Stima parametrica

Stima di Bayes

Stima di Bayes

⇨ A differenza dell'approccio ML, in cui supponiamo θ come fissato ma sconosciuto, *l'approccio di stima Bayesiana* dei parametri considera θ come una variabile aleatoria.

⇨ In pratica il processo di learning Bayesiano *stima un modello implicitamente*, ossia non restituisce un vettore di parametri θ visibile, ma *una distribuzione su di esso*, data dal training set disponibile D

⇨ ATTENZIONE: differenza tra **classificazione** Bayesiana e **stima** Bayesiana

36

Stima di Bayes: il problema

- Goal: stimare la densità a posteriori $P(\omega_i|x)$ che sta alla base della classificazione Bayesiana
- è necessario conoscere:
  - Le probabilità a priori $P(\omega_i)$
  - Le densità condizionali $P(x|\omega_i)$
- Quantità sconosciute ma stimabili dal training set
  - Sia D il set totale di campioni: il nostro compito si trasforma così nella stima di $P(\omega_i|x,\mathbf{D})$

37

Stima di Bayes: il problema

- Dato il set di training D, la posterior diventa (con il teorema di Bayes):

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

- Osservazioni:
  - $P(\omega_i|D) \Rightarrow P(\omega_i)$ (prior note)
  - $D = D_1, D_2, \ldots, D_c$ con i campioni in D_i appartenenti a ω_i
  - D_i non da informazioni sui parametri di $p(\mathbf{x}|\ \omega_j, D)$ se $i \neq j$.

38

Stima di Bayes: il problema

⇨ Quindi: si può lavorare con ogni classe indipendentemente, ossia

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

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

39

Stima di Bayes: il problema

⇨ semplificare la notazione portandola a **c diverse istanze dello stesso problema**, ossia:

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

PROBLEMA DEL LEARNING BAYESIANO: *Si usa un set di campioni D, estratti secondo la distribuzione sconosciuta p(**x**), per determinare p(**x**|D)*

40

Stima di Bayes: la stima

⇨ (alla lavagna)

41

Stima di Bayes: la stima

Riassumendo:

⇨ Quello che si fa è stimare $p(\mathbf{x}|D)$ tramite l'ausilio di un modello di parametri implicito $\boldsymbol{\theta}$.

⇨ si *esplicita* il calcolo di $p(\mathbf{x}|D)$, per stimare $p(\mathbf{x})$, *convertendo il problema di stima di una densità di probabilità a quello di stima di un vettore di parametri*.

$$p(\mathbf{x} \mid D) = \int p(\mathbf{x}, \boldsymbol{\theta} \mid D) d\boldsymbol{\theta}$$
$$= \int p(\mathbf{x} \mid \boldsymbol{\theta}, D) p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta}$$

42

Stima di Bayes: la stima

Riassumendo (2):

⇨ per ipotesi, la selezione di **x** è indipendente dai campioni di training *D*, dato **θ,**

$$p(\mathbf{x} \mid D) = \int p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta}$$

⇨ Pertanto la distribuzione p(**x**) è completamente conosciuta quando conosco il vettore dei parametri **θ**

⇨ L'obiettivo diventa stimare la posterior dei parametri p(**θ** |D)

⇨ data la stima p(**θ**|D) posso estrarre il massimo (Maximum A Posteriori – MAP) oppure posso fare la media pesata

43

Stima di Bayes: la stima

Vantaggi:

⇨ Stima più accurata: questo approccio ***permette di tenere conto dell'effetto di tutti i modelli,*** descritti dal valore della funzione integrale***, per tutti i possibili modelli.***

$$p(\mathbf{x} \mid D) = \int p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta}$$

Svantaggi:

⇨ stimare la posterior dei parametri non è sempre banale

⇨ integrare in tutto lo spazio dei parametri può essere "difficile" o "intrattabile

⇨ occorre definire i priori p(**θ**)

44

Esempio: caso Gaussiano

⇨ Utilizziamo le tecniche di stima Bayesiana per calcolare la densità a posteriori p($\boldsymbol{\theta}$ |D) e la densità p(**x**|D) per il caso in cui

$$p(\mathbf{x} \mid \boldsymbol{\theta}) \equiv p(\mathbf{x} \mid \boldsymbol{\mu}) \approx N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

⇨ Gaussiana univariata (unidimensionale) di media 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

45

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($\boldsymbol{\theta}$ |D) sia trattabile, cioè abbia la stessa forma della condizionale

⇨ Questo semplifica di molto l'analisi

⇨ Esempio: gaussiana per gaussiana, dirichlet per multinomiale

46

Esempio: caso Gaussiano

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

⇨ 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.

47

Esempio: caso Gaussiano

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

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

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

48

⇨ Riarrangiando, si ottiene

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

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

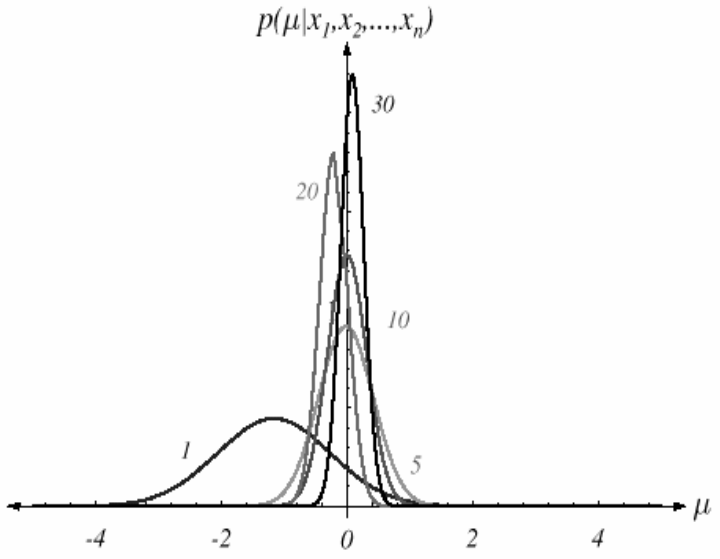
$$\sigma_n^2 = \frac{\sigma_0^{\,2}\sigma^2}{n\sigma_0^{\,2}+\sigma^2}$$

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

σ_n^2 misura l'incertezza della nostra scelta.

49

Esempio: caso Gaussiano



50

Esempio: caso Gaussiano

Terzo passo: stima della densità condizionale p(x|D), che in notazione esatta, ricordiamo, è $P(x|\omega_i, \boldsymbol{D}i)$

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

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

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

Esempio: caso Gaussiano

⇨ dove

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

52

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 $\hat{\sigma}^2$

53

Stima di Bayes: in generale

⇨ Riassumendo ed estendendole al caso generale, le formule principali viste sono:

$$p(\mathbf{x} \mid D) = \int p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta}$$

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

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

⇨ Si noti la somiglianza con l'approccio ML, con la differenza che qui non si cerca il max puntuale

54

Conclusioni: Bayes vs ML

- ML restituisce una stima puntuale $\hat{\boldsymbol{\theta}}$, l'approccio Bayesiano una distribuzione su $\boldsymbol{\theta}$ (più ricca, tiene conto di tutti i possibili modelli)
- 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, $p(\boldsymbol{\theta}|D)$ converge ad una funzione delta)
- In Bayes occorre stimare i prior (svantaggio o vantaggio -- e.g. modelli sparsi)

55

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

57

Stima non parametrica

- ⇨ Idea: stimare la pdf andando ad analizzare le singole regioni dello spazio
  - ⇨ mi interessa p(x=x0), vado a considerare la regione attorno ad x0 ed effettuo una stima a partire da quella regione
- ⇨ Esempio: istogramma
  - ⇨ suddivido lo spazio in regioni di larghezza uniforme
  - ⇨ dato un insieme di punti D campionato dalla distribuzione che devo stimare, per ogni regione conto il numero di punti che ci cade dentro
  - ⇨ questa rappresenta la stima non parametrica della pdf

58

Stima non parametrica

⇨ Più formalmente (alla lavagna)

59

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:

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

⇨ 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!

60

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 (Esempio istogramma)

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

61

Parzen Windows

⇨ Assumiamo che la regione R sia un ipercubo di lato h (in uno spazio d-dimensionale)

$$V = h^d$$

⇨ Possiamo ottenere una forma analitica per K, il numero di punti che cadono nella regione R, definendo la seguente funzione

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

che rappresenta un ipercubo di lato unitario centrato nell'origine (window function)

62

Parzen Windows

⇨ La funzione

$$\gamma\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$$

vale 1 se **x**i cade nell'ipercubo di lato h centrato in **x**, zero altrimenti

⇨ Il numero k di punti che stanno nell'ipercubo di lato h (la regione R) centrato in **x** è quindi

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

63

Parzen Windows

⇨ Sostituendo nella formula di prima otteniamo

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

⇨ Questo suggerisce un metodo più generale per stimare le densità di probabilità

- ⇨ Permettere più tipi di funzioni window
- ⇨ La stima della pdf è ottenuta come la media di queste funzioni di **x** e $\mathbf{x}_i$
- ⇨ In altre parole ogni campione contribuisce alla stima della pdf in un punto **x**
- ⇨ Il contributo è diverso a seconda della distanza da **x**

64

Parzen Windows

⇨ Per avere una p(x) valida occorre che:

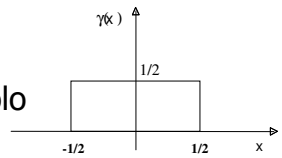
$$\gamma(\mathbf{u}) \geq 0$$
$$\int \gamma(\mathbf{u}) d\mathbf{u} = 1$$

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

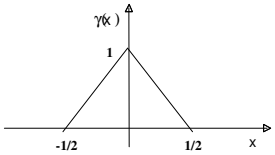
65

Esempi di funzioni potenziali

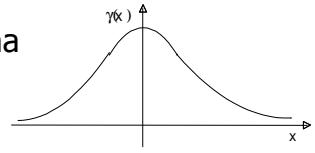
1) $\gamma(\mathbf{x}) = \begin{cases} 0{,}5 & |\mathbf{x}| \leq 1 \\ 0 & |\mathbf{x}| > 1 \end{cases}$ Rettangolo



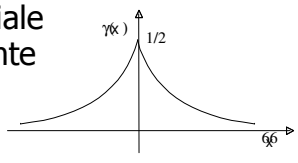
2) $\gamma(\mathbf{x}) = \begin{cases} 1-|\mathbf{x}| & |\mathbf{x}| \leq 1 \\ 0 & |\mathbf{x}| > 1 \end{cases}$ Triangolo



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

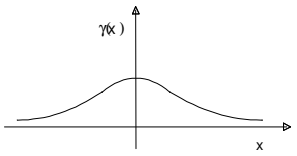


4) $\gamma(\mathbf{x}) = \frac{1}{2} e^{-|\mathbf{x}|}$ Esponenziale decrescente

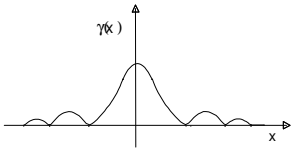


66

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



67

Effetto dell'ampiezza h

⇨ NOTA: solo i punti "vicini" ad x influiscono sul calcolo della p(x)

⇨ h determina l'ampiezza della finestra di interesse, cioè definisce in qualche modo il concetto di vicinato

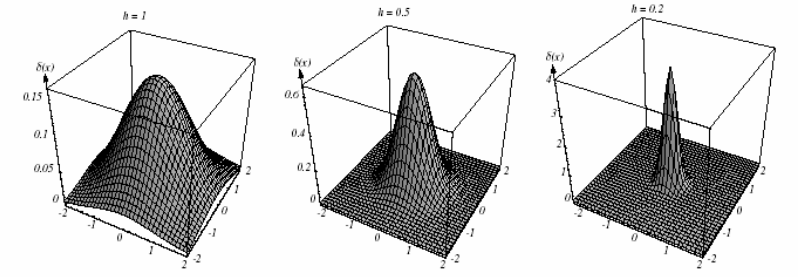


FIGURE 4.3. Examples of two-dimensional circularly symmetric normal Parzen windows for three different values of h. Note that because the $\delta(\mathbf{x})$ are normalized, different vertical scales must be used to show their structure. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern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

68

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

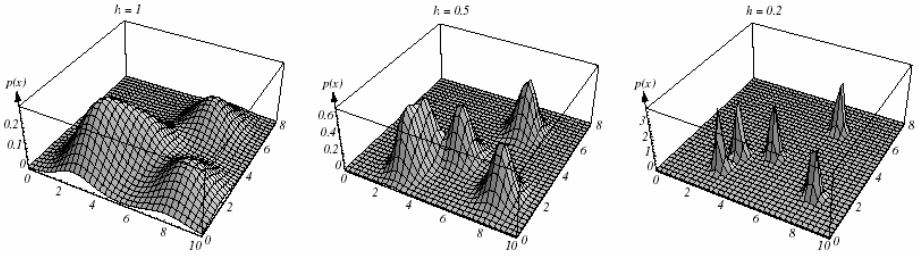


FIGURE 4.4. Three Parzen-window density estimates based on the same set of five samples, using the window functions in Fig. 4.3. As before, the vertical axes have been scaled to show the structure of each distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

K-Nearest Neighbor

Secondo metodo per stimare non parametricamente p(x)

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

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

⇨ 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

70

K-Nearest Neighbor

- Presentazione:
  i. come funziona e su che intuizione si basa
  ii. in che senso rappresenta il classificatore di Bayes nel caso di stima non parametrica delle posterior
- Funzionamento (molto semplice):
  - sia X un insieme di esempi etichettati (il training set, ogni punto ha la sua classe)
  - dato un punto x da classificare, si calcola l'insieme U dei K punti dell'insieme X più vicini ad x secondo una determinata metrica
  - Si calcola la classe Ck più frequente all'interno dell'insieme U
  - Si assegna x a Ck

71

K-Nearest Neighbor

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

(alla lavagna)

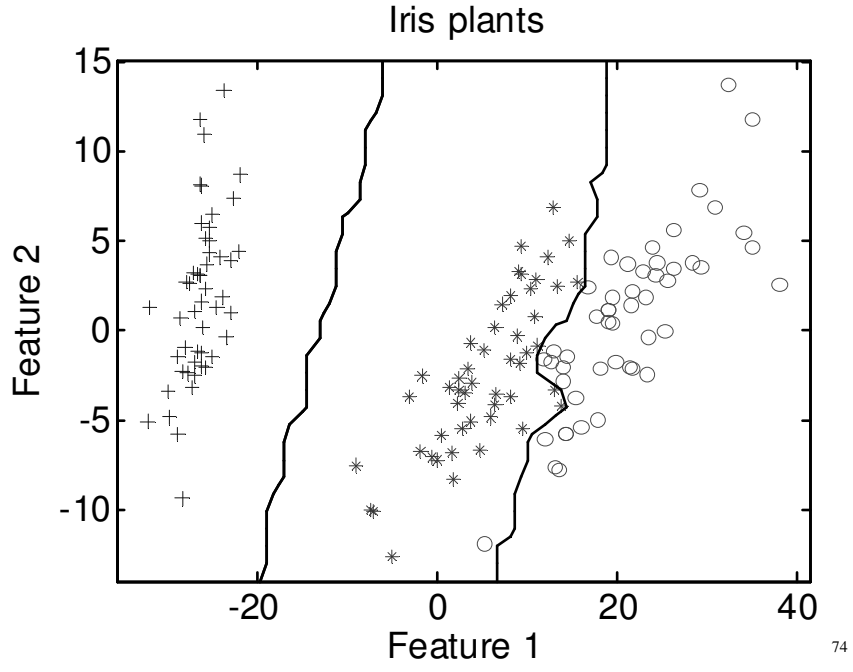
72

K-Nearest Neighbor

VANTAGGI

- ⇨ tecnica semplica 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)

73



74

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 $k \cong \sqrt{N}$

75

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)

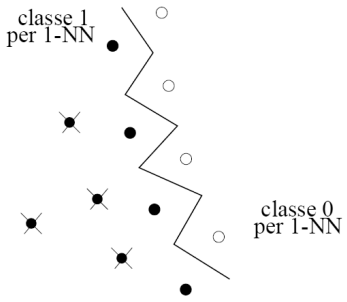
⇨ si provano diversi valori e si tiene quello che funziona meglio

⇨ Metodi locali: si decide guardando la regione dove si sta operando (ad esempio guardando il K che funziona meglio localmente)

76

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



77

K-Nearest Neighbor: note finali

- ⇨ Editing: rimuovere tutti i punti che non vengono classificati correttamente dall'algoritmo
  - ⇨ chiaro che così facendo non si eliminano tutti gli errori (i punti eliminati potrebbero essere cruciali per la classificazione di altri punti)

78