MATHEMATICS (LM39)

(Lecce - Università degli Studi)

Teaching DATA MINING

GenCod A004898

Owner professor Massimo CAFARO

Teaching in italian DATA MINING

Teaching DATA MINING

SSD code ING-INF/05

Reference course MATHEMATICS

Course type Laurea Magistrale

Credits 6.0

Teaching hours Front activity hours: 42.0

For enrolled in 2018/2019

Taught in 2018/2019

Course year 1

Language ENGLISH

Curriculum PERCORSO COMUNE

Location Lecce

Semester Second Semester

Exam type Oral

Assessment Final grade

Course timetable https://easyroom.unisalento.it/Orario

BRIEF COURSE DESCRIPTION

The course provides a modern introduction to data mining, which spans techniques, algorithms and methodologies for discovering structure, patterns and relationships in data sets (typically, large ones) and making predictions. Applications of data mining are already happening all around us, and, when they are done well, sometimes they even go unnoticed. For instance, how does the Google web search work? How does Shazam recognize a song? How does Netflix recommend movies to its users? The principles of data mining provide answers to these and others questions. Data mining overlaps the fields of computer science, statistical machine learning and data bases. The course aims at providing the students with the knowldedge required to explore, analyze and leverage available data in order to turn the data into valuable and actionable information for a company, for instance, in order to facilitate a decision-making process.

REQUIREMENTS

Calculus. Probability and Statistics. Linear Algebra. Programming skills.

COURSE AIMS

Knowledge and understanding. The course describes methods and models for the analysis of large amounts of data. Students must have a solid background with a broad spectrum of basic knowledge related to data mining:

- the students must have the basic cognitive tools to think analytically, creatively, critically and in an inquiring way, and have the abstraction and problem-solving skills needed to cope with complex systems;
- they must have solid knowledge of data mining models and methodologies;
- they must be able to work on large data collections, including heterogeneous and produced at high speed data, in order to integrate them - in particular by knowing how to manage their origin and quality - and to carry out in-depth thematic analyses, drawing on this knowledge to improve the decision-making process.

Applying knowledge and understanding. After the course the student should be able to:

- describe and use the main data mining techniques;
- understand the differences among several algorithms solving the same problem and recognize which one is better under different conditions;
- tackle new data mining problems by selecting the appropriate methods and justifying his/her choices;
- tackle new data mining problems by designing suitable algorithms and evaluating the results;
- explain experimental results to people outside of statistical machine learning or computer science.

Making judgements. Students must have the ability to process complex and/or fragmentary data and must arrive at original and autonomous ideas and judgments, and consistent choices in the context of their work, which are particularly delicate in the profession of data scientist. The course promotes the development of independent judgment in the appropriate choice of technique/model for data processing and the critical ability to interpret the goodness of the results of the models/methods applied to the datasets under examination.

Communication. It is essential that students are able to communicate with a varied and composite audience, not culturally homogeneous, in a clear, logical and effective way, using the methodological tools acquired and their scientific knowledge and, in particular, the specialty vocabulary. Students should be able to organize effective dissemination and study material through the most common presentation tools, including computer-based ones, to communicate the results of data analysis processes, for example by using visualization and reporting tools aimed at different types of audiences.

Learning skills. Students must acquire the critical ability to relate, with originality and autonomy, to the typical problems of data mining and, in general, cultural issues related to other similar areas. They should be able to develop and apply independently the knowledge and methods learnt with a view to possible continuation of studies at higher (doctoral) level or in the broader perspective of cultural and professional self-improvement of lifelong learning. Therefore, students should be able to switch to exhibition forms other than the source texts in order to memorize, summarize for themselves and for others, and disseminate scientific knowledge.

TEACHING METHODOLOGY	The course aims to provide students with advanced tools for data analysis, through which to extrapolate relevant information from large datasets and guide the related decision-making processes. The course consists of frontal lessons using slides made available to students via the Moodle platform, and classroom exercises. The frontal lessons are aimed at improving students' knowledge and understanding through the presentation of theories, models and methods; students are invited to participate in the lesson with autonomy of judgement, by asking questions and presenting examples. The exercises are aimed at understanding the algorithms and models presented.
ASSESSMENT TYPE	Oral exam. During the exam the student is asked to illustrate theoretical topics in order to verify his/her knowledge and understanding of the selected topics. The student must demonstrate adequate knowledge and understanding of the issues presented or indicated, applying in a relevant manner the theories and conceptual models covered by the study programme.
OTHER USEFUL INFORMATION	**Office Hours** By appointment; contact the instructor by email or at the end of class meetings.
FULL SYLLABUS	Introduction. Map-Reduce. Mining data streams. Frequent Items. Frequent Itemsets and association rules. Mining similar items and Locality-Sensitive Hashing. Graph analysis. Link analysis and PageRank. Clustering. Recommendation systems. Mining Social-Network Graphs. Dimensionality reduction. Classification.
REFERENCE TEXT BOOKS	Mining of Massive Datasets J. Leskovec, A. Rajaraman and J. Ullman Freely availableonline: http://www.mmds.org Data Mining and Analysis M. J. Zaki and W. Meira Freely available online: http://dataminingbook.info