Bias negli algoritmi (tecnico)

Oggigiorno, il termine “bias” viene fortemente utilizzato in svariati campi del Machine Learning e della statistica, dove assume significati differenti pur ritrovando in seguito applicazioni strettamente connesse tra loro.

Per spiegare cos’è un bias e perché esso sia così fondamentale nell’ambito del Machine Learning, è dapprima necessario introdurre ed elargire il concetto di Machine Learning. Il primo a coniare tale termine fu Tom Mitchell, il quale fornì la definizione seguente: «si dice che un programma apprende dall’esperienza E con riferimento a alcune classi di compiti T e con misurazione della performance P, se le sue performance nel compito T, come misurato da P, migliorano con l’esperienza E». Appare chiaro dunque, che il Machine Learning fonda i suoi punti cardine su una serie di concetti chiave che si possono così definire:

* *E – Esperienza*, ovvero il risultato di un processo di apprendimento che fornisce una maggiore velocità nella computazione di decisioni, rispetto alla velocità garantita dalla previa e minore esperienza.
* *T – Task*, ovvero il compito da eseguire.
* *P – Probabiltà*, ovvero la possibilità che il software riesca a completare il compito con successo.

Dunque, il Machine Learning, definito anche apprendimento automatico, è una tecnica di acquisizione di informazioni che, attraverso lo studio delle caratteristiche di dati noti meglio definiti come “training data”, ne riproduce i comportamenti ponendosi come obiettivo l’emulazione di questi ultimi su dati nuovi ed ignoti. Le funzioni matematiche o più in generale algoritmiche che consentono di compiere questo processo vengono definite come “target functions”.

Il termine “bias”, fu introdotto anch’esso per la prima volta da Mitchell nel 1980 con il seguente significato: “l’insieme di assunzioni che il classificatore utilizza per predire dati futuri in output, basandosi su dati in ingresso ad esso noti”. Esistono differenti tipologie di bias: se definito come *assoluto*, esso rappresenta l’ipotesi legittimata dall’algoritmo, che la target function da apprendere sia parte di un set designato e specifico di funzioni; qualora invece esso sia definito come *relativo*, l’algoritmo assume come ipotesi che la target function appartenga con maggiore probabilità ad un set di funzioni piuttosto che ad un altro.

Nell’ambito statistico, il bias rappresenta l’errore persistente o sistematico che si prevede venga compiuto dall’algoritmo di apprendimento quando quest’ultimo opera su dati noti.

Per esplicare tale concetto, supponiamo che sia la target function da determinare; a tale scopo, si applica l’algoritmo di apprendimento ripetutamente su un dataset contenente diversi dati utilizzati come campione. Al termine del processo, l’algoritmo restituisce per ciascuna entrata del dataset, un’ipotesi in merito al valore che assume, ad esempio . Essendo riferito a ciascun campione in esame (), i valori saranno rispettivamente rappresentati come ed . Computando una media tra tutti i valori di ottenuti dal complesso del dataset utilizzato, si ottiene un valore medio di ipotesi definito come .

Giunti al termine di tali considerazioni, si può definire come bias statistico di un algoritmo l’errore ottenuto nel valore ipotizzato medio:

Sarebbe errato pensare che il concetto di bias adottato nell’ambito del Machine Learning sia completamente disgiunto dalla definizione applicata in ambito statistico. Per analizzare tale relazione, è necessario considerare le caratteristiche che ciascuna tipologia di bias possiede:

* I bias *assoluti*, possono essere:
  + Inappropriati: nello spazio delle ipotesi di tale bias non sono presenti approssimazioni ragionevoli della target function.
  + Appropriati: nello spazio delle ipotesi sono contenute approssimazioni plausibili.
* I bias *relativi*, possono essere:
  + Molto forti: anche se questi ultimi non possono escludere approssimazioni ragionevoli per il valore assunto dalla target function, preferiscono adottare ipotesi con valori più inconsistenti.
  + Molto deboli: non consentono all’algoritmo di focalizzarsi su ipotesi più appropriate, mentre preferiscono invece che quest’ultimo consideri più ipotesi possibili.

Nella tabella, viene esposta la dipendenza esistente tra bias assoluto, bias relativo e bias statistico.

|  |  |  |
| --- | --- | --- |
| ML bias | | Statistical bias |
| Assoluto | Relativo | Valore |
| appropriato | molto forte | basso |
| appropriato | ok | basso |
| appropriato | molto debole | elevato |
| inappropriato | molto forte | basso |
| inappropriato | ok | moderato |
| inappropriato | molto debole | elevato |

Se il bias assoluto dell’algoritmo è inappropriato, il suo bias statistico possiede un valore elevato. Tuttavia, qualora il bias relativo fosse molto forte, potrebbe ugualmente verificarsi il caso in cui il bias statistico sia elevato sebbene il bias assoluto risulti appropriato.

Si può concludere dunque, che nell’ambito dell’apprendimento automatico così come della statistica, sia necessario conoscere la tipologia di bias che affligge l’algoritmo ed in particolar modo il valore che esso assume, al fine di determinare la veridicità dell’operato compiuto dall’algoritmo.