



JNOTES

CREATIVE NOTES



CATEGORIZATORI GENETRATIVI

I MODELLI GENETRATIVI (FARANO DI CATTURARE LA DISTRIBUZIONE: CONJUNTA DEI DATI E DELLE ETICHETTE, CLASSI)

NEL MONDANZO IN LUI AFFRONTANO UN PROBLEMA DI CLASSIFICAZIONE A INSIEME CHIUSO ADDIANO:

- UN PATIENTE O OSSERVAZIONE x_t CHE VOLGONO CLASSIFICARE COME APPARTENENTE A UNA DELLE N CLASSI POSSIBILI
- UN MODELLO PROBABILISTICO CHE CONSIDERA x_t COME UNA REALIZZAZIONE DI UNA VARIABILE CASUALE X_t
- UN ETICHETTA DI CLASSE SCONOSCIBILE CHE PUÒ ESSERE DESCRITA DA UNA RV $C_t \in \{1..n\}$

[SOLAMENTE SI SCEGLIE DI RAPPRESENTARE LE CLASSI CON NUMERI INTRI CONSECUTIVI SENZA CHE CUIVA È IL SIGNIFICATO SENZIALE, ASSEGNATO A CLASSENA ETICHETTA. PER PROBLEMI SIMILI, SPesso SI UTILIZZA LA CODIFICA {0,1} X_t SCARICANDO MATRIMONIALE]

Regola di Decisione di Bayes

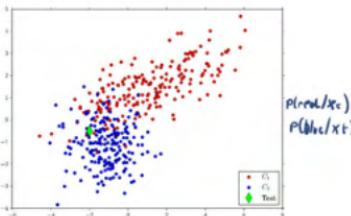
Secondo la regola di decisione ottimale di Bayes, dovremo assegnare l'osservazione x_t alla classe con la più alta probabilità a posteriori [è quindi così per lui in termini che un'osservazione appartenga alla classe $c_t = \arg\max_c P(C_t=c_t | X_t=x_t)$ e dovrà aver osservato le caratteristiche x_t . [probabilità a priori + likelihood]

Esempio: x_t rappresenta le caratteristiche estremate da un insieme

Le etichette rappresentano l'etichetta appurata ($c_1=1, c_2=2, c_3=3, \dots$) →

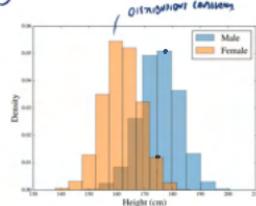
Per ogni etichetta si calcola $P(x_t=c_t | x_t=x_t)$ cioè la probabilità che la classe dell'individuo sia c_t dopo aver osservato le caratteristiche di x_t e avere preso rispetto al dataset di classificazione [se il dataset è fatto male potrai avere solo di una probabilità che dicono ...]

A binary example



What class is the green sample from?

A univariate, binary example: forensics — infer the gender from the height



What's the gender of a 174 cm tall suspect?

Questo esempio si può intuire che se è possibile che sia maschio

IL CLASSIFICATORE PUÒ TENERE (UNO ANCHE) DELI PARI PROBABILITÀ $P(c_t=c_t)$

IL TEOREM DI BAYES È ALI BASE DEL VALORE DEL PROBABILITÀ A POSTERIORI
 LA PRIMA IPOTESI CHE VIENE FATTA PERÒ È CHE I SAMPLI SIANO INDEPENDENTI E INFORMATIVI
 NELLE DISTRIBUZIONI IL CHE SIGNIFICA CHE OGNI SAMPLIO DI TEST segue LA STESSA DISTRIBUZIONE DI
 PROBABILITÀ CONCERNITA DEI DATI E DELLE CLASSI (SIMPLICIO).

QUINDI DATA LA COPPIA (SAMPLIO) DI TEST t :

(X_t, C_t) segue la stessa distribuzione di probabilità (X, C) che è stata utilizzata dai dati di training,
 attraverso l'utilizzo della MAXIMUM LIKELIHOOD.

Possiamo quindi definire $f_{X,C}(x, c)$ ovvero la funzione di probabilità concernita delle R.V.
 $X + C$ ossia tutta la relazione probabilistica tra le classi e le caratteristiche x .

QUINDI DATA UN SAMPLIO DI TEST X_t E UNA IPOTESI DI CLASSE C X LE ASSUNZIONI FATTE (I.H.)
 POSSONO VEDERSI IN FORMA CONCISA

$$f_{X,C}(x_t, c) = f_{X,C}(x_t, c)$$

POICHÉ Sono IN UN PROBLEMA DI CLASSIFICAZIONE A INSERIMENTO CONSUMO (TUTTE LE POSSIBILI CLASSI SONO NOTE E IL LAVORO DEVE APPARTENERE AD UNA DI QUESTE). APPLICANDO DATI PIAZZAMENTO STA PROBABILITÀ

$$P(C=c_t | X_t=x_t) = \frac{f_{X,C}(x_t, c)}{\sum_{c \in C} f_{X,C}(x_t, c)}$$

$$\sum_{c \in C} f_{X,C}(x_t, c) = f_X(x_t)$$

DOVE IL DENOMINATORE SONO TUTTI LE POSSIBILI CLASSI $\{ \}$ È UN VALORE COSTANTE CHE NORMALIZZA LA
 LA PROBABILITÀ A POSTERIORI GARANTENDO CHE SOMMABILE TIME $\forall c \in C$ IL VALORE SIA 2.

DI PIÙ È LA DENSITÀ DI PROBABILITÀ MARGINALE DOVE $X=x_t$ VALORE FISSO E COME TUTTE LE PROBABILITÀ

RICONOSCE CONDIZIONATA NO DATI

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \quad \forall y \in S(Y) \quad f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

LA DENSITÀ CONCERNITA PUÒ ESSERE SCRITA ANCHE COSE $f_{X,C}(x_t, c) = f_{X|C}(x_t|c)P_C(c)$

① LA DENSITÀ CONDIZIONATA $f_{X|C}(x_t|c)$ DEFINITE COME SONO DESCRITTI I DATI RISPETTO AD OGNI CLASSE (SE CONFERMATO X_t È UNO DATO IN CLASSE C CHE SONO DISTINTI DA DATI RISPETTO A QUESTA (L'ARTELLA COME È DESCRUITO SE SI TRATTI DI MOLTI QUESTI È UN LIKELIHOOD CHE AFFRONTA DATI DI TRAINING)

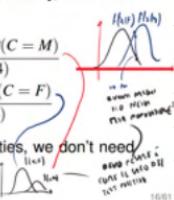
② LA PROBABILITÀ A PRIORI DELLA CLASSE $P_C(c)$ CHE rappresenta LA FREQUENZA RELATIVA ALLA CLASSE
 NEL LIVELLO APPLICATIVO (QUESTO CHE INDIVIDUALEMENTE SPAGNA SU DATI DI CLASSIFICAZIONE [NO PIÙ POSSIBILI FORME NEL NUOVO CONSET?]) ESSA AGISCE DA PFLUENTONE

We need to compute the class **posterior** probability, which depends also on the class **prior** probability

$$P(C = M|X = 174) = \frac{f_{X|C}(174|M)P(C = M)}{f_X(174)}$$

$$P(C = F|X = 174) = \frac{f_{X|C}(174|F)P(C = F)}{f_X(174)}$$

If we want to just compare the two probabilities, we don't need to compute the normalization term $f_X(174)$



A NOI BASTA IL NUMERATORE PER LA LIQUIDAZIONE E DEFINITAMENTE È SEMPRE UGUALE

Ricorda —

RELOVA DI DATES

Ci permette di invertire la condizione, la formula è

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)}$$

Quello che avremo voluto costruire con il nostro modello è una cosa che ci permetta di calcolare $f_{X|C}(x|c)$ [una sua probabilità sarà data come input alla nostra funzione]

Dopo che neanche noi conosciamo la distribuzione dei sample isolati la funzione di verosimiglianza per trovarla rispetto ad ogni classe

In generale quindi vogliamo costruire un modello parametrico per la class-conditional probability che ci calcoli la funzione dei parametri al probabilità condizionata

$$f_{X|C}(x|c) = f_{X|C}(x|c, \theta)$$

(CONSIDERO I PARAMETRI
che si sa già la classe
(x, c) corrisponde)

Nel caso causando θ contenuti il numero degli errori è la misura di covariogramma
"Se sapiamo che un campione appartiene alla classe c quale distribuzione di probabilità seguono le sue caratteristiche?"

DICHIARAZIONE PROBABILITÀ A PRIORI

$P_C(c)$ rappresenta la prior probabilità ed è indipendente dai parametri del modello, essa dipende specificamente dall'applicazione e rappresenta la nostra conoscenza o aspettativa sulla frequenza delle classi min di osservare i dati

In un approccio frequentista è la frequenza relativa attiva (probabilità) di classificare un nel flusso di dati che il nostro classificatore dovrà emettere [quasi sempre in base al dataset di valutare]

È per questo che il nostro modello non si basa su questa

GAUSSIAN CLASSIFIER

Dopo aver scoperto la densità condizionata in densità condizionata (likelihood) + probabilità a priori, il passo successivo consiste nel calcolo della verosimiglianza $f_{X|C}(x|c)$

PER DATI CONTINUI ASSUMONO VALORI IN R. E LA SCELTA DEL MODELLO DIPENDE DALLA NATURA DEI DATI STESSI IN QUESTO CASO FAGLIANO L'IPOTESI CHE:

I DATI DI CIASUNA CLASSE POSSONO ESSERE EFFICACIAMENTE ADDOTTATI DA UNA DISTRIBUZIONE GAUSSIANA (NORMALE)

MULTIVARIANZA (cioè UTILE E SPECIALE X DIVERSI MOTIVI):

• FORMALISMO TEORICO: TEOREMA DEL LIMITE CENTRALE

• TRATTABILITÀ MATEMATICA:

• PARAMETRIZZAZIONE PARZIALE: DEFINITA DA MEDEA E COVARIANZA (ECONOMIA NEL CASO MULTIVARIATO)

QUINDI (I DATI CONDIZIONATI AGLI STESI SEGUONO UNA DISTRIBUZIONE GAUSSIANA (NORMALE)):

$$(X_t | C_t = c) \sim (X_t | C_t = c) \sim N(\mu_c, \Sigma_c)$$

VICEL N DISTRIBUZIONE CONDIZIONATA DELLE PERTURB. $X = \{X_1, X_2, \dots\}$ È UNA QUADRUPA (MULTIVARIANZA) CON

$\mu_C = VETTORE$ DELLE MEDIE, SPACCIO X CLASSE

$\Sigma_C = MATEMATICA$ DI COVARIANZA SPECIFICA PER CLASSE

SE CONSIDERIAMO QUESTI PARAMETRI PER Ogni CLASSE, POTEMMO COLLOCARME APROXIMATIVAMENTE $f_{X|C}(x_t | c) = N(x_t | \mu_c, \Sigma_c)$

Dove $N(x_t | \mu_c, \Sigma_c)$ rappresenta densità di probabilità di una distribuzione normale con media μ_c e matrice di covarianza Σ_c

NELLA REALITÀ NOI NON CONOSCiamo I VALORI DEI PARAMETRI DEL MODELLO: $\theta = \{(\mu_1, \Sigma_1), \dots, (\mu_n, \Sigma_n)\}$

QUESTI PARAMETRI DEVO NO ESSERE STIMATI DAI DATI. A NOSTRA DISPOSIZIONE ABBEDO UN INSIGNE DI ADDETTAMENTO ETICHETTATO

$$D = \{(x_1, c_1), (x_2, c_2), \dots, (x_n, c_n)\}$$

DOVE:

$X = \{x_1, \dots, x_n\}$ SONO I CAMPIONI OSSERVATI

$C = \{c_1, \dots, c_n\}$ SONO LE LABELS DI CIASUNA OBSERVATION $c_i \in \{1, \dots, k\}$

L'ASSUNZIONE CAUTELARE PER OTTENERE DI STIMARE I PARAMETRI È CHE ORI I PARAMETRI DEL MODELLO θ , LE OBSERVATION. R.V. SIANO INDEPENDENTI E IDENTICAMENTE DISTRIBUITI (iid)

$$[(x_i, c_i) \perp\!\!\!\perp (x_j, c_j)] | \theta \quad \forall i, j \quad (x_i, c_i | \theta) \perp\!\!\!\perp (x_j, c_j | \theta)$$

I CAMPIONI DI TEST E ADDETTAMENTO SONO INDEPENDENTI TRA DI LORO (CONDIZIONALMENTE AL PARAMETRO)

I CAMPIONI DI TEST SONO INDEPENDENTI DA QUELLI DI ADDETTAMENTO

TUTTI I CAMPIONI DI TEST E ADDETTAMENTO SEGUONO LA STEMA DISTRIBUZIONE [RIVARO DALLA STIMA ASSUNZIONE]

Nota: X_i È IL VETTORE PERTURBAZIONE E c_i È LA CLASSE SPECIFICA ASSOCIAA AL SENSO DA CUI POSSONO STIMARE IL R.V. θ . θ È UN VETTORE CON PARAMETRI DIVISI PER Ogni CLASSE

DEFINIZIONE DATO CHE POSSO UNA DISTRIBUZIONE GAUSSIANA PER $X|C$, ora la iid. DEI DATI DELL'ECOLE DE SINGAPORE HANNO A UNO UNICO VETTORE DI X :

$$(X_t | C_t = c, \theta) \sim (X_t | C_t = c, \theta) \sim (X_t | C_t = c, \theta) \sim N(\mu_c, \Sigma_c)$$

[STO CONSIDERANDO SOLO
UNA CLASSE C]

PARAMETRI CONO DELL'ECOLE SONO

$$\theta = \{(\mu_1, \Sigma_1), \dots, (\mu_n, \Sigma_n)\}$$

PER OTTO CLE
VOLGO IL CAMPIONE
CONSIDERANDO X
DENTRO UNO SULL'ALTRI
DEI VALORI

Appocio Enfouantis

SEGUENDO L'AMMOLIO FREQUENZIATO, VUOLONO CALCOLARE QUANTO VADA STIMA PUNTUAZIONE DELL'AMMOLIO. DANNI LA STIMA IN USOFONO X CALCOLARE.

$$f_{X_t|G_t}(x_t|c) \approx N(x_t|\mu_c, \Sigma_c)$$

$$f_{X|C} \left(\begin{bmatrix} x_0 \\ x_5 \end{bmatrix} | m \right) = N \left(\begin{bmatrix} x_0 \\ x_5 \end{bmatrix} | u_m, Z_m \right)$$

IL PIEMONTE PIÙ COLONIZZATO PER FALE GIÀ È LA STIMA DI MASSIMA VENOSOGLIANZA (PIEMONTE LAKWOOD ESTIMATIVO).
NLE. TROVIAMO I VALORI OBIETTIVI DEL MIGLIORAMENTO IN PROSPETTIVA DI OSSERVARE I DATI DI ADDESTRAMENTO

$$L(0) = f(x_1, \dots, x_n, c_1, \dots, c_m | 0) \quad (x_1, \dots, x_n, c_1, \dots, c_m | 0)$$

(PROM AL PIAZZA È SLANTA COSÌ, TUTTO (X, ω) DI UN SISTEMA SPATIALE IN DISTORSIONE. O) (ANALOGO RIFERIMENTO SOSPETTO CHE ENTROGLI VALORI GIA' IN DISTORSIONE. QVELLA DISTORSIONE SONNO LE NOSESE IPOTESI.)

DATA IN I.B.D OFI 5971

$$C(0) = \prod_{i=1}^n f_{x_i < 10}(x_i, c_i | 10)$$

- Ricchezza di organismi con contributo NO BATES

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}, \quad y \in S(Y)$$

7 *com de p
650000
na pris*

IN PWD

14 days —
After 14 Days

$$f_{\pi}(x) \circ (\pi/x) = \underline{f(x)\pi} / (\pi x) \circ \pi$$

PROBLEMA IN DIFESA DI PROSPETTIVE CONDIVISE CON IL MESSAGGIO DI CONFIDENZIALITÀ.

Densità convolare per le caratteristiche ora in classe

$$f(x|c, \theta) = N(x|\mu_c, \sigma_c)$$

PROBABILITÀ A PROVA DI UNA CLASSE CHE DIPENDE DALL'APPLICAZIONE $P(\cdot)$

Q U I M O I X U S S I M B O L O D (S A T R I A N O S)

$$f_{x, q_0}(x, c/10) = f_{x/c, 0}(x/c, 0) \cdot P_c(c) = \mu(x/c, \tau_c) P_c(c)$$

$$L(\theta) = \prod_{i=1}^n f_{X_i|C_i}(\mathbf{x}_i | c_i, \theta) = \prod_{i=1}^n f_{\mathbf{X}|C, \theta}(\mathbf{x}_i | c_i, \theta) P(c_i) = \prod_{i=1}^n N(\mathbf{x}_i | M_{c_i}, \Sigma_{c_i}) P(c_i)$$

È IL NUOVO TOTALE DI CAMPIONI NEL DATASET

X; i' IL VENUS OFIE GATTAFANTUR OFEL'i-FEND WATWOF

Ci è in classe dell'iesmo capitolato

Mc: È IL VENTRE DELLE APOFE PER INCHIUSCI

Σ_{ci} è la matrice di covarianza per la classe ci.

Per semplificare i calcoli lavoriamo con i numeri

$$l(0) = \log L(0) = \sum_i \log N(x_i | \mu_{c,i}, \Sigma_i) + \sum_i \log p(c_i)$$

DOI: 10.4236/ojs.v1101301

$\chi_1 \cap \dots \cap \chi_n \cap \dots \cap \chi_{n-1} \cap \dots \cap \chi_1$
 $\log N(\nu_{\text{cal},n}, S_n)$
 $\log N(\nu_{\text{exp}}, S_d)$

$$l(\theta) = \sum_{c=1}^C \sum_{i: x_i \in c} \log N(x_i | \mu_c, \Sigma_c) + \xi$$

μ è il vettore delle medie di classe

Σ somma su tutti i vettori i che appartengono alla classe c

{ escludere i termini relativi alle probabilità a priori delle classi che non differiscono da 0 (le nostre iniziali) non ci interessa

Osserviamo quindi che la log-likelihood si scomponga in una somma di termi, uno per ogni classe

$$l(\theta) = \sum_{i: x_i \in c} l_c(\mu_c, \Sigma_c) + \xi, \quad l_c(\mu_c, \Sigma_c) = \sum_{i: x_i \in c} \log N(x_i | \mu_c, \Sigma_c)$$

Quindi possiamo minimizzare $l(\theta)$ ricavandone separatamente ciascun termine $l_c(\mu_c, \Sigma_c)$ ovvero parlarne della classe c i parametri μ_c, Σ_c di una classe non influenzando le variazioni delle altre classi.

[per spiegazione ottimizzazione reale minima di probabilità]

$$\text{allora } \log N(x | \mu, \Sigma) = -\frac{D}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)$$

Dove:

D è la dimensione del vettore x

$|\Sigma|$ è il determinante della matrice di covarianza

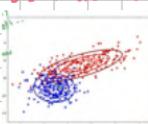
Σ^{-1} è l'inversa della matrice di covarianza

Saranno finite e più convenienti lavorare con la matrice di precisione

$$\Lambda = \Sigma^{-1}$$

$$\log N(x | \mu, \Sigma) = -\frac{D}{2} \log 2\pi + \frac{1}{2} \log |\Lambda| - \frac{1}{2} (x - \mu)^T \Lambda (x - \mu)$$

Nota: il termine $(x - \mu)^T \Lambda (x - \mu)$ è uno scalare positivo o zero. Più grande al quadrato il vettore x e la media μ , questo termine è costante = 0 e definisce una superficie nello spazio delle caratteristiche nel caso bidimensionale è un ellisse



L'allungamento e la forma dell'ellisse sono dati dai valori dei quovettori e autovettori di Σ

Nel nostro classificatore ad ogni classe è associata una propria famiglia di ellissi concentriche, ciascuna corrispondente ad un diverso valore di probabilità.

I punti più lontani sono più probabili i più estremi meno

$$l_c(\mu_c, \Sigma_c) = h + \frac{N_c}{2} \log |\Lambda_c| - \frac{1}{2} \sum_{i: x_i \in c} (x_i - \mu_c)^T \Lambda_c (x_i - \mu_c)$$

h è una costante che include il termine $-\frac{D}{2} \log 2\pi$

N_c è il numero di campioni nella classe

Utilizziamo le proprietà delle matrici dato lo scatene (forma quadratica)

$$l_c(\mu_c, \Sigma_c) = \sum_{i: x_i \in c} \log N(x_i | \mu_c, \Sigma_c) = \sum_{i: x_i \in c} \left[-\frac{D}{2} \log 2\pi + \frac{1}{2} \log |\Lambda_c| - \frac{1}{2} (x_i - \mu_c)^T \Lambda_c (x_i - \mu_c) \right]$$

$$\begin{aligned} \text{N.B. } & \frac{N_c}{\text{numero classi}} - \frac{N_c D}{2} \log 2D + \frac{N_c}{2} \log |\Lambda_C| - \frac{1}{2} \sum_{i \in C} (x_i - \mu_C)^T \Lambda_C (x_i - \mu_C) \\ & \downarrow \end{aligned}$$

$$h = -\frac{N_c D}{2} \log 2D \quad [\text{costante}]$$

$$L_C(\mu_C, \Sigma_C) = h + \frac{N_c}{2} \log |\Lambda_C| - \frac{1}{2} \sum_{i \in C} (x_i - \mu_C)^T \Lambda_C (x_i - \mu_C)$$

ESEMPIO: IL TERMINE $(x_i - \mu_C)^T \Lambda_C (x_i - \mu_C)$
 $(x_i - \mu_C)^T \Lambda_C (x_i - \mu_C) = x_i^T \Lambda_C x_i - x_i^T \Lambda_C \mu_C - \mu_C^T \Lambda_C x_i + \mu_C^T \Lambda_C \mu_C$.
 DIO CHE Λ_C È SIMMETRICA [SOMMA DI QUADRATI] $x_i^T \Lambda_C \mu_C = \mu_C^T \Lambda_C x_i$ QUINDI

$$(x_i - \mu_C)^T \Lambda_C (x_i - \mu_C) = x_i^T \Lambda_C x_i - 2 \mu_C^T \Lambda_C x_i + \mu_C^T \Lambda_C \mu_C$$

$$L_C(\mu_C, \Sigma_C) = h + \frac{N_c}{2} \log |\Lambda_C| - \frac{1}{2} \sum_{i \in C} [x_i^T \Lambda_C x_i - 2 \mu_C^T \Lambda_C x_i + \mu_C^T \Lambda_C \mu_C]$$

$$L_C(\mu_C, \Sigma_C) = h + \frac{N_c}{2} \log |\Lambda_C| + \frac{1}{2} \sum_{i \in C} x_i^T \Lambda_C x_i + \mu_C^T \Lambda_C \sum_{i \in C} x_i - \frac{1}{2} \sum_{i \in C} \mu_C^T \Lambda_C \mu_C$$

IL TERMINE $\mu_C^T \Lambda_C \mu_C$ NON DIPENDE DA I DATI $x_i^T \Lambda_C x_i$ È UNA COSTANTE QUADRATICA, DA UNO SOTTRAWE ALTRI POSSONO CONSIDERARSI IN TRIVELLA

$$L_C(\mu_C, \Sigma_C) = h + \frac{N_c}{2} \log |\Lambda_C| - \frac{1}{2} \operatorname{Tr}(\Lambda_C \sum_{i \in C} x_i x_i^T) + \mu_C^T \Lambda_C \sum_{i \in C} x_i - \frac{N_c}{2} \mu_C^T \Lambda_C \mu_C$$

La log-likelihood è poi scritta DIPENDE DA 3 STATISTICHE FONDAMENTALI

$Z_C = N_c$ Numero di campioni nella classe C

$F_C = \sum_{i \in C} x_i$: SOMMA DI TUTTI I VETTORI DELLE CARATTERISTICHE DELLA CLASSE C

$S_C = \sum_{i \in C} x_i x_i^T$: SOMMA DEI PRODOTTI CIRCONFERENZIALI VETTORI DELLE CARATTERISTICHE.

Per un campione di classe C, queste 3 statistiche (Z_C, F_C, S_C) CONTINNO GRAVIAMENTE TUTTA L'INFORMAZIONE NECESSARIA PER STIMARE I PARAMETRI $\mu_C \in \mathbb{R}^n$.

(Invece, non si sapeva di più con i dati originali)

INFINE PASSEREMO LA (10) INVECEGGI I VALORI OTTENUTI DI $\mu_C \in \mathbb{R}^n$.

MISURAZIONI DISPERSE ALLA POCIA

PARTENDO DA NERON

$$L_C(\mu_C, \Lambda_C) = h + \frac{N_c}{2} \log |\Lambda_C| - \frac{1}{2} \sum_{i \in C} x_i^T \Lambda_C x_i + \mu_C^T \Lambda_C Z_C x_i - \frac{N_c}{2} \mu_C^T \Lambda_C \mu_C$$

Calcoliamo la derivata rispetto a μ_c e la proiettiamo sul vettore a 0

Non: $\frac{\partial \text{Lc}(\mu_c)}{\partial \mu_c} = 0 \Rightarrow \sum_{i \in c} x_i - N_c \mu_c = 0$ se è simmetrico

$$\nabla_{\mu_c} \text{Lc}(\mu_c, \lambda_c) = \lambda_c \sum_{i \in c} x_i - N_c \lambda_c \mu_c = 0$$

$$\mu_c = \frac{1}{N_c} \sum x_i$$

μ_c è la media campionaria dei dati, appartenenti alla classe c

Possiamo alla Σ_c

Stiamo lavorando con la matrice di precisione $\Lambda_c = \Sigma_c^{-1}$, poi invertendo il risultato.

Possiamo dalla formula alternativa della likelihood

$$\text{Lc}(\mu_c, \lambda_c) = N + \frac{N_c}{2} \log |\Lambda_c| - \frac{1}{2} \text{Tr} \left(\Lambda_c \sum_{i \in c} (x_i - \mu_c)(x_i - \mu_c)^T \right)$$

Non: $\frac{\partial \log |\Lambda_c|}{\partial \lambda_c} = (\Lambda_c^{-1})^T = \Lambda_c^T$

$\frac{\partial \text{Tr}(\Lambda_c)}{\partial \lambda_c} = \Lambda_c^T$

Perché simmetrica, $\Lambda_c^T = \Lambda_c$, mentre $\Lambda_c^{-1} \in \mathbb{Z}$

$$\nabla_{\Lambda_c} \text{Lc}(\mu_c, \lambda_c) = \frac{N_c}{2} \Lambda_c^{-T} - \frac{1}{2} \sum_{i \in c} (x_i - \mu_c)(x_i - \mu_c)^T =$$

$$= \frac{N_c}{2} \sum_{i \in c} (x_i - \mu_c)(x_i - \mu_c)^T = 0$$

$$\Rightarrow \Sigma_c = \frac{1}{N_c} \sum_{i \in c} (x_i - \mu_c)(x_i - \mu_c)^T$$

Quindi, è utile alla matrice di covarianza campionaria rispetto ai dati appartenenti alla classe c .

Non in tutte le Σ deve essere definita positiva, se è così allora questa soluzione è effettivamente il massimo globale oltretutto verosimiglianza. $[x_i \neq 0]$

Stimate queste quantità dei dati di addestramento, possiamo utilizzarle in modellazione di un nuovo punto x_t per la classe c

$$f_{x_t | c}(x_t | c) = f_{x_t | c}(x_t | c) = N(x_t | \mu_c, \Sigma_c)$$

Dato che

$$P(c=c_t | x_t) = f_{x_t | c}(x_t | c) P(c=c_t)$$

$$\sum_{c=c_t} f_{x_t | c}(x_t | c) [\text{Dato che ha null probability il denominatore ha poco importanza}]$$

RATIO (CLASSIFICATORE BINARIO)

NEL CONTESTO BINARIO ASSUMO SPESO 2 CLASSI, CHE SOVENTE DENOTANO COSE h_1 (IPOTESI POSITIVA) E h_0 (IPOTESI NEUTRALE). [IN SICURA DELLE ETICHETTE I MARMORI]

LA REGOLA DI DECISIONE BAYESIANA RIMANE LA STESSA: ASSEGNA UNAMPIONE DI TEST x_t ALLA CLASSE CON PROBABILITÀ A POSTERIORI PIÙ ALTA. CONTINUAMENTE

$$\begin{aligned} P(C=h_1/x_t) &\text{ MAU CHE } x_t \text{ APPARTIENE ALLA CLASSE } h_1 \\ P(C=h_0/x_t) & " h_0 \end{aligned}$$

SOMMARE è CONVENIENTE COMBINARE IL MAPPOATO TRA LE 2 PROBABILITÀ

$$r(x_t) = \frac{P(C=h_1/x_t)}{P(C=h_0/x_t)} \quad \begin{array}{l} \text{Se } r > 1 \text{ h}_1 \text{ è PIÙ PROBABILE DI } h_0 \in [1, +\infty] \\ \text{ASIMMETRICO SOLITAMENTE SI OPERA CON UNA MAPPIAZIONE LOGARITMICA} \end{array}$$

$$\log(r(x_t)) = \log\left(\frac{P(C=h_1/x_t)}{P(C=h_0/x_t)}\right) \quad \begin{array}{c} \text{Se } r > 0 \\ \text{h}_1 \quad \text{h}_0 \end{array} = 0 \text{ PUNTO DI INDIFFERENZA, DECISIONE CASUA SU ALTRI CRITERI}$$

La vera eleganza del ratio è nel farsi quando rischiavano il mappoato di probabilità a posteriori con le medesime difficoltà di classificazione e della stessa priorità:

$$\begin{aligned} \log r(x_t) &= \log\left(\frac{P(C=h_1/x_t)}{P(C=h_0/x_t)}\right) = \log \frac{f_{x|C}(x_t/h_1) / f_{x|C}(x_t/h_0)}{f_{x|C}(x_t/h_0) / f_{x|C}(x_t/h_1)} = \log \frac{f_{x|C}(x_t/h_1) P(C=h_1)}{f_{x|C}(x_t/h_0) P(C=h_0)} \\ &= \log \frac{f_{x|C}(x_t/h_1)}{f_{x|C}(x_t/h_0)} + \log \frac{P(C=h_1)}{P(C=h_0)} \end{aligned}$$

IL LOG-MAPPOATO È UNA SOMMA DI 2 CONTRIBUTI:

$$\text{- IL LOG-MAPPOATO DI VEROBILITÀ} = \ln r(x_t) = \ln \frac{f_{x|C}(x_t/h_1)}{f_{x|C}(x_t/h_0)}$$

DIREMO CHE NOSTRO AGIOLO, QUANDO I DATI OSSERVATI SUPPORTANO L'IPOTESI h_1 , SUSPENGO

$$\text{- IL LOG-MAPPOATO} \cdot \log \frac{P(C=h_1)}{P(C=h_0)} \text{ RISPECCHIA LA NORMA CONOSCIMENTO E ASPETTATIVA PRIMA DI OSSERVARE I DATI}$$

IN UN PROBLEMA BINARIO SI PUÒ PARAHETIZZARE IN PROBABILITÀ A PRIORI, IN PARTICOLARE

$$P(C=h_1) = \pi$$

$$P(C=h_0) = 1 - \pi$$

IL LOG-ODDS AVERGHIAMO CHIÈSE AL VIVO:

$$\log \frac{\pi}{1-\pi}$$

L'UNA A SUA VOLTA PUÒ ESSERE INTERPRETATA CON UNO SCORI CHE INDICANO

QUANTO PUNTAMENTE I SOGLI SUPORTANO L'IPOTESI h_1 RISPETTO AD h_0

- VALORI PIÙ ALTI HANNO MAGGIOR SUPPORTO VERSO h_1
- VICEVERA' h_0
- O MAIORI CHE ERRORE SONO ULTIMAMENTE POSSIBILI

In DECISIONE FINALE SI HA PENO PER

$$\text{log}(x|x) = \log \frac{f_{x|x}(x|x)}{f_{x|x}(x|H_0)} + \log \frac{\pi}{1-\pi} \geq 0$$

$$\log \frac{f_{x|x}(x|x)}{f_{x|x}(x|H_0)} \geq -\log \frac{\pi}{1-\pi}$$

CIOÈ OPIENO DI PUNTAZIONE DELLA PANIER PROBABILITÀ OVVERO CALCOLO CONOSCIMENTO CHE APPARTENGONO A PRATICI SUL DATI DA CLASSIFICARE (CONTESTO APPUNTITIVO)

$$-\log \frac{\pi}{1-\pi} = t = \text{threshold}$$

- COSÌ FALMDO IO POSSO SVILUPPARE UN SISTEMA CHE CALCOLA ACCUMULATIVAMENTE LA $\text{llr}(x)$, INFORMATIVAMENTE DI UN APPUNTITO FINALE
- ADATTARE IL SISTEMA A DIVERSI NUOGLI APPUNTITI, SPERIMENTANDO ADDIFLANDO LA SOLITA AI DECISIONE SENZA DOVER RIADATTARNE IL NUOGLIO.

INTERPRETAZIONE CONFIDENZIALE CLASSIFICAZIONE

NEL CASO DI UN CLASSIFICATORE GAUSSIANO CHIUSO CON PARAMETRI $(\mu_1, \lambda_1^{-1}), (\mu_0, \lambda_0^{-1})$ POSSANO CALCOLARE IL $\text{llr}(x)$

CONCETTO

$$\text{llr}(x) = \log \frac{N(x|h_1)}{N(x|h_0)} = \log \frac{N(x|\mu_1, \lambda_1^{-1})}{N(x|\mu_0, \lambda_0^{-1})}, \text{ QUESTA ESPRESSIONE FALMDO}$$

ACCUMULI MANIPOLAZIONI ALLEGGERITE PUÒ ESSERE RISUITA CON UNA FORMA QUADRATICA IN X:

$$\text{llr}(x) = x^T A x + x^T b + c$$

DOVE:

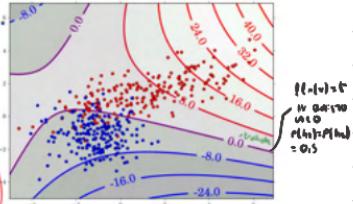
$$A = \frac{1}{2}(\lambda_1 - \lambda_0)$$

$$b = \lambda_1 \mu_1 - \lambda_0 \mu_0$$

$$c = -\frac{1}{2}(\mu_1^T \lambda_1 \mu_1 - \mu_0^T \lambda_0 \mu_0) + \frac{1}{2}(\log |\lambda_1| - \log |\lambda_0|)$$

DEMO CIOÈ QUESTO CI FA CAPIRE CHE IL CONFINE DI DECISIONE (DOVE $\text{llr}(x)=t$) SARÀ IN GENERALE UNA FORMA QUADRATICA

Binary problem — decision boundaries



LA FORMA DELLA SUPERFICIE DI PRELISIUNE DIPENDE dalle RELAZIONI TRA LE MATRICI DI COVARIANZA DELLE 2 CLASSE.

① QUANDO $H_1 \neq H_2$ IL TERMINE QUADRATICO H È NON NULLO. IL CONCONE DI DECISIONE È UNA SUPERFICIE QUADRATICA L'ENFATIZZANTE, CHE IN 2D CORRISPONDE AD UNA CURVA: ELLISSE, IPERBOLA, PARABOLA.

$$H > 0 \quad \text{e/o} \quad H < 0$$

- ② QUANDO $H_1 = H_2$ IL TERMINE H SI ANNULLA E LA FUNZIONE DI DECISIONE DIVENTA LINEARE IN x . IN QUESTO CASO IL CONCONE DI DECISIONE È UN IPERPLANO CHE APPARE COME UNA LINEA RETTA IN 2D.

Multiclass Problem

$C \in \{h_1, h_2, \dots, h_n\}$, LORE ORO HA AVUTO USO UNA TABELLA DI PROBABILITÀ A POSTERIORI DI TUTTI GLI ELIMINATORI PER RISOLVERE IL PROBLEMA DI CLASSIFICAZIONE.

$$P(C=h_i|x_t) = \frac{f_{x_t}(x_t|h_i) P(h_i)}{\sum_{h \in \{h_1, h_2, \dots, h_n\}} f_{x_t}(x_t|h)}$$

$$\underset{h \in \{h_1, h_2, \dots, h_n\}}{\operatorname{argmax}} f_{x_t}(x_t|h) P(h)$$

IL RISULTATO DI UNA DECISIONE OTTIMALE CONSIGLIA POKO NELLO SCEGLIERE UNA PIÙ ALTA PROBABILITÀ A POSTERIORI.

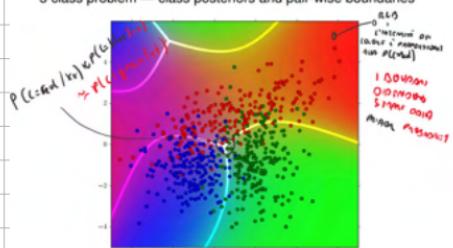
$$C^* = \underset{h}{\operatorname{argmax}} P(h|x_t)$$

DATO CHE IL DECISIONNATORE È LO STESSO PER TUTTE LE CLASSI SI PUÒ SEMPLIFICARE E SCRIVERE:

$$C^* = \underset{h}{\operatorname{argmax}} f_{x_t}(x_t|h) P(h) = \underset{h}{\operatorname{argmax}} [\log f_{x_t}(x_t|h) + \log P(h)]$$

ANCHE QUI SE PARMIAMO UNA PARTE INTUITIVA AL MODELLO DATA PARTE DIPENDENTE DALL'APPLICAZIONE.

3 class problem — class posteriors and pair-wise boundaries



COME APPARTENIMENTI A UNA DETERMINATA CLASSE.

NEL CASO MULTICLASSE, I CONCONE DI DECISIONE SONO DETERMINATI DALLE REGIONI DOVE 2 CLASSI HANNO UN'UNICA PROBABILITÀ A POSTERIORI, mentre tutte le altre classi hanno probabilità inferiori.

IL CAMPO PER IL PROBLEMA A 3 CLASSI MOSTRA COME LE REGIONI DI DECISIONE SIANO DEFINITE DA SEMPREGLI ALI CAVO QUADRATICHE Ogni REGIONE RAPPRESENTA NELLA LINEA DI PUNTI CHE VENGONO CLASSIFICATI

APPLICAZIONI PRATICHE

IL MODELO GAUSSIANO COMPLETO ABBINDE LA STIMA DI NUEROSI PARAMETRI:

- PER OGNI CLASSE DOBBIANO STIMARE UN VETTORE DELLE MEDIE μ_{c_i} (OMMEREI)
- E UNA MATELLA DI COVARIANZA Σ_c CON $[D \times (D+1)]/2$ PARAMETRI INDEPENDENTI
INDIVIDUALE

QUANDO D È ELEVATO RISPETTO AL NUMERO DI SAMPLI DISPONIBILI, LA STIMA DELLA MATELLA DI COVARIANZA DIVENTA AMBITOLAMENTE COMPLESSA. CON POCI DATI LE STIME POSSONO ESSERE INSTABILI O ADDIRITTURA SINGOLI (NON INVERTIBILI) CONFRONTANDO IL CLASSIFICATORE.

MENTRE IL PESO COMPUTAZIONALE DIVENTA ADDIRITTURA ELEVATO. PER QUESTO SI ADOTTANO DELLE STRATEGIE CHE PERMETTANO DI SEMPLIFICARE IL TUTTO.

NAIVE BAYES

L'APPROCCIO NAIVE BAYES INTRODUCE UNA SEMPLIFICAZIONE SIGNIFICATIVA MA POTENTE: LE DIVERSE COMPONENTI DEL VETTORE DELLE CARATTERISTICHE SONO CONSIDERATE CONDIZIONALMENTE INDEPENDENTI DENTRO CLASSE. $x_{c_1} \perp\!\!\!\perp x_{c_2} | C$
CIÒ SIGNIFICA

$$f_{X|C}(x|C) \propto \prod_{j=1}^D f_{X_{c_j}|C}(x_{c_j}|C)$$

$$X = \begin{bmatrix} x_{c_1} \\ x_{c_2} \\ \vdots \\ x_{c_D} \end{bmatrix} \quad X_{c_j} \text{ è la j-esima componente}$$

DOVE x_{c_j} RAPPRESENTA LA j-ESIMA COMPONENTE DEL VETTORE X (DA NON CONFONDERE CON x_j CHE INDICA IL j-ESIMO CARATTERE DEL DATASET)

NAIVE = INLEGNO, L'ASSUNZIONE DI INDEPENDENZA CHE FAGLIO È SPesso UNA SEMPLIFICAZIONE PERMETTENDO CHE LE CARATTERISTICHE SONO CONSIDERATE COMPLETAMENTE INDEPENDENTI ES. PIÙ o ALTRE

CIÒ SIGNIFICA AVERE UNA CORR=0

L'ASSUNZIONE DI NAIVE NON È LEGATA A UNA SPECIFICA FAMIGLIA DI DISTRIBUTIONI.

NAIVE BAYES GAUSSIANO

NELL'IPOTESI GAUSSIANA MODELLANO OGNI COMPONENTE CONDIZIONALE $f(x_{c_j}|C)$ COME UNA GAUSSIANA UNIVARIATA

$$f_{X_{c_j}|C}(x_{c_j}|C) = N(x_{c_j}|C; \mu_{c_j}, \sigma_{c_j}^2)$$

- PER Ogni CLASSE, ABBIANO ANCORA D MEDIE μ_{c_i}
- E ABBIANO SOLO D VARIANZE σ^2 DATO CHE NON C'È CORRELAZIONE INVECE DI $(D \times D)/2$

APPLICANDO IL MODELLO X INDIVIDUALE LA FUNZIONE DI LIKELIHOOD ABBIANO WE

$$L(C) \propto \prod_{i=1}^n \prod_{j=1}^D N(x_{i,j}|C; \mu_{c_j}, \sigma_{c_j}^2)$$

$$L(C) = \sum_{c=1}^n \sum_{i=1}^n \sum_{j=1}^D \log N(x_{i,j}|C; \mu_{c_j}, \sigma_{c_j}^2) =$$

$$\text{CALCOLANDO PFA CONDIZIONATE} = \sum_{c=1}^n \sum_{i=1}^n \sum_{j=1}^D \log N(x_{i,j}|C; \mu_{c_j}, \sigma_{c_j}^2)$$



NON È TUTTO, POSSO ESSERE
CONSIDERATO UNO STILE DI SCRITTURA

$\lambda_i(\{x_i^{(1)}\})/n$
 $\log \left(\frac{\lambda_i(\{x_i^{(1)}\})}{\lambda_i(\{x_i^{(2)}\})} \right)$
 $\log \left(\frac{\lambda_i(\{x_i^{(1)}\})}{\lambda_i(\{x_i^{(2)}\})} \right)$

PER ACCORDARE I DATI CON IL MODELLO

QUESTA FORMULAZIONE CI MOSTRA CHE POSSANO OTTENERE UN LOG-VEROLOGIANTZA INDEPENDEMENTE PER OGNI UNA COMPOUNTE j E OGNI UNA CLASSE c . LE STIME DI MASSA VENDONIGLIANZA RISULTANTI QUINDI SONO

$$M_{c,j} = \frac{1}{N_c} \sum_{i|c_i=j} x_{i,j} \quad , \quad \sigma_{c,j}^2 = \frac{1}{N_c} \sum_{i|c_i=j} (x_{i,j} - M_{c,j})^2$$

QUESTE SONO LE PROFILI E LE VARIANZE COVARIANZE DI OGNI COMPONENTE, VALUTATE SEPARATAMENTE PER OGNI CLASSE.

OGNI FEATURES CHE COMPOSTE IL PROFILO SONO ASSOCIATE CON UNA RV X , DENTRO CHIUSI \mathbf{x} DO XIL.

SPOSTANDOSI QUINDI SUL GENERALE, NEL PROFILO MULTIVARIATO E' CORE SE STANO CONSIDERANDO IL PROSPETTICO IN CUI Z E' DIAGONALE.

$$f_{XIL}(x_{il}) = \prod_{j=1}^J N(x_{ilj} | M_{c,j}, \sigma_{c,j}^2) = N(\mathbf{x}_{il}, \mathbf{Z}_c) \rightarrow \text{MVN}$$

Dove M_c E' IL VETTORE DEGLI AVEI COMPONENTI E' COMPOSTO Mentre \mathbf{Z}_c E' UNA MATRICE DIAGONALE.

$$\mathbf{M}_c = \begin{bmatrix} M_{1,c} \\ M_{2,c} \\ \vdots \\ M_{J,c} \end{bmatrix} \quad \mathbf{Z}_c = \begin{bmatrix} \sigma_{1,c}^2 & & & \\ & \sigma_{2,c}^2 & & \\ & & \ddots & \\ & & & \sigma_{J,c}^2 \end{bmatrix}$$

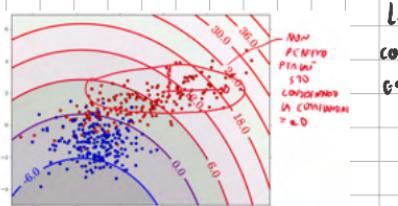
QUESTA EQUIVALENZA NON E' GENERALMENTE VERA PER ALTRE FAMIGLIE DI DISTRIBUZIONI. MA E' SPECIFICA DEL MODELLO GAUSSIANO.

INTERPRETAZIONE GRAFICA

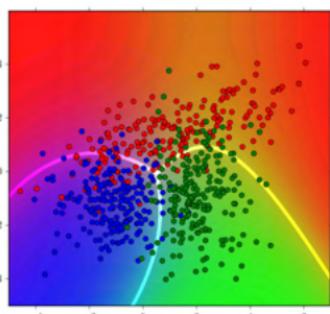
POICHÉ LE MATRICI DI COVARIANZA SONO DIAGONALI, GLI ASSI PRINCIPALI DELLE ELISIOPHI SONO PERFETTAMENTE ALLINEATI CON GLI ASSI COORDINATI DEL NOSTRO SPAZIO DELLE CARATTERISTICHE.



NON ESSENDO COMPLEMENTARI NELL'ESEMPIO NON E' OBBLIGO IN QUANTO PENA CHE I DATI NON SIANO IN RELAZIONE TRA LORO ANCHE SE LO SONO.



LE ELISI DI COMBINAZIONI INVOLTI SAMMANO TUTTE ALLEGATE CON GLI ASI COORDINATI (COSÌ CHIARITO), E I COMBINATI DI DECISIONI, PUR GESSO CHE QUESTE CURVE HANNO VINCOLI SUL LORO ORIENTAMENTO.



TITOLO COVARIANCE

ASSUMONO CHE TUTTE LE CLASSI CONDIVIDANO LA STESSA MATRICE DI COVARIANZA.

$$\mathbf{p}_{\text{MC}} = N(\mathbf{x}| \mu_{\text{MC}}, \Sigma)$$

- OGNI CLASSE HA IL SUO VETTORE DELLE MEDIE MC SPECIFICO
- LA MATELLE DI COVARIANZA Σ È UNA STESSA PER TUTTE LE CLASSI

QUESTA ASSUNZIONE SI DASA SULL'IDEA CHE LE DIVERSE CLASSI DIFFERISCONO SOLO INIZIAMENTE PER LA LORO POSIZIONE NELLO SPAZIO DELLE CARATTERISTICHE (MEDIA) MENTRE LA STRUTTURA DI VARIABILITÀ (COVARIANZA) SIA SIMILE TRA LE CLASSI.

$$\mathbf{x}_{c,i} = \mu_c + \boldsymbol{\varepsilon}_i \quad \boldsymbol{\varepsilon}_i \sim N(0, \Sigma^*)$$

Dove $\boldsymbol{\varepsilon}_i$ rappresenta un rumore gaussiano indipendente dalla classe. In altre parole immagina che ogni osservazione sia generata prendendo la media della sua classe e aggiungendo un rumore con la stessa distribuzione per tutte le classi

Di solito si usa quando abbiamo alta dimensionalità o pochi campioni

VANTAGGI



- INVECE DI STIMARE UNA MATELLE DI COVARIANZA SEPARATA X OGNI CLASSE (N MATELLI) NE STIMANO SOLO UNA RIDUCENDO IL NUMERO DI PARAMETRI DA $N(D \times (D+1))/2$ A $D(D+1)/2$
- POSSONO USARE TUTTI I CAMPIONI DEL DATASET (NON SOLO QUELLI DI UNA SINGOLA CLASSE) PER STIMARE LA MATELLE DI COVARIANZA CONDIVISA. (SEMPRE CHE POSSANO TUTTI STIMA Z IN SESTE FARLO) OBTENENDO UNA STIMA + ROBUSTA
- LE PROBABILITÀ DI OTTENERE UNA MATELLE SINGOLARE O MULTICORRELATA SI RIDUCE DRAMMATICAMENTE

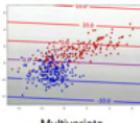
APPLICANDO IL CLASSIFICATORE, NEL FRAMEWORK DELLA MAXIMUM LOG-LIKELIHOOD LE STIME DEI PARAMETRI PER QUESTO APPROCCIO SONO

$$\mu_c^* = \frac{1}{N_c} \sum_{i \in c} \mathbf{x}_i$$

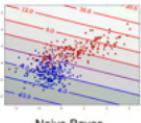
$$\Sigma^* = \frac{1}{N - C} \sum_{i \in c} (\mathbf{x}_i - \mu_c)(\mathbf{x}_i - \mu_c)^T$$

Dove N è il numero totale di campioni $N = \sum_{c=1}^C N_c$

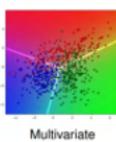
- CORI DELL'APPROCCIO VENGONO STIMATE COME PRIMA, MA LA MATELLE DI COVARIANZA È UNA MATELLA PONDERATA DELLE MATELLI DI COVARIANZA DI CIASUNA CLASSE PONTE PER IL NUMERO DI CAMPIONI IN CIASUNA CLASSE.



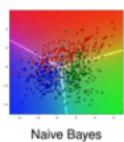
Multivariate



Naive Bayes



Multivariate



Naive Bayes

IN QUESTO CASO, I CONFINI DI DECISIONE DIVENTANO LINEARI.

PER UN PROBLEMA DIVARICO

$$\text{llr}(x) = \log \frac{f_{\text{MC}}(x|\mu_1)}{f_{\text{MC}}(x|\mu_0)} = \log \frac{N(x|\mu_1, \Lambda^{-1})}{N(x|\mu_0, \Lambda^{-1})}$$

ESAMINANDO QUESTA EQUAZIONE POICHE' I TERMINI QUOTIENTICI SI CANCELLANO A VILLENDRA.

$$\text{llr}(x) = x^T b + c$$

Dove

$$b = \Lambda(\mu_1 - \mu_0)$$

$$c = -\frac{1}{2} (\mu_1^T \Lambda \mu_1 - \mu_0^T \Lambda \mu_0)$$

IN UNO SPAZIO BIODIMENSIONALE, IL CONCONE È UNA LINEA RETTA, IN TRE DIMENSIONI UN PIANO, IN DODICI DIMENSIONI UN IPERPIANO.

L'ORIENTAMENTO DI QUESTO CONCONE DIPENDE DA Λ

PONCON IN TUTTE COVARIANZE è LOA

Riconoscono che nell'LOA centrano in direzione in cui massimizza il QUOTIENTE DI RAYLEIGH GENERALIZZATO:

$$\frac{\mathbf{w}^T S_w \mathbf{w}}{\mathbf{w}^T S_b \mathbf{w}}$$

DOVE:

$$S_w = \Lambda^{-1} \text{ è la matrice di dispersione dentro le classi}$$

$$S_b = (\mu_1 - \mu_0)(\mu_1 - \mu_0)^T \text{ è la matrice di dispersione tra le classi}$$

Per risolvere il problema in maniera più "semplice" applichiamo la trasformazione $x' = \Lambda^{-1} x$, questa trasformazione è chiamata WHITENING perché non modifica la varianza in tutte le dimensioni

Nota: se moltiplichiamo i dati $x' = \Lambda x$, le matrici di dispersione diventano $ASwA^T + A^T S_b A$

$$S'_w = \Lambda^T S_w \Lambda = \Lambda^T \Lambda^{-1} = I \quad (\Lambda \text{ è simmetrica})$$

$$S'_b = \Lambda^T S_b \Lambda = \Lambda^T (\mu_1 - \mu_0)(\mu_1 - \mu_0)^T \Lambda^T$$

(risolvendo un problema di classificazione binario)

S'_b è una matrice di rango 1 (con un solo autovettore non nullo) perche' ha una forma speciale: è il prodotto esterno di un vettore con se stesso, $v = \Lambda^T(\mu_1 - \mu_0)$; allora $S'_b = v v^T$, Λ^T è simmetrica.

S'_b ha un autovettore non nullo pari a $\|v\|^2$

il cui corrispondente autovettore è v stesso o un suo multiplo

Gli autovettori sono le soluzioni di $\lambda \mathbf{v} = \mathbf{w}$ (scostano dall'origine) o del massimo normale le nostre soluzioni $\mathbf{w}^\perp = \mathbf{v}$ [normalizzati] è in direzione ottimale nello spazio trasformato.

Proiettando i punti sulla nuova direzione (\mathbf{w}^\perp)
 $\mathbf{w}^\perp \mathbf{x}^\perp = \mathbf{w}^\perp \Lambda^{\frac{1}{2}} \mathbf{x}$

$$\mathbf{w}^\perp = \mathbf{w}^\perp \Lambda^{\frac{1}{2}}$$

$$\mathbf{w} = (\Lambda^{\frac{1}{2}})^T \mathbf{w}^\perp = \Lambda^{\frac{1}{2}} \mathbf{w}^\perp = \frac{\Lambda^{\frac{1}{2}} \mathbf{v}}{\|\mathbf{v}\|} = \frac{\Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}} (\mu_1 - \mu_0)}{\|\mathbf{v}\|} = \frac{\Lambda (\mu_1 - \mu_0)}{\|\mathbf{v}\|}$$

$$\mathbf{w}^\perp \mathbf{x}^\perp = \frac{(\mu_1 - \mu_0)^T \Lambda \mathbf{x}}{\|\mathbf{v}\|} = \frac{\mathbf{w}^\perp \Lambda (\mu_1 - \mu_0)}{\|\mathbf{v}\|}$$

$\|\mathbf{v}\|$ è un fattore di scala che non influenza sulla classificazione

Nella tesa covarianza la $\text{Pr}(x) = \log \frac{f(x|\mu_1)}{f(x|\mu_0)} = \mathbf{x}^T b + c = \mathbf{x}^T \Lambda (\mu_1 - \mu_0) + c$

c è costante, dipende dalle rette e dalla matrice di covarianza in comune.

Non ha significato, il termine b è simile mai ottenuti negli LOA
 emanati finiscono nel proiettare i dati sulla stessa funzione $\Lambda(\mu_1 - \mu_0)$
 Inoltre emanati i risultati si basano sull'assunzione che tutte le classi abbiano stessa
 matrice Σ .

Considerazioni finali

Il multivariante Gaussian Model, è anche noto come QDA, perché i confronti di classificazione sono quantitativi.

Le famiglie di classificatori gaussiani definiti sono quindi:

- QDA: (modelli gaussiani completi)
- LOA: (cov covarianza tesa) modellato con dati limitati
- Naive Bayes: modelli semplici ideali per dati con alta dimensionalità

Le slide finali presentano importanti considerazioni pratiche sull'uso dei classificatori gaussiani:

Riduzione della dimensionalità: Tecniche come PCA possono semplificare la stima dei parametri riducendo la dimensionalità. Inoltre, PCA può eliminare dimensioni con varianza molto piccola (ad esempio, pixel che sono sempre bianchi in tutte le immagini del dataset MNIST).

Scelta del modello:

I modelli multivariati completi funzionano meglio se abbiamo dati sufficienti per stimare affidabilmente le matrici di covarianza

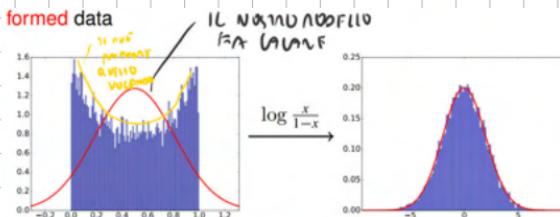
Naive Bayes semplifica la stima, ma può funzionare male se i dati sono fortemente correlati

I modelli con covarianza condivisa possono catturare correlazioni, ma potrebbero funzionare male quando le classi hanno distribuzioni molto diverse

Trasformazione dei dati: Se un modello gaussiano non è adeguato per i nostri dati, possiamo:

Utilizzare una distribuzione diversa più appropriata

Applicare trasformazioni ai dati (come il logaritmo o la trasformazione logit) per renderli più simili a una distribuzione gaussiana



Risultati empirici: I risultati sul dataset MNIST mostrano che:

Il modello gaussiano completo (senza vincoli) ottiene i migliori risultati, suggerendo che le classi hanno matrici di covarianza significativamente diverse

L'assunzione Naive Bayes non è adeguata in questo caso, a causa delle forti correlazioni entro le classi

PCA aiuta a ridurre la complessità: passando da 100 a 50 dimensioni si ottiene un guadagno significativo, ma riducendo troppo (a 9 dimensioni) i risultati peggiorano

Il modello con covarianza condivisa combinato con LDA raggiunge le stesse prestazioni del modello

MNIST — Error rates for Gaussian classifier

Classifier	PCA (100)	PCA (50)	PCA (9)	PCA + LDA (100 → 9)
Naive Tied Gaussian	13.7%	14.4%	25.0%	12.3%
Tied Gaussian	12.3%	12.6%	23.7%	12.3%
Naive Gaussian	12.2%	12.3%	23.4%	11.4%
Gaussian	4.3%	3.6%	12.2%	10.2%

È IL VALORE PMLINE LO SPETTRO SU CUI VENGONO PROLIFERATI I DATI USANDO LOA, NELL'UNIVERSO TUTTE LE INFORMAZIONI NECESSARIE AL TIPO GAUSSIAN (CHE PROFUMA I DATI SULLO STESSO CAPITO)

