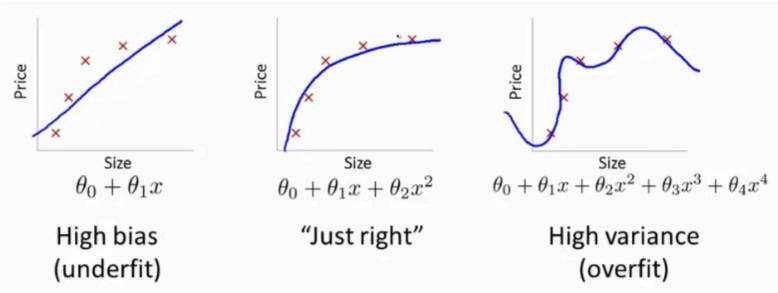
3. FITTING -- adattamento

giovedì 24 ottobre 2024 19:00

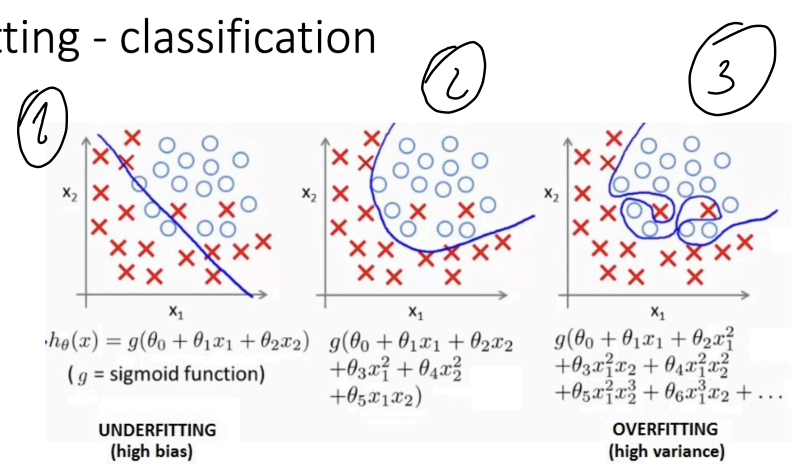
il fitting è strettamente legato al concetto di bias e variance, due elementi cruciali per valutare le performance di un modello e trovare un equilibrio tra l'accuratezza e la generalizzazione del modello.

Fitting - regression



- Underfitting (sottostima): accade quando il modello è troppo semplice per catturare la complessità dei dati. Il modello non è in grado di adattarsi adeguatamente ai dati di addestramento e ha scarso rendimento sia sui dati di addestramento sia sui dati di test. L'underfitting è spesso associato a un elevato bias.
- Overfitting (sovrastima): accade quando il modello è eccessivamente complesso e si adatta perfettamente ai dati di addestramento, anche ai dettagli e al rumore. Questo causa una scarsa generalizzazione ai nuovi dati e si traduce in scarsa performance sui dati di test. L'overfitting è associato a un'alta varianza.

Fitting - classification



1. La linea di separazione è rappresentata da una semplice funzione lineare. Questo modello è troppo semplice e non riesce a separare efficacemente le due classi.

• La scelta di una funzione lineare causa un elevato bias, cioè una distorsione sistematica che impedisce al modello di adattarsi ai dati di addestramento.

2. Buon fitting (equilibrio tra bias e varianza) - Grafico centrale

• La linea di separazione è curvata, utilizzando una funzione polinomiale di grado intermedio. Questa forma permette al modello di adattarsi meglio ai dati di addestramento senza però inseguire dettagli superflui.
• Questo modello riesce a rappresentare correttamente la distribuzione delle due classi con un buon equilibrio tra bias e varianza.

3. Overfitting (alta varianza) - Terzo grafico a destra

• La linea di separazione è estremamente complessa e segue perfettamente i dati di addestramento, includendo anche il rumore (punti isolati che non rappresentano una vera tendenza).
• Questo tipo di modello è troppo specifico e ha un'elevata varianza, cioè tende a variare molto se cambiano i dati di addestramento, riducendo la sua capacità di generalizzazione.

OVERFITTING AND UNDERFITTING

In machine learning, l'obiettivo è creare un modello che sia in grado di generalizzare bene, cioè che possa applicare ciò che ha appreso dai dati di addestramento per fare previsioni accurate su nuovi dati, mai visti prima. La capacità di generalizzare è fondamentale per rendere un modello efficace e utilizzabile in situazioni reali.

Ecco una spiegazione dei concetti chiave:

1. Generalizzazione

• Definizione: La generalizzazione è la capacità di un modello di funzionare bene su dati nuovi, mai visti prima, dopo aver appreso da un set di dati di addestramento.
• Obiettivo: In machine learning, si cerca di sviluppare modelli che possano imparare i pattern o le tendenze di fondo dai dati di addestramento, in modo da applicare

questi concetti a nuovi punti dati. L'obiettivo non è semplicemente "memorizzare" i dati di addestramento, ma imparare dai trend per fare previsioni accurate anche su dati futuri.

2. Underfitting

• Definizione: L'underfitting si verifica quando il modello è troppo semplice per catturare la complessità dei dati di addestramento.
• Conseguenze: Un modello in underfitting ha difficoltà a imparare correttamente dai dati e, di conseguenza, ottiene pessimi risultati sia sui dati di addestramento sia sui dati di test.
• Motivo: Di solito, ciò accade perché il modello ha una struttura troppo rigida o lineare, che non riesce a rappresentare le variazioni e le relazioni presenti nei dati.

3. Overfitting

• Definizione: L'overfitting si verifica quando il modello è troppo complesso e si adatta perfettamente ai dati di addestramento, inclusi anche i dettagli e il rumore.
• Conseguenze: Un modello in overfitting funziona bene sui dati di addestramento, ma ha difficoltà a generalizzare sui nuovi dati, risultando in performance scarse sui dati di test.
• Motivo: In questo caso, il modello "memorizza" i dati di addestramento piuttosto che "imparare" dai trend. Ciò porta a una scarsa capacità di adattarsi a nuovi esempi, poiché il modello ha appreso dettagli specifici che potrebbero non essere rappresentativi di nuovi dati.

BIAS AND VARIANCE

Il bias (o distorsione) rappresenta la deviazione sistematica dell'errore del modello. Indica quanto il modello si discosta, in media, dal valore reale o dalla funzione "vera" che vorremmo approssimare.

$$Bias\left(h(x), f(x)\right) = \left(E[h(x)] - f(x)\right)^2$$

$h(x)$ = PREVISIONE DEL MODELLO

$f(x)$ = FUNZIONE REALE DEL MODELLO

VARIANCE | VARIANZA

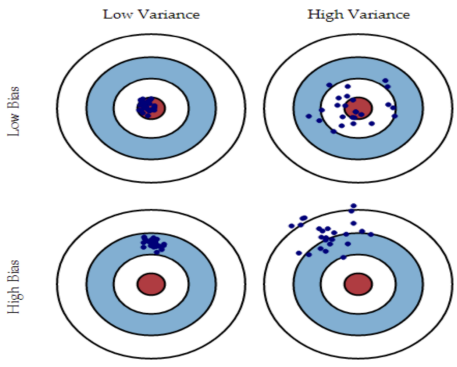
La variance rappresenta la media delle variazioni dei valori predetti dal modello rispetto alla loro media. Indica quanto le previsioni del modello fluttuano quando si addestra il modello su dati diversi.

$$Var(h(x)) = E\left(h(x) - E[h(x)]\right)^2$$

$E[h(x)] =$ è il valore medio delle previsioni del modello.

Interpretazione: Se la variance è elevata, significa che il modello è troppo complesso e sensibile ai dettagli del dataset di addestramento, portando a overfitting. Il modello si adatta troppo bene ai dati di addestramento e perde la capacità di generalizzare su nuovi dati.

Un buon modello ha bisogno di bilanciare bias e variance. Se il bias è troppo alto, il modello non sarà in grado di catturare la struttura dei dati (underfitting). Se la variance è troppo alta, il modello si adatta troppo al set di dati di addestramento e non riesce a generalizzare bene (overfitting).

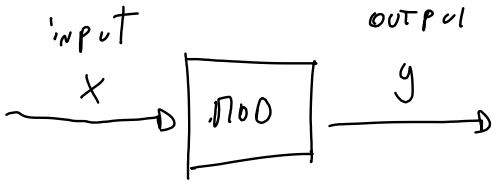


ERRORE ATTESO

Diamo la def di valore atteso.

$$\text{VALORE ATTESO} = E[x] = \sum_{m} p_i \cdot x_i$$

Pi = probabilita associate a x $i:1$



$y = x + \varepsilon$ ε = Noise

Noi abbiamo H(x) la funzione che vogliamo approssimare a f(x).

Def

L'errore quadratico atteso, misura quanto, i valori, in media i valori predetti da H(X) si discostano da i valori reali di y

$$Err(f(x)) = E[(y - h(x))^2]$$

Bias and Variance trade-off (cont.d)

Obiettivo: Stimare l'aspettativa dell'errore rispetto a tutti i possibili dataset D
Dettaglio: Vogliamo calcolare il Mean Squared Error (MSE) per ogni ipotesi e poi prendere la media complessiva.

Problema: Calcolare questa media su tutti i dataset infiniti non è fattibile.

Soluzione: Calcoliamo il valore atteso del MSE su tutti i dataset

Generalization Error (GER): L'aspettativa così calcolata si chiama Generalization Error (GER) e rappresenta l'errore medio atteso su tutti i dataset:

$$GER = E_D[MSE]$$

$h^{(D_i)}(x)$ = Funzione ipotesi del dataset D

N.B $y^{(i)} = f(x^{(i)}) + \varepsilon^{(i)}$

$$MSE\left(h^{(D;)}(x)\right) = E_x\left[h^{(D;)}(x) - f(x)\right]^2$$

.

MSE = rappresenta la singola ipotesi rispetto la funzione obiettivo. L'MSE cattura la differenza media tra i valori previsti dal modello e i valori reali, elevata al quadrato per dare più peso agli errori maggiori e per evitare cancellazioni tra errori positivi e negativi.

GENERALIZZANDO :

$$GER = E_D\left[MSE\right]$$

Tale eq. rappresenta l'aspettativa del Mean Squared Error (MSE) rispetto a tutti i possibili dataset.

ANDANDO A SOSTITUIRE!

$$GER = E_D\left[E_x\left[\left(h^{(D)}(x^{(i)}) - f(x^{(i)})\right)^2\right]\right]$$

$$= E_x\left[E_D\left[\left(h^{(D)}(x^{(i)}) - f(x^{(i)})\right)^2\right]\right]$$

SFRUTTANDO LA PROPRIETA' DI LINEARITA' DELL'ASPETTATIVA.

$$E\left[E\left[f(x,y) \mid x\right]\right] = E\left[f(x,y)\right]$$

$$= E_x\left[E_D\left[f(x,D)\right]\right] = E_D\left[E_x\left[f(x,D)\right]\right]$$

In altre parole, se calcoliamo l'aspettativa condizionata rispetto a una variabile (ad esempio possiamo successivamente calcolare l'aspettativa rispetto a tutte le variabili. Questa proprietà permette di cambiare l'ordine dell'aspettativa.

$$+ E_D\left[\left(h^{(D)}(x^{(i)}) - f(x^{(i)})\right)^2\right]$$

Si parte da qui , poi si definisce una nuova quantita che vada a rappresentare la stima migliore di $f(x^{(i)})$

$$\bar{h}(x^{(i)}) = E_D\left[h^{(D)} \mid x^{(i)}\right]$$

Si aggiunge e si sotrai h tagliato all'interno dell'equazione stessa

$$E_D\left[\left(h^{(D)}(x^{(i)}) - \bar{h}(x^{(i)})\right) + \bar{h}(x^{(i)}) - f(x^{(i)})\right)^2]$$

e B

Attraverso la proprietà del completamento del quadrato.

$$(e+b)^2 = a^2 + b^2 + 2ab$$

$$E_D\left[\left(h^{(D)}(x^{(i)}) - \bar{h}(x^{(i)})\right)^2\right] + E_D\left[\left(\bar{h}(x^{(i)}) - f(x^{(i)})\right)^2\right] + E_D\left[2\left(h^{(D)}(x^{(i)}) - \bar{h}(x^{(i)})\right)\left(\bar{h}(x^{(i)}) - f(x^{(i)})\right)\right]$$

e/1 b/2 2ab/3

Ora guardando attentamente il primo termine (A)
È uguale alla definizione di varianza

$$E_D\left[\left(h^{(D)}(x^{(i)}) - \bar{h}(x^{(i)})\right)^2\right] = Var\left(h^{(D)} \mid x^{(i)}\right)$$

(1)

Il secondo termine rappresenta il BIAS al quadrato

(2) $$E_D\left[\left(\bar{h}(x^{(i)}) - f(x^{(i)})\right)^2\right] = \left(\bar{h}(x^{(i)}) - f(x^{(i)})\right)^2$$

(3) $$2\left(\bar{h}(x^{(i)}) - f(x^{(i)})\right) E_D\left[h^{(D)}(x^{(i)}) - \bar{h}(x^{(i)})\right]$$

Costanti porto fuori

$$= 2\left(\bar{h}(x^{(i)}) - f(x^{(i)})\right)\left(E_D\left[h^{(D)} \mid x^{(i)}\right] - E_D\left[\bar{h}(x^{(i)})\right]\right)$$

$= 0$

rappresenta la deviazione della previsione del modello dalla media delle previsioni. Prendendo l'aspettativa rispetto ai dati D, otteniamo zero, perché la media della deviazione dalla media è, per definizione, zero.

$$2\left[\left(\bar{h}(x^{(i)}) - h(x^{(i)})\right) \times 0 = 0\right]$$

QUINDI ABBIAMO DIMOSTRATO CHE

$$MSE\left(h^{(D)}(x^{(i)})\right) = Bias^2\left(h^{(D)}(x^{(i)})\right) + VAR\left(h^{(D)}(x^{(i)})\right)$$

Il concetto di trade-off tra bias e varianza è fondamentale per comprendere come ridurre l'errore di generalizzazione di un modello di machine learning, ovvero l'errore che il modello commette su dati nuovi, non visti durante l'addestramento.

Per ridurre l'errore di generalizzazione, dobbiamo ridurre sia il bias che la varianza:

Bias: rappresenta quanto una stima media del modello si avvicina alla verità (valore reale). Un alto bias indica che il modello è troppo semplice e non riesce a catturare la complessità del fenomeno, indipendentemente dal dataset utilizzato per l'addestramento. In pratica, un modello con bias elevato fa molti errori anche sui dati di addestramento, il che è tipico dei modelli sottodimensionati (underfitting).

Varianza: rappresenta quanto le ipotesi (cioè le stime) del modello cambiano a seconda del dataset di addestramento. Un'alta varianza significa che il modello, se addestrato su diversi dataset, genera ipotesi molto diverse tra loro. Questo succede perché il modello è troppo complesso: si adatta in modo eccessivo ai dati di addestramento (overfitting), diventando poco capace di generalizzare su nuovi dati.

L'ideale sarebbe mantenere bias e varianza entrambi bassi, in modo che il modello sia sia accurato sia stabile. Tuttavia, c'è un problema: bias e varianza tendono a comportarsi in modo opposto rispetto alla complessità del modello:

Aumentando la complessità del modello, il bias tende a ridursi (perché il modello diventa più capace di adattarsi ai dati), ma la varianza aumenta (perché il modello diventa più sensibile alle specificità del dataset di addestramento). Diminuiendo la complessità del modello, la varianza si riduce (il modello è meno influenzato dai dettagli specifici del dataset), ma il bias aumenta (perché il modello non riesce a catturare la complessità del fenomeno).

Queste due immagini illustrano il concetto di trade-off tra bias e varianza e l'effetto della complessità del modello sull'errore di predizione sia sui dati di addestramento (training set) sia su nuovi dati (test set). Vediamo le due immagini nel dettaglio:

Prima Immagine: Bias e Varianza rispetto alla Complessità del Modello
Asse orizzontale: rappresenta la complessità del modello. I modelli a bassa complessità (a sinistra) sono semplici e spesso troppo rigidi, mentre quelli ad alta complessità (a destra) sono più flessibili e capaci di adattarsi ai dati.

Asse verticale: rappresenta l'errore.
Curva rossa (Bias²): quando la complessità è bassa, l'errore di bias è elevato. Questo accade perché il modello è troppo semplice e non riesce a catturare bene il comportamento dei dati (fenomeno di underfitting). Man mano che la complessità aumenta, il bias si riduce, poiché il modello riesce a catturare meglio la struttura dei dati.
Curva blu (Varianza): con l'aumentare della complessità, la varianza cresce. Ciò accade perché un modello molto complesso si adatta eccessivamente ai dati di addestramento, diventando sensibile a variazioni nel dataset. Questo porta al fenomeno di overfitting, in cui il modello ha un'elevata variabilità e non generalizza bene su nuovi dati.
Curva nera (Errore Totale): rappresenta la somma tra bias e varianza. La forma a "U" invertita indica che c'è un punto ottimale di complessità del modello, dove l'errore totale è minimo. Questo punto rappresenta il miglior compromesso tra bias e varianza (il "trade-off ottimale").
Conclusione della prima immagine: La complessità ottimale del modello è quella che minimizza la somma del bias e della varianza, portando all'errore totale minimo.

Seconda Immagine: Training Set e Test Set rispetto alla Complessità del Modello
Asse orizzontale: rappresenta, di nuovo, la complessità del modello.
Asse verticale: rappresenta l'errore di predizione.
Curva azzurra (Training Sample): mostra l'errore sui dati di addestramento. Con l'aumentare della complessità, il modello riesce a fittare sempre meglio i dati di addestramento, quindi l'errore diminuisce.
Curva rossa (Test Sample): rappresenta l'errore su un nuovo campione di dati, non visto durante l'addestramento. Con complessità bassa, il modello ha un alto bias e quindi un errore elevato. Con l'aumentare della complessità, l'errore su nuovi dati prima diminuisce, ma poi inizia a crescere nuovamente quando il modello diventa troppo complesso e inizia a sovradattarsi ai dati di addestramento.
Zone della seconda immagine:

Alta varianza e basso bias: a destra, con modelli molto complessi, vediamo che l'errore sul test set cresce, segnalando overfitting.

Alto bias e bassa varianza: a sinistra, dove il modello è troppo semplice, vediamo un errore alto su entrambi i set (underfitting).
Conclusione della seconda immagine: L'obiettivo è scegliere una complessità del modello tale che l'errore sul test set sia minimizzato. Questo punto ottimale si trova dove la curva rossa raggiunge il valore minimo, rappresentando un buon compromesso tra adattamento ai dati di addestramento e capacità di generalizzazione.