Stima dei Parametri

Corso di Apprendimento Automatico

Laurea Magistrale in Informatica

Nicola Fanizzi

Dipartimento di Informatica
Università degli Studi di Bari

20 gennaio 2009

Sommario

- Introduzione
- Stima dei parametri di massima verosimiglianza
- Stima dei parametri bayesiana

Introduzione I

In un contesto Bayesiano, si potrebbe progettare un classificatore ottimo conoscendo:

- $p(\omega_i)$ (prob. a priori)
- $p(x|\omega_i)$ (densità condizionate)

Sfortunatamente, raramente si ha una informazione completa.

Progettare un classificatore a partire da un campione di esempi:

- Nessun problema con la stima della prob. a priori
- I campioni sono spesso troppo piccoli per la stima delle densità condizionate (grandi dimensioni dello spazio delle feature)

Introduzione II

- L'informazione a priori sul problema

  Es. Una densità $p(x|\omega_i)$
  $p(x|\omega_i) \sim N(\mu_i, \Sigma_i)$ è caratterizzata da 2 parametri

- Tecniche di stima:
  Massima verosimiglianza (*Maximum-likelihood, ML*) e Bayesiana
- Risultati pressochè identici, ma gli approcci sono diversi

Introduzione III

- Nella stima ML, i parametri sono considerati *fissati* ma sconosciuti
  - Parametri migliori ottenuti massimizzando la probabilità di ottenere i campioni osservati
- Nella stima bayesiana, i parametri sono visti come *variabili aleatorie* dalla distribuzione sconosciuta
  - L'osservazione di esempi cambia la distribuzione a posteriori, con la stima dei valori dei parametri
  - Effetto: assottigliamento della densità sui veri valori dei parametri

In entrambi gli approcci,
si usa $p(\omega_i|x)$ come regola di classificazione

Stima di massima verosimiglianza I

- Buona proprietà di convergenza al crescere del campione di esempi
- Tecnica più semplice d'ogni altra alternativa

Principio generale
Si assuma di avere c classi e un dataset $D = D_1 \cup D_2 \cup \cdots \cup D_c$ di esempi indipendenti e identicamente distribuiti (i.i.d. se riguardati come var. aleatorie)

- per denotare la dipendenza dal parametro, si scrive

$$p(x|\omega_j) \equiv p(x|\omega_j, \theta)$$

es. $p(x|\omega_j, \theta) \sim N(\mu_j, \Sigma_j)$ con:
$\theta = (\mu_j, \Sigma_j) = (\mu_j^1, \mu_j^2, \ldots, \sigma_j^{11}, \sigma_j^{22}, cov(x_j^n, x_j^m), \ldots)$

Stima di massima verosimiglianza II

- Usare l'informazione degli esempi di training per stimare $\theta = (\theta_1, \theta_2, \ldots, \theta_c)$, dove ogni θ_i è associato con una categoria $(i = 1, 2, \ldots, c)$
- Supponendo che $D = \{x_1, x_2, \ldots, x_n\}$, per l'indipendenza degli esempi

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

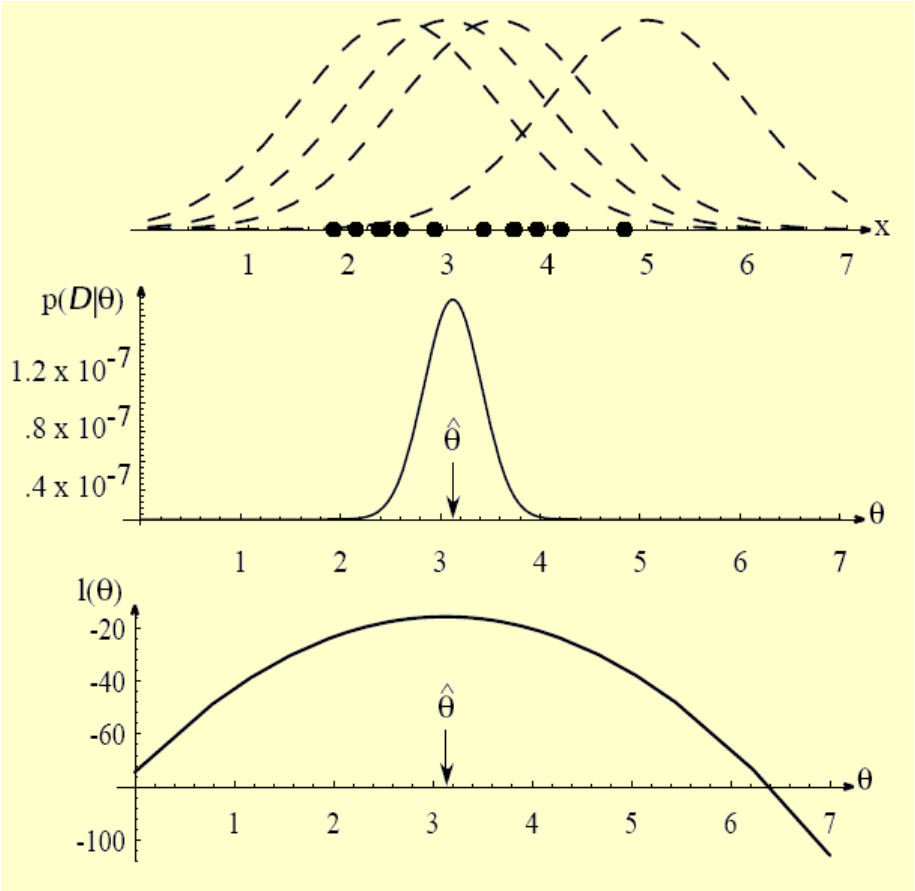
verosimiglianza di θ rispetto all'insieme di esempi

- La stima ML di θ è, per definizione, il valore $\hat{\theta}$ che massimizza $p(D|\theta)$

"Valore di θ che meglio si accorda con il campione di training realmente osservato"

Stima di massima verosimiglianza III

Distribuzioni candidate (linee tratteggiate)
relative a punti tratti da una Gaussiana di media sconosciuta:
$p(D_j|\mu)$ in funzione della media e log-likelihood



con tanti esempi, la funzione di likelihood tende restringersi

Stima ottimale I

- Sia $\theta = (\theta_1, \ldots, \theta_p)^t$ e sia ∇_θ l'operatore di gradiente

$$\nabla_\theta = \left[\frac{\partial}{\partial \theta_1}, \ldots, \frac{\partial}{\partial \theta_p} \right]$$

- Si definisce $l(\theta)$ come funzione di *log-verosimiglianza* (log-likelihood)

$$l(\theta) = \ln p(D|\theta)$$

- Nuova formulazione del problema:

  determinare θ che massimizza la log-likelihood

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

Stima ottimale II

Condizioni necessarie per l'ottimizzazione:

$$\nabla_\theta l = 0$$

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

Una soluzione $\hat{\theta}$ potrebbe essere

- un vero massimo globale,
- un minimo/massimo locale o
- un flesso (raramente)

Bisogna anche controllare gli estremi dell'insieme di definizione della funzione

Stima MAP (maximum a posteriori)

Gli stimatori *maximum a posteriori* (MAP) cercano il valore di θ che massimizzi $p(D|\theta)p(\theta)$ o anche $l(\theta) + \ln p(\theta)$

- Si può vedere uno stimatore ML come uno stimatore MAP per una densità a priori uniforme
- Uno stimatore MAP cerca il picco (*moda*) di una densità a posteriori
- Svantaggio:
  con trasformazioni non lineari arbitrarie dello spazio del parametro la densità cambia così come la soluzione

Stima Bayesiana I

Apprendimento Bayesiano per problemi di classificazione

- Nella stima ML θ è supposto prefissato
  nella stima Bayesiana θ è una variabile casuale
- Nella classificazione Bayesiana il calcolo delle probabilità a posteriori $P(\omega_i|x)$ è fondamentale
- Scopo: calcolare $P(\omega_i|x, D)$
  dato il campione D, la formula di Bayes permette di scrivere:

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

Stima Bayesiana II

Notando che

- $p(x, D|\omega_i) = p(x|\omega_i, D)P(\omega_i|D)$
- $p(x|D) = \sum_{j=1}^{c} p(x, \omega_i|D)$
- $p(\omega_i|D) = p(\omega_i)$ ottenuti dal campione di training

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

Semplificando: c problemi della forma:
usare un insieme D di esempi con distribuzione $p(x)$ per determinare $p(x|D)$

Stima della densità a posteriori I

Il calcolo di $p(x|D)$ è applicabile ad ogni situazione nella quale una densità sconosciuta sia parametrizzabile

Assunzioni di base

- Si assume nota la forma di $p(x|\theta)$, ma non il parametro
- La conoscenza su θ si assume contenuta in una densità a priori $p(\theta)$
- Il resto della conoscenza è contenuto in un insieme D di n variabili casuali $x_1, x_2, \ldots, x_n$ che segue $p(x)$

Stima della densità a posteriori II

Problema di base
Calcolare la densità a posteriori $p(\theta|D)$ *per derivarne poi* $p(x|D)$ (migliore approssimazione di $p(x)$ con i dati disponibili)

Si può scrivere:

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

ma $p(x, \theta|D) = p(x|\theta, D)p(\theta|D)$, quindi

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

L'integrale si calcola tramite metodi numerici (es. Monte Carlo)

Caso generale I

Abbiamo visto che

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

Usando la formula di Bayes:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$$

Per l'assunzione di indipendenza:

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

Caso generale II

Osservazioni

- Se $p(\theta|D)$ ha un picco per il valore $\hat{\theta}$ con $p(\hat{\theta}) \neq 0$ e non cambia molto in un suo intorno,
  allora $p(D|\theta)$ ha anche essa un picco nello stesso punto
- Quindi sarà approssimativamente $p(x|D) \simeq p(x|\hat{\theta})$,
  risultato che si otterrebbe usando la stima ML come se fosse il valore reale:
- Se il picco di $p(D|\theta)$ è rilevante,
  allora l'influenza della densità a priori si può ignorare

Approccio ricorsivo-incrementale I

Separiamo i campioni per classi,
indicando esplicitamente la cardinalità: $D^n = \{x_1, \ldots, x_n\}$

Per $n > 1$ tramite l'eq. $p(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$:

$$p(D^n|\theta) = p(x_n|\theta)p(D^{n-1}|\theta)$$

Sostituendo nelle relazioni precedenti:

$$p(\theta|D^n) = \frac{p(x_n|\theta)p(\theta|D^{n-1})}{\int p(x_n|\theta)p(\theta|D^{n-1})d\theta}$$

Notare che si può partire da $p(\theta|D^0) = p(\theta)$ e continuare calcolando $p(\theta|x_1), p(\theta|x_1, x_2), \ldots$

Approccio ricorsivo-incrementale II

Parametri / statistiche sufficienti

- Per calcolare $p(\theta|D_n)$ si preservano tutti gli esempi in D_{n-1}
- Per alcune distribuzioni pochi parametri associati con $p(\theta|D_{n-1})$ contengono *tutta l'informazione necessaria*
- La sequenza di densità converge ad una funzione *delta di Dirac* centrata sul valore vero del parametro: si dice in tal caso che $p(x|D)$ è identificabile

Differenze

I metodi visti finora convergono solo asintoticamente dati molti esempi

- La stima ML è preferibile in termini di *complessità* (ricerca di minimo contro integrazione multi-dimensionale) e di *interpretabilità* (singolo modello contro media pesata di modelli)
- L'info a priori è da assumere parametrica $p(x|\hat{\theta})$ per la stima ML, quella bayesiana $p(x|D)$ sfrutta invece tutta l'informazione disponibile
- Per questo, se $p(\theta|D)$ è irregolare o asimmetrica, $p(x|D)$ sarà molto variabile a seconda dei metodi (problemi di bias e varianza)

Errori

Il classificatore determina in base alla densità a posteriori la classe che massimizza la probabilità d'appartenenza

Possibili errori:

- **errore di indistinguibilità** densità $p(x|\omega_i)$ che si sovrappongono per alcuni valori di i.
  Ineliminabile: dipende dal problema
- **errore di modello** occorre informazione sul dominio per la scelta del modello corretto
- **errore di stima** dovuto alla limitatezza del campione;
  si attenua aumentando gli esempi

Problematiche I

Dimensionalità

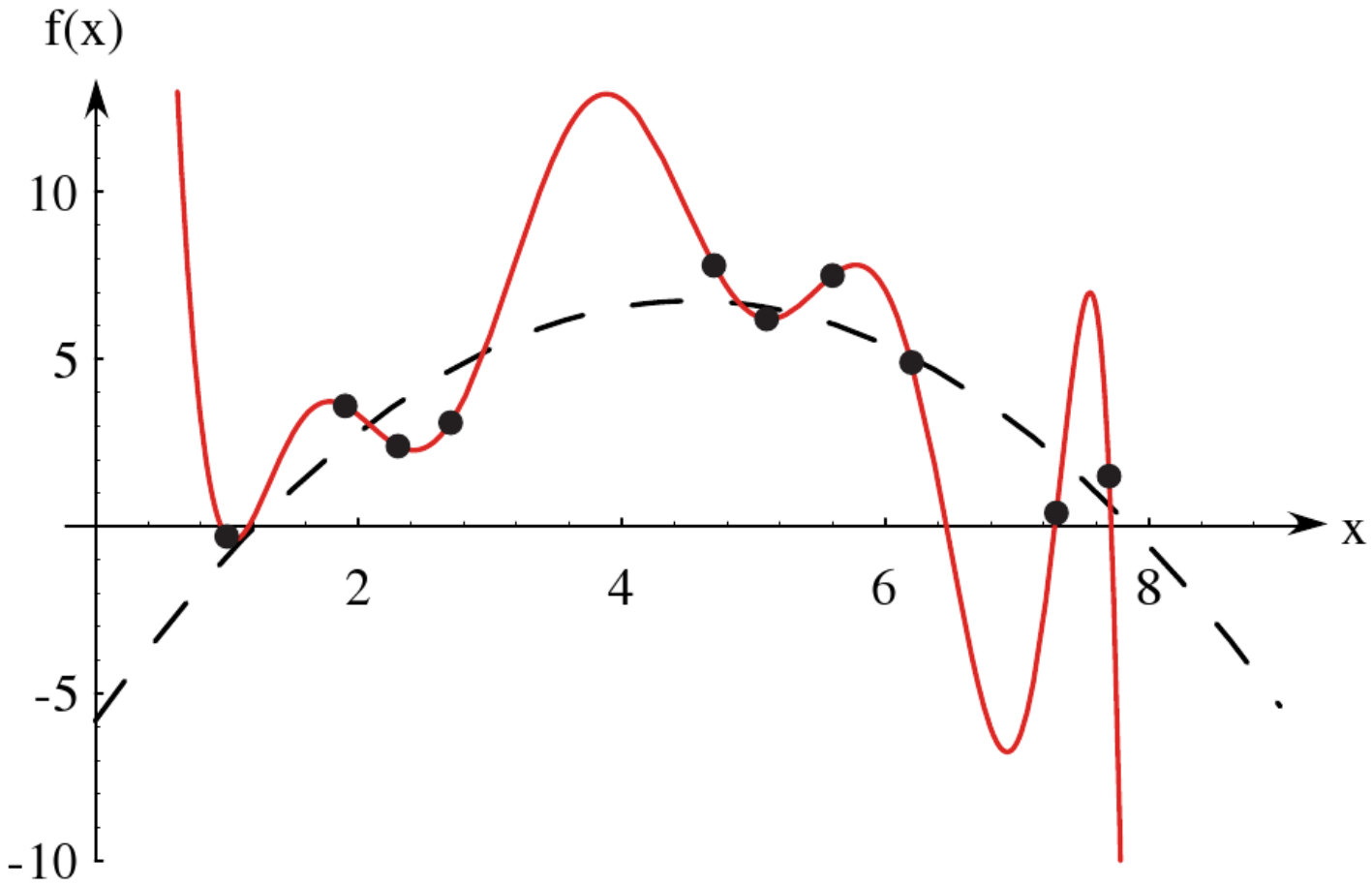
- Problemi che coinvolgono 50 o 100 caratteristiche (binarie)
- L'accuratezza predittiva dipende dalla dimensione e del numero dei dati di training
- Le feature più utili sono quelle la cui differenza tra le medie è grande relativamente alla deviazione standard
- In pratica, oltre un certo punto, l'aggiunta di altre feature porta a peggiorare la performance: modello sbagliato

Problematiche II

Evitare il fenomeno dell'overfitting

- riduzione della dimensionalità conservando solo le feature rilevanti o combinando più feature
- condivisione della matrice di covarianza tra le varie classi
- la matrice può essere sottoposta ad un meccanismo di soglia in modo da eliminare correlazioni *accidentali*

Esempio



parabola con l'aggiunta di errore gaussiano

Problematiche IV

- Si parte con un modello polinomiale (10 deg grado), per poi livellare (*smoothing*) o semplificare il modello, eliminando i termini di grado maggiore
- NB: a volte anche *una retta* potrebbe avere prestazioni superiori!
- Questo in genere aumenta l'errore di training ma abbassa quello sugli esempi di test

Credits

- R. Duda, P. Hart, D. Stork: *Pattern Classification*, Wiley