STATISTICAL LEARNING PER I BIG DATA

a cura della prof.ssa Anna Maria Paganoni – Statistica

Descrizione del corso

"I keep saying the sexy job in the next ten years will be statisticians. [...] The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a hugely important skill in the next decades. [...] Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it." [Hal Varian, Google's Chief Economist, in an interview to The New York Times, Jan. 2009]

Lo statistical learning, o apprendimento statistico dai dati, costituisce un insieme un set di strumenti essenziali per acquisire conoscenze da insiemi vasti e complessi di dati (big data) emersi in campi che vanno dalla biologia alla finanza, dal marketing all'astrofisica, dalle scienze sociali alla medicina negli ultimi vent'anni. In questo corso si introdurranno in modo critico alcune tecniche di statistical learning e di studio dell'incertezza, cercando di sfatare i principali pregiudizi e luoghi comuni sulla statistica.

Organizzazione

Il corso è suddiviso in 4 moduli; durante ogni modulo gli studenti avranno accesso a diversi **materiali didattici online**, tra cui videolezioni, slide, schede di esercitazioni/laboratori informatici. Gli studenti saranno in contatto costante con docenti e tutor del Politecnico. Inoltre, grazie a un **forum online** potranno lavorare insieme agli altri iscritti all'interno di una classe virtuale. È prevista infine anche la partecipazione a **webinar**, tenuti direttamente da un docente del corso. Alla fine dei 4 moduli, coloro che avranno completato le attività del corso riceveranno un **attestato** di partecipazione e un **badge digitale**, da allegare al proprio cv.

Struttura del corso

Modulo 1 – Descrivere e raccontare i dati

Nel primo modulo si introdurrà e discuterà le varie tipologie di dati statistici ed i principali strumenti per sintetizzare, descrivere e rappresentare una distribuzione di dati univariata e bivariata: indici di posizione, di dispersione, di correlazione, profondità, grafici a barre, a torta, istogrammi e boxplot. Si tratterà come comunicare il contenuto informativo di un dataset, facendo particolare attenzione ai possibili errori nella comunicazione. Si introdurrà il concetto di campione statistico e di dimensione campionaria discutendo il relativo concetto di rappresentatività, per arrivare a introdurre le basi della modellazione matematica dell'incertezza, del calcolo delle probabilità e dell'inferenza statistica.

Modulo 2 – Condurre un'analisi descrittiva su un dataset

Nel secondo modulo si imparerà ad implementare i concetti teorici introdotti nel modulo precedente mediante l'utilizzo di R, un software libero dedicato alle analisi statistiche (https://cran.r-project.org/). Si imparerà ad importare un dataset, a controllare la coerenza della matrice dei dati, a calcolare i principali indici di sintesi, e a costruire le opportune visualizzazioni dei dati, gestendo anche casi di confronto tra più popolazioni o dataset multivariati. Le analisi verranno svolte su dataset reali, scelti per imparare ad utilizzare gli strumenti statistici appresi per rispondere a semplici domande di ricerca sul fenomeno che ha generato i dati oggetto di studio.

Modulo 3 – Costruire un modello statistico

Nel terzo modulo si imparerà il significato di modello statistico, allo scopo di spiegare – mediante lo studio della regressione lineare semplice – la variabilità di una risposta aleatoria in funzione di una variabile dipendente (covariata). Si studierà la formulazione matematica del modello, come stimarne i parametri relativi, come quantificare e modellare la variabilità residua, come utilizzare tale modello per fare previsioni, quantificando l'incertezza e l'affidabilità di tali previsioni. Si introdurranno opportuni indici per valutare la bontà del modello proposto (goodness of fit) per capire quanto sia ragionevole pensare che i dati statistici siano stati generati in accordo al modello proposto.

Modulo 4 – Condurre un'analisi inferenziale su un dataset

Nel terzo modulo si imparerà ad implementare e adattare un modello di regressione lineare semplice ad un dataset reale, utilizzando il software R (https://cran.r-project.org/). Si capirà come interpretare l'output automatico fornito dal software, come rappresentare l'adattamento del modello ai dati, e come selezionare modelli eventualmente più adeguati. Anche in questo caso si punterà l'attenzione sull'interpretazione del modello statistico, sul significato dei parametri stimati e sulla loro significatività statistica, sempre allo scopo di aumentare la conoscenza del ricercatore sul fenomeno che ha generato i dati oggetto di studio.

Note biografiche

Anna Maria Paganoni – Ha conseguito il dottorato in Matematica presso l'Università degli Studi di Milano ed è in servizio al Politecnico di Milano dal 1998, dove attualmente è professore ordinario di Statistica. Ricopre il ruolo di Coordinatore del Corso di Studi in Ingegneria Matematica dove insegna Statistica e Modelli e Metodi dell'Inferenza Statistica. Delegata del Rettore al Data Analytics coordina un gruppo di ricerca per l'analisi e le banche dati del Politecnico di Milano. I suoi recenti interessi di ricerca si concentrano sui metodi statistici per la classificazione e il pattern recognition in ambito clinico, sui metodi statistici non parametrici per l'inferenza su dati ad alta dimensionsionalità, sullo studio di grandi database amministrativi, sui modelli per l'analisi di dati funzionali, sui modelli a effetti misti non parametrici per l'analisi di dati con struttura gerarchica.