UNIVERSITÀ DEGLI STUDI DI PALERMO

DEPARTMENT	Ingegneria
ACADEMIC YEAR	2021/2022
MASTER'S DEGREE (MSC)	CYBER-PHYSICAL SYSTEMS ENGINEERING FOR INDUSTRY
INTEGRATED COURSE	BIG DATA AND ANALYTICS - INTEGRATED COURSE
CODE	21506
MODULES	Yes
NUMBER OF MODULES	2
SCIENTIFIC SECTOR(S)	ING-INF/05, ING-INF/03
HEAD PROFESSOR(S)	LA CASCIA MARCO Professore Ordinario Univ. di PALERMO
OTHER PROFESSOR(S)	TINNIRELLO ILENIA Professore Ordinario Univ. di PALERMO LA CASCIA MARCO Professore Ordinario Univ. di PALERMO
CREDITS	9
PROPAEDEUTICAL SUBJECTS	
MUTUALIZATION	
YEAR	1
TERM (SEMESTER)	2° semester
ATTENDANCE	Not mandatory
EVALUATION	Out of 30
TEACHER OFFICE HOURS	**LA CASCIA MARCO** Monday 15:00 17:00 Microsoft Teams Codice: wztkv0u **TINNIRELLO ILENIA** Monday 9:00 12:00 Ufficio del docente, presso il DEIM, secondo piano.

DOCENTE: Prof. MARCO LA CASCIA

PREREQUISITES	The course is self-consistent. However, it is recommended to have some basics of probability theory, linear algebra, programming and databases.
LEARNING OUTCOMES	Knowledge and understanding The student will be able to understand and analyze data classification problems and to apply concepts and methodologies for extracting useful information from the data. In particular, she/he will understand the behavior of Bayesian, linear and non-linear classifiers and Markov-based classifiers, as well as explore performance of different clustering and feature extraction algorithms. The student will have knowledge of the calculation models, architectures and infrastructures necessary for processing Big Data and will be familiar with issues related to Big Data analysis. Applying Knowledge The student will be guided to implement some reference algorithms for applications on real datasets. She/he will be stimulated to extrapolate the concepts and the methods presented in the course for some specific case studies in order to apply them (and relevant design considerations) to different application scenarios and, in particular, to the analysis of audio and video signals. The student will be able to design efficient software systems for processing Big Data. Judgements The student will be able to compare different approaches for data analysis according to the characteristics of the data sets and application scenarios. She/he will be also able to generalize the concepts and the methods acquired within the course and to related them to other statistical tools presented in other disciplines. The student will be able to evaluate suitability and performance of the various algorithmic and architectural solutions for managing Big Data Communication skills The student will learn the ability to rationally communicate her/his knowledge about the concepts and methods of the discipline, with a good level of clearness, fluency and correct use of technical language. In particular, she/he will be able to justify the design choices and the application of specific tools for solving the proposed analysis or synthesis problems. The student will be able to hold conversations on issues related to the design and management of systems for Big Data analysis. Learning skills The student will be able to read autonomously textbooks and scientific literature on machine learning, in order to study in depth approaches not discussed in the teacher-led lessons. The student will be able to understand the operating principles of new Big Data management tools.
ASSESSMENT METHODS	EXAM ORGANIZATION The examination is based on a mandatory written test and an optional oral exam. The oral exam allows to improve the written test evaluation. To take the oral exam, it is required to have at least a sufficient evaluation of the written test. The grade of the written test is given in the range 0-30/30. The minimum grade to pass the test is 18/30. The oral test is evaluated in the range of 0-3/30 to be added to the grade of the written test. The final grade is given by the written test grade (in case the student does not take the oral exam) or by the sum of the written test and oral exam grades. DESCRIPTION OF THE TESTS The written test includes two open questions and three exercises, similar to the examples discussed in the course, in which the student has to apply the concepts and the methodologies presented during the lessons to simple problems of data analysis, system modeling and statistical inference. The written test lasts 2.5 hours. The test is devised to evaluate: - The knowledge and understanding levels of learning concepts and algorithms; - The ability of applying the acquired knowledge to solve autonomously learning problems and system optimizations; - The ability to communicate knowledge, analyses and conclusions, and justify the design choices. The oral exam lasts about 30 minutes. It is based on the presentation of a python project developed autonomously by the student on a case study. The exam allows to assess: - The ability of applying the learning schemes to real problems, by exploiting programming languages and libraries; - The ability to communicate knowledge, analyses and conclusions, with a good level of clearness, fluency and correct use of language; - The ability of reinterpretation of the concepts and interdisciplinary connections, showing evidence for autonomously undertaking further studies or professional

	activity. LEARNING OUTCOMES In order to provide the overall evaluation, we will estimate the results achieved in the following course objectives. Knowledge and understanding: Evaluation of knowledge, understanding and integration of principles, concepts, methods and techniques of the discipline. Applying knowledge: Evaluation of capabilities in applying theoretical and technical knowledge for tackling and solving problems; evaluation of the autonomy level and originality of proposed solutions. Making judgements: Evaluation of logical, analytical and critical abilities for reaching appropriate judgments and decisions, based on available information and data. Communication skills and learning skills: Evaluation of the ability to communicate knowledge, analysis and conclusions, with a good level of clearness, fluency and correct use of language. Evaluation of the capability of reinterpretation and interdisciplinary connection, showing evidence for autonomously undertaking further studies or professional activity. GRADES 30-30 and laude: Excellent. Full knowledge and understanding of concepts and methods of the discipline, excellent analytical skills even in solving original problems; excellent communication and learning skills. 27-29: Very good. Very good knowledge and understanding of concepts and methods of the discipline; very good communication skills; very good capability of concepts and methods applications. 24-26: Good. Good knowledge of main concepts and methods of the discipline; discrete communication skills; limited autonomy for applying concepts and methods for solving original problems. 21-23: Satisfying. Partial knowledge of main concepts and methods of the discipline; satisfying communication skills; scarce judgment autonomy. 18-20: Acceptable: Minimal knowledge of concepts and methods of the discipline; minimal communication skills; very poor or null judgement autonomy. 0-17: Non acceptable: Insufficient knowledge and understanding of concepts and methods of the discipline.
TEACHING METHODS	Teacher-led lessons and exercises; practical exercises in Python.

MODULE
DATA ANALYTICS AND STORAGE

Prof. MARCO LA CASCIA

SUGGESTED BIBLIOGRAPHY

Dispense fornite dal docente / Lecture notes.

AMBIT	20917-Attività formative affini o integrative
INDIVIDUAL STUDY (Hrs)	96
COURSE ACTIVITY (Hrs)	54

EDUCATIONAL OBJECTIVES OF THE MODULE

This module presents more machine learning techniques and fundamentals of Big Data management systems.
In particular, the main objectives of the module are:
1) Learn classification techniques based on linear and nonlinear classifier and on deep learning and learn clustering techniques.
2) Learn the fundamentals of Big Data management systems, programming models for Big Data processing, storage and querying techniques for structured and semi-structured data.
3) Implement simple classification, clustering and data analysis technique using the Python programming language.
4) Evaluate architectural and algorithmic approaches and determine the more appropriate for the problem at hand.

SYLLABUS

Hrs	Frontal teaching
4	Linear classifiers: the perceptron and training techniques. Gradient technique. Cost functions. Training for non-separable classes. Generalizations to multi-class classifiers. Classifiers with activation functions different from threshold functions: the logistic regressor. Linear regressors and bias/variance dilemma.
4	Non linear classifiers: multi-stage neural networks. The XOR problem. Backpropagation algorithm. Feature space transformation. Decision-trees. Combination of multiple non-linear classifiers.
6	Introduction to Deep Learning: CNN, Autoencoder, LSTM, GAN, Graph Neural Networks.
4	Clustering: K-Means and fuzzy K-Means. Selection of the clusters number.
4	Introduction to Big Data: terminology, main features and application examples. Main issues of big data management: memory occupancy, processing speed, computational complexity, errors in data, data accuracy, data compression.
4	Big Data pre-processing: error handling, filtering, transformation, integration. Dimensionality reduction: Principal-Component Analysis, Singular-Value Decomposition.
2	Programming models for Big Data processing: MapReduce.
6	Data models. Relational and non relational models. Storage and querying of Big Data. NoSql databases.
2	Infrastructures for Big Data analytics. Apache Spark.

Hrs	Practice
3	Linear and nonlinear classifiers.
3	Deep learning.
3	Clustering.
3	Pre-processing and dimensionality reduction.
3	Systems for storage and querying of Big Data
3	Developing a complete pipeline to analyze big data.

MODULE
MACHINE LEARNING

Prof.ssa ILENIA TINNIRELLO

SUGGESTED BIBLIOGRAPHY	
AMBIT	20917-Attività formative affini o integrative
INDIVIDUAL STUDY (Hrs)	48
COURSE ACTIVITY (Hrs)	27

EDUCATIONAL OBJECTIVES OF THE MODULE

This module presents data-driven techniques for classification problems, based on statistical approaches and models with memory.
More into details, the main objectives of the module are:
1) Learn techniques for estimating the probability density or the empirical distribution function of the features used for the classification;
2) Apply Bayesian inference in classification problems, including memory-based processes;
3) Develop reinforcement-learning techniques.

SYLLABUS

Hrs	Frontal teaching
4	Introduction to the course. Fundamentals of probability and random variables.
8	Classification: hypothesis representation, decision regions, cost functions. Bayesian classifiers. Multi-variate Gaussian distributions. Estimates of probability densities for continuous and discrete features: parametric and non-parametric estimation. Maximum likelihood estimation.
8	Discrete-time Markov processes: transition matrices, steady-state equilibrium conditions, probability limit distributions. Application examples: Google PageRank algorithm. The Viterbi algorithm for classification problems with memory. Markov decision processes and Reinforcement Learning.

Hrs	Practice
7	Introduction to Python and the scikit learn library. Applications of all course concepts to real-world problems of classification and implementation examples.