

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

– 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 $\Pr(\omega_i | X)$ 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

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

ove

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

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

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

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

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 $\Pr(\omega_i | X) - \Pr(\omega_j | X) = 0$, $\forall i, j \in \{1, \dots, m\}, i \neq j$.

– 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 sugli 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

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

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

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

– 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 $\Pr(\theta | X)$. Dal *teorema di Bayes* si ha

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

da cui segue che

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

– 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 $\Pr(X | \omega_i)$, per $i \in \{1, \dots, m\}$. In generale, affinché si possa avere una buona stima della funzione di densità 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

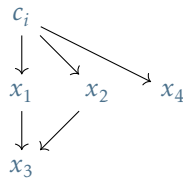
$$\Pr(X | \omega_i) = \prod_{j=1}^n \Pr(x_j | \omega_i)$$

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

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

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