Algoritmi Avanzati
A.A. 2017–2018, primo semestre
Traccia delle lezioni

Mauro Brunato

Versione 2017-09-17

Caveat lector

Lo scopo principale di questi appunti è quello di ricostruire quanto detto a lezione. Queste note non sono complete, e la loro lettura non permette, da sola, di superare l'esame. Le fonti utili a ricostruire un discorso coerente e completo sono riportate alla pagina web del corso, dov'è disponibile anche la versione più recente di queste note:

`https://disi.unitn.it/~brunato/AA/`

Alcune fonti per approfondimenti sono indicate nelle note a pie' di pagina di questo documento.

Si suggerisce di confrontare la data riportata sul sito web con quella che appare nel frontespizio per verificare la presenza di aggiornamenti.

Changelog

2017-09-16

Versione iniziale:

- Introduzione, scopo del corso

Indice

Parte I

Appunti di teoria

Capitolo 1

Introduzione alla Data Science

1.1 La Data Science

Immagazzinare dati è fin troppo facile, e visti i costi ormai irrisori lo fanno tutti. Il problema è farne uso: i dati grezzi non sono immediatamente comprensibili, bisogna estrarre informazioni (fruibili da persone o da algoritmi) a fini migliorativi.

Questo è lo scopo della Data Science[1]; con sfumature diverse, si parla anche di "Business analytics", "Business Intelligence". Anche le discipline della Statistica e della Ricerca Operativa occupano lo stesso ambito, ma i loro nomi sono spesso associati a tecniche classiche e non sempre la loro definizione include il machine learning, che costituisce invece il nucleo del nostro corso.

In Figura 1.1 troviamo alcune "parole chiave" che definiscono le varie discipline e operazioni che portano dalla raccolta dati iniziale al risultato finale.

Il modus operandi più consueto consiste nell'utilizzare i dati per creare un modello (statistico, matematico, informatico) degli stessi da utilizzare poi per acquisire conoscenza sul sistema in esame, prendere decisioni, modificare processi, eccetera. Spesso, il risultato finale di tale processo sono nuovi dati che possono essere utilizzati per affinare il modello, e così via. Vedremo vari esempi nel resto del corso, che si concentrerà principalmente sulla fase dell'analisi dei dati e del machine learning.

[1] `https://en.wikipedia.org/wiki/Data_science`.

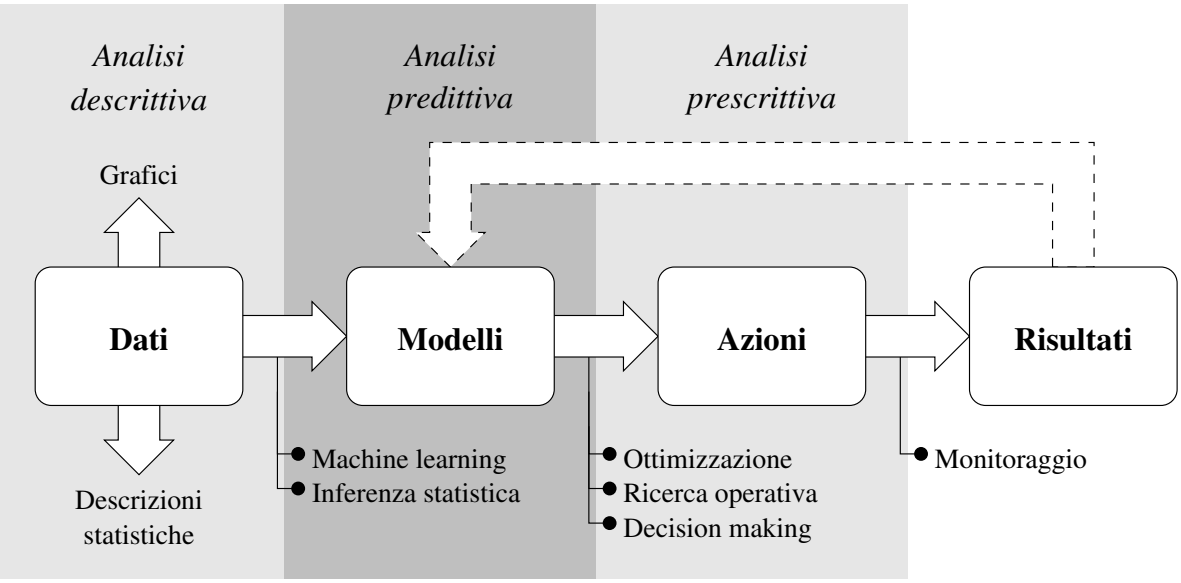

Figura 1.1: Alcune parole chiave

1.2 Scopo del corso

Esistono decine, se non centinaia, di pacchetti software e di librerie in grado di importare dati, costruire modelli, effettuare previsioni. Uno sviluppatore può creare una rete neurale, addestrarla e utilizzarla in un suo programma senza nemmeno sapere come funziona, semplicemente utilizzando le API di una libreria e trattando i diversi algoritmi come "scatole nere". Insomma, non è necessario conoscere un algoritmo di machine learning per poterlo utilizzare, come non è necessario conoscere i protocolli di rete in dettaglio per realizzare un'applicazione web, e si possono ordinare gli elementi di un vettore senza conoscere gli algoritmi di ordinamento.

Nonostante ciò, le tecniche che studieremo in questo corso vanno applicate con estrema attenzione, perché sono soggette a errori estremamente perniciosi, ma non sempre facili da evitare. Il corso sarà in parte dedicato a studiare le metodologie utili a evitare o a minimizzare l'impatto dei trabocchetti più comuni.

Ecco alcuni esempi di trabocchetto tratti dalla statistica classica. Più tardi ne vedremo anche alcuni specifici del machine learning.

1.2.1 Significatività di un esperimento: il p-value

Sappiamo che la Statistica offre strumenti per capire se un certo risultato è significativo o meno. Il principio generale è: più campioni misuriamo, minori sono gli errori che commettiamo nello stimare le grandezze statistiche della popolazione. Un metodo molto usato in ambito sperimentale è il cosiddetto "p-value". Formulata un'ipotesi H, ed effettuato un esperimento consistente nella misurazione di alcuni campioni di una popolazione, diciamo che l'esperimento conferma l'ipotesi se il suo "p-value" è al di sotto di una soglia, detta *significatività*, da noi stabilita a priori. Il p-value esprime la probabilità di ottenere risultati simili a quelli effettivamente ottenuti nel caso in cui H sia falsa[2].

Supponiamo, ad esempio, di sospettare che una certa moneta sia truccata in modo da offrire sempre il lato "testa" in un lancio. La nostra ipotesi H è "Questa moneta è truccata". Il nostro esperimento consiste di $N = 10$ lanci, e otteniamo sempre testa. Il p-value dell'esperimento risponde alla domanda "quale sarebbe la probabilità di ottenere dieci teste se H fosse falsa, cioè se la moneta *non* fosse truccata?" La risposta è ovviamente $p = 2^{-10} < 1/1000$. Con un p-value così basso possiamo concludere che il nostro risultato non è dovuto al caso, e che quindi la nostra ipotesi è confermata.

Cosa succede, però, se ogni italiano esegue questo test su una moneta di propria scelta? Anche se nessuna delle monete fosse truccata, ben 60000 italiani (circa uno su mille) concluderebbero che la loro moneta lo è. Dovremmo togliere dalla circolazione quelle sessantamila monete?

Si consideri che nella maggior parte degli esperimenti scientifici, soprattutto quando il costo degli esperimenti è elevato (si pensi a medicina, genetica, psicologia), la soglia di significatività è molto maggiore di 1/1000 (spesso ci si accontenta di 1/20, cioè del 5%). Se un esperimento consiste nella verifica di 20 diverse ipotesi, e ci si accontenta di una significatività del 5%, è possibile che una delle ipotesi venga confermata erroneamente, spesso per ingenuità, talvolta intenzionalmente (in tal caso, si parla di *p-hacking*[3]).

1.2.2 Correlazioni spurie

Un effetto collaterale dell'abbondanza di dati è la facilità con la quale, in assenza di un'ipotesi precisa da verificare, sia sempre possibile trovare serie di dati in apparente dipendenza l'uno dall'altro nonostante la completa estraneità[4].

[2] `https://en.wikipedia.org/wiki/P-value`.

[3] Un esempio: la cioccolata accelera il dimagrimento (`https://io9.gizmodo.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800`). Più in breve: `https://xkcd.com/882/` — vedere `https://explainxkcd.com/882/` per chiarimenti.

[4] `http://www.tylervigen.com/spurious-correlations`

1.2.3 Il paradosso di Simpson

In alcuni casi, nonostante l'ipotesi da verificare sia precisa e ogni test statistico dica che i risultati sono significativi, può accadere che la conclusione sia errata semplicemente perché le nostre informazioni sono incomplete. Un esempio è il cosiddetto "paradosso di Simpson", nel quale l'assenza di una variabile esplicativa causa la formulazione di un'ipotesi opposta rispetto alla realtà dei fatti[5].

[5]Si veda `https://en.wikipedia.org/wiki/Simpson%27s_paradox`, in particolare i primi due esempi.