UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

CONSIGLIO INTERCLASSE IN MATEMATICA

General information		**Academic year 2022-2023**
Academic subject	**Numerical Methods in Data Science**	
Degree programme	Mathematics (L-35)	
Programme year	Third year	
Term	Second semester (February 28, 2023 – May 30, 2022)	
European Credit Transfer and Accumulation System credits (ECTS)	7	
Language	Italian	
Attendance	Not compulsory	

Lecturers		
Name and surname	Flavia Esposito (instructor of record)	Nicoletta Del Buono
E-mail	flavia.esposito@uniba.it	Nicoletta.delbuono@uniba.it
Telephone		+39 080 544 2711
Department and office	Department of Mathematics, room 17 second floor	Department of Mathematics, room 23 second floor
Virtual meeting room	Microsoft Teams: access code 687hs39	
Web page	https://www.dm.uniba.it/members/esposito	https://www.dm.uniba.it/members/delbuono
Office hours	By appointment via email	By appointment via email

Syllabus	
Learning objectives	Acquisition of basic numerical techniques for data science application. Acquisition of basic knowledge for exploratory Data Analysis and matrix methods to deal with problems arising in learning from data
Course prerequisites	The knowledge generally acquired in the L-35 Mathematics degree with particular references to the disciplines of Numerical Analysis I (Calcolo Numerico I) and classical Mathematical Analysis in one and more variables
Course contents	Introduction to Factorisation Methods for Data Science. Linear algebra and learning from structured data. Representation of data through vectors and matrices. Singular value decomposition of a data matrix. Properties of right and left singular vectors and their relation to the fundamental vector spaces generated by a data matrix. The importance of SVD in Data Science: examples. The Karhunen-Loève transformation and its relation to SVD. Principal components and best low-rank approximation of a data matrix. Reduced form of the SVD and the Eckart-Young theorem. Data analysis by SVD factorization of the covariance and correlation matrix. Principal component analysis and equivalence with maximum problem for data variance. PCA geometry. Analysis of nonnegative data. Positive and nonnegative matrices. Perron-Frobenius theorems. Eigenface model. The information vector model (VSM) and Latent Semantic Indexing. Introduction to algebraic formalisation of information retrieval problems. Automatic indexing process, stop-listing and stemming operations. Index-term weighting functions. Application of QR and SVD factorizations to the term-document matrix and their geometric interpretation. Low-rank approximation of semantic space. The query-matching process and similarity measures. Term-terminal comparison and clustering of synonymous terms.

	Mention of relevance feedback mechanisms based on the use of the truncated SVD of the term-document matrix. Models based on eigenvalues and eigenvectors for web information retrieval. The hyperlink structure of the web and its representation through undirected graphs. Concepts of inlink and outlink. HITS and PageRank models for ranking web networks. Construction of Hub and Authority matrices and their properties. Construction of the Google matrix and its properties. Reducibility and graphs. Convergence of the PageRank model and irreducibility of the Google matrix. Introduction to problems on networks and graphs. Examples and problems (Konigsberg bridge problem and house door problem). Definitions of paths and cycles. Eulerian and Hamiltonian paths. Graphs and matrix representations. Adjacency, incidence, degree matrix and Laplacian matrix L_G of a graph. Some spectral properties of the adjacency matrix and the L_G matrix. Eigenvalues and connectivity measures of a graph. Notes on complete graphs and their properties. Minimum cost path problem and Dijkstra's algorithm. Introduction to Machine Learning (supervised and unsupervised problems, introduction to classification and regression with examples, concepts of over and under fitting, bias-variance trade-off), Data structure (vectors, matrices and tensors), exploratory data analysis (EDA), variable types and descriptive statistics (measures of central tendency, measures of variability, quantiles) with related graphical representations (box-plot, bar chart, histograms, scatter plot). Inspection of relationships between variables, definition of Pearson's correlation coefficient. Pre-processing of the data: definition/classification and treatment of missing values, definition of outliers and identification of outliers, transformations, and normalisations of the data. The problem of Curse of Dimensionality and how to solve it with linear factorization techniques, Interpretations of matrix decompositions, Matrix decompositions as recommendation systems (hints). Introduction to Nonnegative Matrix Factorization (NMF). History and examples. Motivations and interpretation of nonnegative factors, part-based representation and learning. Example eigenfaces and comparison between VQ, PCA and NMF, Formalisation of NMF as a matrix optimisation problem. Analysis of the two most commonly used objective functions, probabilistic interpretation of the NMF. Beta and Bregman divergences (definitions and equivalences). Geometric interpretation of the NMF. Nonnegative Rank Factorization. Numerical algorithms for the calculation of the NMF (Multiplicative Updates and Block Coordinate Descent). Convergence theorems and demonstrations. Notes on regularized NMF. Introduction to the clustering problem, useful distances for clustering, divisive and agglomerative hierarchical clustering (representations by dendrogram and link methods). dendrogram and linkage methods: single, complete, average), K-means. Heuristic methods for the choice of the optimal number of clusters in K-means, Equivalence between NMF and K-means with relaxed orthogonality constraint (Theorems and

	demonstrations), K-medoid (hints). Internal indices, External indices and other measures of goodness of clustering algorithms. Introduction to supervised problems. Expected Risk and Empirical Expected Risk. Introduction to regression, under and over-fitting problem in regression. Simple and Multiple Linear Regression, Estimation of coefficients with OLS (Ordinary Least Squares and Normal equations). Gauss-Markov theorem and conditions for the best correct linear estimator. Geometric interpretation of regression. Regression for the study of collinearity and the identification of outliers in data matrices. Measures of evaluation of regression models (SST, SSE, SSR, R^2, adjusted R^2, RMSE, MAE, F statistics). Polynomial regression (hints). Introduction to classification problems. SVD for hand-written digits classification, K-Nearest Neighbors, Tree approaches to classification (Decision Trees and Random Forests). Support Vector Machines, Maximal Margin Classifier, Soft Margin Classifier, Mercer's Theorem, Measures of Classification Goodness, Cross Validation (briefly).
Reference books	-G. Strang, Linear Algebra and Learning from Data, Wellesley-Cambridge Press, 2019 - C. Meyer, Matrix Analysis and Applied Linear Algebra, SIAM, 2003. - I.T. Jolliffe, Principal Component Analysis, Second Edition, Springer, 2002 - A. Cichocki, R. Zdunek, A.H. Phan, S.I Amari, Nonnegative Matrix and Tensor Factorizations, Wiley, 2009 - T. Hastie, R. Tibshirani J. Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, 2009
Additional course materials	Notes and slides provided by the instructor

Work schedule				
	Total	Lectures	Hands-on learning (recitations/laboratories /seminars/other)	Self-study
Hours	175	48	15	112
ECTS credits	7	6	1	

Teaching methods	
	- Lectures conducted with the aid of teaching aids (slides). - Computer-based exercises.

Expected learning outcomes	
Knowledge and understanding	Acquiring some techniques for solving classification and clustering problems. Ability to using numerical codes implementing standard techniques for analyzing real data.
Applying knowledge and understanding	The acquired theoretical and practical knowledge is used in applied mathematical fields and for solving problems in the context of learning from data.
Making judgements	Ability to identify the right numerical techniques to address and numerically solve problems arising from real applications involving big data.
Communication skills	Acquisition of the language and mathematical formalism necessary for the consultation and comprehension of texts, the exposition of acquired knowledge, the description, analysis, and resolution of applied and Data Science problems

Learning skills	Acquisition of an adequate study method, supported by the consultation of texts and the computer implementation of the numerical techniques exposed during the lectures.

Assessment and feedback	
Assessment methods	Oral examination on the syllabus and exercises or project assigned by the lecturer
Evaluation criteria	• *Knowledge and understanding*: Students must demonstrate adequate knowledge of the main topics of the course • *Applying knowledge and understanding*: Students must demonstrate adequate knowledge of the possible applications of the theoretical concepts and possess adequate ability to implement these applications • *Making judgements*: Students must demonstrate adequate autonomy in selecting the most appropriate theoretical concepts for solving practical problems. • *Communication skills*: Students must demonstrate an adequate expository capacity of the studied topic and an adequate capacity in analysis and synthesis • *Learning skills*: Students must demonstrate a good ability to make interdisciplinary connections
Grading policy	The evaluation of the oral examination and the awarding of the final mark will be based on the following learning assessment scale: – Insufficient grade (<18): Fragmentary and superficial knowledge of the contents, errors in the application of the concepts, poor exposition – Grade 18-20: Sufficient but general knowledge of content, simple exposition, uncertainties in the application of theoretical concepts – Grade 21-23: Appropriate but not extensive knowledge of content, ability to apply theoretical concepts, ability to present content in a simple manner – Grade 24-25: Appropriate and extensive knowledge of the content, fair ability to apply the knowledge, ability to present the content in an articulate manner. – Grade 26-27: Accurate and comprehensive knowledge of the content, good ability to apply the knowledge, ability to analyse, clear and correct presentation. – Grade 28-29: Extensive, complete, and thorough knowledge of the content, good application of the content, good analytical and synthesising skills, secure and correct presentation. – Grade 30 and 30 with distinction: very broad, complete, and in-depth knowledge of the content, well-established ability to apply the content, excellent ability to analyse, summarise and make interdisciplinary connections, very good exposition

Additional information	
	-