_2$,... $\mathbf{T}_C$ (disgiunti e indipendenti, ognuno contenente gli elementi di una classe)
  - il goal è stimare $P(\omega_i)$ e $P(x|\omega_i)$ per ogni classe, combinarle per avere la posterior e classificare con la regola di Bayes

Si possono vedere come C istanze diverse dello stesso problema:

Dato *un set di campioni D, estratti secondo la distribuzione sconosciuta p(**x**), l'obiettivo è determinare p(**x**|D) (stima che dipende dal training set D)*

Riassunto

- Stima parametrica: p(x|D) ha una forma nota, e si parametrizza con un vettore Θ (e.g. la media di una gaussiana)

- Due approcci:
  - Maximum Likelihood: si assume p(x|D) = p(x|Θ)
    - I punti vengono da una distribuzione di parametro fissato ma sconosciuto
    - L'obiettivo è stimare Θ
  - Bayes: Θ non è fissato ma è una variabile, la p(x|D) viene stimata tenendo conto della distribuzione su questa variabile

Stima di Bayes: la stima

⇨ (alla lavagna)

Stima di Bayes: la stima

Vantaggi:

- ⇨ Stima più accurata: questo approccio ***permette di tenere conto dell'effetto di tutti i possibili valori dei parametri***

$$p(x|D) = \int p(x|\theta)\, p(\theta|D)\, d\theta$$

Svantaggi:

- ⇨ stimare la posterior dei parametri p(Θ|D) non è sempre banale
- ⇨ integrare in tutto lo spazio dei parametri può essere "difficile" o "intrattabile
- ⇨ occorre definire i priori p(Θ)

Esempio: caso Gaussiano

- Utilizziamo le tecniche di stima Bayesiana per calcolare la densità p(**x**|D) per il caso in cui

$$p(x|\theta) \equiv p(x|\mu) \approx N(\mu, \Sigma)$$

- Gaussiana univariata (unidimensionale) di varianza nota, solo la media μ è sconosciuta

Primo passo: occorre definire un prior sul parametro μ, che rappresenti la conoscenza a priori su μ.

- Ci serve una distribuzione p(μ)
  - Assumiamo una gaussiana $p(\mu) \approx N(\mu_0, \sigma_0^2)$

In pratica μ_0 rappresenta la migliore scelta iniziale per il parametro μ, con σ_0^2 che ne misura l'incertezza

Esempio: caso Gaussiano

NOTA: la scelta del prior è arbitraria, ma:

- ⇨ deve essere fatta (il prior deve essere noto)
- ⇨ di solito si sceglie un prior coniugato
  - ⇨ prior che assicura che la forma della posterior $p(\Theta|D)$ sia trattabile, cioè abbia la stessa forma della condizionale
  - ⇨ Questo semplifica di molto l'analisi
  - ⇨ Esempio: gaussiana per gaussiana, dirichlet per multinomiale

Esempio: caso Gaussiano

Secondo passo: stima della posterior p(Θ|D), a partire da *n* campioni di training D={x_1, x_2,..., x_n}

⇨ Si applica il teorema di Bayes, ottenendo

$$p(\mu \mid D) = \frac{p(D \mid \mu)\, p(\mu)}{\int p(D \mid \mu)\, p(\mu)\, d\mu}$$

$$= \alpha \prod_{k=1}^{n} p(x_k \mid \mu)\, p(\mu)$$

dove α è un fattore di normalizzazione dipendente da D.

Esempio: caso Gaussiano

$$p(\mu|\mathcal{D}) = \alpha \prod_{k=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]}^{p(x_k|\mu)} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]}^{p(\mu)}$$

$$= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{\mu-x_k}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right]$$

$$= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right], \qquad (29)$$

⇨ Riarrangiando, si ottiene

$$p(\mu \mid D) = \frac{p(D \mid \mu)p(\mu)}{\int p(D \mid \mu)p(\mu)d\mu} = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\{-\frac{(\mu - \mu_n)^2}{2\sigma_n^2}\}$$

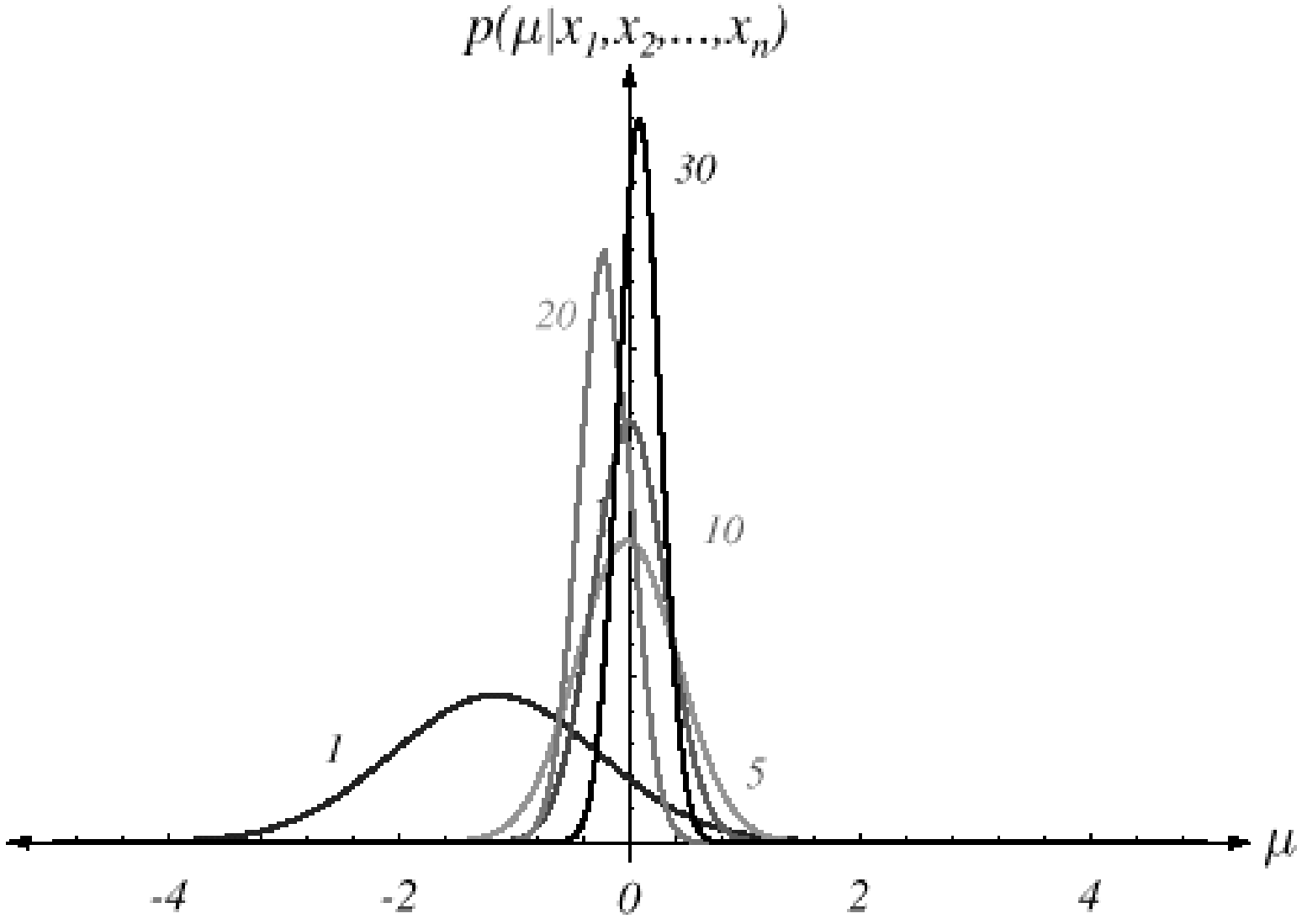
$$\text{dove} \quad \mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right) + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

μ_n rappresenta la nostra migliore scelta per μ dopo aver osservato n campioni.

σ_n^2 misura l'incertezza della nostra scelta.

Esempio: caso Gaussiano



Esempio: caso Gaussiano

Terzo passo: stima della densità condizionale p(x|D)

$$p(x|\mathcal{D}) = \int p(x|\mu)\,p(\mu|\mathcal{D})\,d\mu$$

$$= \int \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]d\mu$$

$$= \frac{1}{2\pi\sigma\sigma_n}\exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right] f(\sigma,\sigma_n), \qquad (36)$$

Esempio: caso Gaussiano

⇨ dove

$$f(\sigma, \sigma_n) = \int \exp\left[-\frac{1}{2} \frac{\sigma^2 + \sigma_n^2}{\sigma^2 \sigma_n^2} \left(\mu - \frac{\sigma_n^2 x + \sigma^2 \mu_n}{\sigma^2 + \sigma_n^2} \right)^2 \right] d\mu.$$

Esempio: caso Gaussiano

- Concludendo, la densità *p(x|D)* ottenuta è la densità condizionale desiderata

$$P(\omega_i \mid \mathbf{x}, D) = \frac{p(\mathbf{x} \mid \omega_i, D)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j, D)P(\omega_j)}$$

che assieme ai prior $P(\omega_i)$ produce le informazioni desiderate per il design del classificatore, al contrario dell'approccio ML che restituisce solo le stime puntuali

$$\hat{\mu} \; e \; \hat{\sigma}^2$$

Conclusioni: Bayes vs ML

- ML restituisce una stima puntuale $\hat{\theta}$, l'approccio Bayesiano una distribuzione su $\boldsymbol{\theta}$ (più ricca, tiene conto di tutti i possibili valori dei parametri)
- Bayes più accurato (in linea di principio), ML più fattibile in pratica
- Inoltre: ML, per un dataset abbastanza grande, produce risultati buoni
  - le stime risultano equivalenti per training set di cardinalità infinita (al limite, p(Θ|D) converge ad una funzione delta)
- In Bayes occorre settare i prior sui parametri (svantaggio o vantaggio?)

Stima non parametrica

Stima non parametrica

- Problema della stima parametrica: si assume che la forma delle densità di probabilità sia nota, ma questa assunzione non può essere fatta in molti problemi di riconoscimento.

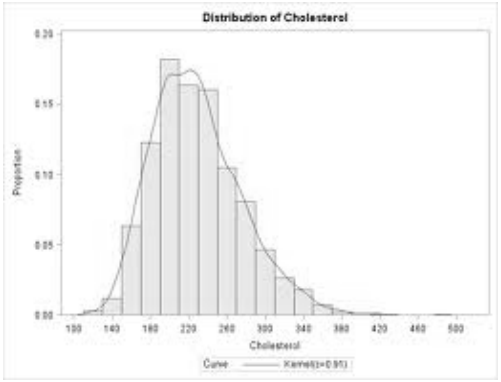
- In particolare, se la scelta della forma parametrica è sbagliata, la stima sarà molto povera
  - Esempio: distribuzione multimodale che viene assunta essere Gaussiana

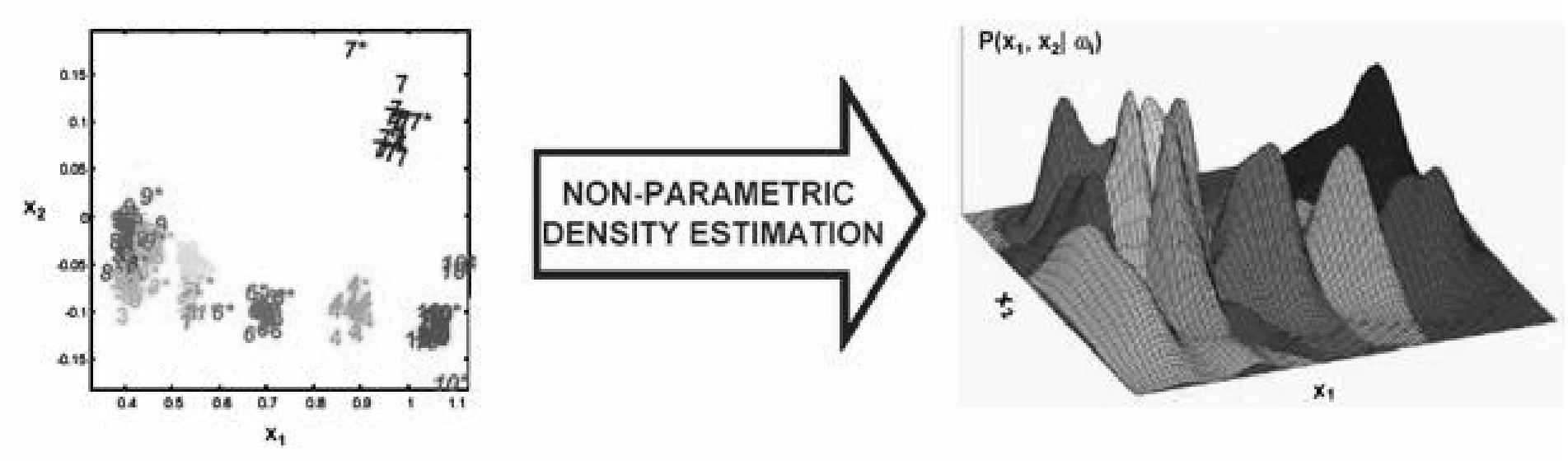
- Soluzione: metodi non parametrici:
  - fanno poche assunzioni (nessuna) sulla forma della pdf da stimare

Stima non parametrica

- Idea: stimare la pdf andando ad analizzare le singole regioni dello spazio
  - Se bisogna stimare p(x=x0), si va a considerare la regione attorno ad x0 e si effettua la stima a partire da quella regione
  - Si può ripetere per tutti i punti dello spazio (o per tutti i punti di interesse)

Stima non parametrica





⇨ Più formalmente (alla lavagna)

Stima non parametrica

Riassumendo:

- Data una regione R di volume V, dati N punti (di cui K cadono nella regione R), si approssima p(x) in quella regione come:

$$p(x) \approx \frac{K}{NV}$$

- NOTA: questa formula deriva da due approssimazioni, la cui bontà dipende da R
  - K/N stimatore di P: migliore per R grande
  - p(x) costante in R: migliore per R piccola
- Scelta di R è quindi cruciale!

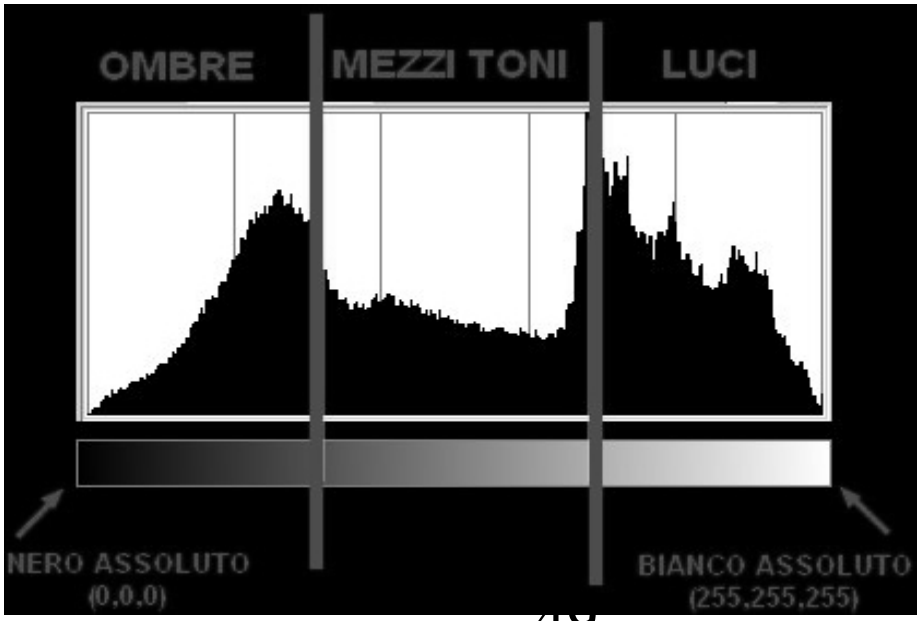
Stima non parametrica

- Come fare a stimare la pdf su tutto lo spazio?
  - Si approssima lo spazio suddividendolo in regioni (uniformi?)
  - Per ogni regione si può calcolare la p(x) con la formula della stima parametrica (la regione ha una p(x) costante)
  - Mettendo assieme tutte queste stime si ha la p(x) totale

(attenzione alla dimensione della regione)

- Esempio:
  - dato un insieme di punti D campionato dalla distribuzione che devo stimare
  - Si suddivide lo spazio in regioni di larghezza uniforme
  - Per ogni regione si conta la frazione di punti di D che appartengono alla regione
  - Queste frazioni caratterizzano la stima non parametrica della pdf

- Esempio: istogramma



Stima non parametrica

Due possibilità per determinare p(x) per ogni possibile x

1. (più intuitiva): si fissa la regione R centrata in x (in particolare si fissa il suo volume V), si calcola K dai dati e si stima p(x)
   - ⇨ più punti ci sono nel volume fissato V, più alta la probabilità
   - ⇨ Parzen Windows

2. (meno intuitiva): si fissa K, si sceglie R in modo tale che contenga K punti attorno ad x, si determina V e si stima p(x)
   - ⇨ più grande è la regione che devo considerare per trovare K punti, più bassa è la probabilità
   - ⇨ K-Nearest Neighbor

Parzen Windows

- Procedimento: dato un training set x1...xN, la probabilità in un generico punto x_0 si stima come segue:
  - Si fissa la regione R centrata in x_0
  - Si calcolano i punti del training set che appartengono alla regione
  - Si determina il volume della regione R
  - Si stima la $p(x_0)$ con la formula

(Dettagli alla lavagna)

$$p(x) \approx \frac{K}{NV}$$

Parzen Windows

- Riassumendo: fissata come regione R un cubo di lato h centrato in un punto x_0, il numero di punti che cade in R è

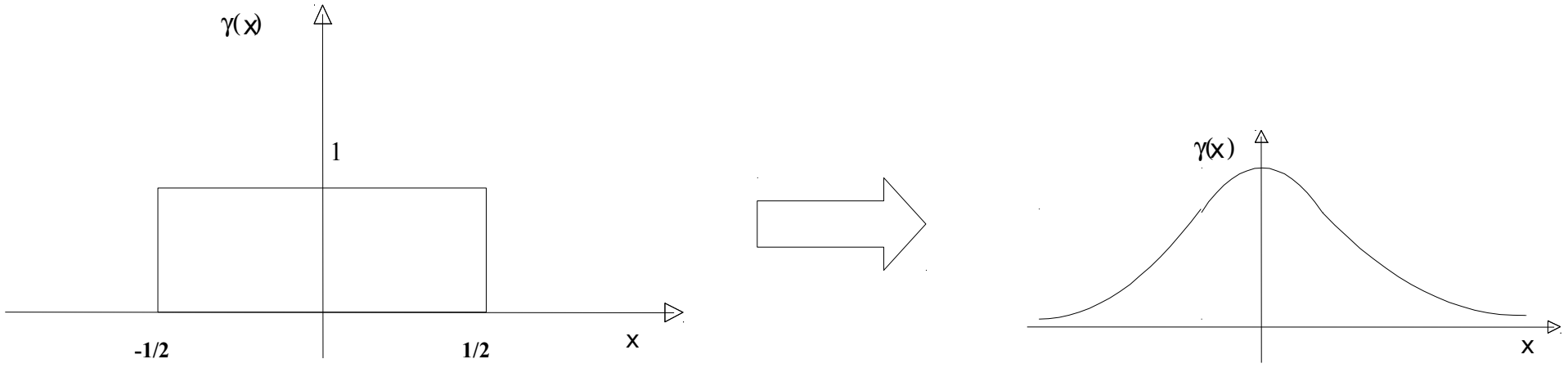
$$k = \sum_{j=1}^{N} \gamma\left(\frac{x_0 - x_j}{h} \right)$$

- E $p(x_0)$ si calcola come

$$p(x_0) = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{h^d} \gamma\left(\frac{x_0 - x_j}{h} \right),$$

Parzen Windows

- Possiamo pensare di usare funzioni gamma più “smooth” (dolci)
- Esempio: caso monodimensionale



Parzen Windows

- La stima della pdf è ottenuta come la media di queste funzioni valutate in ogni punto del training set:
  - Un punto contribuisce a seconda della distanza
- Per avere una p(x) valida occorre che:

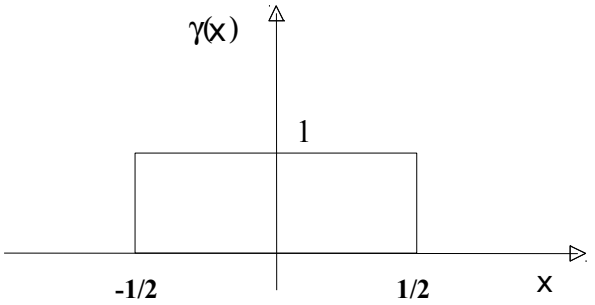
$$\gamma(u) \geq 0$$
$$\int \gamma(u)\, du = 1$$

Ce ne sono di molti tipi, ognuno caratterizzato da un'ampiezza h (l'ampiezza della finestra)

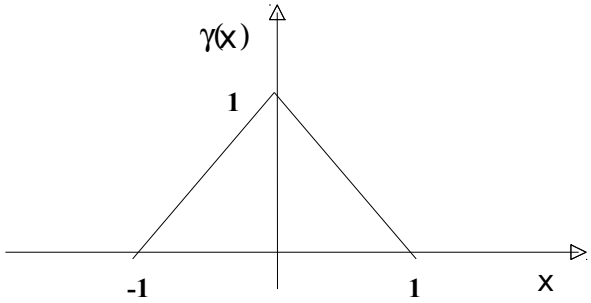
- Tipicamente chiamate FUNZIONI KERNEL!

Esempi di funzioni potenziali

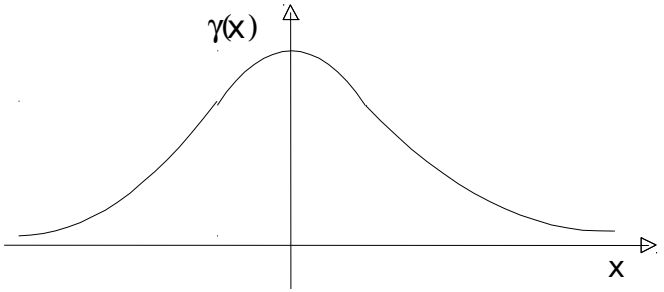
1) $\gamma(x)=\begin{cases}1 & |x|\le 0.5\\ 0 & |x|>0.5\end{cases}$ Rettangolo



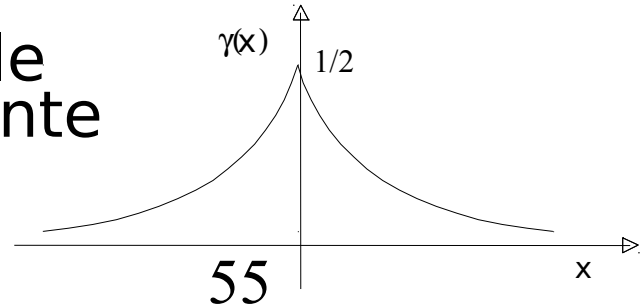
2) $\gamma(x)=\begin{cases}1-|x| & |x|\le 1\\ 0 & |x|>1\end{cases}$ Triangolo



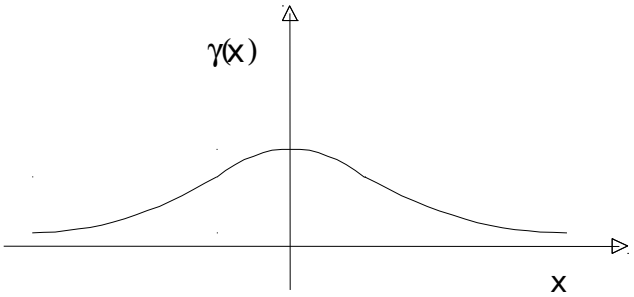
3) $\gamma(x)=(2\pi)^{-\frac{1}{2}}e^{-\left(\frac{x^2}{2}\right)}$ Gaussiana



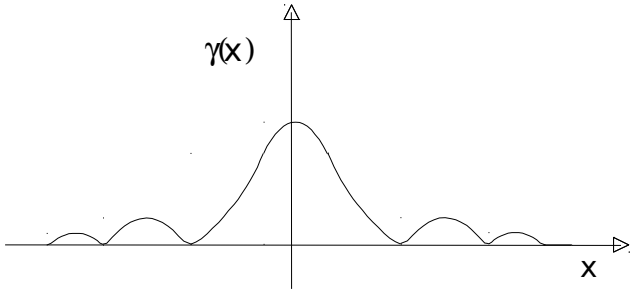
4) $\gamma(x)=\frac{1}{2}e^{-|x|}$ Esponenziale decrescente



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



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



Effetto dell'ampiezza h

- NOTA: solo i punti "vicini" ad x_0 influiscono sul calcolo della $p(x_0)$
- h determina l'ampiezza della finestra di interesse, cioè definisce in qualche modo il concetto di vicinato

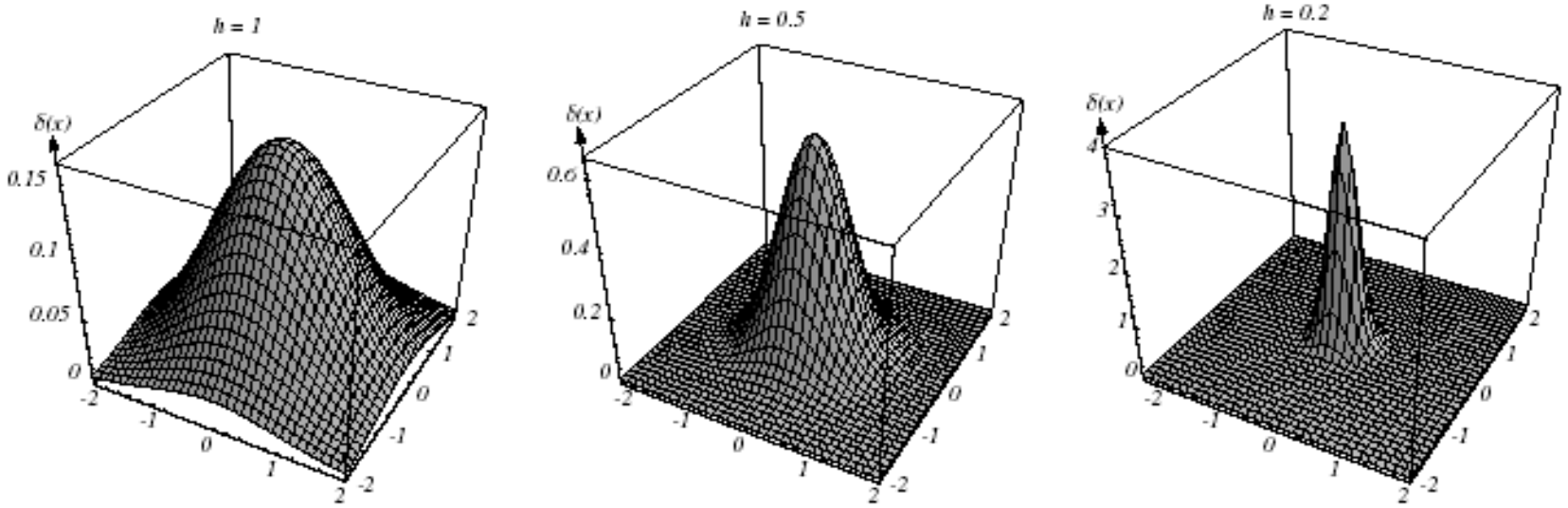


FIGURE 4.3. Examples of two-dimensional circularly symmetric normal Parzen windows for three different values of *h*. Note that because the $\delta(\mathbf{x})$ are normalized, different vertical scales must be used to show their structure. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern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

Effetto dell'ampiezza h

La scelta di h è cruciale

⇨ h troppo grande: molto smooth, tutto più o meno uguale

⇨ h troppo piccolo: un sacco di picchi singoli (dove x=xi)

Occorre trovare un buon compromesso

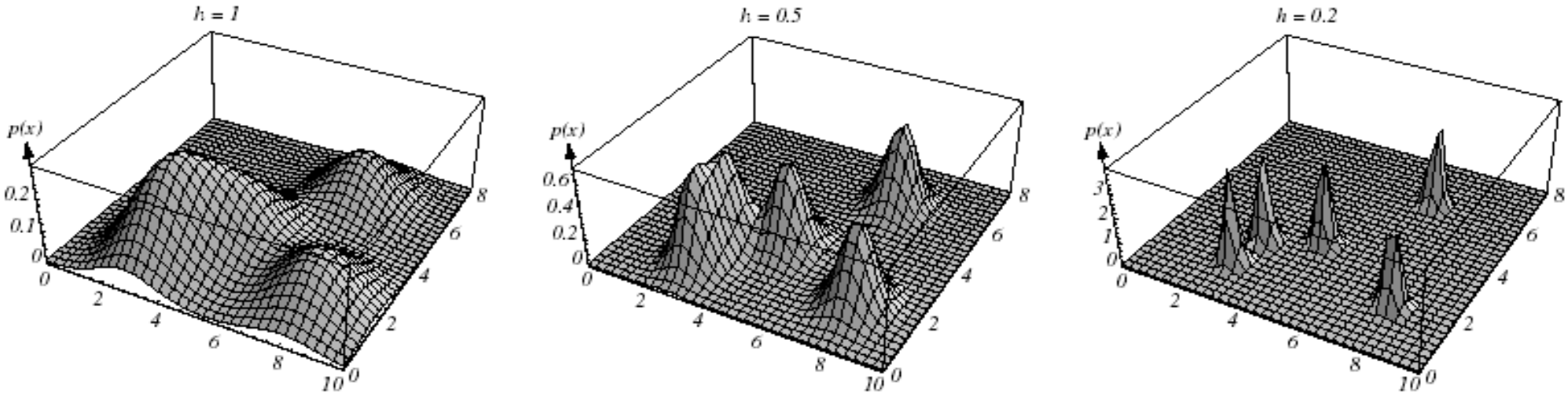


FIGURE 4.4. Three Parzen-window density estimates based on the same set of five samples, using the window functions in Fig. 4.3. As before, the vertical axes have been scaled to show the structure of each distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

K-Nearest Neighbor

Secondo metodo per stimare non parametricamente $p(x_0)$

$$p(x) \approx \frac{K}{NV}$$

si fissa K, si sceglie R in modo tale che contenga K punti attorno ad x_0, si determina V e si stima $p(x_0)$

⇨ Effettuando questa stima non parametrica delle posterior di tutte le classi, e applicando la regola di classificazione di Bayes, si ottiene il classificatore K-nearest Neighbor

K-Nearest Neighbor

K-NN come classificatore di Bayes con stima non parametrica della pdf

(alla lavagna)

K-Nearest Neighbor

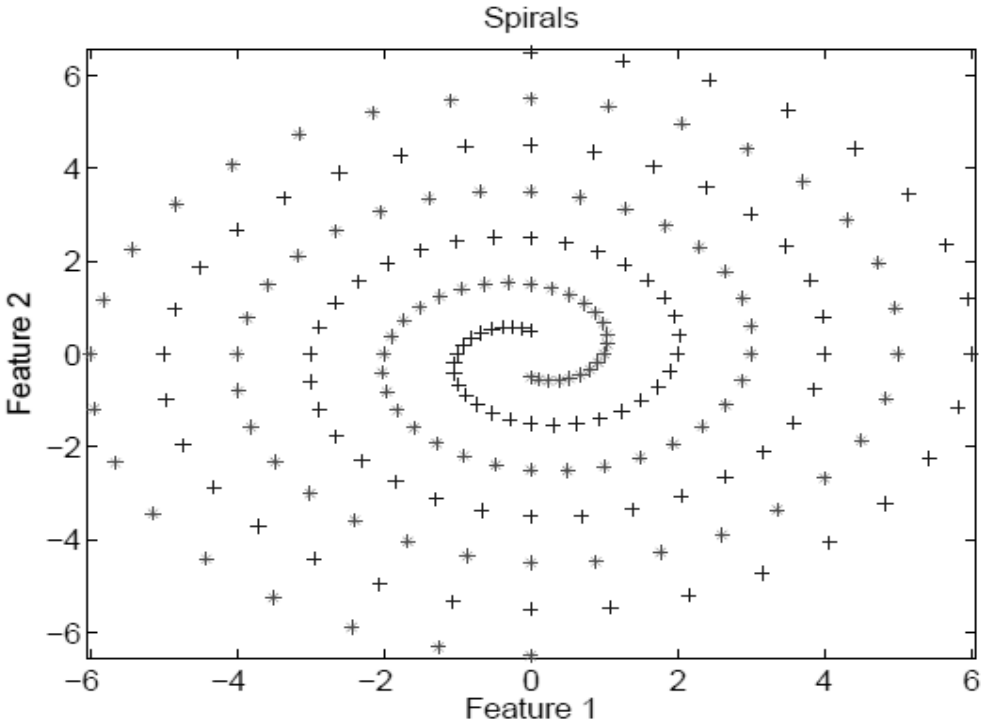
- Riassumendo, il K-NN funziona nel seguente modo:
  - sia X un insieme di esempi etichettati (il training set, ogni punto ha la sua classe)
  - dato un punto x_0 da classificare, si calcola l'insieme U dei K punti dell'insieme X più vicini ad x_0 secondo una determinata metrica
  - Si calcola la classe C_j più frequente all'interno dell'insieme U
  - Si assegna x_0 a C_j

K-Nearest Neighbor

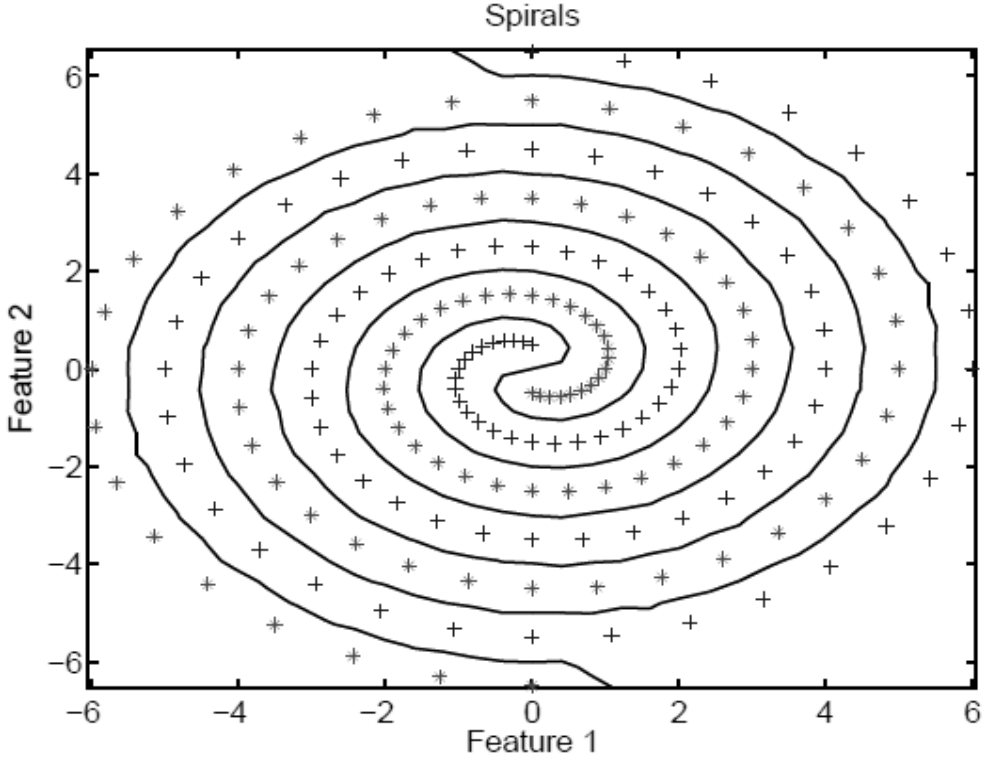
VANTAGGI

- ⇨ tecnica semplice e flessibile
- ⇨ tecnica intuitiva (assume che punti della stessa classe abbiano probabilmente caratteristiche simili, cioè una distanza bassa)
- ⇨ tecnica che funziona anche per dati non vettoriali (basta avere una misura di distanza appropriata)
- ⇨ ragionevolmente accurata (il confine di separazione è comunque non lineare)
- ⇨ ci sono pochi parametri da aggiustare
- ⇨ sono stati dimostrati molti risultati teorici su questa tecnica (asintoticità del comportamento, bounds)

Problema di classificazione difficile!



Confine di decisione di 1-Nearest Neighbor



K-Nearest Neighbor

SVANTAGGI

- ⇨ Tutti i punti del training set devono essere mantenuti in memoria
- ⇨ vengono utilizzati solo pochi punti dello spazio per prendere la decisione (solo K punti)
- ⇨ dipendentemente dalla metrica utilizzata, occorre pre-processare lo spazio
- ⇨ Serve una misura di distanza buona
- ⇨ La scelta di K spesso è cruciale (K = 1 → Nearest Neighbor rule)
  - ⇨ scelta tipica $k \cong \sqrt{N}$

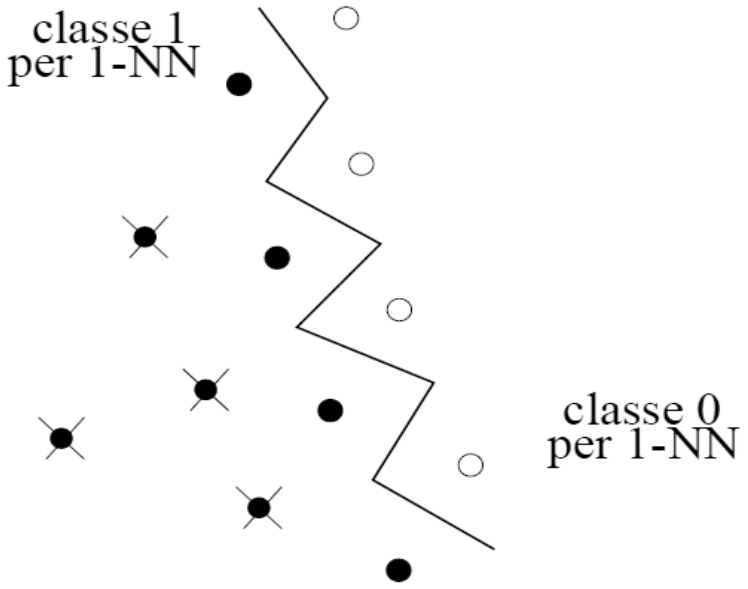
K-Nearest Neighbor: note finali

Determinazione di K

- è equivalente al parametro h di Parzen Windows
  - troppo piccolo si hanno stime troppo rumorose
  - troppo grande si hanno stime troppo grezze
- Metodo per stimare K:
  - crossvalidation sul training set (o su un altro insieme chiamato Validation Set)
    - si provano diversi valori e si tiene quello che funziona meglio
  - Metodi locali: si decide guardando la regione dove si sta operando (ad esempio guardando il K che funziona meglio localmente)

K-Nearest Neighbor: note finali

- Condensing/Editing: metodi per ridurre la dimensionalità del training set (che deve essere mantenuto in memoria)
- Condensing: rimuovere dal training set tutti quei punti che non hanno effetto sul confine di decisione



K-Nearest Neighbor: note finali

- Editing: rimuovere dal training set tutti i punti che non vengono classificati correttamente dall'algoritmo

- PROBLEMA DI QUESTI DUE METODI:

  così facendo non siamo sicuri di aver eliminano tutti gli errori

  - (i punti eliminati potrebbero essere cruciali per la classificazione di altri punti non presenti nel training set)