PROBABILISTI INTERPRETATION OF LINEAR REGRESSION

sabato 19 ottobre 2024 18:22

L'interpretazione probabilistica di un modello di Machine Learning si riferisce al modo in cui possiamo comprendere e descrivere il comportamento di un modello statistico, come quello lineare, utilizzando i concetti della probabilità.

Quindi abbiamo detto che un modello di regressione lineare cerca di fare previsioni o meglio pronosticare i valori di output.

QUINDI ABBIAMO $x^{(i)}$ INPUT / Features

ABBIAMO $y^{(i)}$ OUTPUT

Utilizzando la funzione ipotesi: $h(x) = \theta^T x^{(i)}$

Ipotesi --> In quanto una fetta di problemi hanno un equazione di questo tipo (Ma puo' anche variare) ↴

$$\hat{y} = \theta^T x + \varepsilon^{(i)}$$

$\varepsilon^{(i)}$ = ERRORE ASSOCIATO ALLA PREVISIONE

PRIMA DI PARLARE DI INTERPRETAZIONE PROBABILISTICA DEGLI ERRORI DI UN MODELLO BISOGNA ACCERTARSI CHE:

- ERRORI INDIPENDENTI
- $\varepsilon^{(i)}$ viene modellato come una variabile casuale stiamo considerando eventi che seguono una certa distribuzione probabilistica.
- Si assume che tutti gli errori seguano una distribuzione normale.
  Media 0 : Questo implica che ci si aspetta che l'errore medio sia zero, il che ha senso in quanto, in media, il modello dovrebbe produrre previsioni corrette.
  - Varianza σ^2: La varianza rappresenta la dispersione degli errori attorno alla media. Se σ è piccola, gli

errori saranno concentrati vicino a 0, mentre una grande varianza significa che gli errori sono più dispersi.

$$p(e^{(i)}) = N(0, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(e^{(i)})^2}{2\sigma^2}}$$

Nella $e^{(i)}$ si può sostituire

$e^{(i)} = (x - \mu)$ μ = Media

Distribuzione di probabilità condizionale

input $x^{(i)}$ è un Vettore θ

Vediamo che cosa è regressione lineare con rumore gaussiano,

$$p(y^{(i)}|\mathbf{x^{(i)}};\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y^{(i)} - \theta^T \mathbf{x^{(i)}})^2}{2\sigma^2}} \qquad i = 1 \dots m.$$

Ricordiamo che $e^{(i)} = y^{(i)} - \theta^T x$

VERROSIMIGLIANZA

Def

La verosimiglianza L(θ) rappresenta la probabilità che, dato il modello θ, i dati osservati (X,y) siano stati generati.

$$L(\theta) = p(y | X; \theta)$$

SOSTITUENDO L'ESPRESSIONE CON AL FORMA GAUSSIANA

OTTENIAMO:

$$L(\theta) = \prod \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\left(\frac{y^{(i)} - \theta^t x^{(i)}}{2\sigma^2}\right)^2}$$

L'obiettivo finale sarà appunto quello di massimizzare la funzione di verosomiglianza per ottenere parametri ottimali.
In altre parole, vogliamo trovare i parametri che rendono più probabile osservare i dati forniti.

$$\hat{\theta} = \arg\max_{\theta} L(\theta) = \arg\max \log L(\theta)$$
$$= \arg\max l(\theta) \quad \text{CON} \quad \log L(\theta) = l(\theta)$$

Poiché la funzione di verosimiglianza può essere complessa da massimizzare direttamente, spesso si lavora con il logaritmo della verosimiglianza, dato che il logaritmo è una funzione monotona. Questo significa che massimizzare il logaritmo della verosimiglianza è equivalente a massimizzare la verosimiglianza stessa.

L'idea è trasformare il prodotto di probabilità condizionate, che rappresentano la distribuzione di probabilità dei dati, in una somma, utilizzando il logaritmo naturale.

ABBIATO $$l(\theta) = \log \prod_{i}^{m} p\left(y^{(i)} \middle| x^{(i)}; \theta\right)$$

Il logaritmo trasforma la produttoria il una sommatoria.
Si sfrutta la proprietà del logaritmo che permette di trasformare il logaritmo di un prodotto in una somma dei logaritmi.

$$= \sum_{i=1}^{m} \log \underbrace{\frac{1}{\sigma\sqrt{2\pi}}}_{A} \underbrace{e^{-\frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\sigma^2}}}_{B}$$

Ora il log viene moltiplicato per A E B

$$= \sum_{i=1}^{m} \left(\log \frac{1}{\sigma\sqrt{2\pi}} + \log e^{-\frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\sigma^2}} \right)$$

Ora che succede possiamo semplificare tale equazione ? COME ?

Praticamente le costanti le porto fuori dalla sommatoria fattorizzando il termine 1 con log

$$m \log \frac{1}{\sigma\sqrt{2\pi}} - \sum_{i=1}^{m} \frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\sigma^2}$$

$$= m \log \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2} \cdot \sum_{i}^{m} \left(y^{(i)} - \theta^T x^{(i)}\right)^2$$

TROVATA TALE EQ. ORA DOBBIAMO MASSIMIZZARE

Dobbiamo massimizzare la funzione di verosimiglianza perché l'obiettivo di molti algoritmi di apprendimento automatico, come la regressione lineare nel contesto di questa dimostrazione, è trovare i parametri ottimali θ
θ che rendano il nostro modello il più "verosimile" possibile rispetto ai dati osservati.

$$\max_{\theta} l(\theta) = \max_{\theta} \left(m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i}^{m} \left(y^{(i)} - \theta^T x^{(i)}\right)^2 \right)$$

PONIAMO $m \log \frac{1}{\sigma\sqrt{2\pi}}$ = COSTANTE

$$= \max \left(-\frac{1}{2\sigma^2} \cdot \sum \left(y^{(i)} - \theta^T x^{(i)}\right)^2 \right)$$

STESSA COSA σ^2 VARIANZA

$$= \max \left(-\frac{1}{2} \sum \left(y^{(i)} - \theta^T x^{(i)}\right)^2 \right)$$

$\max = -\min \iff \min = -\max$

Per tale mot.

$$= \min \left(0.5 \cdot \sum_{i=1}^{m} \left(y^{(i)} - \theta^T x^{(i)}\right)^2 \right)$$

MSE

ORA:

$$\text{MSE} \quad \min_{\theta} \left(\frac{1}{2m} \sum_{i=1}^{m} \left(h\left(x^{(i)}\right) - y^{(i)} \right)^2 \right) \quad \text{MSE}$$

$h\left(x^{(i)}\right)$ = PRONOSTICO

$y\left(x^{(i)}\right)$ = VALORE REALE OSSERVATO

m = IL NUMERO DI CAMPIONI

MINIMI QUADRATI

FORMULAZIONE INUSUALE VEROSIMILMENTE MAL

$$\min_{\theta} \left(\frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \theta^T x^{(i)} \right)^2 \right)$$

$\theta^T x$ = PRONOSTICO LINEARE DEL MODELLO

$y^{(i)}$ = Rappresenta il valore osservato.

Problema identico: Come detto nella slide, i due problemi sono identici a livello di ottimizzazione rispetto a θ. La costante $\frac{1}{m}$ o $\frac{1}{2m}$ non influisce sul risultato finale, poiché non cambia la posizione del minimo della funzione da ottimizzare. L'unica differenza è che una delle formulazioni normalizza l'errore rispetto al numero di campioni.

Costante di regolarizzazione: La frase "These problems are identical but a regularization constant that has no influence in the minimization problem" si riferisce al fatto che questo termine costante (ad esempio il fattore m) non influisce sulla posizione del minimo della funzione. Influisce solo sul valore numerico della funzione obiettivo, ma non cambia il valore di θ che minimizza l'errore.

IMPORTANTE, minimizzare il problema dei minimi quadrati nella regressione lineare (come visto a destra) è equivalente a minimizzare l'errore quadratico medio (MSE) (come visto a sinistra). La differenza nei fattori costanti non influisce sull'ottimizzazione rispetto a
θ, perciò possiamo usare indifferentemente una o l'altra formulazione per trovare i migliori parametri del modello che riducono l'errore di previsione.