ANALISI DEI DATI PER LE IMPRESE INDUSTRIALI (3)
La preparazione dei dati: Overfitting e Underfitting, Bias e Varianza

Continuiamo l'argomento della preparazione dei dati, concentrandoci sulla trasformazione dei valori qualitativi e su quattro parole chiave: **Overfitting e Underfitting, Bias e Varianza**.

Abbiamo già visto l'analisi delle distribuzioni, **la necessità di ricondurre i dati stessi ad una unità di misura comune** (la standardizzazione o la normalizzazione), **riconoscere i mancanti** ed eventualmente sostituirli. A volte **occorre eliminare** i valori esterni alle nostre osservazioni.

Feature Engineering

Diversi algoritmi non sono in grado di trattare dati qualitativi (Sì-No, caldo-freddo, …) **insieme ai dati numerici**, dobbiamo quindi ricorrere a questa tecnica chiamata **Feature Engineering**.

Tramite questo **processo il set di dati viene elaborato per convertire i vari tipi** (categorici, letterali: 'stringhe', data e ora, …) **in valori numerici comprensibili per un algoritmo**. Una volta che i vari dati di testo sono stati convertiti, sono pronti per essere inseriti nel modello.

Semplificare il modello

Come accennato precedentemente, **il numero delle osservazioni (righe) deve superare il numero di caratteristiche (colonne) per produrre un modello affidabile. È comunque meglio ridurre per quanto possibile il numero delle caratteristiche stesse** per non appesantire troppo il calcolo. Riprendendo il nostro foglio dati ridotto a poche righe, riportato nell'immagine qui sotto, vediamo che ogni colonna rappresenta una caratteristica, ad esempio un sensore, mentre ogni riga rappresenta una macchina.

Per spiegare perché ci conviene ridurre il numero delle caratteristiche, dobbiamo capire come in realtà funziona il sistema.

I dati nel nostro ipotetico Excel **vengono trasferiti in matrici con un corrispondente numero di righe e colonne.** Gli algoritmi, secondo certe sequenze di calcolo, devono prendere in esame l'interazione di tutti dati presenti. L'ultima colonna "Failure" è l'obiettivo della nostra ricerca.

Machine ID	Sensor_1	Sensor_2	Sensor_3	Sensor_4	Sensor_5	Sensor_6	Sensor_7	Sensor_8	Sensor_9	Sensor_10	Sensor_11	Failure
M_0001	2633	918	4229	13792	23403	13764	6054	1642	721	1659	1666	no
M_0002	9244	22732	15307	8553	5707	2999	1428	1361	1429	810	535	yes
M_0003	3183	28526	8735	4898	3685	2960	2398	1938	1627	1471	1385	no
M_0004	2785	2642	5857	7417	24343	16183	4876	2390	2478	2180	1408	yes

Prima analizziamo i dati con la colonna "Failure" presente, poi proviamo senza "Failure" e confrontiamo i risultati Yes e No con le due iterazioni.

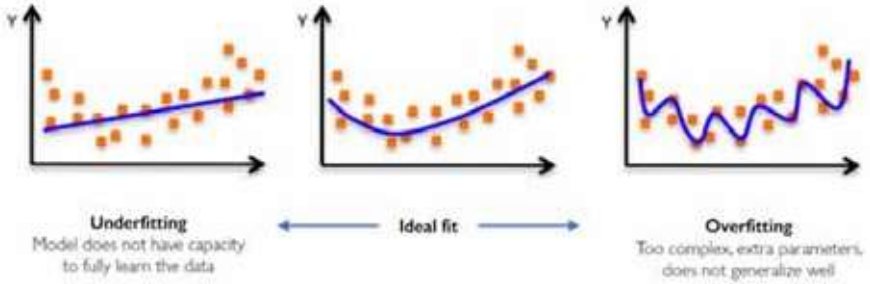
Infine sottoponiamo i nuovi dati per ottenere il risultato atteso.

Matematicamente **se abbiamo "n" caratteristiche, l'algoritmo normalmente cerca una soluzione in "n-1" dimensioni**. Se avessimo soltanto 3 sensori, quindi 3 dimensioni, nel caso di una correlazione lineare si cercherebbe un piano sul quale stesse il maggior numero di punti, ovvero di soluzioni. In questo caso cercheremo un piano perché ha due dimensioni: 3-1 = 2.

È chiaro che se abbiamo 11 parametri, il computer deve trovare soluzioni in un "iperpiano" di grado 10, quindi la cosa si complica parecchio. *(NdA: ho semplificato molto la spiegazione, mi scusino i lettori più rigorosi).*

Overfitting e Underfitting

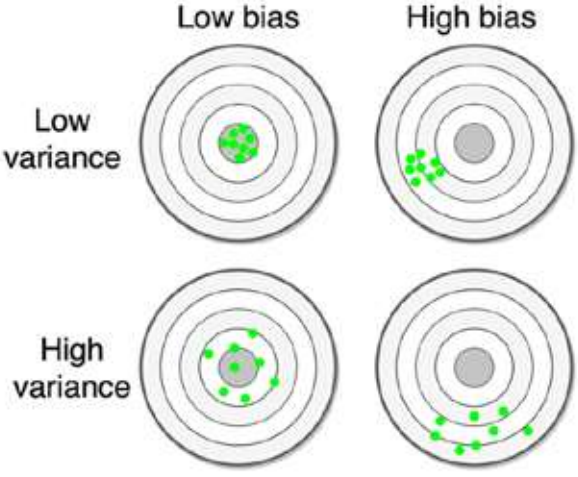
La maggior parte dei modelli nell'apprendimento automatico



dovrebbe **soddisfare due importanti requisiti**: primo, i modelli dovrebbero **rilevare tutte le caratteristiche nascoste nei dati di training** e secondo, dovrebbero **funzionare bene con dati nuovi.**

Un problema che si può verificare con molti parametri è che **il modello si adatti "troppo" bene e sia "troppo" preciso**. Quando gli verranno sottoposti nuovi dati per la classificazione o la regressione, i risultati saranno inferiori alle aspettative perché non riuscirà a generalizzare. **Questo fenomeno si chiama *overfitting*, ovviamente speculare all'*underfitting***, che limita la precisione perché non riesce a creare un modello.

Bias e Varianza
Il bias (scostamento) rappresenta quanto la media di una statistica si discosta dal valore vero.
Usiamo la metafora del gioco delle freccette nel quale, naturalmente, si mira al centro (si veda l'immagine sottostante). **In situazioni di alto bias, abbiamo un giocatore con una pessima vista che lancia sistematicamente tutte le sue freccette in una zona non centrata.** Non è necessariamente un cattivo giocatore, poiché colpisce il bersaglio ma ha bisogno degli occhiali.
La varianza ti dice di quanto i campioni si disperdono dal valor medio o come i dati si distribuiscono intorno al centro.
Tornando al gioco delle freccette**,** un giocatore con alta varianza e basso bias, colpisce parzialmente il suo obiettivo.
I giocatori ad alta varianza e alto bias sono terribili e le loro freccette sono dappertutto.

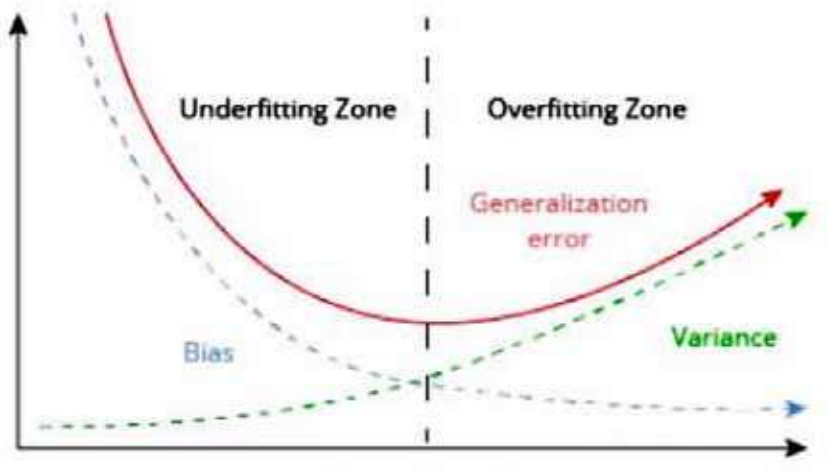


Il compromesso bias-varianza è una proprietà specifica di tutti i modelli di machine learning (supervisionati), che impone un compromesso tra la flessibilità del modello e il comportamento su dati che non ha mai visto.
L'obiettivo finale del machine learning è scegliere un modello che abbia contemporaneamente una bassa varianza e un basso bias.

Riassumendo, **il bias è la tendenza dell'algoritmo ad apprendere costantemente un modello non corretto**, non tenendo conto di tutte le informazioni nei dati: **underfitting.**

Il **bias** viene utilizzato per consentire al modello di apprendimento automatico di apprendere in modo semplificato; quindi gli algoritmi sono più veloci da addestrare e più facili da capire. Tuttavia, ciò fa sì che il modello sia eccessivamente ridotto e quindi non soddisfi i requisiti: **underfitting**.

La varianza è la tendenza dell'algoritmo ad apprendere elementi casuali, indipendentemente dal segnale reale, adattando modelli che seguono troppo da vicino l'errore o il rumore nei dati: **overfitting**.



Maggiore è la **varianza** del modello, più complesso è il modello stesso che può apprendere funzioni più articolate. Tuttavia, se il modello è troppo complesso per il set di dati, un modello con **varianza** elevata fa sì che il modello si adatti in modo eccessivo: **overfitting**.

Immagini: a cura dell'Autore e da Edalab, Andrew Ng, researchgate.net.