Corso: Information Theory and Data Science
CFU: 6
Codice: 80300052
Lingua: inglese
CdS: ICT and Internet Engineering (Laurea Magistrale)
Docente: Mauro De Sanctis

OBIETTIVI FORMATIVI
La teoria dell'informazione classica, introdotta da Claude Shannon nel 1948, è la scienza che consente di quantificare matematicamente il contenuto informativo di un messaggio e il suo trasferimento attraverso un sistema. In particolare, la teoria dell'informazione fornisce gli strumenti matematici per fornire una visione quantitativa del contenuto informativo e una visione qualitativa del trasferimento di informazioni. Il principale contributo di Shannon è stato quello di stabilire i limiti dei sistemi di comunicazione. Una definizione generale di informazione è: nuova conoscenza derivata da studio, esperienza, messaggi, ecc. Pertanto, l'informazione è legata alla conoscenza. Le aziende oggi sono ricche di dati, ma povere di informazioni. Le tecniche di data science possono aiutare le aziende ad acquisire conoscenze dalle enormi quantità di dati a loro disposizione.
Questo corso ha l'obiettivo di fornire una visione unificata di informazione, conoscenza e analisi dei dati, estendendo le applicazioni della teoria dell'informazione oltre la teoria classica delle comunicazioni. In particolare, il corso si concentra su metodi statistici per il data science. Verranno presentate applicazioni di teoria dell'informazione e data science in diverse aree, tra cui comunicazioni digitali, economia, marketing, biologia, medicina, meteorologia.

LEARNING OUTCOMES:
Information Theory, as initiated by Claude Shannon in 1948, is the science that allows to mathematically quantify the information content of a message and its transfer through a system. In particular, information theory provides mathematical tools to deal with a quantitative view of information content and with a qualitative view of information transfer. The main contribution of Shannon was to establish bounds on communication. However, a general definition of Information is: new knowledge derived from study, experience, messages, etc. Therefore, Information is linked to Knowledge.

Companies today are rich in terms of data but poor in terms of useful information extracted from data. Data science techniques can help companies discover knowledge and acquire business intelligence from their massive datasets.
This course will give a unified view of Information, Knowledge and Data Analysis, extending the applications of Information Theory beyond the classical theory of communications. In particular, the course focuses on statistical methods for data analysis. Applications of information theory and data science in several areas will be presented including: digital communications, economics, marketing, biology, medicine, meteorology.

Prerequisiti

Italiano
Teoria della probabilità (o equivalente).

Inglese
Probability Theory (or equivalent).

Programma

Italiano
Elementi di teoria della probabilità: variabili e processi aleatori di tipo continuo e discreto, densità di probabilità, massa di probabilità, valore atteso.
Teoria dell'informazione: concetto di informazione, autoinformazione, entropia di Shannon, misure alternative di entropia, entropia relativa, divergenza di Kullback-Leibler, divergenza di Jensen-Shannon, entropia condizionale, entropia congiunta, informazione reciproca, correlazione totale, entropia differenziale, misure di informazione normalizzate.
Applicazioni al data science: concetto base di data science, definizione di dataset e attributi/feature, train set e test set, tipi di dati, analisi multivariata, descrizione statistica dei dataset, case study, metriche teoriche delle informazioni in attività di data science, preparazione dei dati, pulizia dei dati, discretizzazione degli attributi, riduzione della dimensionalità (Singular Value Decomposition), regole di associazione (unidimensionale e multidimensionale), algoritmi di classificazione (ad es. ID3, C4.5, Bayes, K-NN), alberi di classificazione, rilevamento di anomalie, clustering, addestramento e test di algoritmi, visualizzazione dei dati. Analisi e predizione di serie temporali. Metodi di valutazione degli algoritmi di data science.
Esperimenti informatici: introduzione a Python, progetti in Python con applicazioni della teoria dell'informazione all'analisi dei dati, progetti in Python con applicazioni di algoritmi di data science in diverse aree.

Italiano
Elementi di teoria della probabilità: variabili e processi aleatori di tipo continuo e discreto, densità di probabilità, massa di probabilità, valore atteso.
Teoria dell'informazione: concetto di informazione, autoinformazione, entropia di Shannon, misure alternative di entropia, entropia relativa, divergenza di Kullback-Leibler, divergenza di Jensen-Shannon, entropia condizionale, entropia congiunta, informazione reciproca, correlazione totale, entropia differenziale, misure di informazione normalizzate.
Applicazioni al data science: concetto base di data science, definizione di dataset e attributi/feature, train set e test set, tipi di dati, analisi multivariata, descrizione statistica dei dataset, case study, metriche teoriche delle informazioni in attività di data science, preparazione dei dati, pulizia dei dati, discretizzazione degli attributi, riduzione della dimensionalità (Singular Value Decomposition), regole di associazione (unidimensionale e multidimensionale), algoritmi di classificazione (ad es. ID3, C4.5, Bayes, K-NN), alberi di classificazione, rilevamento di anomalie, clustering, addestramento e test di algoritmi, visualizzazione dei dati. Analisi e predizione di serie temporali. Metodi di valutazione degli algoritmi di data science.
Esperimenti informatici: introduzione a Python, progetti in Python con applicazioni della teoria dell'informazione all'analisi dei dati, progetti in Python con applicazioni di algoritmi di data science in diverse aree.

Modalità di valutazione

- Prova pratica
- Prova orale

Descrizione delle modalità e dei criteri di verifica dell'apprendimento

Italiano

L'esame consiste in un project work. Il project work consiste nell'implementazione di una tecnica di data science (codice sorgente e un documento descrittivo che contiene eventualmente risultati sotto forma di tabelle o figure). Il project work deve essere adeguatamente documentato dallo studente. In fase di esame orale, ogni studente presenta e discute il suo progetto con il docente. Inoltre, le domande orali sono finalizzate a verificare la preparazione dello studente e la sua capacità di stabilire collegamenti critici tra i diversi temi affrontati durante il corso.
Il voto finale dell'esame, espresso in trentesimi, sarà assegnato secondo i seguenti criteri:

- Non idoneo: gravi lacune nella comprensione degli argomenti, scarsa capacità di analisi e giudizio, esposizione incoerente e linguaggio inadeguato.
- 18–21: conoscenze di base acquisite, capacità analitiche limitate e supportate dal docente, linguaggio generalmente corretto.
- 22–25: buona padronanza dei concetti fondamentali, autonomia nell'analisi e uso corretto del linguaggio.
- 26–29: conoscenze solide e ben organizzate, autonomia nella rielaborazione critica, esposizione chiara e corretta.
- 30–30 e lode: preparazione completa e approfondita, riferimenti culturali aggiornati, esposizione brillante e linguaggio preciso.

Inglese

The exam consists in a project work. The project work consists in the implementation of a data science technique (source code and a descriptive document eventually containing results in the form of Tables or Figures). The project work needs to be properly documented by the student. During the oral exam, each student presents and discusses their project with the instructor. In addition, the oral questions are aimed at assessing the student's preparation and his/her ability to make critical connections between the various topics covered throughout the course.
The final exam grade, expressed out of thirty, will be assigned according to the following criteria:

- Fail: serious gaps in understanding the topics, poor analytical and judgment skills, incoherent presentation, and inappropriate language.
- 18–21: basic knowledge acquired, limited analytical skills that require instructor support, generally correct language use.
- 22–25: good grasp of fundamental concepts, autonomous analysis, and correct use of language.
- 26–29: solid and well-structured knowledge, independent critical thinking, clear and accurate expression.
- 30–30 with honors: complete and thorough preparation, up-to-date cultural references, brilliant presentation, and precise language.

Testi adottati

Italiano

Slides fornite dal docente

Inglese
Slides provided by the teacher

Bibliografia di riferimento

E. Cianca, M. De Sanctis, M. Ruggieri, ``Information and Coding: theory overview, Design, Applications and Exercises'', ARACNE Editrice, 2007.
M. J. Zaki, W. Meira, ``Data Mining and Analysis - Fundamental Concepts and Algorithms'', Cambridge University Press, 2014.

Modalità di svolgimento

In presenza

Descrizione della modalità di svolgimento e metodi didattici adottati

Italiano
Il corso si svolge con lezioni frontali tradizionali per l'80% e per il 20% con lezioni pratiche ("hands-on") che richiedono l'utilizzo di un PC da parte dello studente. Nelle lezioni "hands-on" lo studente è guidato dal docente nello svolgimento di attività legate alla teoria dell'informazione e alle tecniche di data science.

Inglese
The course is based on 80% of traditional frontal lectures and on 20% of "hands-on" practical lectures. The "hands-on" lectures require the use of a PC by the student. During the "hands-on" lectures the student is led by the teacher through a set of activities concerning information theory and data science techniques.

Modalità di frequenza

Frequenza facoltativa

Descrizione della modalità di frequenza

Italiano
La frequenza è facoltativa ma fortemente consigliata per poter raggiungere pienamente gli obiettivi formativi.

Inglese

The attendance is not mandatory but it is strongly suggested to achieve all the planned learning outcomes.
##

CONOSCENZA E CAPACITÀ DI COMPRENSIONE:
Al termine del corso lo studente sarà in grado di: guadagnare una conoscenza approfondita della teoria dell'informazione e delle sue applicazioni nelle tecniche di apprendimento statistico; comprendere le sfide che si presentano nei problemi di analisi dei dati.

KNOWLEDGE AND UNDERSTANDING:
After the course, the student will be able to: gain an in depth knowledge of the information theory and its applications to statistical learning techniques; understand the challenges that arise in data science tasks.

CAPACITÀ DI APPLICARE CONOSCENZA E COMPRENSIONE:
Al termine del corso lo studente sarà in grado di: comprendere come applicare metodi di data science statistico a problemi selezionati; dimostrare competenza nella ricerca indipendente.

APPLYING KNOWLEDGE AND UNDERSTANDING:
After the course, the student will be able to: understand how to apply statistical data science methods to selected problems; demonstrate competency in independent research.

AUTONOMIA DI GIUDIZIO:
Durante il corso lo studente verrà incoraggiato a ragionare sull'analisi del problema da risolvere e sui vari compromessi che si devono affrontare nell'analisi dei dati.

MAKING JUDGEMENTS:
During the course the student is pushed to think on the problem analysis and on the trade-offs that rise on data analysis.

ABILITÀ COMUNICATIVE:
Lo studente deve realizzare un progetto accuratamente documentato. In questo modo lo studente migliora le capacità di produrre documentazione tecnica in modo chiaro ed efficace. Inoltre i risultati del progetto vengono presentati al docente, migliorando così la capacità di esposizione orale.

COMMUNICATION SKILLS:
The student performs a project work that needs to be properly documented. By performing the project, the student improves his capacity to produce clear and effective technical documentation. Moreover, the project results are presented to the teacher, so that the student improves his presentation skills.

CAPACITÀ DI APPRENDIMENTO:
Durante il corso verranno forniti dei dataset reali su cui applicare esempi di tecniche di data science. Lo studente apprenderà come elaborare dataset reali usando un PC.

LEARNING SKILLS:

During the course, real datasets will be provided in order to apply examples of data science techniques. The student will learn how to process real datasets using a PC.