MACHINE LEARNING

APPUNTI A CURA DI: RICCARDO LO IACONO

Università degli studi di Palermo

a.a. 2023-2024

Indice.

1 Classificatori Bayesiani **1**
1.1 Superfici decisionali 1
1.2 Stima della densità di probabilità 1
1.3 Classificatori naive 2
1.4 Reti Bayesiane 2

2 Classificatori context-dependent **3**
2.1 Classificatori Bayesiani context-dependent 3

– 1 – Classificatori Bayesiani.

Sia $X = (x_1, \dots, x_n)^T$ un vettore di features. Siano $\omega = (\omega_1, \dots, \omega_m)$ classi distinte e sia $\Pr(\omega_i \mid X)$ la probabilità che ω_i sia la classe di appartenenza di X. Quel che si fa con i classificatori bayesiani, è massimizzare tale probabilità. Per far ciò si definisce la funzione di rischio

$$r = r_1 \Pr(\omega_1) + \cdots + r_n \Pr(\omega_n)$$

ove

$$r_i = \sum_{j=1}^{m} \lambda_{ij} \int_{R_j} \Pr(X \mid \omega_i) \mathrm{d}X$$

con R_j j-esima *superficie decisionale* (si veda la sezione a seguire), λ_{ij} penalità per aver assegnato X a ω_i quando la classe corretta è ω_j.

– 1.1 – Superfici decisionali.

Si parta dal considerare il caso bidimensionale. Sia $X = (x_1, \dots, x_n)^T$ un vettore di features, e siano ω_1, ω_2 le due possibili classi. Allora se rappresentato X su di un piano, è possibile identificare due regioni, siano queste R_1, R_2, tali che

$$X \in \begin{cases} R_1 \iff \Pr(\omega_1 \mid X) > \Pr(\omega_2 \mid X), \\ R_2 \iff \Pr(\omega_2 \mid X) > \Pr(\omega_1 \mid X). \end{cases}$$

Si definisce superficie decisionale una funzione $g(x)$ tale che $\Pr(\omega_1 \mid X) - \Pr(\omega_2 \mid X) = 0$.

Più in generale, supposto $X = (x_1, \dots, x_n)^T$ un vettore di features e $\omega = (\omega_1, \dots, \omega_m)$ possibili classi, una superficie decisionale è una funzione $g_{ij}(x)$ tale che $\Pr(\omega_i \mid X) - \Pr\left(\omega_j \mid X\right) = 0$, $\forall i, j \in \{1, \dots, m\}, i \neq j$.

– 1.2 – Stima della densità di probabilità.

Noto come calcolare la probabilità per ogni classe, resta il problema di come identificare la distribuzione di probabilità dei dati. Si distinguono in questo contesto due approcci:

- *approccio parametrico:* è nota la forma funzionale dei dati, da cui è facile ricavare la distribuzione di probabilità;
- *approccio non-parametrico:* sono noti i valori di alcune features, si può allora stimare la forma funzionale.

Nello specifico a seguito ci si concentra suglia approcci funzionali, in particolare saranno trattati i criteri di *massima verosimiglianza* e *massima probabilità a posteriori*.

– 1.2.1 – Massima verosimiglianza.

Sia supposto $X = (x_1, \dots, x_n)^T$ un vettore di features, con x_i stocasticamente indipendente da $x_j, \forall i \neq j$. Sia inoltre $\Pr(X)$ nota, unicamente dipendente da un qualche parametro ignoto θ; cioè

$$\Pr(X) = \Pr(X \mid \theta) = \prod_{i=1}^{n} \Pr(x_i \mid \theta)$$

Si definisce $\Pr(X \mid \theta)$ verosimiglianza di θ ad X. Segue banalmente

$$\theta_{ML} = \arg\max_{\theta} \left\{ \prod_{i=1}^{n} \Pr(x_i \mid \theta) \right\}$$

– 1.2.2 – Massima probabilità a posteriori.

Il criterio di massimizza verosimiglianza non sempre è applicabile, si procede in questi casi ad applicare il criterio di massima probabilità a posteriori. Per esso, noto $X = (x_1, \dots, x_n)^T$ vettore di features, si deve calcolare θ_{MAP} tale da massimizzare $\Pr(\theta \mid X)$. Dal *teorema di Bayes* si ha

$$\Pr(\theta \mid X) = \frac{\Pr(\theta)\Pr(X \mid \theta)}{\Pr(X)}$$

da cui segue che

$$\begin{aligned}\theta_{MAP} &= \arg\max_{\theta} \{\Pr(\theta \mid X)\} \\ &= \arg\max_{\theta} \left\{ \frac{\Pr(\theta)\Pr(X \mid \theta)}{\Pr(X)} \right\}\end{aligned}$$

– 1.3 – Classificatori naive.

Siano $X \in \mathbb{R}^n$ un vettore di features, $\omega = (\omega_1, \dots, \omega_m)$, e si supponga di dover stabilire $\Pr(X \mid \omega_i)$, per $i \in \{1, \dots, m\}$. In generale, affinché si possa avere una buona stima della funzione di densita sarebbero necessari n^m punti. Se si assume però che x_i e x_j sono stocasticamente indipendenti per ogni $i, j \in \{1, \dots, n\}, i \neq j$, allora

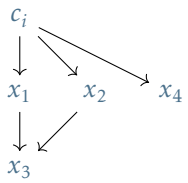
$$\Pr(X \mid \omega_i) = \prod_{j=1}^{n} \Pr\left(x_j \mid \omega_i\right)$$

caso in cui $n \cdot m$ punti risultano sufficienti. Si dimostra che anche nei casi in cui tale indipendenza non sia rispettata, un classificare bayesiano da risultati soddisfacenti.

– 1.4 – Reti Bayesiane.

Come detto in generale i classificatori bayesiani operano bene in molti casi. Vi sono casi però in cui è necessario calcolare le probabilità congiunte in maniera esatta, nascono per tale ragione le reti Bayesiane. Queste procedono come semplici classificatori bayesiani, ma all'occorrenza calcolano opportunamente le probabilità congiunte.

Esempio: sia supposto $X = (x_1, \dots, x_4)$ secondo le relazioni rappresentate dal grafo a seguito riportato.



e siano $c = (c_1, \dots, c_3)$ classi, allora

$$\Pr(c_i \mid X) = \Pr(c_i)\Pr(x_3 \mid x_2 x_1)\Pr(x_4)$$

Più in generale, una rete (o network) bayesiana è un grafo diretto e aciclico i cui nodi rappresentano le variabile.

– 2 – Classificatori context-dependent.

I classificatori discussi sinora, sono utilizzabili in casi in cui i dati sono simultaneamente presenti, ma soprattutto se le classi sono unicamente dipendenti dai valori assunte dalle stesse e da quello delle features. Esistono però scenari in cui tale situazione non si verifica; è pertanto necessario poter definire modelli che apprendono dinamicamente.

– 2.1 – Classificatori Bayesiani context-dependent.

Sia supposto $X = (x_1, \dots, x_n)$ un vettore di features e siano $\omega = (\omega_1, \dots, \omega_m)$ classi. Per quanto detto sinora X è assegnato ad $\omega_i \iff \Pr(\omega_i \mid X) > \Pr\left(\omega_j \mid X\right), \forall i \neq j$. Come detto però, ciò è limitato ai casi in cui vi è una sorta di indipendenza tra le classi. Considerando il caso in cui invece tale indipendenza viene meno, sia

$$\Omega_i = \left\{\omega_{i_j}\right\}_{j \in \{1,\dots,n\}}$$

Da ciò la regola di classificazione Bayesiana può essere riscritta come

$$X \to \Omega_i \iff \Pr(\Omega_i \mid X) > \Pr\left(\Omega_j \mid X\right), \forall i \neq j$$

Ora, affinche il modello possa essere definito context-dependent, è necessario che esso tenga traccia degli stati precedenti del classificatore; per farlo, tra le altre possibilità, vi sono le catene di Markov, per le quali

$$\Pr\left(\omega_{i_k} \mid \omega_{i_{k-1}}, \dots, \omega_{i_1}\right) = \Pr\left(\omega_{i_k} \mid \omega_{i_{k-1}}\right) \tag{1}$$

cioè la dipendenza è ristretta all'ultimo stato della classe.

Definizione: un processo statistico tale da soddisfare l'*Equazione* (1) è detto *processo di Markov*.

– 2.1.1 – Equazioni di Chapman-Kolmogorov.

Vantaggio principale dei modelli basati sulle catene di Markov, è che, attraverso quelle che sono note come *equazioni di Chapman-Kolmogorov*, è possibile determinare lo stato in cui si troverà in futuro (si veda l'esempio a seguire). Nello specifico, partendo dal definire le *probabilità transitorie ad un passo* come

$$p_{ij}(k) = \Pr(X_{k+1} = j \mid X_k = i)$$

ove con X_k si intendo lo stato del classificatore allo stato k, e tali che

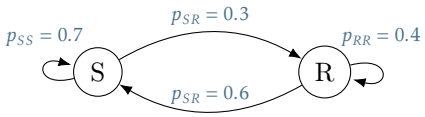
$$\sum_{j=1}^{N} p_{ij}(k) = 1, \forall i, k \in \{1, \dots, M\}.$$

Sfruttando la legge della probabilità totale, l'equazione di Chapman-Kolmogorov permette di definire la probabilità transitoria a n passi, come

$$p_{ij}(k) = \sum_{r=1}^{R} p_{ir}(k, u) p_{rj}(u, k+n), \qquad n \leq u \leq k+n \tag{2}$$

ove $p_{ij}(k, u) = \Pr(X_u = j \mid X_k = i)$.

Esempio: si supponga un modello meteorologico, come quello a seguire.



Si supponga di voler calcolare la probabilità che tra due giorni piova, supposto che oggi vi sia il sole. Dall'*Equazione* (2) segue

$$p_{SR}(d_0, d_2) = p_{SS}(d_0, d_1)p_{SR}(d_1, d_2) + p_{SR}(d_0, d_1)p_{RR}(d_1, d_2) = \cdots = 0.33$$

ove $d_i, i = 0, 1, 2$ indica il numero di giorni da quello attuale.