Introduzione ai Metodi per il Data Science

Introduzione

Benvenuti!
In questa lezione vedremo che cosa significa essere uno scienziato dei dati con un'introduzione alla terminologia di base usata dai professionisti e una panoramica sui tipi di problemi che nascono. Poi presenteremo i tipi di dati esaminando i vari tipi di dati disponibili e i modi in cui possono essere manipolati.
Descriveremo i cinque passi nei quali si sviluppa la scienza dei dati, compresa la manipolazione e la pulizia dei dati che affronteremo più in dettaglio nel seguito della lezione affrontando il problema dell'individuazione degli outlier.
Cominciamo…

Obiettivi della Lezione

L'argomento di questa lezione è la scienza dei dati (Data Science), una disciplina che ha sperimentato una rapida crescita negli ultimi decenni. Proprio per questo motivo ha sollevato un grande interesse sia nei media, sia nel mercato del lavoro.

Gli Stati Uniti nel 2015 nominarono il primo "esperto di scienza dei dati", **Dhanurjay "DJ" Patil** un matematico e informatico americano che ha ricoperto il ruolo di Chief Data Scientist dell'Office of Science and Technology Policy degli Stati Uniti dal 2015 al 2017. Le competenze dei data scientist sono sempre più richieste e la loro applicazione va ben oltre l'attuale mercato del lavoro con competenze in ambito matematico e di programmazione.

Per gli esempi utilizzeremo il software R che, insieme a Python, ha conosciuto negli ultimi anni gli sviluppi maggiori per le tecniche implementate nel campo del Data Science.

Preliminari R

R è un linguaggio di programmazione ad oggetti (vettori, matrici, dataframe, liste) per il calcolo statistico e la grafica, supportato dalla R Foundation for Statistical Computing, il cui riferimento potete trovarlo al link https://www.r-project.org/foundation/). Il linguaggio, iniziato nel 1995 da Ross Ihaka e Robert Gentlemen, attualmente è sviluppato da R Development Core Team.

Gran parte del sistema è scritto nel dialetto R di S (un linguaggio di programmazione, sviluppato da John Chambers (Bell laboratori, che hanno sviluppato anche Unix e C). Per le attività ad alta intensità di calcolo sono implementati parti di codice in C, C ++ e Fortran che possono essere collegati e chiamati in fase di esecuzione.

Dal 1997 R ha conosciuto uno sviluppo internazionale

Oggi, R è disponibile come Software Libero e funziona su un'ampia varietà di piattaforme UNIX, Linux, Windows e MacOS. Dal momento che il software è libero, il codice sorgente è disponibile e può essere modificato.

RStudio e R

Un modo per utilizzare R è quello di sfruttare le possibilità di RStudio che rappresenta un'interfaccia grafica organizzata in 4 pannelli. Una volta installato sul vostro computer R e Rstudio avviando RStudio, una nuova sessione di R viene avviata. Nella configurazione che vedete appaiono cerchiati in blu il software che stiamo utilizzando, ovvero la console di R (pannello in alto a destra) dove possiamo digitare 2+2 ottenendo 4. Nel pannello in alto a sinistra invece abbiamo l'editor, in cui possiamo scrivere le linee di comando che possono essere lanciate con il tasto Run. Nell'esempio mostrato stiamo caricando un dataset da un file esterno con la funzione read.csv. L'oggetto creato (x) possiamo trovarlo nel panel in basso a sinistra (pannello environment). Inoltre, lanciando la funzione plot nell'editor possiamo produrre un grafico che viene visualizzato nel pannello in basso a destra (dove, ad esempio, può essere anche visualizzata la funzione help).

R base e pacchetti (packages)

R è un programma modulare. Con il software, vengono scaricati numerosi pacchetti di base, ma molte altre funzioni possono essere aggiunte grazie a pacchetti e plugins aggiuntivi, disponibili in un apposito sito (repository). I pacchetti sono raccolte di funzioni R, dati e codice compilato in un formato ben definito. Repository pubblici (e privati) vengono utilizzati per ospitare raccolte di pacchetti R. La più grande collezione di pacchetti R è disponibile sul CRAN. Un pacchetto può essere installato dal CRAN usando la funzione install.packages("nomepacchetto"), oppure utilizzando RStudio andando su Strumenti e Installa pacchetti.

CRAN

CRAN è composto da una serie di server mirror distribuiti nel mondo ed è usato per la distribuzione delle versioni del linguaggio R e dei suoi pacchetti. Il mirror migliore per la propria posizione viene scelto automaticamente quando si scarica un pacchetto, ma è sempre possibile selezionarne uno a piacimento. Cliccando su packages compare tutta lista di pacchetti.

Per sapere quali pacchetti sono installati nel proprio sistema si usa la funzione library(). Invece per caricare l'help del pacchetto si digita help(nomepacchetto), ma sul CRAN trovate anche il manuale pdf. Per rimuovere il pacchetto dal sistema si usa remove.packages(nomepacchetto)

Era dell'informazione

I media si occupano sempre più spesso di notizie relative a fughe di dati (data leaks), cybercrimine e del fatto che i dati possono fornire molte informazioni sulla nostra vita. Perché proprio ora? Cos'è che rende la nostra epoca così proficua per le aziende che si occupano di dati?

Nel **Diciannovesimo secolo**, il mondo era nel pieno dell'era industriale. Il genere umano stava esplorando la sua collocazione nel mondo della produzione industriale, fianco a fianco con gigantesche invenzioni meccaniche. Naturalmente l'era industriale ebbe i suoi pro e i suoi contro.

Nel **Ventesimo secolo**, poi, l'uomo, ormai piuttosto abile nella creazione di grandi macchine, inizia a perseguire l'obiettivo di renderle sempre più piccole e veloci. L'era industriale era finita ed era stata sostituita da quella che chiamiamo **Era dell'informazione**.

Era dei dati

Come per l'Era industriale, l'Era dell'informazione ha avuto aspetti positivi e altri negativi. Fra le cose buone annoveriamo straordinari oggetti tecnologici, come gli smartphone e la televisione. Fra le cose cattive, un bel problema per il Ventunesimo secolo è rappresentato dalla disponibilità di troppi dati. È così: l'era dell'informazione, nella sua continua ricerca di generare dati, ha fatto letteralmente esplodere la produzione di dati elettronici. Ognuno di noi contribuisce a comporre questa cifra a ogni tweet, a ogni post su Facebook. Non solo creiamo dati a un livello senza precedenti, ma li consumiamo anche a un tasso sempre più accelerato. Abbiamo costruito macchine sempre più piccole in grado di raccogliere dati in continuazione e ora il nostro compito consiste nel capirne il senso. Questo significa essere nell'era dei dati.

Tipi di dati

Il tipo di dati non solo determina i metodi usati per analizzare ed estrarre i risultati, ma il fatto di sapere che i dati non sono strutturati o magari sono quantitativi può dirvi molto sul fenomeno che state misurando. In particolare, abbiamo dati strutturati (o organizzati) ovvero ordinati in una struttura a righe e colonne, dove ogni riga rappresenta un'unica osservazione e le colonne rappresentano le caratteristiche di tale osservazione. Ad esempio, è quello che succede quando apriamo un file Excel in cui abbiamo una struttura a righe e colonne in attesa di dati organizzati. Tuttavia, software come R e Python consentono di lavorare con dati non strutturati (non organizzati) ovvero con dati in formato libero, normalmente testo o audio grezzo che devono essere analizzati meglio per poter essere organizzati. In questo senso la scienza dei dati è la scienza che consiste nel trarre conoscenza dai dati. Mentre un esperto di scienza dei dati preferirà, probabilmente, avere a che fare con dati strutturati, dovrà comunque essere in grado di occuparsi della massiccia quantità di dati non strutturati.

Dati quantitativi e qualitativi

Nella maggior parte dei casi, quando si parla di **dati quantitativi,** normalmente (ma non sempre) si parla di un dataset strutturato con una rigida struttura a righe e colonne (perché i dati non strutturati difficilmente hanno delle caratteristiche così ben separate), ovvero dati che possono essere descritti tramite numeri su cui è possibile eseguire semplici operazioni matematiche, ad esempio la somma.

I **dati qualitativi**, d'altra parte, non possono essere descritti tramite numeri e per questo non possono essere oggetto di semplici operazioni matematiche. Quindi vengono descritti usando delle categorie e un linguaggio "naturale".

Dati strutturati: Aspetti statistici

In statistica abbiamo diversi tipi di dati che possono essere osservati in forma di **cross-section**: un numero di unità (maggiore di 1 indicato con N), osservate in un solo periodo (T=1).

Sfruttando le differenze tra le unità (persone, imprese, regioni, scuole, ecc.) consentono lo studio delle relazioni tra variabili. Dati in serie storica in cui seguiamo un'unità (N=1) (persona, impresa, regione, scuola, ecc.), in T periodi maggiori di 1 che consentono lo studio dell'evoluzione di un fenomeno nel tempo, anche al fine di prevederne l'andamento futuro.

Per finire abbiamo **dati panel** (o dati longitudinali): un numero di unità N maggiore di 1 osservate in T > 1 periodi. Seguendo nel tempo lo stesso gruppo di entità, è possibile lo studio delle relazioni tra variabili e la loro eventuale evoluzione temporale. In questa lezione presenteremo:

- modelli statistici
- rappresentazioni grafiche
- e metodi di correzioni per dati anomali per serie storiche e cross-section

Dati cross-section e in serie storica

In particolare, i modelli statistici utilizzati sono differenti per i **Dati cross-section** per cui è plausibile considerare un insieme di N dati come realizzazioni di N variabili casuali indipendenti e identicamente distribuite.

Invece, per le **Serie storiche**, le osservazioni non possono essere pensate indipendenti perché è fondamentale il tempo, che ha una direzione e porta la serie storica ad avere una memoria (persistenza) e per cui l'ordine delle osservazioni è rilevante.

I 5 Passi per la scienza dei dati

La scienza dei dati segue un processo strutturato, a passi distinti, che se adottato correttamente, preserva l'integrità dei risultati. I cinque passi fondamentali per la scienza dei dati sono:

1) Porre una domanda interessante
2) Ottenere i dati
3) Esplorare i dati
4) Creare un modello per i dati
5) Comunicare e presentare i risultati

Come già detto, in questa lezione vedremo i modelli per dati in cross-section e serie storica e alcune rappresentazioni grafiche che consentono sia di esplorare i dati, sia di presentare i risultati.

Ottenere i dati: file locale

In R i dati possono essere importati da un file di qualsiasi formato, salvato in locale in qualche directory del nostro computer. In particolare, in questo esempio è mostrato come importare un file con estensione csv (comma separated value) per il quale bisogna fare attenzione ai 4 punti cerchiati in 1, 2, 3, 4, cioè:

1) dove si trova esattamente il file
2) se contiene un'intestazione con il nome delle variabili
3) come sono separate tra di loro le colonne
4) e qual è il separatore per i decimali

Inoltre, nell'esempio viene mostrata anche una funzione in R str che mostra la struttura del dataset (ovvero possiamo sapere quante osservazioni e quante variabili sono presenti).

Ottenere i dati: database

Oppure possiamo scaricare i dati direttamente da internet e in particolare da un database come quello di Eurostat che contiene le informazioni statistiche sulle più importanti variabili macroeconomiche dei Paesi europei. In particolare, come mostrato in questo esempio utilizziamo una libreria chiamata RJSDMX, collegandoci al database Eurostat (ma la funzione getProviders ci mostra anche gli altri database disponibili), scaricando informazioni per Italia e Francia. I dati vengono salvati in un oggetto che chiamiamo dati che contiene delle serie storiche a partire da gennaio 2000. Di queste serie possiamo fare anche un grafico che è mostrato nel pannello plot.

Ottenere i dati: dati non strutturati

Un ultimo esempio mostra le possibilità che abbiamo utilizzando il software R ed in particolare la libreria pageviews, che consente di scaricare quante visualizzazioni ha ricevuto la voce Colosseum sulla versione inglese di wikipedia a cadenza giornaliera, per tipo di visualizzatore e da quale tipo di device (desktop o mobile).

Esplorare i dati

Iniziamo ora a capire come possono essere esplorati i dati in R utilizzando la funzione str, che fornisce informazioni sulla struttura del nostro oggetto che in questo caso si chiama mydata e che contiene 32 osservazioni su cui abbiamo rilevato 11 variabili di tipo numerico (descritte come num).

Prime statistiche sui dati

Inoltre, per avere qualche statistica, R nel pacchetto base consente di calcolare:

- la dimensione del dataset (dim)
- il minimo (min)
- il massimo (max)
- il campo di variazione (range)
- la media
- la mediana

Riepilogo sui dati

La funzione summary calcola alcune delle statistiche di sintesi descritte precedentemente per alcune variabili del nostro dataset (in questo caso le prime tre variabili).

Una prima analisi dei dati

Le tabelle di contingenza sono spesso essenziali per organizzare e riepilogare i dati, in particolare con variabili categoriali o qualitative. Una tabella di contingenza è una tabulazione di conteggi e/o percentuali per una o più variabili.

In R, queste tabelle possono essere create usando table() insieme ad alcune delle sue varianti. In questo caso utilizzando il dataset mtcars stiamo calcolando il numero di macchine classificate secondo il numero di cilindri (variabile cyl) e tipo di trasmissione (variabile am).

Verso un modello dei dati

Muoviamoci adesso verso dei modelli per i dati, introducendo il concetto di **Covarianza**. La covarianza di due variabili x e y in un dataset misura come le due siano linearmente correlate. Una covarianza positiva indicherebbe una relazione lineare positiva tra le variabili e una covarianza negativa indicherebbe il contrario.

Per avere una misura di tale relazione possiamo considerare l'analisi di correlazione (un numero che varia da -1 a 1) che tuttavia ci dice se esiste una associazione tra le variabili (ad esempio una relazione tra consumo e reddito). Se vogliamo sapere come varia il valore di una variabile in conseguenza della variazione di un'altra variabile e analizzare la forma della relazione tra variabili dobbiamo considerare un'analisi di regressione di cui parleremo in dettaglio più avanti.

Anscombe data

Anscombe (1973) ha introdotto un esempio in cui ha costruito un dataset artificiale per illustrare l'importanza di produrre grafici prima di analizzare e costruire un modello.

I quattro dataset sono quasi identici nelle semplici statistiche descrittive, ma ci sono alcune particolarità che portano a risultati particolari se si stima un modello di regressione.

I dati hanno caratteristiche molto diverse e se rappresentati su grafici a dispersione notiamo differenti situazioni che vanno da assenza di correlazione ad una relazione lineare, passando per una relazione quadratica. Questo ci dice dell'importanza di visualizzare i dati prima di applicare vari algoritmi disponibili per costruirne dei modelli.

In generale, le domande a cui è possibile rispondere con l'aiuto di uno scatterplot riguardano la relazione che esiste tra due variabili che può essere studiata attraverso diversi tipi di domande, come ad esempio: **esiste una relazione che può essere descritta da una linea retta (ovvero c'è una relazione lineare)?** Oppure: **esiste una relazione che non sia lineare?**

Qualora il grafico a dispersione delle variabili assomigliasse ad una nuvola, vorrebbe dire che non vi è alcuna relazione tra entrambe le variabili e quindi ci si fermerebbe.

Correlazione

Come mostrato in questo grafico, la forma della nuvola di punti ci consente di capire che tipo di relazione possa esistere tra le variabili e che affronteremo a breve nelle analisi sui dati cross-section.

Conclusioni

Bene, siamo giunti alla fine di questa video lezione.
Ti ricordo che abbiamo introdotto i Metodi per il Data Science.

In particolare abbiamo visto:

- la terminologia di base usata dai professionisti e una panoramica sui tipi di problemi che nascono
- i tipi di dati e i modi in cui possono essere manipolati
- i cinque passi nei quali si sviluppa la scienza dei dati, compresa la manipolazione e la pulizia dei dati

Grazie per l'attenzione!