UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI "M.FANNO"

DIPARTIMENTO DI SCIENZE STATISTICHE

CORSO DI LAUREA IN ECONOMIA

PROVA FINALE

"IL RUOLO DELLA STATISTICA NEL MONDO DEI BIG DATA: STRUMENTI ED APPLICAZIONI"

RELATORE:

CH.MO PROF. TOMMASO DI FONZO

LAUREANDO: TITO SCALVENZI

MATRICOLA N. 1136075

ANNO ACCADEMICO 2018 – 2019

INDICE

Introduzione

Viviamo nell'era dei Big Data, dove la scienza, la tecnologia e le industrie stanno producendo enormi flussi di dati, grazie allo sviluppo dell'informatica e di internet. Oltre alla crescita esplosiva del volume dei dati scambiati, i Big Data hanno anche un'alta velocità, un'elevata varietà e un'elevata incertezza. L'emergere dei Big Data ha un significato epocale per la Statistica. Da un lato, essi compensano gli svantaggi (costi ed errori elevati) delle rilevazioni statistiche, dall'altro lato, introducono sfide computazionali e statistiche uniche. Queste sfide conducono alla nascita di un nuovo paradigma statistico e informatico: gli enormi flussi di dati complessi richiedono velocità di elaborazione sempre maggiori ed una capacità di archiviazione sempre più economica, con lo scopo di fornire delle risposte utili e tempestive in ambiti sia di ricerca che di business, ambienti in cui l'incertezza gioca un ruolo fondamentale. Per questi motivi i Big Data hanno originato nuove sfide nel campo statistico rispetto all'analisi convenzionale dei dati. Tuttavia, questo non significa che l'epoca delle rilevazioni statistiche sia tramontata. Anzi, si può forse affermare il contrario, dato che la ricerca, il raggruppamento e la classificazione di grandi moli di dati richiedono ancora oggi l'ampio utilizzo di metodi statistici sempre più raffinati. Si può quindi affermare che l'avvento dell'era dei Big Data migliora la qualità della statistica, ne riduce i costi, fa sì che essa svolga un ruolo importante in una gamma ancora più ampia di applicazioni.

Questo lavoro, pur non avendo alcuna pretesa di esaustività scientifica, dato che spazia in campi che esulano dal mio percorso formativo in senso stretto, vuole fornire una panoramica sulle caratteristiche salienti dei Big Data e su come tali caratteristiche influiscono sull'evoluzione delle Scienze Statistiche.

Il lavoro è organizzato nel modo seguente:

Nel *primo capitolo*, dopo aver inquadrato il ruolo fondamentale della Statistica come strumento privilegiato a supporto delle decisioni, viene discusso il dibattito secolare riguardo il concetto di probabilità. Tale dibattito risulta utile al fine di introdurre le nuove tendenze che hanno cambiato la statistica classica, come l'aumento d'interesse verso i metodi bayesiani. In particolare, vengono approfonditi alcuni aspetti fondamentali delle reti bayesiane e di come queste abbiano riscontrato un grande successo sia in contesti di ricerca che di business.

Nel *secondo capitolo*, vengono introdotte le tendenze principali che hanno dato luogo alla rivoluzionaria "era dei Big Data"; nel tentativo di dare una definizione di tale termine, vengono analizzate le caratteristiche di queste enormi quantità di dati e le diverse fonti da cui provengono. A fine capitolo viene trattato, in maniera semplificata e schematica, il processo di

estrazione della conoscenza dai database ed in particolare il Data Mining, il quale fa ampio utilizzo di metodi e tecniche statistiche.

Nel *terzo capitolo*, vengono analizzate alcune questioni riguardo al rapporto tra Statistica e Big Data, dal quale nascono numerose sfide ed opportunità per il futuro.

Capitolo 1 – La Statistica contemporanea

La Statistica, come disciplina, ha una storia di oltre 300 anni, la quale può essere suddivisa in tre fasi:

- **Periodo statistico classico**, la cui nascita risale alla metà del XVII secolo.
- **Periodo statistico moderno**, che si è sviluppato dalla fine del XVIII secolo alla fine del XIX secolo.
- **Periodo statistico contemporaneo**, in cui il focus della Statistica si orienta verso la statistica inferenziale (Sprent, 2003).

1.1 Il ruolo della Statistica contemporanea nelle Scienze

Diverse discipline scientifiche utilizzano la Statistica come potente strumento per analizzare dati provenienti sia da esperimenti controllati (principalmente scienze naturali) che da studi osservazionali (principalmente scienze umane). Nel corso del London Workshop, sul futuro delle scienze statistiche (Madigan, 2013), viene affermato che quasi tutte le ricerche scientifiche di oggi richiedono la gestione e il calcolo dell'incertezza, e per questo motivo la statistica, la scienza dell'incertezza, è diventata un partner cruciale per la scienza moderna. Ad esempio, la Statistica ha contribuito all'idea dello studio randomizzato controllato, una tecnica sperimentale che è oggi universalmente usata nella ricerca farmaceutica e biomedica, cosi come in molte altre aree scientifiche. In generale, le osservazioni empiriche sono utilizzate per rafforzare teorie scientifiche, le quali vengono valutate attraverso ipotesi. Tale rafforzamento diventa ancora più decisivo se è possibile trovare delle prove per passare da una prospettiva particolare ad una generale attraverso procedure inferenziali.

1.2 Inferenza e probabilità

L'Enciclopedia Treccani definisce l'inferenza statistica come "un procedimento di generalizzazione dei risultati ottenuti attraverso una rilevazione parziale per campioni (limitata cioè alla considerazione di alcune unità o casi singoli del fenomeno di studio) alla totalità delle unità o casi del fenomeno stesso, sulla base di ipotesi plausibili".

In altre parole, potremmo dire che fare inferenza è generare conoscenza per la popolazione dall'osservazione di un campione, e tale processo è mediato dalla probabilità.

Il concetto di incertezza e probabilità sono strettamente legati. Possiamo definire infatti la probabilità come "un numero compreso tra 0 e 1 che misura il grado di incertezza sul verificarsi

di un evento" (Borra, Di Ciaccio, 2014, p.171). Inoltre, la probabilità presenta un duplice significato: *epistemico*, ossia come incertezza relativa alla limitata conoscenza umana, ed *empirico*, ossia come incertezza intrinseca nei fenomeni. Da questo duplice significato nascono diverse interpretazioni riguardo il significato di probabilità, tra cui: la concezione classica, la concezione frequentista e la concezione bayesiana.

1.3 Il dibattito sul significato di "probabilità"

Nel 1763 Thomas Bayes scrisse il "saggio per risolvere un problema nella dottrina delle probabilità", il quale diede inizio a secoli di dibattito riguardo il significato di probabilità.
Sebbene poco notato a suo tempo, alla fine ha generato un approccio alternativo al ragionamento probabilistico classico che ha davvero raggiunto il suo scopo nell'era dei computer. Infatti, grazie ai progressi dell'informatica moderna, si è verificato lo sviluppo di tutte le tecniche basate sul calcolo numerico. Risulta opportuno sottolineare che, oggi, il dibattito non è solo accademico, perché punti di vista diversi su questa questione portano a metodologie diverse. Perciò, nei paragrafi successivi, verranno brevemente trattati i diversi campi di pensiero, i quali conducono a differenze sostanziali all'approccio inferenziale.

1.3.1 Concezione Classica

La probabilità di un evento è il rapporto tra il numero di casi favorevoli e quelli possibili, supposto che tutti gli eventi siano equiprobabili. Tale definizione, basata su eventi discreti e di numero finito, non è estendibile al caso di variabili continue. Un altro elemento debole è la condizione ideale di perfetta uniformità, secondo cui tutti i possibili esiti sono noti in precedenza e tutti sono ugualmente probabili. Quando un fenomeno presenta un numero infinito di possibilità, la definizione classica conduce a paradossi (paradosso di Bertrand).

1.3.2 Concezione Frequentista

Mentre nella concezione classica la probabilità è stabilita a priori (prima di guardare i dati), nella concezione frequentista è invece ricavata a posteriori, ovvero dall'esame dei dati. La concezione frequentista si basa sulla ripetibilità della prova e venne sviluppata dai primi pionieri come Fisher, Jerzy Neyman e Karl Pearson.
Fisher ha introdotto il p-value. Egli era un sostenitore del ragionamento induttivo per generare nuove ipotesi al fine di sviluppare nuove intuizioni. La sua inferenza si basa sul rifiuto

dell'ipotesi nulla (H_0), ovvero più piccolo è il p-value, più grande è la prova contro l'ipotesi nulla, e questo approccio induttivo viene chiamato test di significatività.
Nell'approccio di Neyman e Pearson, chiamato test d'ipotesi, si assume che ci siano due ipotesi (ipotesi nulla ed ipotesi alternativa), e che sia necessario fare una scelta tra le due. Neyman e Pearson infatti non credevano nell'abilità di misurare la prova da un esperimento singolo.
In sintesi, la concezione Frequentista si presenta come una combinazione dei diversi approcci precedentemente presentati.
Di seguito viene enunciato il postulato empirico del caso: "In un gruppo di prove ripetute più volte nelle stesse condizioni, ciascuno degli eventi possibili compare con una frequenza approssimativa eguale alla sua probabilità; generalmente l'approssimazione migliora quando il numero delle prove cresce" (Borra, Di Ciaccio, 2014 p.184). Tale approccio si basa sul fatto che una probabilità riflette la frequenza con cui un particolare risultato sarà osservato in studi ripetuti dello stesso esperimento. La ripetibilità della prova implica che le condizioni in cui vengono svolte le prove ripetute rimangano inalterate (chiaramente ciò non è sempre possibile). Estrarre delle palle da un'urna o lanciare dei dadi, sono situazioni ideali in cui la stessa procedura può essere ripetuta più volte con risultati incerti.
La concezione Frequentista si basa sul "Principio del campionamento ripetuto", grazie al quale vengono costruite procedure inferenziali che posseggono proprietà ottimali al ripetersi dell'operazione di campionamento.
Tuttavia, le statistiche frequentiste hanno sofferto grandi difetti di progettazione e interpretazione che hanno portato ad una serie di limiti nell'analisi dei problemi della vita reale, per esempio:

- Il valore p e l'intervallo di confidenza dipendono fortemente dalle dimensioni del campione.
- Gli intervalli di confidenza non sono vere e proprie distribuzioni di probabilità.

Entrambe le concezioni finora citate si riferiscono ad una probabilità oggettiva, danno una visione empirica della probabilità, dunque non epistemica.

1.3.3 Concezione Soggettivista

Secondo la concezione soggettivista, "la probabilità di un evento è la misura del grado di fiducia che un individuo coerente attribuisce al verificarsi dell'evento in base alle informazioni in suo possesso" (Borra, Di Ciaccio, 2014 p.186).
La nozione di probabilità soggettiva ha una valenza più ampia rispetto a quella di probabilità frequentista poiché esistono dei casi nei quali all'evento non è associabile uno schema di prove

ripetute e di frequenze osservabili (es. eventi unici o astratti). Inoltre, accade che la valutazione soggettiva della probabilità di un evento utilizzi l'informazione derivante dalle frequenze dell'evento stesso (osservate in diverse prove). La scuola di pensiero soggettivista è recentemente diventata di fondamentale importanza poiché l'approccio inferenziale Bayesiano, il quale appunto utilizza la probabilità a priori, è basato su di essa.

1.4 Inferenza Bayesiana

La Statistica Bayesiana prende il nome dal teorema di Bayes, il quale è di fatto una regola per aggiornare la nostra fiducia in un'ipotesi man mano che raccogliamo nuove prove.
Una versione di esso può essere semplificata nel modo seguente:
Quote posteriori = quote precedenti × likelihood ratio
La teoria Bayesiana si occupa di affermazioni di probabilità che sono condizionate al valore osservato. Questa caratteristica condizionale introduce la principale differenza tra l'inferenza Bayesiana e l'inferenza classica. Nonostante in molte analisi semplici si arrivi a conclusioni superficialmente simili dai due approcci, i metodi bayesiani risultano essere particolarmente adatti a risolvere i problemi decisionali poiché l'inferenza statistica basata sull'approccio bayesiano è più intuitiva rispetto a quella basata sull'approccio frequentista ed essa è in grado di utilizzare tutte le informazioni disponibili. Un altro grande vantaggio dell'interpretazione bayesiana è che può essere utilizzata per modellare la nostra incertezza su eventi che non hanno frequenze a lungo termine. Inoltre, secondo Gelman et al. (1995), le analisi ottenute con i metodi bayesiani possono essere facilmente estese a problemi più complessi. Per questi motivi, negli ultimi anni le tecniche statistiche bayesiane hanno registrato un notevole aumento di interesse a discapito dell'approccio frequentista, ovvero quello delle tecniche classiche (Iannazzo, 2007).

1.4.1 I metodi Bayesiani nell'era dei Big Data

In una nuova panoramica pubblicata nella National Science Review di Pechino (Fan et al., 2014), gli scienziati dell'Università di Tsinghua, Cina, presentano gli ultimi progressi riguardanti i metodi Bayesiani per l'analisi dei Big Data, in cui viene sottolineato che i metodi Bayesiani stanno diventando sempre più rilevanti nell'era dei Big Data per proteggere i modelli ad alta capacità contro l'overfitting e per consentire a tali modelli di aggiornare in modo adattivo la loro capacità. Negli ultimi anni tali metodi sono diventati sempre di più sempre più utilizzati nel campo dell'apprendimento automatico. Infatti, il machine learning basato sui metodi bayesiani ha attirato grande attenzione sia da parte dell'industria che del mondo accademico. In

particolare, le reti Bayesiane hanno riscontrato uno sviluppo significativo. Si tratta di un modello costituito da struttura e da parametri, che vengono utilizzati per descrivere la relazione causale tra le variabili di analisi qualitative e quantitative (Zhang, 2017).

1.4.2 Reti bayesiane

Quando un problema reale è caratterizzato da un elevato numero di variabili, le quali presentano complicate relazioni tra di loro, diventa necessario l'utilizzo di strumenti in grado di gestire l'incertezza in maniera quantitativa quando si vuole inferire informazioni su variabili di interesse a partire da dati o osservazioni, sfruttando tutta la conoscenza a disposizione. Le reti bayesiane rappresentano infatti lo strumento ideale per strutturare i problemi ed analizzare i dati, in particolare quando le relazioni probabilistiche causa-effetto tra variabili sono complicate. Esse permettono di aggiornare in maniera quantitativa le probabilità di tutte le variabili in gioco ogni volta che vengono acquisite nuove informazioni su alcune di esse. Queste nuove informazioni vengono chiamate "evidenze", ovvero valori noti di certe variabili, le quali permettono appunto la modifica della probabilità delle altre variabili. Inoltre, anche quando sono vaghi i rapporti di dipendenza tra le variabili, è possibile costruire la corretta struttura della rete attraverso algoritmi di "structure learning", qualora si abbia a disposizione un'adeguata quantità di dati (ciò è quasi sempre possibile nell'era Big Data).

Più nello specifico, una rete bayesiana consiste in un insieme di variabili dette nodi, le cui relazioni (probabilistiche e deterministiche) sono rappresentate da frecce. Le variabili insieme alle frecce che le congiungono costituiscono un grafico aciclico, ovvero un grafico in cui non è possibile partire da una variabile e tornare sulla stessa seguendo le direzioni delle frecce. Ad ogni variabile condizionata da altre variabili, chiamate "genitori", è associata una tabella di probabilità condizionata che quantifica la dipendenza di tale nodo, detto "figlio", dai nodi genitori.

Numerosi vantaggi rendono le reti bayesiane strumenti fondamentali nella gestione dell'incertezza. Inanzitutto, questo tipo di rete presenta le caratteristiche di versatilità, efficacia e apertura. Infatti, il network bayesiano è una rete di una piattaforma in cui è possibile integrare altre tecnologie intelligenti ed altre tecnologie statistiche.

In secondo luogo, essa è concettualmente semplice e flessibile, infatti, è in grado di trasformare efficacemente i dati in conoscenza attraverso metodi visivi e rappresentazioni intuitive che permettono di analizzare i problemi con le stesse modalità di ragionamento del pensiero umano. Le reti bayesiane permettono anche di integrare probabilità provenienti da fonti diverse, come frequenze in un database, conoscenze teoriche e stime soggettive di esperti. Inoltre, per costruire

una rete bayesiana è sufficiente inserire solamente le probabilità condizionate associate alle variabili collegate tra loro, il che porta ad un grande vantaggio computazionale.
Un altro grande vantaggio consiste nel fatto che la rete viene strutturata a livello locale (individuando le relazioni tra i nodi e stimando le probabilità corrispondenti) ed in seguito vengono aggiornate le probabilità a livello globale, grazie all'utilizzo di software che permettono la gestione della complessa trasmissione delle evidenze nella rete. Inoltre, man mano che si acquisisce nuova conoscenza sul sistema, le informazioni acquisite vengono utilizzate dalla rete per aggiornare le stime delle probabilità iniziali. I Big Data infatti richiedono che i nostri modelli siano adattivi quando gli scenari di apprendimento cambiano. Per esempio, i metodi non parametrici bayesiani (NPB) forniscono strumenti eleganti per affrontare situazioni di questo tipo.
Per concludere, una rete bayesiana è leggibile in ogni sua parte, quindi è possibile: leggere l'incertezza delle conclusioni della rete e trovare le più probabili; controllare sotto quali assunzioni valgono le conclusioni della rete; effettuare analisi di sensibilità, per capire quale accuratezza sulle stime di probabilità è necessaria per le conclusioni di nostro interesse.
Il limite principale dell'applicazione dei metodi bayesiani è il carico computazionale, legato all'integrazione di variabili continue nel modello, il quale è stato superato grazie allo sviluppo di appositi software di calcolo e di efficienti tecniche di stima Monte Carlo. Questo ha permesso alla rete bayesiana di venire sempre più utilizzata in campo statistico, nell'analisi delle decisioni, nell'intelligenza artificiale e così via.

1.4.3 Applicazioni

In generale, come affermato precedentemente, le reti bayesiane possono essere utilizzate come efficaci strumenti di Machine Learning. Esse riescono a prevedere comportamenti futuri in base all'esperienza di quelli passati, ad individuare i fattori decisivi che determinano i valori di una variabile, individuare la categoria ("profiling") a cui appartengono determinate osservazioni.
Data la loro elevata funzionalità, le reti bayesiane vengono utilizzate in numerosi ambiti ed applicazioni. In questo paragrafo verranno forniti alcuni esempi.
In diagnostica, esse permettono di individuare la causa più probabile di alcuni sintomi, in modo più veloce e semplificando notevolmente il processo diagnostico. Nel campo del profiling, i filtri anti-spam riescono ad individuare efficacemente le mail spam grazie a sistemi basati sulle reti. Nel marketing, esse vengono utilizzate per identificare gli elementi di successo della campagna e il profilo dei potenziali consumatori interessati ad un certo prodotto. Nell'e-

commerce, sempre grazie alle reti, è possibile suggerire i prodotti al cliente sulla base degli acquisti effettuati in passato e da clienti con profilo simile.

Nel campo delle telecomunicazioni (Abbondanza et al., 2012), il modello Bayesiano è stato utilizzato per sfruttare al meglio le previsioni sui volumi di traffico nella conduzione delle negoziazioni degli accordi di roaming. Telecom Italia ha sviluppato uno strumento chiamato IRMA (International Roaming Multi-negoziation Algorithm) in grado di valutare costi e ricavi in base a diverse tipologie di accordi trattati dalla società. L'algoritmo è in grado di stimare in tempo reale, con un'incertezza che deriva dalle previsioni dei volumi, le variabili probabilistiche rappresentate da costi e ricavi attesi, permettendo così flessibilità e risparmi di tempo durante la fase di contrattazione. Considerando che la società ha oltre 600 accordi di roaming internazionale in 200 paesi, questo strumento è fondamentale per incrementare l'efficacia nella conduzione delle trattative e per ridurne i tempi di finalizzazione.

Un altro esempio della regola di Bayes è fornito dai programmi di controllo ortografico.

Si supponga, ad esempio, un utente che digita la parola "radom" e il computer deve decidere se intendeva scrivere "random" oppure "Radom", la citta' in Polonia. Consultando il database delle lingue di Google, il computer determina che la parola "random" appare 200 volte più spesso di "Radom" in tutti i documenti. In assenza di qualsiasi altro tipo di informazioni, le "quote precedenti" sono 200:1 a favore di "random". Tuttavia, un programma di controllo ortografico che, semplicemente, si limitava sempre alla parola più comune, cambiava ogni parola in "the". Quindi le probabilità precedenti devono essere modificate in base all'evidenza di ciò che è stato effettivamente digitato. Secondo il modello di errori ortografici di Google, è 500 volte più probabile che venga digitato "radom" se la parola che si intendeva digitare è "Radom" (cosa che avviene con probabilità 0,975) rispetto che venga digitata la parola "random" (probabilità 0,00195). Quindi la likelihood ratio è 1/500, e le probabilità a posteriori diventano (200/1)*(1/500), o 2:5. Così il correttore ortografico non correggerà automaticamente la parola. D'altra parte, se il correttore ortografico avesse saputo che la parola proveniva da un documento di statistica, le probabilità precedenti a favore di "random" sarebbero salite ed il correttore ortografico avrebbe corretto automaticamente la parola. Oppure, se l'utente fosse "approssimativo", la percentuale di probabilità di un errore rispetto all'ortografia corretta aumenterebbe, e di nuovo le probabilità a posteriori si sposterebbero a favore del "random". Questo dimostra quanto facilmente la regola di Bayes sia in grado di incorporare nuove informazioni.

Un altro esempio applicativo di successo dell'approccio Bayesiano è la tecnologia inventata da Paul Viola e Michael Jones nel 2001. Le loro ricerche hanno aiutato le fotocamere digitali a

riconoscere i volti usando il ragionamento Bayesiano, disegnando un piccolo rettangolo intorno a tutto ciò che la macchina della fotocamera pensa che possa essere un volto.

Capitolo 2 – La rivoluzione Big Data

2.1 Datafication

Tre tendenze hanno permesso la rivoluzione dei Big Data:

- Il rapido aumento della quantità di dati disponibili.
- L'accelerazione della capacità di memorizzazione dei dati e della potenza di calcolo a basso costo (legge di Moore).
- L'evoluzione nell'approccio di Machine Learning per analizzare insiemi di dati contorti.

I progressi tecnologici, per quanto riguarda l'informatica e la comunicazione, hanno portato ad una quantità sempre maggiore di dati che possono fornire una sintesi della nostra vita moderna. Datafication, secondo Mayer-Schoenberger e Cukier (2013) è la trasformazione dell'azione sociale in dati quantificati online, i quali consentono il monitoraggio in tempo reale di ogni azione e la conseguente analisi predittiva.

Con l'avvento del Web 2.0, caratterizzato da pagine generate dagli utenti e dalla crescita dei social media, sono stati codificati molti aspetti della vita sociale che non erano mai stati quantificati prima, per esempio, amicizie, interessi, conversazioni casuali, ricerche di informazioni, espressioni di gusti, risposte emotive e così via. Le imprese e le agenzie governative scavano nel numero crescente ed esponenziale di metadati raccolti attraverso i social media e le piattaforme di comunicazione, come Facebook, Twitter, LinkedIn, Tumblr, iTunes, Skype, WhatsApp, YouTube e di posta elettronica gratuita, come gmail e hotmail, per tracciare le informazioni sul comportamento umano: "Ora possiamo raccogliere informazioni che non potevamo prima, che si tratti di relazioni rivelate da telefonate o sentimenti svelati attraverso i tweet" (Mayer-Schoenberger e Cukier, 2013).

Per esempio Facebook ha trasformato attività sociali come "amicizie" e "simpatia" in relazioni algoritmiche; Twitter ha reso popolari i personaggi online delle persone e ha promosso idee creando funzioni di "follower" e "retweet"; LinkedIn ha tradotto reti professionali di dipendenti e persone in cerca di lavoro in interfacce digitali (Van Dijck, 2014).

Dal punto di vista dell'informatica, il numero di transistors (micro oggetti di silicone paragonabili ai nostri neuroni) su un microchip raddoppia ogni due anni, anche se il costo dei computer è dimezzato. Questo fattore si riferisce alla legge di Moore, la quale afferma che la crescita nell'industria dei microprocessori è esponenziale, e ciò significa che la capacità di memorizzazione dei dati e della potenza di calcolo si espanderanno rapidamente nel tempo.

Allo stesso tempo, nel mondo statistico, i progressi riguardanti i metodi dell'inferenza Bayesiana (trattati nel capitolo 1), hanno permesso ai ricercatori di acquisire conoscenze sui consumatori, utilizzando modelli sempre più pertinenti e complessi. La valutazione di grandi insiemi di dati raccolti attraverso le piattaforme dei social media si presenta sempre più come il metodo più scrupoloso e completo per misurare l'interazione quotidiana, superiore al campionamento (infatti, "N=tutti") e più affidabile dell'intervista o del sondaggio (approfondimento sul campionamento nel terzo capitolo). Grandi quantità di dati "disordinati" sostituiscono piccole quantità di dati campionati e, come molti sostengono, la notevole dimensione dei set di dati compensa il loro disordine.

I metadati, che non molto tempo fa venivano considerati un sottoprodotto inutile dei servizi mediati da piattaforme, sono stati gradualmente trasformati in risorse preziose che possono essere estratte, arricchite e riallocate in prodotti preziosi.

2.2 Che cosa sono i Big Data?

A differenza dei dati statistici tradizionali che sono compilati per scopi specifici, i grandi dati sono un sottoprodotto presente nei sistemi aziendali e amministrativi, nei social network, nelle reti sociali e nell'internet delle cose (IoT). La Commissione federale per i Big Data della Fondazione TechAmerica (2012) definisce i big data nel seguente modo: "Big data è un termine che descrive grandi volumi di dati ad alta velocità, complessi e variabili che richiedono tecniche e tecnologie avanzate per consentire la cattura, l'archiviazione, la distribuzione, la gestione e l'analisi delle informazioni." Esistono varie possibili definizioni riguardo questo termine, conseguenza della diversità delle fonti e delle discipline.

2.2.1 Quattro "V"?

Il termine Big Data è spesso caratterizzato dalle "V", le quali si riferiscono ad: alto volume, ad alta velocità e di alta varietà. L'analista Doug Laney ha creato le famose tre V nel 2001. L'elenco delle "V" è cresciuto nel tempo, sottolineando sia le opportunità che le sfide che le aziende e le organizzazioni devono affrontare quando incorporano grandi dati nelle loro operazioni di business esistenti.

Volume: Ovvero la dimensione del set di dati, i quali sono riportati in più terabyte e petabyte. Il 90% dei dati presenti oggi giorno sono stati creati negli ultimi due anni. Ogni giorno vengono accumulati 2,5 quintilioni di bytes (IBM). Un set di dati, per essere considerato big data, si ritiene che debba essere di almeno 1 terabyte. Tuttavia, ciò che può essere considerato big data oggi potrebbe non raggiungere la soglia in futuro perché la capacità di archiviazione potrebbe

aumentare, consentendo così di acquisire serie di dati ancora più grandi. Infatti, si presume che l'universo digitale di dati raccolti aumenterà da trilioni di gigabyte a 44 zettabyte entro la fine del 2024.

Velocità: L'alta velocità si riferisce alla velocità con cui i dati sono creati, elaborati e memorizzati. La rapida diffusione di dispositivi digitali come smartphone e sensori ha portato ad una velocità, nella creazione di dati, senza precedenti. Il tasso globale di traffico su Internet è stato stimato essere di 50.000 GB al secondo. Ogni minuto vengono caricate 72 ore di filmati su youtube, 216.000 post di instagram e vengono spedite 204.000.000 di email. Ciò comporta una crescente esigenza di analisi in tempo reale e di pianificazione basata sull'evidenza. Wal-Mart, ad esempio, elabora più di un milione di transazioni all'ora.

Varietà: L'alta varietà si riferisce alla gamma e alla complessità dei tipi di dati e delle fonti. Basta pensare all'eterogeneità strutturale presente in un database. Infatti, esistono tre diverse forme di dati: strutturati, semi-strutturati e non strutturati. I dati strutturati, che costituiscono solo il 5% di tutti i dati esistenti, si riferiscono ai dati tabulari trovati nei fogli di calcolo o nei database relazionali.

Il tipo di dati che aumenta più rapidamente sono i dati non strutturati. Testi, immagini, audio e video sono esempi di dati non strutturati, i quali rappresentano il 90% dei dati generati. Tuttavia, a volte mancano dell'organizzazione strutturale richiesta dalle macchine per l'analisi. Il formato dei dati semi-strutturati, un ibrido tra i dati completamente strutturati e non strutturati, non è conforme a standard rigorosi. Un tipico esempio di dati semi-strutturati è Extensible Markup Language (XML), un linguaggio testuale per lo scambio di dati sul Web. I documenti XML contengono tag di dati definiti dall'utente che li rendono leggibili dalla macchina.

Veracità: IBM ha coniato questo termine come quarta V, che rappresenta l'inaffidabilità insita in alcune fonti di dati. Il termine si riferisce al rumore e alla parzialità dei dati come una delle maggiori sfide per dare valore e validità ai grandi dati. Ad esempio, sebbene i sentimenti dei clienti nei social media siano di natura incerta, poiché implicano un giudizio umano, contengono informazioni preziose. Quindi la necessità di trattare dati imprecisi e incerti è un altro aspetto dei big data, che viene affrontato utilizzando strumenti e analisi sviluppati per la gestione e l'estrazione di dati incerti.

Valore: Sulla base del termine introdotto da Oracle, i big data sono spesso caratterizzati da una "densità di valore" relativamente bassa. Ovvero, i dati ricevuti nella forma originale hanno un valore relativamente basso rispetto al loro volume. Tuttavia, può essere comunque ottenuto un alto valore grazie all’analisi di tali dati: questa è proprio una delle principali sfide nell’era dei Big Data.

2.2.2 Classificazione per tipologie di fonti

La "datafication" è onnipresente e la quantificazione dei dati è in continua crescita: i dati sono generati sempre più spesso, in qualsiasi luogo, in qualsiasi momento e con qualsiasi mezzo. Gli odierni dispositivi e sistemi di datafication sono ovunque, coinvolti e correlati al nostro lavoro, studio, intrattenimento, ambiente socio-culturale, e dispositivi e servizi personali.

Di conseguenza, un'altra possibilità per classificare i big data è ricorrere a una definizione che identifica i big data in base alla loro fonte. Sono state identificate dieci categorie di big data (Buono et al. 2017), organizzate nel modo seguente:

Dati di mercati finanziari: La New York Stock Exchange (NYSE) ha iniziato la raccolta di questi dati nel 1992. Questi dati infra-giornalieri forniscono informazioni dettagliate che potrebbero essere utilizzate per l'analisi dell'efficienza, della volatilità, della liquidità, della determinazione dei prezzi e delle aspettative dei mercati.

Dati di pagamenti elettronici: I pagamenti elettronici comprendono le transazioni con carte di credito e di debito, bonifici, addebiti diretti, assegni, ecc. La più utilizzata è quella delle transazioni avviate dal titolare della carta, vale a dire i pagamenti con carta di credito e di debito. I pagamenti con carte sono considerati come una categoria di Big Data a causa dell'alta frequenza delle operazioni. Vengono concluse migliaia di transazioni nel corso della giornata e, con il rapido aumento del commercio elettronico, anche durante la notte. Le statistiche evidenziano che le carte sono la principale forma di pagamento non in contanti (Capgemini e BNP Paribas, 2016). I pagamenti con carta includono sia gli acquisti online che offline nei POS, il che li rende molto utili per seguire il comportamento dei consumatori, tra cui le vendite al dettaglio. Infatti, in base ai dati, è possibile tenere traccia degli acquisti delle persone e delle famiglie.

Dati di telefoni cellulari: La raccolta di dati dalle funzioni di base di un telefono cellulare, cioè ricevere e fare telefonate e messaggi di testo, fornisce informazioni sulla densità della popolazione, l'ubicazione, lo sviluppo economico di particolari aree geografiche e l'uso dei trasporti pubblici. Grazie alla grande diffusione dei telefoni cellulari e l'enorme volume di dati raccolti, i dati di telefoni cellulari sono categorizzati come Big Data e rappresentano uno strumento importante per le analisi statistiche. Ovviamente, gli operatori di rete sono la principale fonte di dati per la telefonia mobile in quanto tutte le informazioni vengono inviate e ricevute attraverso la loro rete. Nel complesso, i dati dei telefoni cellulari forniscono informazioni dettagliate sul comportamento umano e possono quindi essere utili anche nelle scienze sociali. Un esempio applicativo è uno studio condotto da Ricciato et al. (2015) che

stimano la distribuzione della densità di popolazione a partire dai dati della telefonia mobile in rete in Europa.

Dati di sensori: I dati di sensori si riferiscono principalmente a qualsiasi tipo di output di un dispositivo che rileva e risponde a fonti di input dall'ambiente fisico. Il rapido sviluppo di Internet ha permesso la raccolta e la distribuzione di dati provenienti da sensori e ha dato origine alla cosiddetta "Internet of Things", la quale comprende qualsiasi elemento che può essere incorporato con sensori. L'accesso a Internet permette infatti alla "cosa" di trasmettere automaticamente i dati attraverso le reti e memorizzarli su cloud pubblici e database. Gartner, una società leader nella ricerca e nella consulenza informatica, prevede 25 miliardi di oggetti connessi entro il 2020. Questo aumento considerevole dl numero di sensori è dovuto al calo dei loro costi e allo sviluppo di nuove tecniche di fabbricazione. Ad esempio, la startup mCube sta utilizzando un nuovo metodo di fabbricazione per creare sensori di movimento che sono più piccoli di un granello di sabbia.

Questa vasta rete di sensori, genererà grandi quantità di nuovi dati che dovranno essere analizzati. I sensori misureranno il pianeta, trasformando il mondo fisico in flussi di dati. Secondo McKinsey and Company, i dati rappresenteranno una grande porzione di valore all'interno del mondo IoT. Ad esempio, i sensori possono essere utilizzati per monitorare il numero di persone che entrano nei negozi o il numero di veicoli commerciali che si spostano lungo percorsi specifici, o l'intensità di utilizzo dei macchinari; queste informazioni potrebbero essere rilevante per l'analisi di variabili come le vendite al dettaglio o le esportazioni.

Dati di immagini satellitari: Le immagini satellitari consistono in immagini della Terra o di altri pianeti raccolti dai satelliti. I satelliti sono gestiti sia da governi che da imprese di tutto il mondo. Secondo l'Index of Objects Launched into Outer Space, gestito dall'Ufficio delle Nazioni Unite per gli Affari Esteri (UNOOSA), all'inizio dell'anno 2019, erano 4 987 i satelliti in orbita intorno al pianeta, con un incremento del 2,68% rispetto alla fine di aprile 2018. Le immagini dei satelliti hanno molte applicazioni in meteorologia, oceanografia, agricoltura, silvicoltura, geologia, intelligence, guerra e altro. Recentemente, le immagini satellitari hanno attirato l'interesse anche degli economisti. Per esempio, le foto di case con tetti metallici possono indicare la transizione dalla povertà; luci notturne possono mostrare la crescita economica e il monitoraggio dei camion; e i movimenti delle consegne possono essere utilizzate per la previsione sulla produzione industriale.

Dati di prezzi scanner: I dati di prezzi scanner consistono in dati di scansione dei codici a barre forniti principalmente dai rivenditori. Tali prezzi sono sottoposti a scansione giornaliera e ciò consente una misurazione ad alta frequenza nel commercio al dettaglio.

Dati di prezzi online: Lo sviluppo di internet ha dato origine allo shopping online, permettendo ai prezzi online di essere utilizzati come sostituto o supplemento dei prezzi offline. La raccolta di dati online è chiamata web scraping.

I dati scanner e i dati di prezzi online sono uno strumento potenzialmente utile per previsioni sull'inflazione e sulle vendite al dettaglio.

Dati di ricerca online: I dati di ricerca online consistono in ricerche per particolari parole chiave sul web. L'utente inserisce una parola chiave (o una frase) nel campo di ricerca del motore di ricerca, il quale restituisce all' utente le informazioni che si riferiscono a tale parola chiave. I motori di ricerca elaborano le informazioni in tempo reale, grazie ad un algoritmo.

I dati di ricerca su Internet sono una fonte molto promettente di Big Data strutturati. La letteratura fornisce ampie prove scientifiche che i dati di ricerca sul web hanno capacità predittive in vari campi come l'economia, la finanza, la politica e la salute. Questo non è sorprendente, in quanto questo tipo di dati viene prodotto (ovvero, inserito sul web) direttamente dagli esseri umani e quindi è in grado di riflettere in maniera ottimale il comportamento degli agenti.

Dati testuali: Sono compresi tutti i tipi di dati che forniscono informazioni sotto forma di testo. Un esempio tipico di fonti di dati testuali sono gli archivi di giornali. Tuttavia, ci sono due difficoltà associate agli archivi: la prima è che gli archivi offrono immagini scansionate o fotografie di pagine di giornale, quindi è necessaria una trasformazione per creare una banca dati testuale; la seconda difficoltà è il linguaggio, in quanto i giornali nazionali usano il linguaggio ufficiale, ovvero la lingua di ogni paese.

Dati sui social media: I social media, come nel caso dei dati di ricerca online, riflettono le attività e le relazioni umane. Le discussioni o i post su Facebook, Twitter o Instagram includono una varietà di argomenti di carattere personale. Pertanto, sarebbe ragionevole supporre che, questo tipo di dati possa avere un ruolo chiave nella previsione di variabili economiche e sociali.

2.3 Il processo di scoperta della conoscenza dai database

Il processo di scoperta della conoscenza si riferisce essenzialmente all'identificazione di modelli da un set di dati pre-elaborato che siano:

Validi: I modelli sono in grado di incorporare anche nuovi dati con una data certezza, cioè sono generalizzabili (es. overfitting).

Utili: I modelli dovrebbero permettere di agire sull'oggetto dell'analisi, cioè poter essere azionabili, per esempio implementazione, costi di manutenzione, facilità d'uso.

Inaspettati: I modelli non dovrebbero essere ovvi per il sistema, ma al contrario interessanti. Inoltre essi dovrebbero mantenere un certo grado di equilibrio tra fiducia e scoperta.
Comprensibili: Gli umani dovrebbero essere in grado di interpretare tale modello, evitando che esso diventi una cosiddetta "black box".
Il processo si suddivide principalmente in due categorie di analisi:

- **Analisi predittiva** (apprendimento supervisionato): si tratta di prevedere il futuro sulla base di modelli appresi dai dati passati, avendo a disposizione set di dati etichettati. Essa si suddivide a sua volta in problemi di Classificazione (categoriale), i quali si avvalgono di variabili categoriche, e problemi di Regressione (continua), i quali si avvalgono di variabili numeriche.
- **Analisi descrittiva** (apprendimento non supervisionato): ovvero la descrizione di modelli all'interno di set di dati in cui nessuna etichettatura è richiesta. Alcuni esempi sono clustering, regole di associazione e regole di sequenza.

Un obiettivo puramente predittivo si concentra sull'accuratezza nella capacità predittiva e l'utente potrebbe non interessarsi se il modello rifletta o meno la realtà, fintanto che esso abbia un potere predittivo (ad esempio, un modello che combina gli attuali indicatori finanziari in modo non lineare per prevedere il futuro tasso di cambio dollaro-euro).
Un obiettivo puramente descrittivo si concentra sulla comprensione della generazione dei dati sottostanti e tale modello è interpretato come un riflesso della realtà (ad es. Un modello che mette in relazione le variabili economiche e demografiche ai risultati scolastici utilizzati come base per raccomandazioni sociali al fine di provocare un cambiamento).
Nella pratica, la maggior parte delle applicazioni richiedono un certo grado di entrambi gli obbiettivi, ovvero sia modellazione predittiva che descrittiva.
Il processo KDD (Knowledge Discovery in Database), al fine di trovare i modelli validi, si basa su metodi appartenenti a vari campi, tra cui la Statistica applicata, l'apprendimento automatico e le reti neurali, ed è riassumibile nei seguenti passaggi (Fayyad et al., 1996):

1. **Apprendere il campo di applicazione**: include le conoscenze pregresse rilevanti e gli obiettivi dell'applicazione.
2. **Creazione di un set di dati di destinazione**: include la selezione di un set di dati oppure concentrarsi su un sottoinsieme di variabili o campioni di dati (su cui la scoperta deve essere eseguita).
3. **Pulizia e pre-elaborazione dei dati**: comprende le operazioni di base, come la rimozione del rumore o degli outliers, la decisione della strategia per affrontare i valori mancanti o sconosciuti.

4. **Riduzione e proiezione dei dati**: include la ricerca di caratteristiche utili per rappresentare i dati, a seconda dell'obiettivo, e l'utilizzo della riduzione dimensionale o metodi di trasformazione per ridurre il numero di variabili in esame.
5. **La scelta della funzione di data mining**: include la scelta dello scopo del modello derivato dall'algoritmo di data mining (ad esempio, summarization, classificazione, regressione e clustering).
6. **La scelta dell'algoritmo o degli algoritmi di data mining**: include la scelta del metodo o dei metodi da utilizzare per la ricerca di modelli nei dati, ad esempio decidere quali modelli e parametri possono essere appropriati (ad esempio, modelli per dati categorici sono diversi dai modelli sui vettori su reali) e far corrispondere un particolare metodo di data mining ai criteri generali del processo KDD (ad esempio, l'utente può essere più interessato a comprendere il modello piuttosto che prediligere le sue capacità predittive).
7. **Data mining**: include la ricerca di modelli di interesse in una particolare forma rappresentazionale o un insieme di tali rappresentazioni, comprese le regole o gli alberi di classificazione, regressione, clustering, modellazione delle sequenze, dipendenza e analisi delle linee.
8. **Interpretazione**: include l'interpretazione dei modelli scoperti e la possibile visualizzazione dei modelli estratti, la rimozione di modelli ridondanti o irrilevanti, e la traduzione di quelli utili in termini comprensibili dagli utenti.
9. **L'utilizzo della scoperta**: include integrare tale scoperta nel sistema delle prestazioni (performance system), intraprendere azioni basate su di essa (o semplicemente documentare e segnalare alle parti interessate), nonché verificare e risolvere potenziali conflitti con precedenti scoperte.

Le organizzazioni sfruttano i dati memorizzati in questi laghi di dati utilizzando per lo più due approcci (Sambasivan et al., 2018). Il primo è un approccio top-down in cui un utente ha una richiesta (query) specifica o un'ipotesi da testare. Ad esempio, un'azienda di vendita al dettaglio potrebbe essere interessata a trovare una spesa aggiuntiva di 1000 dollari per il marketing digitale che aumenterà in modo significativo le entrate. Tali domande o ipotesi sono formulate a priori, prima di iniziare l'indagine. Questo tipo di analisi funziona quando gli utenti sanno esattamente cosa cercare all'interno del set di dati.

Il secondo approccio, invece, viene utilizzato per scoprire intuizioni che gli utenti non hanno cercato esplicitamente di determinare ma che vorrebbero scoprire dai soli dati, e viene chiamato Data mining.

2.3.1 Data Mining

Il data mining è un processo di scoperta di conoscenza all'interno dei database. Ovvero, da un enorme quantità di dati, incompleti, rumorosi, sfocati e casuali vengono estratte informazioni implicite che le persone non conoscono in anticipo e sono informazioni potenzialmente utili, non scontate (Wang e Han, 2016). Un'ampia varietà e numero di algoritmi di Data Mining sono descritti in letteratura, dai campi della Statistica, del pattern recognition, del machine learning e dei database. In particolare, l'algoritmo di data mining è composto da un insieme di tre elementi:

1. **Il modello** Un modello contiene parametri che sono da determinare in base ai dati. Esso è formato da due fattori rilevanti: la *funzione del modello* (ad esempio, classificazione o clustering) e la *forma rappresentazionale del modello* (ad esempio, una funzione lineare di variabili multiple o una funzione Gaussiana di densità di probabilità).
2. **Il criterio di preferenza** Rappresenta una base per determinare la preferenza di un modello (o un insieme di parametri) rispetto ad un altro, a seconda dei dati forniti. Il criterio è di solito una qualche forma di funzione di bontà di adattamento del modello ai dati, a volte temperato da un termine lisciante per evitare l'overfitting. Tipicamente, sono presenti un criterio quantitativo esplicito incorporato nell'algoritmo di ricerca (ad esempio, il criterio della massima probabilità di trovare i parametri che massimizzano la probabilità dei dati osservati) ed un criterio implicito che riflette il pregiudizio soggettivo dell'analista in termini di quali modelli sono inizialmente scelti per la considerazione.
3. **L'algoritmo di ricerca** La specificazione di un algoritmo per trovare particolari modelli e parametri. Trovare i parametri migliori è spesso ridotto ad un problema di ottimizzazione (ad esempio, trovare il massimo globale di una funzione non lineare nello spazio dei parametri).

Capitolo 3 – Statistica e Big Data

Indubbiamente la più grande sfida e opportunità che gli statistici di oggi devono affrontare è la crescita dei Big Data, la quale offre nuove opportunità nel campo dell'analisi dei dati.
Gli statistici hanno avuto molto da dire nel mondo del Data Science, poiché sono loro che hanno effettivamente coniato il termine stesso e hanno promosso l'aggiornamento della Statistica in favore di tale disciplina. Anche se diverse comunità possono condividere punti di vista contrastanti su quali discipline siano coinvolte nel Data Science, la Statistica, l'Informatica e la Scienza dell'Informazione sono i tre campi che vengono tipicamente visti come discipline chiave nello sviluppo del Data Science. In sintesi, la visione della Statistica è quella di un campo che si trova di fronte ad abbondanti nuove fonti di dati e problemi impegnativi da risolvere. Un'analisi statistica che sia valida per i Big Data sta diventando sempre più importante.

3.1 Problemi, sfide ed opportunità

3.1.1 La determinazione delle logiche di business e la comprensibilità dei modelli

L'estrazione fine a sé stessa non è possibile e può portare a risultati erronei. Definire chiaramente i problemi aziendali e determinare gli scopi di tale estrazione di dati, insieme all'uso della conoscenza del dominio, sono di fondamentale importanza per la pratica del Data Mining ed in generale per l'intero processo di estrazione di conoscenza dai dati. Occorre porre maggiore enfasi sull'interazione uomo-macchina, piuttosto che sull' automazione totale, poiché molti metodi e strumenti non sono interattivi ed in grado di supportare gli utenti. Al contrario, è necessario lo sviluppo di strumenti per la visualizzazione, l'interpretazione e l'analisi dei modelli scoperti. Tali strumenti interattivi possono consentire a molti utenti di trovare soluzioni pratiche riguardo a problemi del mondo reale, molto più rapidamente rispetto ad esseri umani o computer che operano in modo indipendente (concetto di integrazione). Per esempio, gli approcci bayesiani usano le probabilità precedenti come un modo per codificare la conoscenza passata e sono allo stesso tempo facilmente interpretabili dall'essere umano.

3.1.2 La distribuzione della popolazione diventa quella effettiva

I metodi tradizionali statistici di raccolta dati si basano principalmente su indagini, come questionari, interviste telefoniche o rapporti statistici. Tuttavia, l'accuratezza di questi metodi non può essere garantita ed il costo è piuttosto elevato. Nell'era dei Big Data, grandi quantità di dati diventano facilmente reperibili, ad un costo drasticamente ridotto e caratterizzati da un'accuratezza sempre maggiore. Infatti, grazie alla completezza dei grandi campioni di dati,

viene garantita l'accuratezza dei risultati statistici, riducendo l'errore umano nel processo statistico. Mentre il tradizionale processo inferenziale statistico si basa sulla teoria della distribuzione della popolazione, la quale viene indotta tramite inferenza sul campione rispettando le leggi probabilistiche, con i Big Data è possibile applicare la teoria probabilistica direttamente sull' effettiva distribuzione della popolazione. Non è quindi necessario inferire le caratteristiche generali della distribuzione, ma vengono direttamente utilizzate la distribuzione effettiva e le caratteristiche effettive della popolazione.

3.1.3 Il problema dei dati mancanti

Questo problema è particolarmente accentuato all'interno delle banche dati aziendali. Alcuni attributi importanti possono essere omessi se il database non è stato concepito per la scoperta di conoscenza. I dati mancanti possono derivare da errori dell'operatore, effettivi errori di sistema e di misura, oppure da una revisione nel tempo del processo di raccolta dei dati (ad esempio, vengono misurate nuove variabili, ma sono state considerate irrilevanti pochi mesi prima). Le possibili soluzioni includono strategie statistiche più sofisticate per identificare variabili e dipendenze nascoste.

3.1.4 La gestione di dati e conoscenze in continuo cambiamento

I dati in rapida evoluzione (non stazionari) possono rendere invalidi i modelli precedentemente scoperti. Inoltre, le variabili misurate in un certo database applicativo possono essere modificate, cancellate o incrementate con nuove misurazioni nel corso tempo. Le possibili soluzioni includono metodi incrementali per l'aggiornamento dei modelli e trattare il cambiamento come un'opportunità di scoperta, utilizzandolo per cercare modelli di cambiamento.

3.1.5 Rumore, false relazioni ed effetto "look everywhere"

Mentre gli scienziati passano da un approccio basato su ipotesi ad un approccio basato sui dati, il numero di scoperte errate (ad esempio, geni che sembrano essere collegati ad una malattia, ma in realtà non lo sono) è destinato ad aumentare, a meno che non vengano prese precauzioni specifiche. La statistica classica fornisce metodi per analizzare i dati quando il numero di variabili (p) è piccolo e il numero di osservazioni (n) è grande. Al contrario, in molte applicazioni dei Big Data, si presenta il caso inverso, ovvero p è più grande di n. In entrambe le situazioni, l'obiettivo è quello di sviluppare un modello che descriva come una variabile

risultato sia correlata alle altre variabili e determinare quali variabili sono importanti per caratterizzare tale relazione. Il modello dipende da parametri, uno per ogni variabile, che quantificano la relazione. L'adattamento del modello ai dati comporta la stima dei parametri a partire dai dati e la valutazione dell'evidenza che essi siano diversi da zero (ovvero che sia una variabile rilevante). Quando p è più grande di n, il numero di parametri è enorme rispetto alle loro informazioni presenti nei dati. Perciò, migliaia di parametri irrilevanti appariranno statisticamente significativi se si utilizzano statistiche su un numero ridotto di osservazioni n.
Nella Statistica classica, se i dati contengono qualcosa che ha una probabilità di accadere, possiamo essere sicuri che non è lì per caso. Ma se si guarda in mezzo un milione di "posti", come nel mondo dei Big Data, improvvisamente non è così insolito fare una scoperta del tipo "uno su un milione". Il caso non può più essere liquidato come spiegazione. Gli statistici lo chiamano effetto "look-everywhere", ed è un grosso problema che colpisce gli approcci scientifici basati sui dati, piuttosto che sulle ipotesi. Gli statistici hanno già trovato diversi modi per affrontare l'effetto "look-everywhere". La maggior parte dei set di dati ha solo poche forti relazioni tra le variabili, e tutto il resto è rumore, in inglese "noise". Perciò la maggior parte dei parametri semplicemente non dovrebbe avere importanza. Un modo per renderli irrilevanti è assumere che tutti i parametri, tranne pochi, siano uguali a zero. Alcuni progressi degli ultimi anni hanno reso questo metodo particolarmente promettente per estrarre una piccolissima quantità di informazioni significative da un'enorme quantità di dati. Una di queste tecniche si chiama minimizzazione L1, o LASSO, e venne inventata da Robert Tibshirani nel 1996. La minimizzazione L1, per esempio, nel campo dell'elaborazione delle immagini, permette l'estrazione di un'immagine a fuoco nitido da dati sfumati o rumorosi.

3.1.6 La presenza di tipologie di dati differenti

Le statistiche tradizionali ritengono che i dati derivino principalmente da valori numerici di test, esperimenti o sondaggi. Nell'era dei big data, possono essere presi come oggetto di ricerca statistica non solo quantità misurabili come i dati strutturati, ma possono essere presi anche dati semi-strutturati e non strutturati, i quali non possono essere misurati da relazioni quantitative.
I Big Data non sono solo grandi, sono complessi e si presentano in forme diverse da quelle a cui sono abituati normalmente gli statistici, ad esempio immagini o reti, come visto nella parte introduttiva sui big data. Una sfida importante si basa nel riuscire a trattare, attraverso i consoni strumenti, queste numerose tipologie di dati differenti.

3.1.7 Problemi di scala, il fattore “tempo” ed il ruolo dell’informatica

La statistica classica è sempre stata fatta in modalità offline; molte delle teorie si sono sviluppate in un'epoca (i primi del 1900) in cui il “mondo online” non esisteva nemmeno.
Il tempo di analisi era essenzialmente illimitato. I dati erano comunque piccoli, quindi lo statistico non ha mai dovuto pensare se i calcoli fossero stati fatti in modo efficiente. Oggi è diverso, i database sono caratterizzati da un gran numero di attributi e variabili. Per esempio, le aziende web fanno i loro soldi cercando di prevedere le reazioni degli utenti e indurre gli utenti a determinati comportamenti, ad esempio cliccare su un annuncio sponsorizzato da un cliente. Per prevedere il tasso di risposta è necessario costruire un modello statistico che consideri una grande quantità di n (milioni di click) e una grande quantità di p (migliaia di variabili, per esempio dove posizionare la pubblicità all’interno della pagina). Ciò potrebbe anche essere fatto con l’utilizzo delle tecniche statistiche classiche, poiché il numero di n è comunque superiore al numero di p. Tuttavia, l'azienda Web, a cui potrebbe interessare questo tipo di analisi, potrebbe avere solo millisecondi per decidere come rispondere a un determinato clic dell'utente. Inoltre, il modello deve costantemente cambiare per adattarsi a nuovi prodotti e nuovi clienti.
Per affrontare il problema del tempo, gli statistici hanno iniziato ad adattarsi adottando le idee degli informatici, per i quali la velocità è sempre stata un problema ed una priorità. L'obiettivo in alcuni casi può non essere quello di fornire una risposta perfetta, ma di fornire una buona risposta in tempi rapidi, progettando nuovi algoritmi che presentino un trade-off tra accuratezza teorica e velocità. Infatti, molti dei più diffusi algoritmi per l'analisi statistica non si scalano molto bene e funzionano lentamente su set di dati di grandi dimensioni. Gli attuali metodi di apprendimento automatico in generale richiedono ancora una notevole esperienza umana nell'ideare caratteristiche, priorità, modelli e algoritmi appropriati. Ancora molto lavoro deve essere fatto per rendere il ML più ampiamente utilizzato ed, infine, diventare una parte comune degli strumenti quotidiani nel Data Science.
Allo stesso tempo gli statistici devono comunque continuare a pensare da statistici. Infatti, come sottolineato più volte nei capitoli precedenti, gli statistici sono in grado di comprendere l’incertezza, sono qualificati in modo unico per fare inferenza e sono in grado di estrarre connessioni tra variabili di dati al fine di determinare quali sono reali e quali sono false. Per esempio, trovare modelli falsi che non sono generalmente validi, attraverso soluzioni possibili che includono algoritmi molto efficienti, metodi di campionamento approssimativi, grandi elaborazioni parallele (parallel processing), tecniche di riduzione dimensionale e integrazione di conoscenze pregresse.

3.1.8 Il rischio di overfitting

Quando un algoritmo cerca i parametri migliori per un particolare modello utilizzando un insieme di dati, potrebbe sovraccaricare i dati, con conseguenti scarse prestazioni del modello sui dati di prova (dati test). Possibili soluzioni includono la validazione incrociata, la regolarizzazione e altri tipi di sofisticate strategie statistiche.

3.1.9 Privacy e riservatezza

Questa è probabilmente l'area di maggiore preoccupazione pubblica per i Big Data, e gli statistici non possono permettersi di ignorarla. I dati possono essere resi anonimi per proteggere le informazioni personali, ma non esiste una sicurezza perfetta.

3.1.10 Educazione statistica

Per quanto riguarda l'educazione statistica, il numero di studenti è raddoppiato nel corso degli ultimi dieci anni. Tuttavia, è necessario che gli studenti vengano esposti maggiormente a dati che non possono essere analizzati usando i metodi statistici tradizionali, al fine di prepararli al mondo Big Data di cui inevitabilmente faranno parte. Inoltre, risulta opportuno, ai fini dell'apprendimento, avvicinare fin da subito gli studenti a dati provenienti dal mondo "reale", in modo che essi vedano l'applicabilità dei tali dati a problemi reali. Infine, sarà sempre più indispensabile che gli studenti di statistica ricevano un'adeguata preparazione informatica con lo scopo di sviluppare le competenze informatiche necessarie per l'applicazione dei loro strumenti.

Conclusioni

L’avvento dei Big Data ha un significato epocale per la Statistica, poiché migliora la qualità del metodo statistico, ne abbassa i costi e ne amplia la gamma di applicazioni. Infatti, l’elevato volume dei dati può portare ad una migliore accuratezza e a maggiori dettagli dell’analisi; l’alta velocità della loro elaborazione può portare a stime statistiche più frequenti e più tempestive e l’elevata varietà di dati raccolti può offrire opportunità statistiche in nuovi settori, espandendone cosi l’applicabilità. Tuttavia, i set di dati sono così grandi e complessi che i tradizionali metodi di elaborazione dei dati diventano insufficienti per poterli acquisire, memorizzarli e analizzarli. Infatti, molti metodi statistici tradizionali, efficienti per i dati a bassa dimensione risultano inadatti all’analisi di dati ad alta dimensione.

I Big Data, inoltre, generano caratteristiche uniche che non sono condivise dai set di dati originali, pertanto al fine di progettare procedure statistiche efficaci per l'esplorazione e la previsione di Big Data, dobbiamo affrontare alcune importanti sfide:

- l'alta dimensionalità porta ad accumulo di rumore e alla determinazione di false relazioni tra parametri;
- l'alta dimensionalità combinata con l'ampia dimensione del campione crea problemi quali gli alti costi di calcolo, l’overfitting e l’instabilità algoritmica;
- i grandi campioni nei Big Data sono tipicamente aggregati da fonti diverse, in momenti diversi ed utilizzando tecnologie diverse.

Per gestire tali criticità, nasce l’esigenza di sviluppare nuovi metodi statistici e metodi computazionali in grado di gestire i Big Data. Tale sfida può consistere nel dover riadattare la Statistica classica attraverso nuove idee, sviluppando tecniche in grado di far fronte al rapido cambiamento. Lo sviluppo dei metodi bayesiani, trattato nel primo capitolo, rappresenta un chiaro esempio di risposta a tale esigenza: essi permettono di proteggere i modelli contro l’overfitting, di analizzare i dati quando le relazioni probabilistiche causa-effetto tra variabili sono complesse e di integrare nel modello informazioni provenienti da fonti diverse. La diffusione dei metodi bayesiani è stata possibile in gran parte grazie allo sviluppo di appositi software di calcolo che hanno risolto il problema del carico computazionale, legato all’integrazione di variabili continue nel modello. Infatti, la scalabilità dei metodi statistici sia per l'alta dimensionalità che per le grandi dimensioni del campione dovrebbe essere presa in seria considerazione nello sviluppo di procedure statistiche.

In un’era in cui i modelli devono riuscire a trarre le conclusioni in tempi assai contenuti e devono costantemente cambiare per adattarsi a nuove esigenze (ad esempio nuovi prodotti e nuovi clienti) risulta fondamentale bilanciare l'accuratezza statistica e l'efficienza

computazionale. L'obiettivo in alcuni casi può non essere quello di fornire una risposta perfetta, ma di fornire una buona risposta in tempi rapidi, progettando nuovi algoritmi che presentino un trade-off tra accuratezza teorica e velocità. Gli statistici hanno infatti iniziato a condividere l'approccio degli informatici, per i quali la velocità è sempre stata un problema e allo stesso tempo una priorità. Questo cambiamento di paradigma ha portato a progressi significativi nello sviluppo di algoritmi veloci e scalabili in grandi quantità di dati ad alta dimensionalità.

Infine, un'ulteriore sfida riguarda l'approccio all'insegnamento delle Scienze Statistiche, stante la necessità di individuare nuovi metodi di analisi dei dati volti ad utilizzare efficacemente le risorse fornite dai Big Data. Per esempio, diventa necessario integrare la natura non campionaria che è propria dei Big Data nel sistema statistico classico, con l'obbiettivo di ottenere un campione che rappresenti correttamente la popolazione. Ciò potrà avvenire ad esempio esponendo maggiormente gli studenti a dati che non possono essere analizzati usando i metodi statistici tradizionali, avvicinandoli fin da subito a dati provenienti dal mondo "reale", e facendo in modo che ricevano un'adeguata preparazione interdisciplinare, che combini competenze specifiche per l'ambito d'interesse con abilità statistico-matematiche ed informatiche.

Bibliografia

- ABBONDANZA, I.M. et al., 2012. *Approccio bayesiano nelle contrattazioni di roaming wholesale internazionale.* Notiziario tecnico: TelecomItalia.
- BORRA, DI CIACCIO, 2014. *Statistica: metodologie per le scienze economiche e sociali.* McGraw-Hill Education.
- BUONO, D., et al., 2017. *Big data types for macroeconomic nowcasting.* Eurostat Review on National Accounts and Macroeconomic Indicators.
- CAPGEMINI e BNP PARIBAS, 2016. *World Payments Report*. In: BUONO, D., et al., 2017. *Big data types for macroeconomic nowcasting*.
- FAN, J., et al., 2014. *Challenges of Big Data analysis.* National Science Review, Volume 1, p.293–314.
- FAYYAD, U., PIATETSKY-SHAPIRO, G., SMITH, P., 1996. *The KDD Process for Extracting Useful Knowledge from Volumes of Data*, Vol. 39, No. 11.
- GELMAN, A., et al., 1995. *Bayesian Data Analysis*. Chapman & Hall/CRC.
- IANNAZZO, S., 2007. *Le tecniche statistiche bayesiane e la loro applicazione in modelli di simulazione probabilistica.* Farmeconomia e percorsi terapeutici p.*5-13.*
- LANEY, D., 2001. *3-D data management: Controlling data volume, velocity and variety Application Delivery Strategies by META Group Inc.* p. 949.
- MADIGAN, D., et al., November 11 and 12, 2013. *A report of the London workshop on the future of statistical sciences.* London.
- MAYER-SCHOENBERGER, V., CUKIER, K., 2013. *Big Data. A Revolution that will transform how we live, work, and think.* London: John Murray Publishers.
- RICCIATO, F., et al., 2015. *Estimating population density distribution from network-based mobile phone data.* European Commission, JRC technical report.
- SAMBASIVAN, R., et al., 2018. *A Bayesian Perspective of Statistical Machine Learning for Big Data.*
- SPRENT, 2003. *Modern medical statistics: A practical guide.* Journal of the Royal Statistical Society: p.183-198.
- TECHAMERICA FOUNDATION'S FEDERAL BIG DATA COMMISSION, 2012. *Demystifying big data: A practical guide to transforming the business of Government.*
- VAN DIJCK, 9 Maggio 2014. *Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology.* University of Amsterdam, The Netherlands.

- WANG, C., HAN, D., 2016. *Data Mining Technology Based on Bayesian Network Structure Applied in Learning*. College of information engineering, Huanghuai University, Henan, China. International Journal of Database Theory and Application, Vol.9, No.5 pp.267-274.
- ZHANG, J., 2017. *The Development and Application of Bayesian Networks Used in Data Mining Under Big Data.*

Sitografia

- Enciclopedia Treccani online. *Definizione di inferenza statistica*. Disponibile su: http://www.treccani.it/enciclopedia/inferenza-statistica/. [Ultimo accesso: 25/05/2019]