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

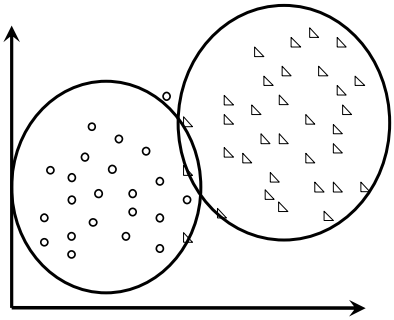
RIASSUNTO:

- ⇨ 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)

4

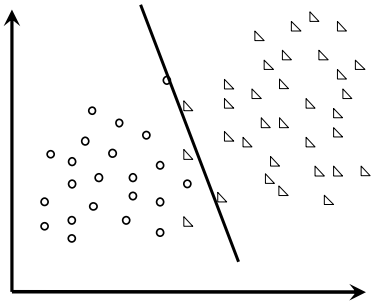
Approcci alla classificazione

Generativi: un modello per ogni classe



$$\tilde{y} = \arg\max_y P(y \mid \mathbf{x})$$

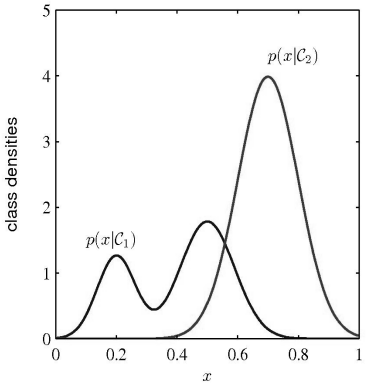
Discriminativi: modellano direttamente il confine

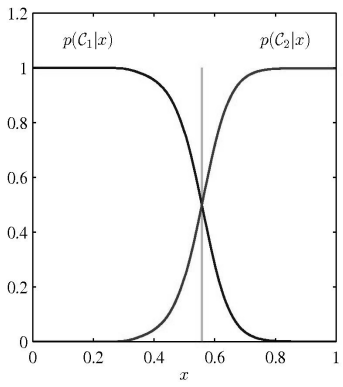


$$\tilde{y} = sign\,(\mathbf{w} \cdot \mathbf{x} + b)$$

5

Approcci alla classificazione





6

Generative vs Discriminative

Generative

- ⇨ Can handle missing data or partially labeled
- ⇨ A new class K+1 can be added incrementally by learning its class-conditional density p(X|K+1) independently of all the previous classes

Discriminative

- ⇨ Flexibility of these models is used in regions of input space where the posterior probabilities differ significantly from 0 to 1 whereas generative approaches model details of the distribution of X which may be irrelevant for determining the posterior probabilities

7

Generative vs Discriminative

- ⇨ Can readily handle compositionality (e.g., faces with glasses and/or hats...) whereas standard discriminative models need to see all combinations of possibilities during training

- ⇨ Tipically very fast at making prediction for new data points, while generative models often require iterative solution
- ⇨ Better predictive performance since these models are trained to predict the class label rather than the joint distribution of input vectors and target

8

Generative vs Discriminative

- ⇨ Number of training examples logaritmic, rather than linear, in the number of the parameters
- ⇨ Models are relatively easy to train
- ⇨ Best to use a generative approach if confidence in the model correctness is high

- ⇨ As the number of training examples increase, discriminative methods catch up, and then overtake the performance of generative
- ⇨ Models often lack the elegant probabilistic concepts of priors structure and uncertainty
- ⇨ Harder to train because of the simultaneous consideration of all classes
- ⇨ Each task to be solved needs a different model and a new training session

Stima delle probabilità

Le probabilità sono sconosciute, occorre stimarle dal training set!

- ⇨ Stime parametrica: si conosce la forma della pdf, se ne vogliono stimare i parametri
  - ⇨ esempio gaussiana, stimo la media
- ⇨ Stime non parametriche: non si assume nota la forma, la pdf è stimata direttamente dai dati
  - ⇨ esempio istogramma
- ⇨ Stime semi-parametriche: ibrido tra le due – i parametri possono cambiare la forma della funzione
  - ⇨ esempio Neural Networks

10

Stima parametrica

Introduzione

Introduzione

- Per creare un classificatore ottimale che utilizzi la regola di decisione Bayesiana è necessario conoscere:
  - Le probabilità a priori $P(\omega_i)$
  - Le densità condizionali $p(\mathbf{x} \mid \omega_i)$
- Le performance di un classificatore dipendono fortemente dalla bontà di queste componenti

- NON SI HANNO PRATICAMENTE MAI TUTTE QUESTE INFORMAZIONI!

12

⇨ Più spesso, si hanno unicamente:

- ⇨ Una vaga conoscenza del problema, da cui estrarre vaghe probabilità a priori.
- ⇨ Alcuni pattern particolarmente rappresentativi, training data, usati per addestrare il classificatore (spesso troppo pochi!)

⇨ La stima delle probabilità a priori di solito non risulta particolarmente difficoltosa.

⇨ La stima delle densità condizionali è più complessa.

13

⇨ Assunto che la conoscenza, benché approssimativa, delle densità a priori non presenta problemi, per quanto riguarda le densità condizionali le problematiche si possono suddividere in:

1. ***Stimare*** la ***funzione sconosciuta*** $p(\mathbf{x} \mid \omega_j)$
2. ***Stimare*** i ***parametri sconosciuti*** della ***funzione conosciuta*** $p(\mathbf{x} \mid \omega_j)$

   Per es., stimare il vettore $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ se

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

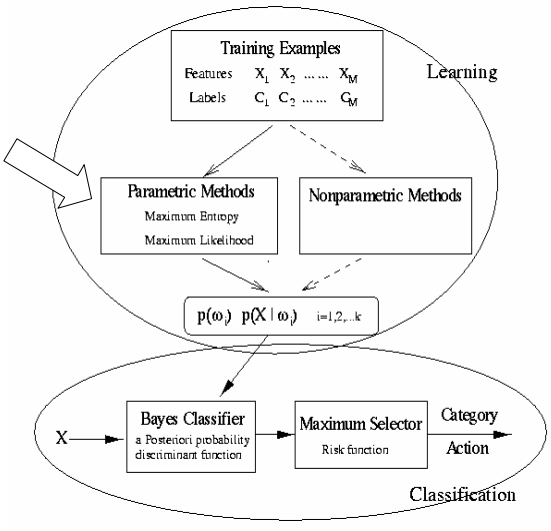
14

Stima dei parametri

- ⇨ Il secondo punto risulta di gran lunga più semplice (sebbene complesso!), e rappresenta un problema classico nella statistica.
- ⇨ Trasferito nella pattern recognition, un approccio è quello di
  - ⇨ stimare i parametri dai dati di training
  - ⇨ usare le stime risultanti come se fossero valori veri
  - ⇨ utilizzare infine la teoria di decisione Bayesiana per costruire un classificatore
- ⇨

15

Uno sguardo d'insieme



16

Stima dei parametri – Probabilità a priori

⇨ Supponiamo di avere un insieme di n dati di training in cui ad ogni pattern è assegnata un'etichetta d'identità (ossia conosco per certo a quale stato ω_j appartiene il pattern k-esimo)

⇨ ➔ problema di learning dei parametri supervisionato

⇨ Allora

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

dove n_i è il numero di campioni con etichetta ω_i

⇨ Questa facile operazione non è di grande utilità, perchè le probabilità a priori, in pratica, non sono così utili, se confrontate alle densità condizionali.

17

Stima dei parametri – Probabilità condizionale

⇨ Supponiamo di avere c set di campioni $D_1, D_2, ..., D_c$ (uno per ogni classe c) tracciati indipendentemente in accordo alla densità $p(x|\omega_j)$

⇨ Assumiamo che $p(x|\omega_j)$ abbia forma parametrica conosciuta

⇨ Il problema di stima dei parametri consiste nello stimare i parametri che definiscono $p(x|\omega_j)$

⇨ Per semplificare il problema, assumiamo inoltre che:

⇨ i campioni appartenenti al set D_i non diano informazioni relative ai parametri di $p(x|\omega_j)$ se $i \neq j$.

18

Stima dei parametri – Due approcci

Quindi: il problema può essere formulato come:

- ⇨ Dato un set di training D={x_1, x_2, ..., x_n} (estratti indipendentemente)
- ⇨ p(x|ω) è determinata da $\boldsymbol{\theta}$, che è un vettore rappresentante i parametri necessari
  (p.e., $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ se $p(\mathbf{x} \mid \omega) \approx N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$)
- ⇨ Vogliamo trovare il migliore $\boldsymbol{\theta}$ usando il set di training.

⇨ Esistono due approcci

- ⇨ Stima Maximum-likelihood (ML)
- ⇨ Stima di Bayes

19

Stima dei parametri – Due approcci

⇨ Approccio Maximum Likelihood

- ⇨ I parametri sono quantità fissate ma sconosciute
- ⇨ La migliore stima dei loro valori è quella che massimizza la probabilità di ottenere i dati di training

⇨ Approccio Bayesiano

- ⇨ I parametri sono variabili aleatorie aventi determinate probabilità a priori
- ⇨ Le osservazioni dei dati di training trasformano queste probabilità in probabilità a posteriori modificando la stima dei veri valori dei parametri.

20

Stima dei parametri – Due approcci

- ⇨ Intuitivamente:
  - ⇨ l'approccio ML è una stima "puntuale"
  - ⇨ l'approccio Bayesiano stima i parametri con un'intera distribuzione (su cui si possono mettere anche i prior)
- ⇨ I risultati dei due approcci, benché proceduralmente diversi, sono qualitativamente simili.

21

Stima parametrica

Stima Maximum Likelihood

Approccio Maximum Likelihood

Punto di partenza:

⇨ la likelihood del training set **D**: P(**D**|θ)

⇨ Vista come funzione di **θ**, P(**D**|θ) viene chiamata ***likelihood*** *di* θ *rispetto al set di campioni* ***D.***

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

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

23

Approccio Maximum Likelihood

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

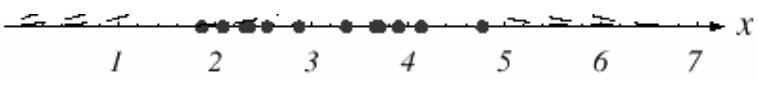
⇨ per comodità spesso si massimizza log(p(**D**|**θ**))

⇨ Ricordiamo l'assunzione che **θ** è fissato ma sconosciuto

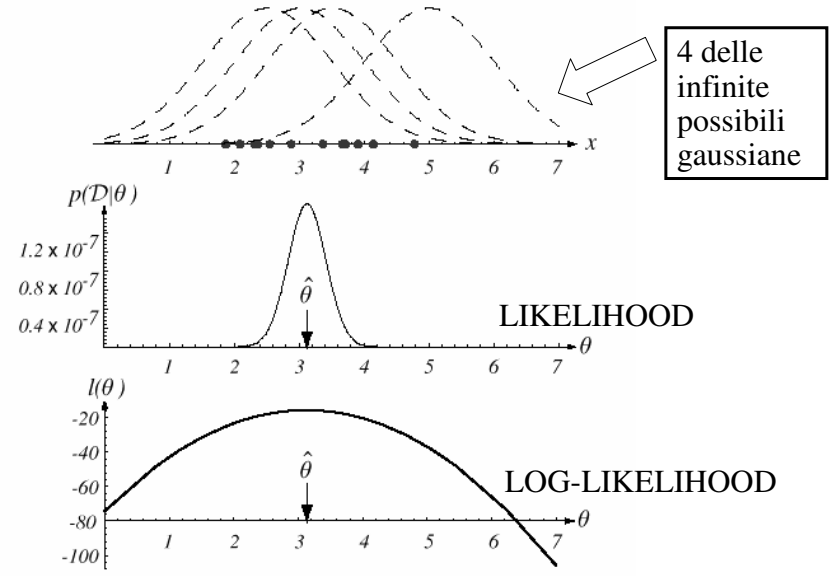
Esempio:

⇨ Punti di training 1-d generati da una densità gaussiana di varianza fissata ma media sconosciuta

⇨ Goal: stimare i parametri (cioè la media)



24



la stima ML corrisponde al valore che "supporta" meglio (è in accordo) con le osservazioni (il training set D)

25

Approccio Maximum Likelihood

NOTE IMPORTANTI:

⇨ La likelihood p(**D**|θ) è funzione di θ, mentre la densità condizionale p(x|θ) 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***

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

26

Approccio Maximum Likelihood

⇨ Se il numero di parametri da stimare è *p,* sia $\boldsymbol{\theta}=(\theta_1 \ldots \theta_p)^t$ e

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

⇨ Lo scopo è di ottenere quindi il vettore

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

⇨ in cui la dipendenza sul data set **D** è implicita

27

Approccio Maximum Likelihood

⇨ Per ricavare il max:

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

$$\Downarrow$$

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

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

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

28

Approccio Maximum Likelihood

⇨ Formalmente, una volta trovato il set di parametri che rende nulla la derivata, è necessario controllare che la soluzione trovata sia effettivamente un massimo globale, piuttosto che un massimo locale o un flesso o peggio ancora un punto di minimo.

⇨ Bisogna anche controllare cosa accade ai bordi degli estremi dello spazio dei parametri

⇨ Applichiamo ora l'approccio ML ad alcuni casi specifici

29

Maximum Likelihood: caso gaussiano

⇨ Consideriamo che i campioni siano generati da una popolazione normale multivariata di media μ e covarianza Σ.

⇨ Diversi casi:

- ⇨ monodimensionale, media sconosciuta varianza nota (alla lavagna)
- ⇨ monodimensionale, media e varianza sconosciuta (alla lavagna)
- ⇨ multidimensionale

30

ML: gaussiana multivariata

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

⇨ Per semplicità, consideriamo il caso in cui solo la media $\boldsymbol{\mu}$ sia sconosciuta. Consideriamo quindi il punto campione $\mathbf{x}_k$ e troviamo:

$$\ln p(\mathbf{x}_k \mid \boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d |\boldsymbol{\Sigma}|\right] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

$$\Downarrow$$

$$\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k \mid \boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

31

ML: gaussiana multivariata

⇨ Identificando $\boldsymbol{\theta}$ con $\boldsymbol{\mu}$ si deduce che la stima Maximum-Likelihood di $\boldsymbol{\mu}$ deve soddisfare la relazione:

$$\sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_k - \hat{\boldsymbol{\mu}}\right) = 0$$

⇨ Moltiplicando per $\boldsymbol{\Sigma}$ e riorganizzando la somma otteniamo

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$$

che non è altro che la semplice ***media*** degli esempi di training, altresì indicata con $\hat{\boldsymbol{\mu}}_n$ per indicarne la dipendenza dalla numerosità del training set.

32

ML: gaussiana multivariata

Caso media e varianza sconosciute

⇨ Il caso multivariato si tratta in maniera analoga con più conti. Il risultato è comunque:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k \qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$$

33

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

34