Società Italiana di Statistica

30 marzo 2018

STATEMENT DELLA SIS SU STATISTICA, SCIENZA DEI DATI E BIG DATA

Dopo il convegno SIS Statistics and Data Science del giugno 2017 a Firenze, abbiamo ritenuto opportuno fare il punto sul ruolo della Statistica nella produzione ed analisi dei Big data e nella Scienza dei dati.

Infatti, l'uso massivo delle tecnologie informatiche per la generazione, memorizzazione, analisi e trattamento dei Big data e nella Scienza dei dati non garantisce di per sé la loro qualità statistica e non deve privilegiare solo gli aspetti pratici e commerciali ma perseguire soprattutto finalità scientifiche.

Offriamo questo documento alla discussione comune e al lavoro del Gruppo SIS Statistics and Data Science, costituito proprio in occasione del convegno del 2017, con l'obiettivo di incoraggiare la collaborazione tra discipline e lo sviluppo di metodi di produzione e di analisi dei dati che siano efficaci per la conoscenza scientifica dei fenomeni.

La Statistica è anche Scienza dei dati in continua evoluzione.

La Statistica è una disciplina scientifica che studia i metodi di raccolta, di organizzazione, di analisi, d'interpretazione e presentazione dei dati, ed è riconosciuto il suo ruolo preminente nel progresso della conoscenza in qualsiasi campo del sapere. Già da oltre un secolo, ha sviluppato anche metodologie per gestire ed analizzare una grande mole di dati, come quelli rilevati in occasione dei Censimenti e di indagini campionarie su vasta scala o presenti in archivi amministrativi, definendo i principi che stabiliscono l'attendibilità dei dati raccolti (profilo di errore) e l'attendibilità dei risultati delle analisi.

Lo sviluppo dei metodi per l'analisi statistica e demografica è sempre stato affiancato dalle tecnologie di memorizzazione e di calcolo più moderne per l'epoca, che restituissero analisi efficaci e complesse dei dati trattati: interazione di fonti, integrazione di dati, fusione di dati, data mining, analisi dei dati funzionali, analisi multivariate, simulazioni, algoritmi di stima, e così via, non sarebbero stati possibili senza un continuo dialogo e scambio con gli studiosi di Informatica. Tanto che già nella seconda metà del secolo scorso si parlava di data analysis e data driven analysis, di learning by data, di data science e ci si interrogava sulla validità delle conoscenze raggiunte.

Lo sviluppo della Scienza dei dati nell'era dei Big data

La Scienza dei dati, Data Science, o data-driven science, è l'insieme di principi metodologici e tecniche multidisciplinari volti ad estrarre conoscenza dai dati. Ingredienti fondamentali sono la Statistica, insieme all'Informatica e alle tecniche per il trattamento informatico dei dati e ai domini applicativi di interesse.

L'attenzione alla Scienza dei dati è oggi accentuata dall'esplosione del volume, della varietà e della velocità dei cosiddetti Big data.

Essi sono, ad esempio, ottenuti dalle tracce digitali lasciate dall'uso di telefoni mobili, dai GPS, tablet, lettori di carte magnetiche, di codici a barre, telecamere, satelliti, Internet, come pure dalla registrazione a siti per l'erogazione di servizi amministrativi e la partecipazione a social network. Laboratori pubblici e privati raccolgono

e gestiscono grandi masse di dati per finalità scientifiche e di sviluppo tecnologico e anche per controllare processi, di carattere industriale o istituzionale.

I dati così prodotti riguardano i più diversi ambiti delle scienze naturali e sociali, sono big anche per complessità (testi, immagini diagnostiche e satellitari, segnali…) e possono fornire informazioni su eventi, comportamenti e relazioni finora poco visibili.

Non è però chiaro secondo quali paradigmi la cosiddetta Scienza dei dati potrebbe e dovrebbe analizzarli per portare beneficio alla società e alla vita di tutti, come Melchiorre Gioia auspicava per la Statistica due secoli fa.

Big data o semplicemente dati

Molte tipologie di Big data non sono una novità, basti pensare al già citato esempio sui Censimenti. Ma anche se di una nuova tipologia, i Big data sono semplicemente dati, e devono rispondere a requisiti di qualità con un profilo di errore noto e controllabile, almeno probabilisticamente.

La loro qualità statistica non è però garantita dall'uso massivo delle tecnologie informatiche per la loro generazione, memorizzazione, analisi e trattamento.

La modalità di generazione dei cosiddetti Big data può essere così complessa e multiforme che spesso tali dati sono incompleti, selettivi e distorti, nonostante il loro volume. Un alto numero di osservazioni non garantisce assenza di selettività e neppure una esaustiva osservazione del fenomeno. Alta dimensionalità dei dati, velocità di raccolta e di registrazione unite al rischio di errori variabili e sistematici rendono necessari nuovi e robusti approcci di analisi statistica.

Infatti, nei dati comunque generati c'è una componente di incertezza e casualità. Tale casualità - indotta dalle modalità di produzione dei dati, dal disegno degli esperimenti, dalla natura o dall'errore di misura o semplicemente dall'errore umano - è importante perché la sua rappresentazione tramite modelli probabilistici permette ai ricercatori di studiare il processo sottostante che genera i dati e di quantificare l'incertezza specificando modelli statistici adeguati.

Inoltre, i Big data non sono in alcun modo sostitutivi dei dati tradizionali. Ci sono molti fenomeni per i quali i Big data non sono rilevabili. Lo stesso progresso tecnologico, che ha portato ai Big data, rende migliori, più veloci, versatili ed affidabili anche le tradizionali raccolte dei dati con interviste da cellulare, via Web o Internet oppure tramite interrogazioni di dati pubblici e/o amministrativi.

È necessario investire, anche nella produzione di statistica ufficiale, nell'integrazione di diverse fonti di dati usando un cosiddetto approccio multi-source. La tradizionale inferenza basata su disegni di campionamento probabilistici caratteristica di molte indagini su larga scala potrà essere integrata con l'inferenza assistita e/o basata su modelli e anche con gli strumenti dell'inferenza algoritmica, caratteristica dei metodi di machine learning.

Invece che fissare l'attenzione solo sulle innovazioni dei Big data, è il momento di considerare le innovazioni di tutte le tipologie di dati usando tutte le fonti sia tradizionali sia nuove al fine di avere un quadro quanto più possibile chiaro del fenomeno studiato per produrre nuova conoscenza.

Il futuro della Scienza dei dati

Lo sviluppo della Scienza dei dati è possibile solo se l'insieme delle competenze statistiche, delle tecniche per il trattamento informatico dei dati e dell'Informatica saranno usate non solo per finalità pratiche e commerciali ma anche per finalità scientifiche.

Big data, dati amministrativi, di laboratori per controlli di processo, sono di grande interesse per finalità istituzionali, imprenditoriali, di marketing e di vendita e attualmente oggetto prevalente delle applicazioni di data

analytics. Queste devono saper affrontare in modo statistico problemi come eterogeneità, accumulazione e propagazione di errori, correlazioni spurie, e endogeneità incidentali.

È importante che i risultati ottenuti siano riproducibili da altri ricercatori con altre – diverse – fonti di dati. Il ragionamento statistico può aiutare in questo ambito a distinguere causazione e correlazione e così di identificare gli interventi che modificherebbero gli esiti del fenomeno studiato.

È necessario definire un profilo di errore totale, distinguendo le varie possibili fonti di errore nel processo produttivo dei dati al fine di valutare l'attendibilità e validità delle conoscenze ottenute.

La specificazione di questo profilo distingue l'uso statistico scientifico dei dati, e quindi anche dei Big data, dagli altri usi.

L'auspicio è che la ricerca in questo campo riguardi oltre agli aspetti computazionali maggiormente quelli più prettamente statistici, per rendere il ragionamento statistico più efficiente e avere informazioni utili per prendere decisioni in base alla nuova tipologia di dati.

Scienza dei dati o Scienze dei dati?

La base comune della Scienza dei dati è costituita dai futuri paradigmi di errore e protocolli di analisi statistica dei dati e di machine learning, e anche dalle conoscenze informatiche sul management dei data base, e sui sistemi distribuiti e paralleli. Gli statistici devono sviluppare queste conoscenze lavorando a stretto contatto da un lato con gli sviluppatori di software e gli informatici e dall'altro con gli esperti dei domini applicativi.

Infatti, l'analisi dei dati a fini scientifici richiede una conoscenza dettagliata e specifica dei singoli campi di applicazione, e le metodologie di generazione dei dati e della loro analisi si prefigurano diverse al cambiare dei settori applicativi. Per questo motivo è opportuno usare il plurale e parlare di Scienze dei dati o di Data Sciences.

Del resto, la ricchezza di informazioni che caratterizza l'attuale offerta di dati e di metodologie statistiche permette di fornire il giusto dato e la metodologia appropriata per il target d'interesse.

Ad esempio, quando il focus è sui mercati finanziari, quotazioni e transazioni su tali mercati e le metodologie di previsione possono essere il giusto dato e l'analisi statistica appropriata; nel caso di studi giuridici, i testi dei documenti ed il loro corpus sono i dati e l'analisi testuale costituisce un utile strumento di Fraud Detection; i post sui social network e la Sentiment analysis sono la base per il calcolo di indicatori di vari fenomeni: dalla felicità della popolazione alla reputazione online della propria azienda.

La riflessione finale è che occorre una efficace collaborazione tra ambiti disciplinari, sia per definire ed attivare adeguati percorsi formativi sia per svolgere ricerca sui temi di frontiera prima indicati. Ciò consentirà lo sviluppo delle Data Sciences e l'applicazione rigorosa del linguaggio e dei principi della Statistica alla generazione e all'analisi dei Big data per ottenere dati di cui si conosca la qualità, che siano riproducibili, e consentano di effettuare analisi scientifiche.

Una prima bozza di questo documento, predisposta dal Presidente della SIS è stata discussa con i seguenti Soci: Francesco Billari, Università Bocconi, Paolo Giudici, Università degli Studi di Pavia, Michele La Rocca, Università degli Studi di Salerno, Fulvia Mecatti, Università degli Studi di Milano-Bicocca, Antonietta Mira, Università dell'Insubria e Università della Svizzera Italiana a Lugano, Piercesare Secchi, Politecnico di Milano, Nicola Torelli, Coordinatore del Gruppo SIS SDS, Università degli Studi di Trieste; e nell'ambito del Consiglio Direttivo della SIS.