Analisi discriminante

LDA, QDA e altro

N. Torelli

2024

Università di Trieste

Introduzione

Introduzione

- L'analisi discriminante lineare è una tecnica di classificazione classica la cui versione originale si deve a R. Fisher
- L'approccio dell'analisi discriminante per la classificazione si propone di ottenere un modello per la distribuzione $P(X|Y)$ delle variabili di input X in ciascuna delle classi
- Si usa poi il teorema di Bayes per ottenere $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$
- In un primo approccio si potrebbero usare distribuzioni normali per ogni classe che conduce **all'analisi discriminante lineare** (o quadratica).
- L'approccio è tuttavia abbastanza generale e potrebbero anche essere utilizzate altre distribuzioni.
- Qui consideremo l'approccio classico in cui si ipotizza una distribuzione gaussiana (multivariata eventualmente) in ciascuna classe.

Analisi Discriminante Lineare (LDA)

Analisi discriminante e teorema di Bayes

Consideriamo la probabilità di classificare Y nella classe k

$$Pr(Y = k|X = x) = \frac{Pr(X = x|Y = k) \times Pr(Y = k)}{Pr(X = x)}$$

- Per impostare il problema in vista dell'analisi discriminante riscriviamo:

$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

- dove $f_k(x) = Pr(X = x|Y = k)$ è la densità di X nella classe k, ad esempio potremmo assumere che per ogni classe vi sia una distinta distribuzione gaussiana.
- $\pi_k = Pr(Y = k)$ è la probabilità marginale di appartenere alla classe k *a-priori*.

Classificazione secondo la massima densità

- Una nuova osservazione viene semplicemente classificata in base alla densità, guardando dove è più alta, se le probabilità a priori sono tutte uguali.
- Se le a-priori sono diverse, occorre prendere in considerazione anche loro e

Analisi discriminante lineare con una sola variabile ($p = 1$)

- Consideriamo quindi la Gaussiana per la distribuzione di X nella classe k

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

  μ_k è la media e σ_k^2è la varianza per le osservazioni della classe k.
- Una prima assunzione semplificatrice è quella di porre $\sigma_k^2 = \sigma^2$ ovvero le varianze sono tutte uguali nei diversi gruppi.
- Sostituendo questa densità nella formula vista precedentemente derivata dal teorema di Bayes, si ottiene

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}}$$

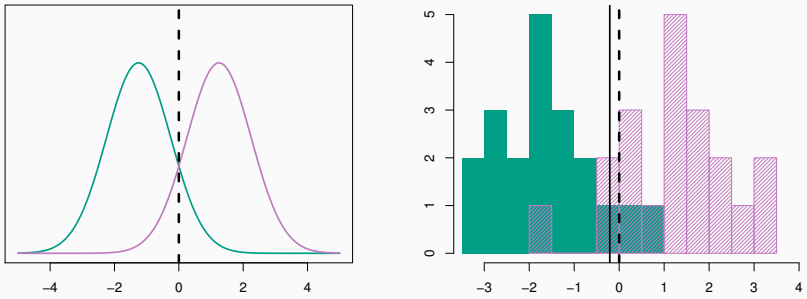
Funzioni discriminanti

- Per classificare un individuo per cui $X = x$, dobbiamo individuare per quale k si ha il più elevato $p_k(x)$.
- Dopo alcuni passaggi, considerando il logaritmo e semplificando escludendo i termini che non dipendono da k si può verificare che questo è equivalente a assegnare x alla classe che ha il valore più elevato del seguente punteggio discriminante

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k)$$

- Si noti che tale punteggio è una funzione lineare di x
- se, ad esempio, $K = 2$ e $\pi_1 = \pi_2 = 1/2$ allora la regola di decisione è di classificare come $k = 1$ se $X > \frac{\mu_1+\mu_2}{2}$

Esempio



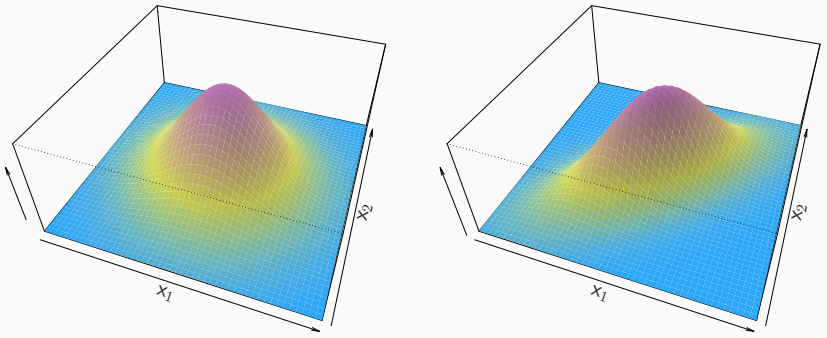
- Nella parte sinistra della figura immaginiamo di conoscere i parametri delle due gaussiane ($\mu_1 = -1.5, \mu_2 = 1.5, \pi_1 = \pi_2 = 0.5, \sigma^2 = 1$).
- Nella parte destra si rappresentano degli istrogrammi con dati provenienti dalle due gaussiane. I questo caso sostituiamo ai veri parametri ignoti le loro stime e poi stimiamo la funzione discriminante

Stima dei parametri

- $$\hat{\pi}_k = \frac{n_k}{n}$$
- $$\hat{\mu}_k = \frac{1}{n_k} \sum_{1:y_i=k} x_i$$
- $$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{1:y_i=k} (x_i - \hat{\mu}_k)^2$$

  che risulta come media ponderata delle varianze stimate in ciascuna classe

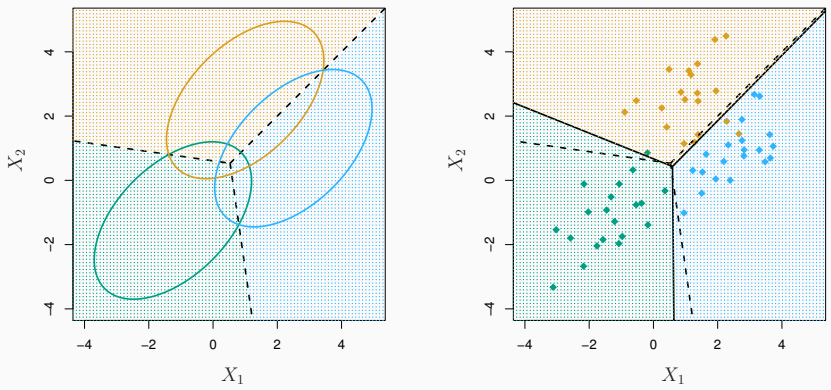
Analisi discriminante con $p = 2$



$$f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu))}$$

con funzione discriminante: $\delta_k(x) = x^T \mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T \mathbf{\Sigma}^{-1}\mu_k + log(\pi_k)$
che ha ancora la forma di una funzione lineare del tipo
$c_0 + c_1 x_1 + ... + c_p x_p$

Esempio con 3 classi



- Ove $\pi_k = 1/3$
- Le linee tratteggiate sono i confini di decisione di Bayes. Se le conoscessimo avremmo il miglior classificatore. In realtà avendo i dati stimiamo le linee solide che implicano maggiori errate classificazioni

Dalla funzione discriminante alle probabilità

- Una volta stimata la funzione $\hat{\delta}_k(x)$, si possono calcolare le probabilità

$$\hat{Pr}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}$$

- Classificare nella classe k in cui $\hat{\delta}_k(x)$ è più elevato equivale a classificarle nella classe per cui $\hat{Pr}(Y = k|X = x)$ è massima.
- Se $K = 2$ si classifica nella classe 2 se $\hat{Pr}(Y = 2|X = x) \geq 0.5$, in caso contrario nella classe 1

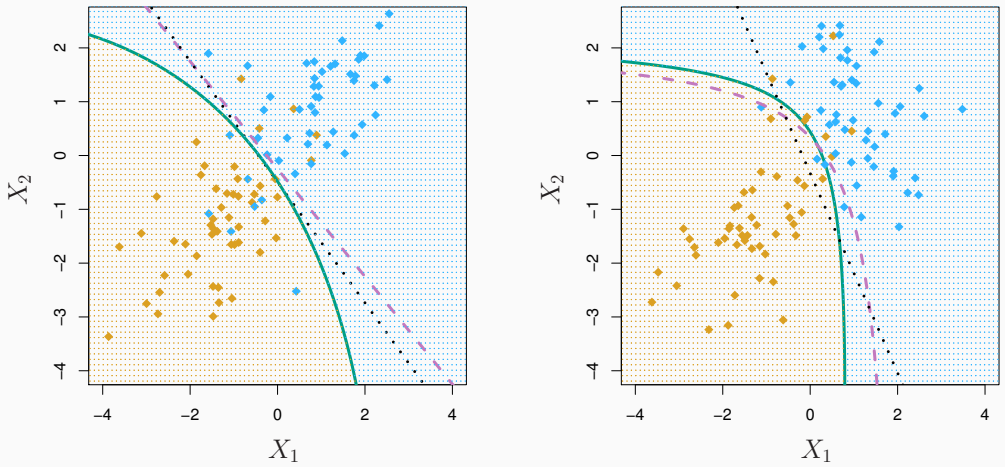
Altre forme di analisi discriminante

Oltre la LDA

- $$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

  - Se $f_k(x)$ sono Gaussiane multivariate, con la stessa matrice di covarianza in ogni classe, si ottiene l'analisi discriminante lineare (LDA).
- Se si fanno scelte diverse per $f_k(x)$, otteniamo classificatori diversi.
  - Se ad esempio per $f_k(x)$ si fa ancora l'ipotesi Gaussiana multivariate ma con $\mathbf{\Sigma}_k$ diverso in ciascuna classe, si ottiene l'analisi discriminante quadratica.
  - Con $f_k(x) = \prod_{j=1}^{p} f_{jk}(x_j)$ (modello di indipendenza condizionata) in ogni classe si ottiene l'approccio Naive Bayes. Se utilizziamo Gaussiane questo significherebbe che $\mathbf{\Sigma}_k$ è diagonale.
  - Altre forme di analisi si possono ottenere proponendo modelli di densità specifici per $f_k(x)$ inclusi approcci nonparametrici (come il metodo del nucleo per stimare le singole marginali delle X).

Analisi discriminante quadratica (QDA)



L

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}_k^{-1}(x-\mu_k)) + log(\pi_k) - \frac{1}{2}\log|\Sigma|_k$$

Poichè vi sono diverse Σ_k per gruppo, la funzione discriminante è quadratica

Naive Bayes

- Un notevole semplificazione si ha se si assume che, all'interno di ciascuna classe, le variabili x siano indipendenti, così la densità multivariata può essere espressa nella classe utilizzando le marginali (che possono anche essere delle gaussiane)
- Molto utile se p è grande e i metodi come la QDA o anche la LDA diventano di difficile applicazione.
  - Nel caso della Naive Bayes Gaussian si assume quindi che in ogni classe $\mathbf{\Sigma}_k$ è diagonale:
  - permette di usare vettori di variabili esplicative misti (qualitativi and quantitativi). Se X_j è qualitativo, basta sostituire a $f_{kj}(xj)$ la sua funzione di probabilità definita per ogni categoria della variabile discreta.
- Naive Bayes ha assunzioni più forti ma nonostante questo garantisce spesso buoni risultati della classificazione.
- Naive Bayes risulta estremamente utile nel caso di p molto grande
- Inoltre è possibile proporre approcci nonparametrici come, ad esempio, il metodo del nucleo per stimare le singole marginali delle X.

Classificazione con KNN

Classificazione con K-Nearest Neighbour

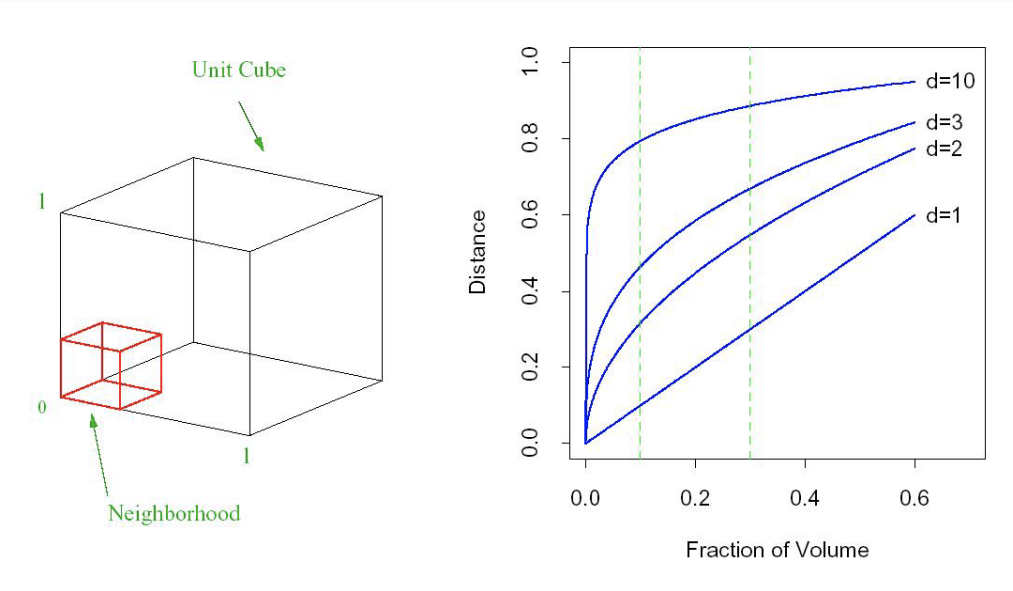
- Conviene richiamare la classificazione KNN (K-Nearest Neighbour) che è un approccio non-parametrico per stimare localmente attorno a uno specifico valore x quale sia il gruppo k cui corrisponde la maggiore probabilità $Pr(Y = k|X = x)$.
- Con K-NN, fissato un valore intero K per un valore qualsiasi x_0 del training set,
  - si individuano i K valori più prossimi a x_0, chiamiamo questo insieme L_0.
  - si stima $\hat{Pr}(Y = k|X = x_0) = \frac{1}{K} \sum_{i \in L_0} I(y_i = k)$ ovvero la frazione di elementi di tipo k nell'intorno dei K punti più vicini
- Sulla base di tali stime si classifica l'osservazione x_0 nella classe per cui la probabilità stimata è più elevata

La trappola della dimensione (the curse of dimensionality)

- Tecniche non parametriche di lisciamento o approssimazione locale (ad esempio, metodo del nucleo, KNN) funzionano bene con un numero non troppo limitato di osservazioni ma soprattutto con un basso numero di dimensioni p.
- Se p aumenta per tali metodi si presenta un problema noto come "la maledizione (o trappola) della dimensionalità" *curse of dimensionality*
- Quando la dimensione p aumenta i punti osservati si disperdono molto rapidamente nello spazio a p dimensioni.
- fissato x_0, un intorno di x_0 di raggio fisso includerà un numero trascurabile di osservazioni al crescere di p.
  - se il raggio è piccolo non troverò nessun vicino
  - i vicini più vicini in realtà nello spazio a p-dimensioni sono spesso molto distanti
  - per bilanciare tale effetto la dimensione del campione n dovrebbe crescere in modo esponenziale con p
- Se vi è un obiettivo inferenziale (stimare la funzione di densità) le functioni in uno spazio multidimensionale sono difficili da interpretare e rappresentare.
- Procedure per problemi a alta dimensionalità devono quindi essere più strutturati

Esempio: Curse of dimensionality

- Si immaginino i dati distribuiti uniformemente in un cubo d-dimensionale di lato unitario
- La frazione del volume che si considera prendendo un cubo d-dimensionale del medesimo lato imporrebbe di aumentare il bordo così da vanificare il concetto di stima locale



Considerazioni finali

Regressione logistica e LDA

- Nel caso di classificazione binaria, si noti che per LDA

$$log\left(\frac{p_1(x)}{1-p_1(x)}\right) = log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1x_1 + ... + c_px_p$$

  che ha quindi la stessa forma di un modello di regerssione logistica.
- La differenza risiede nella diversa procedura di stima dei parametri.
  - La regressione logistica usa la verosimiglianza basata sulla specificazione di $Pr(Y|X)$. le X sono quantità non stocastiche quindi
  - LDA usa la verosimiglianza basata sulla distribuzione congiunta $Pr(X, Y)$.
  - Nonostante tali differenze, nella pratica i risultati sono spesso molto simili.
- Si noti che la regressione logistica può anche adattarsi a linee di separazione non lineari come accade per la QDA. E' sufficiente includere i termini quadratici nel modello.

Alcune considerazioni conclusive

- La regressione logistica è la tecnica più popolare per la classificazione, specialmente quando $K = 2$.
- LDA è più utile e facile da utilizzare quando n è piccolo, anche quando $K > 2$, le classi sono ben separate, e l'assunzione di gaussianità è ragionevole.
- Naive Bayes è molto utile quando p è molto grande.

Cosa si può fare in R

LDA e QDA

- Una versione delle procedure di LDA e QDA è implementata nel package `MASS`, rispettivamente `lda()` e `qda()`
- La sintassi è simile a quella di altri modelli predittivi (tipo `lm` o `glm`) per cui basta mettere la formula tipo `y~x1+x2...` e specificare il data frame su cui stimare il modello
- Si può poi usare la funzione `predict` in modo analogo per ottenere una lista con tre elementi
  - `class`, `posterior` e `x`

KNN

- Si può utilizzare la funzione `knn()` nel package `class`
- La sintassi è diversa da quella usata nel caso di `lm()`

Naive Bayes

- Il package `e1071`, che contiene anche una procedura per KNN, contiene la funzione `naiveBayes()`