

Analisi dei Dati

Cristiano Varin

Università Ca' Foscari

Laurea in Informatica 2022/23

Introduzione

Il corso fa parte del curriculum **Data Science**

Più precisamente fa parte di un grappolo di insegnamenti di statistica all'interno del curriculum:

- **Probabilità e statistica** (anno 2, primo semestre)
- **Analisi dei dati** (anno 2, secondo semestre)
- **Analisi predittiva** (anno 3, primo semestre)

Questi corsi vi permettono di apprendere i metodi statistici di base che sono utilizzati nell'analisi dei dati

I corsi seguono un preciso ordine che

- parte dalla teoria della **probabilità**,
- passa per l'**inferenza statistica** e
- si conclude con l'utilizzo di modelli statistici per risolvere problemi di **previsione**

Big data e sport

massachusetts institute of technology

MITnews

engineering science management architecture + planning humanities, arts, and social sciences campus press video connect

Sports analytics: a real game-changer

2013 MIT Sloan Sports Analytics Conference underscores how statistics keep changing the way sports are played — and changing minds in the industry.

Peter Dizikes, MIT News Office

today's news

March 4, 2013

As 2,700 people settled into their seats on Friday morning at the MIT Sloan Sports Analytics Conference, author Michael Lewis surveyed the scene from the dais and reminded everyone of how this massive annual event got started.

"This is Daryl's class," Lewis said.

Under the sea
MIT senior Grace Young's love of marine robotics will lead her to spend up to a month underwater this semester, collecting data and teaching classes over Skype to help save the oceans.

Seeing through silicon
October 2, 2013

Keeping 'digital storefronts' fresh
October 2, 2013

Droplets get a charge out of jumping
October 2, 2013

similar stories


The MIT Sloan Sports Analytics Conference took place March 1-2.
PHOTO: L. BARRY HETHERINGTON

multimedia


Election forecaster Nate Silver (left) spoke with Daryl Morey MBA '00, general manager of the NBA's Houston Rockets and co-founder of the conference.
PHOTO: L. BARRY HETHERINGTON

related

MIT Sloan Sports Analytics Conference

ARCHIVE: "How numbers can reveal hidden truths about sports"

Big data e sport



Moneyball

Il film racconta la storia di Billy Beane, il manager degli Oakland Athletics

Gli Oakland Athletics sono una delle squadre professionalistiche di baseball della Major League

Dopo una stagione sfortunata, Beane dovette ricostruire la squadra con un budget molto ridotto

Beane divenne famoso perché riuscì a costruire una squadra molto competitiva acquistando giocatori scartati dalle altre squadre

Come ci riuscì? Usando la statistica!

Per rifondare la squadra Beane si fece aiutare da Paul De Podest

De Podest era un economista che analizzò per Beane i dati di centinaia di giocatori di baseball per vari decenni (**sabermetrics**)

Moneyball

Le analisi statistiche di De Podest rilevarono che:

- i scopritori di talenti trascuravano le statistiche che potevano prevedere i buoni giocatori
- le squadre più ricche tendevano a collezionare stelle senza pensare alla loro compatibilità

Il metodo statistico di De Podest permise di assemblare una squadra molto competitiva

Gli Oakland Athletics rischiarono di vincere la stagione spendendo una frazione delle altre squadre

Oggi, non c'è una squadra professionista americana che non abbia un gruppo di **data analysts**

Lo **sport analytics** viene usato non solo per reclutare giocatori, ma anche per guidare le strategie di gioco

Big data e sport



Predictive analytics per il marketing



Le donne incinte sono clienti ideali

Target è una delle principali società di vendita al dettaglio americane

Target utilizza metodi statistici per identificare le donne nel secondo trimestre di gravidanza

Perché Target è interessata alle donne incinte?

Gli esperti di marketing sanno che la grande maggioranza dei clienti sono molto abitudinari

I clienti tendono ad acquistare sempre la stessa marca di biscotti: è abbastanza difficile convincerli a provare un'altra marca (e quindi ottenere del profitto)

Tuttavia, ci sono alcuni periodi nella vita di una persona in cui vi è maggiore disponibilità a cambiare le abitudini di acquisto

Qual è il più importante di questi periodi? L'arrivo di un bambino!

Target prevede le gravidanze

Target vuole identificare le donne nel secondo trimestre di gravidanza prima di altre società di vendita al dettaglio

Il gruppo di analisi predittiva di Target ha sviluppato un modello di previsione della gravidanza molto accurato

Il modello di previsione della gravidanza è stato costruito sui dati storici di acquisto delle donne che hanno avuto bambini

Quali sono gli ingredienti del modello di previsione della gravidanza?

Gli acquisti tipici di una donna incinta attorno alle 20 settimane di gestazione, ad esempio integratori di calcio, magnesio e zinco

I coupon

Target attira nei suoi negozi le donne incinte identificate dal suo modello inviando loro coupon di vestiti per bambini, pannolini, culle, ...

 <p>1.50 off up & up® baby formula</p> <p><input type="checkbox"/> select to print</p>	 <p>\$1 off Aveeno Baby Eczema Therapy lotion Enter code T06NWKX to redeem at Target.com</p> <p><input type="checkbox"/> select to print</p>	 <p>\$1 off 5-pk or larger Gerber Onesies Enter code T06U3QWJ to redeem at Target.com</p> <p><input type="checkbox"/> select to print</p>
 <p>\$3 off when you buy any ONE AVENT Multi-pack BPA-Free Baby Bottles</p> <p><input type="checkbox"/> select to print</p>	 <p>2.25 off when you buy any TWO PediaSure® products</p> <p><input type="checkbox"/> select to print</p>	 <p>2.50 off when you buy any Pull-Ups Training Pants & Pull-Ups Flushable Moist Wipes</p> <p><input type="checkbox"/> select to print</p>

Questa strategia di marketing basata sull'analisi dei dati ha avuto molto successo facendo guadagnare tantissimo a Target

How Companies Learn Your Secrets

By CHARLES DUHIGG FEB. 16, 2012

A man walked into a Target outside Minneapolis and demanded to see the manager ‘My daughter got this in the mail!’ he said. ‘She’s still in high school, and you’re sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?’

(...) The manager apologized and then called a few days later to apologize again.

(...) On the phone, though, the father was somewhat abashed. ‘I had a talk with my daughter,’ he said. ‘It turns out there’s been some activities in my house I haven’t been completely aware of. She’s due in August. I owe you an apology.’

La fine della storia

Le donne iniziarono a reagire male quando scoprirono che Target sapeva che erano incinte

Se una donna incinta pensa di essere stata spiata, non utilizzerà i coupon!

Qual è stata la reazione di Target?

Smettere di prevedere le gravidanze?

No, Target incominciò ad inviare un mix di coupon con cose che le donne incinte non comprerebbero mai, così da dare l'impressione di non spiarle!

Fonti:

- Charles Duhigg (2012). [How Companies Learn Your Secrets](#). The New York Times Magazine
- Kashmir Hill (2012). [How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did](#). Forbes

Obiettivo e programma del corso

Il corso ha l'obiettivo di insegnarvi gli strumenti di base che vengono utilizzati nell'analisi dei dati

Programma:

- 1.** concetti di base
- 2.** stima puntuale
- 3.** stima intervallare
- 4.** verifica d'ipotesi
- 5.** dipendenza

Saranno privilegiate le **idee** e i **principi** piuttosto che i 'dettagli matematici' (che però non verranno neanche trascurati del tutto!)

I metodi verranno illustrati con **casi studio** che sorgono in ambito informatico, tecnologico, scientifico, biomedico, economico ed aziendale

Lezioni e ricevimento

Il corso consiste di 24 lezioni:

Giorno	Ora	Aula
Martedì	12.15 - 13.45	Aula 1
Mercoledì	15:45 - 17:15	Aula 2

Ricevimento studenti in presenza e via Zoom:

- Martedì 10:00 - 12:00
- Zoom ID riunione 810 4234 9468
<https://unive.zoom.us/j/81042349468>

Il ricevimento a distanza va richiesto in anticipo
scrivendo a cristiano.varin@unive.it

Vi invito a controllare gli avvisi per eventuali variazioni
dovute a riunioni o altri impegni

Contatti

Il mio ufficio (Z.A11) si trova al primo piano dell'ala A dell'edificio Zeta

Per accedere all'ala A dovete prendere le scale a sinistra dell'entrata principale dello Zeta

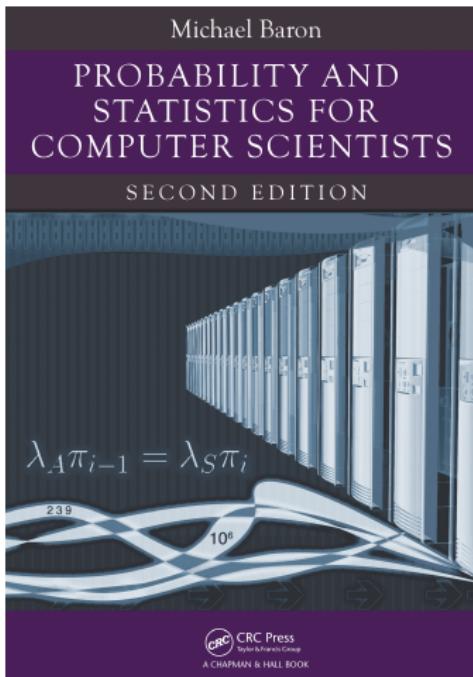
Mi potete contattare via email scrivendo a

cristiano.varin@unive.it

Vi chiedo, però,

- di scrivermi esclusivamente per questioni organizzative
- mentre se avete dubbi su lezioni o esercizi, sono disponibile a ricevimento e **non via email**

Libro di testo



Il libro è disponibile nella biblioteca BAS anche in formato elettronico

Materiali didattici disponibili su Moodle

Pagina Moodle del corso organizzata in sezioni:

- **Lezioni**
 - diario del corso
 - unità didattiche
- **Esercizi**
- **Laboratori con R**
- **Dati**
- **Casi studio**
- **Esame**
 - esempi risolti della prova d'esame
 - formulario e altri materiali utili per l'esame
 - template per la soluzione dell'esame

Riceverete un'email tramite il forum di Moodle:

- ogni volta che carico dei nuovi materiali
- ogni volta che correggo dei materiali pubblicati

Perché R?



Ci sono diverse ottime ragioni per usare R:

- potente e flessibile
- multipiattaforma: Linux, OS X, Unix, Windows
- gratuito! potete scaricarlo da

<https://cran.r-project.org>

- popolare: probabilmente il più diffuso software statistico al mondo per la didattica e la ricerca
- open source: con aggiornamenti continui effettuati da una grande comunità di statistici, scienziati e professionisti
- moltissimo materiale disponibile online: manuali, Stackoverflow, Rseek, blogs, ...

Prerequisiti

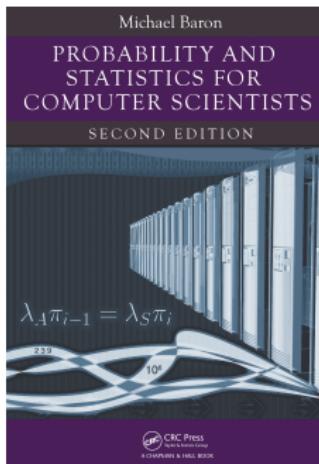
Per seguire [Analisi dei dati](#) è necessaria una buona padronanza degli argomenti del corso di [Probabilità e statistica](#)

Argomenti di probabilità dati per assunti:

- probabilità condizionata e indipendenza
- variabili casuali
- valore atteso, varianza, covarianza e correlazione
- principali variabili casuali discrete
- principali variabili casuali continue
- Teorema del limite centrale
- Legge dei grandi numeri

Prerequisiti

Nel caso abbiate difficoltà con i prerequisiti di probabilità potete anche leggere i capitoli 2, 3 e 4 del libro di testo di questo corso



- Chapter 2 'Probability'
- Chapter 3 'Discrete Random Variables and Their Distributions'
- Chapter 4 'Continuous Distributions'

Esame

Prova scritta per valutare le conoscenze teoriche e la capacità di applicare la teoria a problemi reali

Per lo svolgimento dell'esame è necessario dotarsi di un **computer portatile**. Qualora vi fossero difficoltà in merito, si è pregati di contattare il docente

Formato:

- quattro esercizi
- suddivisi in **due blocchi** di due esercizi con una breve pausa in mezzo
- ogni esercizio vale 8 punti
- l'esame è superato con:
 - un punteggio complessivo di **almeno 18 punti**
 - prendendo **almeno 9 punti in ciascuno dei due blocchi**
- punteggi superiori a 30 corrispondono a 30 e lode

Come si svolge l'esame

L'esame si svolge tramite la piattaforma [Moodle](#)

All'inizio dell'esame, il docente chiede di aggiornare la pagina Moodle del corso in cui appare il testo dell'esame e comunica la password per accedervi

Le soluzioni devono essere scritte in [R Markdown](#) riportando i principali conti svolti in [R](#) e le principali formule scritte in un linguaggio simile a [LaTeX](#)

Verrà reso disponibile un [template](#) da usare per scrivere le soluzioni degli esercizi

Inoltre verranno resi disponibili e discussi in classe degli esempi di prova d'esame risolti

Il tempo a disposizione per risolvere l'esame è un'ora per ogni blocco di due esercizi (due ore in totale)

Verranno concessi cinque minuti aggiuntivi per blocco per caricare le soluzioni in Moodle

Materiali ammessi all'esame

Durante l'esame non è ammesso l'uso di libri, appunti o supporti elettronici (diversi dal computer usato per la prova stessa)

È però ammesso l'uso di un formulario che contiene tutte le formule usate nel corso che verrà reso disponibile tramite la piattaforma Moodle

Durante l'esame il computer può essere usato esclusivamente per accedere alla pagina [Moodle](#), svolgere i conti in [R](#) e scrivere le soluzioni in [R Markdown](#)

Qualsiasi altro uso del computer e in particolare la navigazione in rete, l'accesso alla posta elettronica o a social networks non è ammesso pena l'esclusione immediata dall'esame

Prova intermedia

Verrà svolta una prova intermedia poco dopo la metà del corso

La prova intermedia:

- riguarda gli argomenti della prima metà del corso
- corrisponde ai **primi due esercizi dell'esame** da svolgersi in metà del tempo dell'esame
- viene superata con **almeno 9 punti**

Chi supera la prova intermedia e decide di far valere il punteggio ottenuto:

- può saltare la prima parte del **primo appello**
- ottiene un punteggio nel **primo appello** pari alla somma del punteggio della prova intermedia e il punteggio ottenuto negli ultimi due esercizi del **primo appello**

Il punteggio ottenuto nella prova intermedia **non** può essere utilizzato in appelli successivi al primo

Domande?

