Versione documento: 22 febbraio 2022, h16:26

Università degli Studi di Salerno
Laurea in Statistica per i Big Data

ANALISI E VISUALIZZAZIONE DEI DATI (PT I)

Pietro Coretto
pcoretto@unisa.it

SYLLABUS

Sommario. La "Parte I" del corso ha ad oggetto sia gli aspetti concettuali e metodologici dell'analisi esplorativa dei dati (EDA), sia gli strumenti e i metodi di elaborazione dei dati, calcolo numerico, e programmazione (ECP). La "Parte I" del corso è dedicata all'analisi di distribuzioni univariate. La componente ECP del corso è basata basa sull'ambiente di sviluppo R, essa è strettamente funzionale all'EDA. Nella parte iniziale del corso i due aspetti sono trattati separatamente. Successivamente le due componenti (EDA e ECP) saranno presentate in modo integrato. Di seguito si riporta il cronoprogramma dettagliato del corso.

Programma Dettagliato

Nozioni introduttive (lezione 1). Statistica, informatica e dati – nozioni introduttive: popolazione, campione, campionamento, inferenza, EDA – primi esempi di data set – elementi di un data set: unità statistica, variabili/features – espressione di variabili: livelli e labels – categorie di variabili rispetto alla loro espressione: quantitative, ordinali, nominali – categorie di data sets rispetto all'indicizzazione delle unità: cross-section, serie temporali, serie spaziali, serie spazio-tempoarali, dati longitudinali – esperimenti e studi osservazionali – fasi e componenti dell'EDA.

Linguaggio R: tipi numerici e principali strutture dati (lezioni 2–3). L'ambiente di calcolo R – operazioni aritmetiche – operazioni logiche – funzioni matematiche di uso comune – funzioni logiche di uso comune – assegnazione di variabili ed uso della memoria – variabili locali, globali, environments e workspace – tipi di dati – funzioni – strutture dati e ADS (abstract data structures) – vettori (1D-arrays) – attributi e metadati di un vettore – codifica delle variabili non numeriche (la funzione `factor()`) 2D-arrays – 3D-arrays – arrays di dimensione qualsiasi – operazioni su arrays – matrici e algebra lineare – liste – data.frame – subsetting/slicing di un ADS – metadati più comuni di un ADS.

Linguaggio R: inputs/outputs (lezioni 4–5). Tecniche elementari di storage e codifica dei dati (ascii, csv, tsv) – lettura di dati in formato ascii – data-import e formattazione dei dati – RData altri formati proprietari (xls, ods, mat, dta, etc) – scrittura dei dati in

formato RData e ascii – interazioni elementari con OS – tecniche elementari per manipolare stringhe – gestione dell'output della console.

Linguaggio R: device grafico (lezione 5). introduzione al device grafico – primi esempi di rappresentazioni grafiche – parametri grafici generici – rappresentazione dei colori e palettes – overlays e grafici multipli – aspect ratio, scaling, e risoluzione – grafica raster (bitmat), vettoriale, lossy, lossless – scrittura di grafici in formati comuni (pdf, jpg, png, etc) – ottimizzazione dei formati grafici in funzione dell'utilizzo.

Linguaggio R: programmazione (lezioni 6–7). Strutture di controllo (if, if-else, if-else, if-esle if-else, while, for, repeat) – breakers (stop, break, etc.) – scripts – stile di programmazione (commenti descrittivi, buona sintassi, etc) – approfondimenti sulla rappresentazione numerica (elementi della lista `.Machine`, effetti del "*rumore di troncamento*") – programmazione funzionale – costruzione di una funzione – output complessi e passaggio di argomenti – classe di un oggetto, metodi e programmazione ad oggetti – costruzione di "*print method*" e "*plot method*" – *environment* globale vs *function environment* – funzioni nascoste e funzioni primitive – introduzione al debugging.

Distribuzione, funzione quantile e densità (lezioni 8–10). Raw data e visualizzazione mediante stripchart – distribuzione di dati univariata (DU) – caratteristiche rilevanti di DU – frequenze relative e assolute – windowing dei dati continui – costruzione di DU aggregate per frequenze – stripchart, barplot, e piechart – massa e densità dei dati – misure di densità – misura di densità mediante istogramma – rappresentazione grafica dell'istogramma – frequenze cumulate – funzione di distribuzione empirica cumulata (ECDF) - quantile di livello α – approssimazione dei quantili basata sull'ECDF – ordinamento dei dati e ranks – principali tecniche di ordinamento in R – ordinamento gerarchico di un data set – approssimazione dei quantili basata sullo smoothing dei dati ordinati.

Centralità e dispersione di una dispersione (lezioni 11–12). Weighting dei dati e somme pesate – weighting basato sulla distribuzione empirica – media empirica, mediana e moda – serie di quantili – quartili e Tukey 5-numbers – decili e percentili – trasformazioni dei dati – trasformazioni lineari come operazioni di shift e scaling – trasformazioni lineari e unità di misura – effetti delle trasformazioni lineari sulle misure di – dispersione (scatter, spread) – deviazioni rispetto ad un punto – deviazione totale quadratica e assoluta – centralità della media rispetto alle deviazioni quadratiche – centralità della mediana rispetto alle deviazioni assolute – variabilità – varianza campionaria, varianza non distorta, deviazione standard, IQR, ampiezza del range – decomposizione della varianza nei momenti – proprietà delle misure di variabilità – effetti delle trasformazioni lineari sulle misure di dispersione – standardizzazione dei dati.

Simmetria, modello normale standard e code (lezione 13). Distribuzione dei dati limite ($n \to \infty$) – continuità al limite della densità e dell'ECDF – modello normale standard – proprietà di forma del modello normale standard – simmetria ed asimmetria – simmetria e misure di centralità – simmetria e quantili – misure di asimmetria – code di una distribuzione – decadimento esponenziale delle code nel modello normale – Normal Q-Q-plot e confronto con le code normali – Q-Q-plot per il confronto tra distribuzioni qualsiasi – eccesso di curtosi.

Contaminazione e robustezza (lezioni 14–15). Contaminazione dei dati – processi di contaminazione vs code pesanti/lunghe – nozione intuitiva di robustezza – *finite sample replacement breakdown point* – breakdown point asintotico – calcolo del breakdown point

per media e mediana – misure di centralià e dispersione robuste (trimean, medie troncate, MAD, etc) – *Tukey fences* e classificazione dei punti regolari, outliers, e gross-outliers. – boxplot – Tukey fences e boxplot.

Kernel density e smoothing (lezione 15). Discontinuità dell'istogramma – nozione di bandwidth, e bandwidth dell'istogramma – bandwidth ottimale di Scott e Freedman-Diaconis – kernel smoothing e funzioni kernel – kernel density – effetto della scelta della funzione kernel e del bandwidth – principali metodi di selezione data-driven per il bandwidth.

Esame e Valutazione del Profitto

- L'esame consiste in una prova scritta in laboratorio ed un colloquio orale.
- La prova scritta è valutata con un massimo di 30 punti allocati su diversi quesiti.
- Generalmente durante il colloquio orale si procede alla correzione della prova scritta, ed alla discussione di eventuali punti da chiarire, ed il voto finale è dato dalla somma dei punti conseguiti.
- Tuttavia, i punteggi maggiori o uguali a 27/30 dovranno essere confermati con una discussione orale più approfondita.

Software

Le lezioni si basano sull'uso del linguaggio di programmazione `R`[1] sviluppato e distributo con licenza GPL attraverso il *Comprehensive R Archive Network* (CRAN)[2]. Per questo corso è necessario installare `R-base-package` (comunemente detto `R`), ed un buon IDE.

Installazione di `R-base-package`

- **linux:** https://cran.r-project.org/bin/linux/
  Selezionare la distro appropriata e seguire le istruzioni.
- **macOS:** https://cran.r-project.org/bin/macosx/
  Nella sezione "*Latest Release*" sono disponibili diversi file `R-x.y.z.pkg`, dove `x.y.z` sono gli identificativi della versione ordinati a partire dalla più recente. Verificare quale è la versione più recente compatibile con la versione di `macOS` in uso. Scaricare il corrispondente file `*.pkg` e seguire la procedura di installazione. Sin consiglia inoltre l'installazione di `XQuartz`[3].
- **Microsoft Windows:** https://cran.r-project.org/bin/windows/base/
  Scaricare il file `*.exe` dal link "*Download R x.y.z for Windows*", dove x.y.z sono solitamente gli identificativi della versione corrente. Eseguire il file `*.exe` scaricato (possibilmente con permessi di amministratore), e seguire la procedura guidata di installazione.

[1] Vedi: https://it.wikipedia.org/wiki/R_(software)
[2] Vedi: https://en.wikipedia.org/wiki/R_package#Comprehensive_R_Archive_Network_(CRAN)
[3] Disponibile al seguente link: https://www.xquartz.org/

Installazione dell'IDE
Vi sono tantissime alternative. Durante il corso, e nei laboratori del campus destinati alle attività didattiche, utilizzeremo `RStudio`:

- https://rstudio.com/products/rstudio/download/#download

Bibliografia Consigliata

Libro di testo. I due testi riportati di seguito coprono abbondantemente il materiale delle lezioni.

Iacus, S. M., e G. Masarotto (2007). *Laboratorio di statistica con R*. McGraw-Hill.

Wickham, H., e G. Grolemund (2016). *R for data science: import, tidy, transform, visualize, and model data*. O'Reilly Media, Inc. (Anche disponibile in versione elettronica: https://r4ds.had.co.nz/)

Letture consigliate. La seguente lista include una selezione di articoli scientifici che trattano il problema della visualizzazione dei dati con un approccio non tecnico. Si consiglia la lettura, sebbene il seguente materiale non sarà oggetto di verifica.

Gelman, A., C. Pasarica, e R. Dodhia (2002). Let's Practice What We Preach. *The American Statistician*, 56(2), 121–130. https://doi.org/10.1198/000313002317572790

Goodman, A. A., e M. A. Borkin (2018). New Thinking on, and with, Data Visualization. arXiv:1805.11300, http://arxiv.org/abs/1805.11300'>arXiv:1805.11300.

Cleveland, W. S., w R. McGill (1985). Graphical Perception and Graphical Methods for Analyzing Scientific Data. *Science* 229(4716): 828–33. https://doi.org/10.1126/science.229.4716.828.