2. REGRESSIONE LOGISTICA

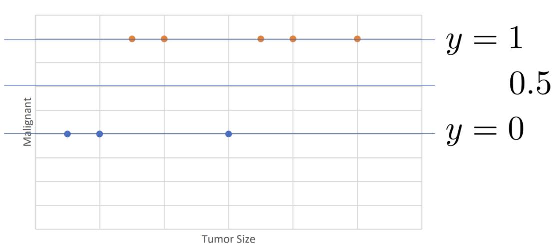
martedì 22 ottobre 2024 18:45

La regressione logistica è un modello statistico utilizzato per predire la probabilità che un'osservazione appartenga a una di due categorie possibili (binaria) sulla base di una o più variabili indipendenti. È particolarmente utile quando la variabile dipendente è dicotomica, ad esempio, "successo" o "fallimento", "sì" o "no", "malato" o "sano", ecc.

La regressione logistica è molto usata in vari campi, come:

- Medicina: Predire la probabilità di una malattia in base ai sintomi e ai fattori di rischio.
- Marketing: Stimare la probabilità che un cliente effettui un acquisto in base al suo comportamento passato.
- Finanza: Valutare il rischio di insolvenza di un cliente in base a dati finanziari.

Classifier threshold



Essenzialmente quando si parla di Regressione Logistica

L'OUTPUT deve essere per forza $0 \le h(x) \le 1$

IPOTESI RAPPRESENTAZIONE
FUNZIONE SIGMOIDE

La funzione ipotetica per la regressione logistica è indicata con

$h_\theta(x) = \theta^T x$ che rappresenta la previsione del modello basato sui parametri θ sui DATI X

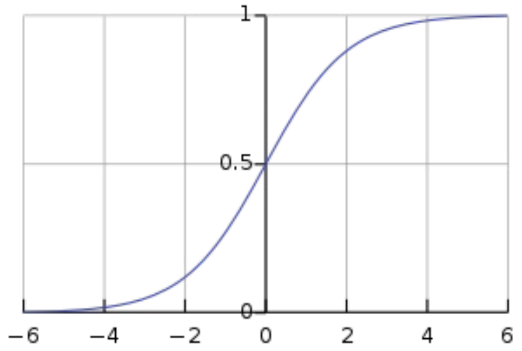
Ricordiamo che $\theta_T X$ E' UNA COMBINAZIONE LINEARE.

$\boxed{\text{FUNZIONE SIGMOIDE}}$

ORA che succede noi sappiamo $h_\theta x \in (-\infty, +\infty)$

Def: $\frac{1}{1+e^{-z}}$ $\quad z = \theta^T x$

$= \frac{1}{1+e^{-\theta^T x}} \rightarrow$ Def $h_\theta(x) = g(\theta^T x)$



ORA FACCIAMO UN ESEMPIO:

$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2) = g(-3 + x_1 + x_2)$

$\theta_0 = -3, \quad \theta_1 = 1, \quad \theta_2 = +1$

$h_\theta(x) \geq 0.5$?

Lo posso rappresentare con la sigmoide

$h_\theta(x) = \frac{1}{1+e^{-g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}} \geq 0.5$

Affinche io possa rappresentare o assegnare 1 o 0

Se $-3 + x_1 + x_2 \geq 0$ ALLORA ASSEGNA 1 a y

$= y = 1$

INTERPRETAZIONE PROBABILISTA SIGMOIDE

Rappresentiamo la stima probabilistica :

$h_\theta(x) = p(y=1 \mid x; \theta)$

Facendo l esempio dei tumori come nelle slide

$h_\theta(x) = 0.7$ tumore benigno

Allora $h_\theta x = 0.3$ tumore maligno

Generalizzo $p(y=0 \mid x;\theta) = 1 - p(y=1 \mid x;\theta)$

La regressione logistica non solo prevede una classe, ma restituisce anche una probabilità associata a quella previsione, rendendo il modello particolarmente utile per problemi in cui l'incertezza è importante, come nella diagnosi medica.

COST FUNCTION | FUNZIONE DI COSTO

L'obiettivo della regressione logistica è trovare i parametri θ che meglio spiegano i dati di addestramento. Questo viene fatto minimizzando una funzione di costo, che misura quanto le previsioni del modello differiscono dai valori effettivi.

Si desidera che la soluzione del problema di minimizzazione della funzione di costo corrisponda a risolvere il problema di massima verosimiglianza, cioè trovare i parametri θ che massimizzano la probabilità (verosimiglianza) dei dati osservati. Poiché l'output della regressione logistica può essere solo 0 o 1 (stiamo affrontando una distribuzione di Bernoulli), possiamo formulare la funzione di verosimiglianza per il modello. Sotto l'assunzione che i campioni siano indipendenti e seguano la stessa distribuzione, possiamo scrivere la funzione di verosimiglianza come segue:

$$L(\theta) = L(\theta; X; y) = p(y \mid X; \theta) = \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta)$$

Allora:

$$\begin{cases} p(y^{(i)}=1 \mid x^{(i)};\theta) = h_\theta(x^{(i)}) \\ p(y^{(i)}=0 \mid x^{(i)};\theta) = 1 - h_\theta(x^{(i)}) \end{cases}$$

Nella forma compatta:

Sfrutta le proprietà delle potenze $e^0 = 1$

[illegible]

$$\boxed{p(y^{(i)} | x^{(i)}; \theta) = h_\theta(x^{(i)})^{y} \cdot (1 - h_\theta(x^{(i)}))^{1-y}}$$

QUINDI se $y^{(i)} = 1$ — ESP

$$p(y^{(i)} | x^{(i)}; \theta) = h_\theta(x^{(i)})^{y^{(i)}} = h_\theta(x^{(i)})$$

E l'altro termine diventa 0

Se $y^{(i)} = 0$ Il termine

$$(1 - h_\theta(x^{(i)}))^{1-y^{(i)}} = 1 - h_\theta(x^{(i)})$$

$$L(\theta) = \prod_{i=1}^{m} h_\theta(x^{(i)})^{y^{(i)}} \cdot (1 - h_\theta(x^{(i)}))^{1-y^{(i)}}$$

COSTO FUNZIONE FORMA LOGARITMICA

RICORDIAMO $\ell(\theta) = \log L(\theta)$ SOST.

$$\ell(\theta) = \log \prod_{i=1}^{m} h_\theta(x^{(i)})^{y^{(i)}} \cdot (1 - h_\theta(x^{(i)}))^{1-y^{(i)}}$$

$$= \sum_{i=1}^{m} \left(y^{(i)} \log h_\theta(x^{(i)}) + (1 + y^{(i)}) \log (1 - h_\theta(x^{(i)})) \right)$$

Questa è chiamata log-verosimiglianza. La motivazione principale per utilizzare il logaritmo è la semplificazione: i prodotti complicati diventano somme, rendendo più semplice la derivazione della funzione di costo.

Quando si lavora con l'apprendimento automatico, in genere si preferisce minimizzare una funzione di costo piuttosto che massimizzare una verosimiglianza.

$$\boxed{J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)})) \right)}$$

FUNZIONE DI ERRORE DI ENTROPIA INCROCIATA

$$\hookrightarrow \hat{\theta} = \arg\min J(\theta)$$

Dobbiamo trovare valori theta che minimizzino la nostra funzione Costo.

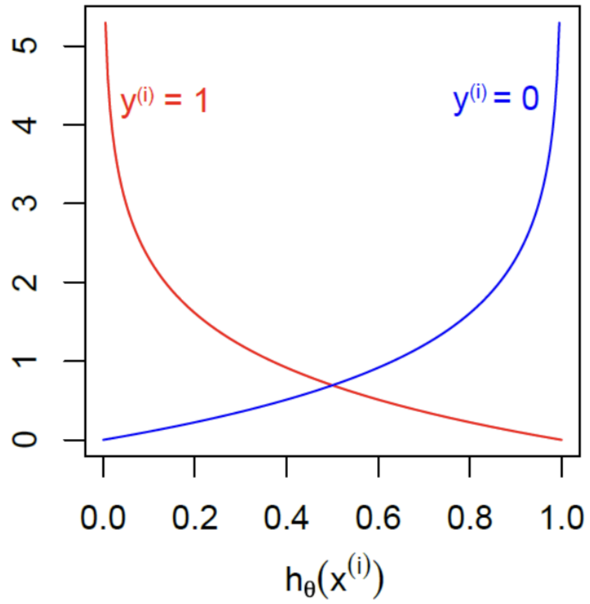
ERRORI | REGRESSIONE LOGISTICA

$$e^{(i)} = -y^{(i)} \log h_\theta(x^{(i)}) - (1-y^{(i)}) \log(1 - h_\theta(x^{(i)}))$$

Se $y^{(i)} = 1$

$\rightarrow e^{(i)} = -\log h_\theta(x^{(i)})$

SE H è vicino allo zero allora ERRORE ALTO
SE H è alto quindi vicino a 1 ERRORE BASSO



Se $y^{(i)} = 0$ $\quad e^{(i)} = -\log(1 - h_\theta(x^{(i)}))$

H è basso cioe 0 allora l'errore è basso
H è alto cioe vicino a 1'errore tende a +inf