3 · 4 LUGLIO 2024

Data Science e Statistica Uifficiale: Opportunità e Criticità

Marco Alfò

Dipartimento di Scienze Statistiche,
Sapienza Università di Roma

Daniela Cocchi, Brunero Liseo, Maria Giovanna Ranalli

Conferenza Nazionale di **Statistica**

La statistica ufficiale nel tempo dell'Intelligenza Artificiale

#CNStatistica15

Indice

Al di là di (quasi) inutili definizioni..

La Data Science rappresenta un'evoluzione fondamentale

Offre nuovi strumenti e metodologie per la raccolta, l'analisi e la diffusione di informazioni cruciali per la società,

Produce evidenti benefici:

- aumento del volume e varietà dei dati
- analisi predittive più approfondite ed efficaci
- produzione di statistiche «smart» e in tempo reale
- comunicazione più efficace e coinvolgente
- notevoli applicazioni, già sperimentate, in statistica ufficiale

Data Science e Statistica Ufficiale

Progetti (inter)nazionali

- Statistiche in tempo reale sul mercato del lavoro, web scraping e text analysis su offerte di lavoro
- Stime dell'inflazione, dati scanner supermercati e tecniche di machine learning per la stima dell'indice dei prezzi al consumo
- Censimenti «virtuali», registri individuali e fonti amministrative
- Fiducia economica, economia digitale, sentiment analysis, web scraping, dati social media
- Stima della povertà, statistiche agricole immagini satellitari
- Previsioni economiche, tecniche di machine learning per stime in tempo reale
- Mobilità e turismo, dati roaming e di telefonia mobile
- e molti altri…

Opportunità

- **Qualità e tempestività delle statistiche ufficiali**
- Integrazione di diverse tipologie di dati
- **Integrazione di diverse fonti di dati**
- Dettaglio e granularità delle analisi
- Sviluppo di nuovi indicatori e misure
- Accuratezza tecniche di previsione
- Rivisitazione processo raccolta dati
- Comunicazione più efficace e diretta

Opportunità

Qualità e tempestività

- Utilizzo di dati raccolti in tempo reale o quasi-reale, ad es. dati transazioni, sensori, IoT, etc.)
- Automazione del processo di raccolta ed analisi dei dati, anche a livelli fini di dettaglio
- Identificazione (e correzione?) automatica di dati inconsistenti, anomali, errori, etc
- Analisi mancate risposte, integrazione con fonti amministrative, metodi di machine learning per l'imputazione (multipla?)
- Riduzione ritardo di pubblicazione, frequenza di aggiornamento più rapida
- Riduzione carico rispondenti
- …

Opportunità

Integrazione di diverse fonti di dati

- Possibilità di «appaiare» più fonti informative grazie a tecniche più efficaci da un punto di vista computazionale
- Utilizzo di dati strutturati (es fonti amministrative) e non strutturati (immagini, testi, frasi, sensori, social media, etc)
- Copertura più ampia del fenomeno, ad un livello di dettaglio maggiore
- Riduzione del costo di rilevazione
- Validazione incrociata delle informazioni
- Gestione più generale dell'informazione mancante
- Molte le sfide a livello metodologico, tecnologico, etico, legale, normativo…

Criticità

- **Integrazione di fonti di dati eterogenee**
- **Rappresentatività e distorsione**
- Nuove competenze e formazione del personale
- Questioni etiche e di privacy nel trattamento dei dati
- **Machine learning, il problema del «black box»**
- Comunicazione di informazioni complesse al pubblico
- …

Criticità

Integrazione di fonti di dati eterogenee

- Differenze nei formati dei dati, nelle definizioni, nella qualità, nel livello di aggregazione e nella frequenza di rilevazione
- Coerenza temporale
- Linkage e scalabilità
- Sovra- e sotto-copertura, differenze negli universi di riferimento
- Metodologie di inferenza da dati integrati
- Gestione dell'incertezza
- Tracciabilita' e **Riproducibilità**
- …

Criticità

Rappresentatività e Distorsione

- Sotto- o sovra-rappresentazione di gruppi specifici della popolazione di riferimento
- Campionamento non probabilistico
- Distorsione da selezione, auto-selezione, demografico, tecnologico, etc.
- Equità ed inclusività, rischio di ampliamento disuguaglianze esistenti
- Documentazione, trasparenza e comunicazione
- Difficoltà di quantificazione della distorsione e nell'utilizzo di strumenti di correzione, post-stratificazione e calibrazione
- Necessità di sviluppo competenze, formazione
- …

Criticità

Machine Learning, il problema del «black box»

- Il termine «black box» si riferisce a metodi di decisione/previsione di cui è difficile comprendere il funzionamento
- Tali metodi possono produrre risultati accurati, ma il modo in cui pervengono a tali risultati può non essere chiaro
- Interpretabilità, Trasparenza, Replicabilità sono argomenti da considerare attentamente perché legati a
- Responsabilità, Conformità normativa e Fiducia Pubblica
- Selezione di tecniche interpretabili, interpretazione post-hoc (tramite apposite misure), **explainable** AI, validazione esterna da metodi «tradizionali»
- Documentazione, formazione e competenze trasversali
- …

3 · 4 LUGLIO 2024

grazie per l'attenzione

NOME COGNOME

Qualifica | Ente

CNS15ª

Conferenza Nazionale di **Statistica**

La statistica ufficiale nel tempo dell'Intelligenza Artificiale

#CNStatistica15