ANALISI DEI DATI PER LE IMPRESE INDUSTRIALI (7)

Regressione logistica

In un articolo precedente abbiamo affrontato, fra gli algoritmi supervisionati, la regressione lineare.
Ricordiamo che la regressione lineare è uno degli algoritmi più semplici e viene utilizzata principalmente per risolvere i problemi di regressione, ovvero definire un parametro numerico, come il prezzo di un appartamento.
L'obiettivo della regressione lineare è trovare la retta che meglio si adatta ai dati per prevedere il risultato del problema.

Oggi ci occupiamo della **regressione logistica**.
Anche se si chiamano entrambe regressioni bisogna fare **attenzione a non confonderle**:

- Regressione lineare: la regressione riguarda la **previsione di un risultato numerico** continuo, trovando le correlazioni tra variabili dipendenti e indipendenti.
- Regressione logistica: la classificazione riguarda la **previsione di un'etichetta**, identificando a quale categoria appartiene un oggetto, ad esempio 0 o 1.

Regressione logistica

La regressione logistica viene utilizzata principalmente per risolvere problemi di classificazione binaria, quindi normalmente differenziare due classi come 0-1, funziona-non funziona, caldo-freddo...
La regressione logistica **stima la probabilità di appartenenza** a una determinata classe.
Ad esempio, si determina la probabilità che un'e-mail sia spam o non spam, in base ad un set di dati. Poiché il risultato è la probabilità di un evento, il suo valore è limitato tra 0 e 1.
Nella regressione logistica la somma ponderata dei dati di ingresso viene elaborata tramite una funzione di attivazione chiamata "**sigmoide**", che mappa i valori tra 0 e 1. Il risultato è una curva a forma di S invece di una linea retta, come nella lineare (si veda fig. sotto).
Come il modello di regressione lineare, anche quello di regressione logistica calcola una somma ponderata dei dati di ingresso. Invece di fornire direttamente il risultato, questo algoritmo lo elabora tramite **la funzione logistica, chiamata anche logit**.

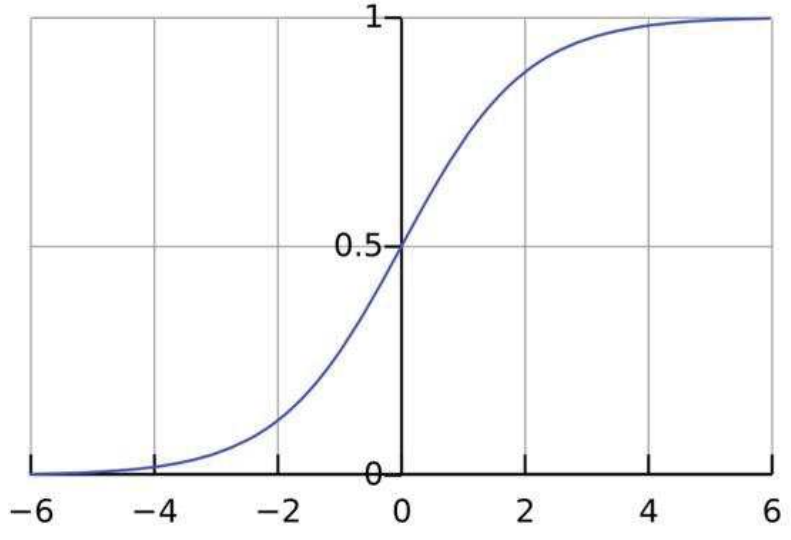
ODDS

I cosiddetti "ODDS", molto comuni nei paesi anglosassoni perché usati nelle scommesse (pensate alle corse dei cavalli), sono il rapporto tra probabilità di successo e probabilità di fallimento.
Se la probabilità di successo è p, quella di fallimento sarà **1-p**, quindi gli odds saranno p/**(1 — p)**.

Si rappresenta la **funzione logistica logit** come **logaritmo naturale degli odds**:

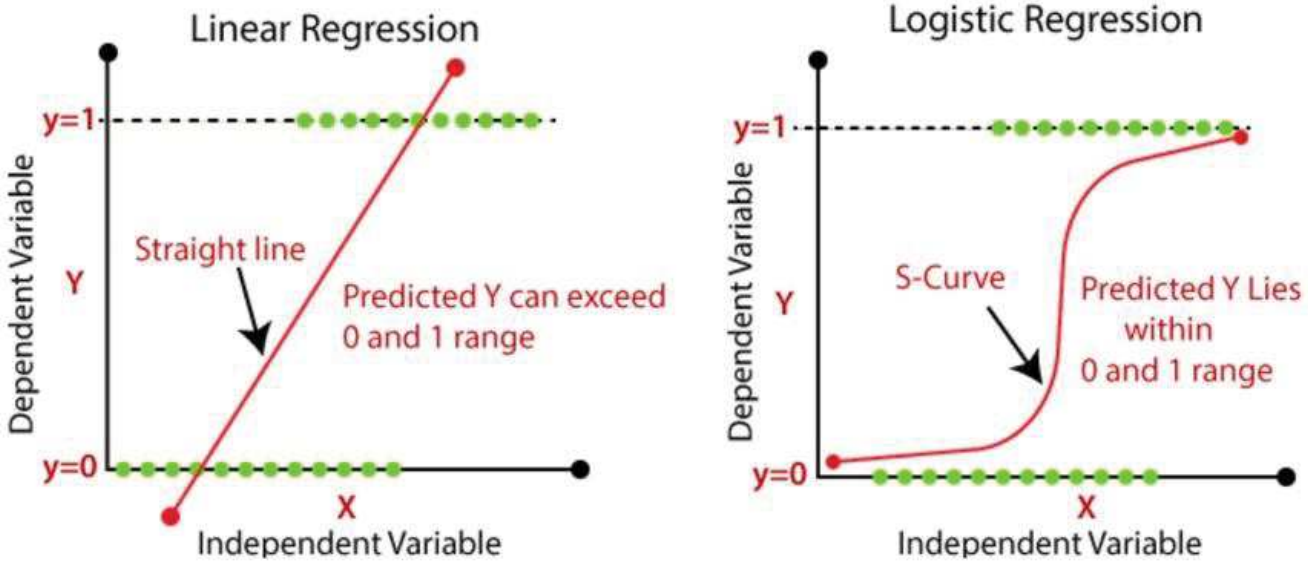
$$\text{Logit Function} = \log\left(\frac{p}{1-p}\right)$$

La funzione logistica per sua natura restituisce solo valori tra 0 e 1 per la variabile dipendente.
Una volta che il modello ha stimato la probabilità di appartenenza ad una classe, può darci facilmente la sua previsione. Se la probabilità stimata è superiore al 50% (o 0,5), il risultato è etichettato come 1, se la probabilità è inferiore al 50%, il risultato è etichettato come 0.
A seconda del risultato atteso possiamo comunque modificare la soglia a vantaggio di una o dell'altra classe. Ad esempio, se prevediamo



una probabilità di pioggia e vogliamo essere sicuri di non bagnarci, spostiamo la soglia a 0,7 invece che a 0,5…
La regressione logistica ha una notevole "resistenza" nei confronti degli "outliers", ovvero i valori estremi che possiamo trovare nei nostri dati. Riesce a gestirli bene, a differenza della regressione lineare che ne viene influenzata perché si modifica la pendenza della retta.
Vediamo le principali differenze fra le due regressioni:

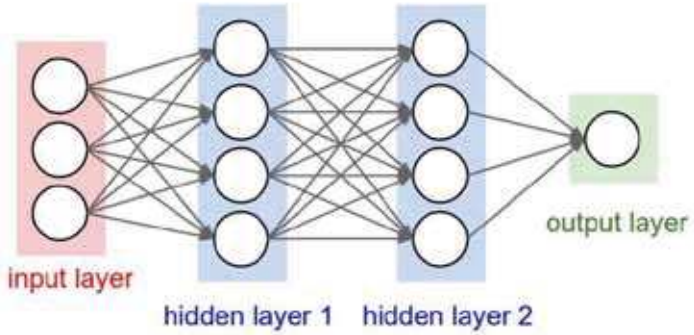
Regressione lineare	Regressione logistica
Utilizzata per prevedere la variabile dipendente numerica continua.	Utilizzata per prevedere la variabile dipendente categorica.
I risultati devono essere un valore continuo, come il prezzo e l'età…	I risultati devono essere valori categorici come 0 o 1, Sì o No…
La relazione tra la variabile dipendente y e la variabile indipendente x deve essere lineare.	Non è necessario che la relazione sia lineare tra le variabili dipendenti e indipendenti.
Usata per risolvere i problemi di regressione.	Usata per risolvere i problemi di classificazione.
Si utilizza la linea retta che si adatta meglio per prevedere i risultati e che si avvicina di più ai punti rappresentativi dei dati.	Utilizziamo la curva S (Sigmoide) per classificare i risultati previsti.
Non è necessario stabilire un valore di soglia.	È richiesto un valore di soglia, normalmente 0,5.
La regressione lineare presuppone la distribuzione statistica normale (gaussiana) della variabile dipendente.	La regressione logistica assume la distribuzione binomiale della variabile dipendente.



Source: javatpoint

La regressione logistica **è molto utilizzata**, vediamo alcuni esempi:

- **Rilevamento delle frodi**: i modelli possono aiutare a identificare anomalie nei dati, che sono segnali di possibili frodi. Grandi organizzazioni hanno iniziato a adottarla anche per eliminare falsi account utente.
- **Previsione delle malattie**: in medicina, questo approccio è molto utilizzato per prevedere la probabilità di patologie o malattie per una specifica popolazione.
- **Previsione di perdita di dipendenti o di clienti**: specifici comportamenti possono essere indicativi di problematiche all'interno di un'organizzazione. Ad esempio le risorse umane potrebbero voler sapere se ci sono dipendenti con alto potenziale che sono a rischio di licenziarsi. In alternativa le vendite potrebbero essere interessate a quali dei clienti sono a rischio di abbandono. Ciò può indurre ad approntare una strategia preventiva di fidelizzazione per evitare perdite.



Un altro aspetto interessante della regressione logistica è che l'algoritmo **viene utilizzato anche nelle reti neurali**, ad esempio nella classificazione, come "funzione di attivazione", per dare una **previsione della classe di appartenenza nell'output layer**, livello che ci definisce il risultato.

Immagini da:
http://towardsdatascience.com
http://www.kdnuggrts.com
http://www.javatpoint.com