

**FAC.SIMILE SCHEDA DI INSEGNAMENTO CON LE INDICAZIONI SPECIFICHE PER
CIASCUN CAMPO DA COMPILARE (DA PREDISPORRE SU CARTA INTESTATA DEL
DIPARTIMENTO/SCUOLA)**

CORSO DI STUDIO: Laurea Magistrale in Bioinformatica (LM-8)

ANNO ACCADEMICO: 2023-2024

DENOMINAZIONE DELL'INSEGNAMENTO: Metodi Numerici per la Bioinformatica

Principali informazioni sull'insegnamento	
Anno di corso	<i>1 anno</i>
Periodo di erogazione	<i>1 semestre (dal 02.10.2023 al 19.01.2024)</i>
Crediti formativi universitari (CFU/ETCS):	<i>6 CFU</i>
SSD	<i>MAT/08</i>
Lingua di erogazione	<i>Italiano</i>
Modalità di frequenza	<i>Facoltativa (fortemente consigliata)</i>

Docente	
Nome e cognome	<i>Del Buono Nicoletta e Flavia Esposito</i>
Indirizzo mail	<i>nicoletta.delbuono@uniba.it e flavia.esposito@uniba.it</i>
Telefono	<i>0805442711</i>
Sede	<i>Stanza n 24, Il piano, Dipartimento di Matematica</i>
Sede virtuale	<i>cuzrciw</i>
Ricevimento	<i>Su appuntamento, da concordare per e-mail; in presenza o in remoto</i>

Organizzazione della didattica			
Ore			
Totali	Didattica frontale	Pratica (laboratorio, campo, esercitazione, altro)	Studio individuale
<i>150</i>	<i>32</i>	<i>30</i>	<i>88</i>
CFU/ETCS			
<i>6</i>	<i>4</i>	<i>2</i>	

Obiettivi formativi	<i>Acquisizione delle metodologie numeriche classiche e delle tecniche di algebra lineare per l'analisi di dati biomedici strutturati</i>
Prerequisiti	<i>Conoscenze di base di matematica coerenti con quelle richieste per l'ammissione al corso di laurea</i>
Metodi didattici	<i>- Didattica frontale condotta con l'ausilio di supporti didattici (slide). - Esercitazioni al calcolatore con il supporto di codici in linguaggi open source forniti dal docente e utilizzo di dati di benchmark presenti nei repository online</i>

<p>Risultati di apprendimento previsti</p> <p>DD1 Conoscenza e capacità di comprensione</p> <p>DD2 Conoscenza e capacità di comprensione applicate</p> <p>DD3-5 Competenze trasversali</p>	<p><i>Al termine dell'insegnamento lo studente dovrà essere in grado di formalizzare matematicamente dei problemi di analisi dei dati di tipo biomedico e identificare la tecnica numerica più appropriata per rispondere alle domande di ricerca relative ai dati disponibili</i></p> <p>DD1: <i>conoscenza e capacità di comprensione;</i></p> <ul style="list-style-type: none"> • <i>Formalizzazione matematica dei problemi di analisi dei dati</i> • <i>Approcci e tecniche numeriche per l'analisi dei dati</i> • <i>Comprensione e utilizzo di codici numerici per l'analisi di insiemi di dati biologici</i> <p>DD2: <i>capacità di applicare conoscenza e comprensione;</i></p> <ul style="list-style-type: none"> • <i>Conoscenza e utilizzo dei meccanismi di base dell'algebra lineare per l'analisi dei dati biomedici.</i> • <i>Comprensione e implementazione dei flussi di lavoro per l'analisi numerica dei dati biomedici.</i> <p>DD3: <i>capacità critiche e di giudizio</i> <i>Le attività di esercitazione e la presentazione e discussione di specifici casi di studio durante le lezioni frontali permetteranno l'acquisizione di capacità critiche e di giudizio per analizzare e interpretare i dati di tipo biomedico rappresentati come matrici numeriche</i></p> <ul style="list-style-type: none"> • <i>Autonomia di giudizio</i> <p><i>Al termine dell'insegnamento lo studente dovrà essere in grado di</i></p> <ul style="list-style-type: none"> • <i>Comprendere la struttura del dato proposto e identificare la metodologia più appropriata per descriverlo</i> • <i>Individuare la tecnica adatta a rispondere ai quesiti posti</i> • <i>Gestire la presenza di outlier, missing data e saper gestire eventuali normalizzazioni</i> • <i>Implementare un workflow per l'analisi del dato</i> <p>DD4: <i>capacità di comunicare quanto si è appreso.</i> <i>La discussione di specifici casi di studio durante le lezioni frontali e le esercitazioni guidate su dati di benchmark forniti dai docenti permetteranno l'acquisizione delle opportune capacità di comunicazione e trasmissione dei concetti appresi</i></p> <ul style="list-style-type: none"> • <i>Abilità comunicative</i> <p><i>Al termine dell'insegnamento lo studente dovrà essere in grado di</i></p> <ul style="list-style-type: none"> • <i>Spiegazione delle modalità di identificazione dei metodi numerici più appropriati per l'analisi dei dati biologici</i> • <i>Identificazione e spiegazione della tecnica numerica più appropriata per rispondere alle domande di ricerca relative ai dati disponibili</i> • <i>Capacità di comunicare in modo efficace le modalità di implementazione del flusso di lavoro scelto per l'analisi dei dati biomedici da analizzare</i> <p>DD5: <i>capacità di proseguire lo studio in modo autonomo nel corso della vita.</i> <i>Durante le lezioni frontali e le esercitazioni saranno fornite le indicazioni relative a pacchetti software open source e repository (certificati) di dati di tipo biomedico utili a verificare le conoscenze acquisite e a proseguire autonomamente nello studio di nuovi dataset biomedici.</i></p> <ul style="list-style-type: none"> • <i>Capacità di apprendere in modo autonomo</i> <p><i>Al termine dell'insegnamento lo studente dovrà essere in grado di</i></p> <ul style="list-style-type: none"> • <i>Spiegare i concetti acquisiti e sviluppare la capacità di scegliere la tecnica numerica più appropriata per l'analisi dei dati in esame.</i> • <i>Utilizzare un formalismo matematico appropriato per descrivere l'analisi dei dati.</i>
--	--

<p>Contenuti di insegnamento (Programma)</p>	<p><i>-Algebra Lineare per la bioinformatica: concetti di vettore, matrice, operazioni vettoriali e matriciali. Concetti di norme matriciali di spazi vettoriali, spazi di matrici, spazi vettoriali generati. Basi di spazi vettoriali e di matrici. Sistemi lineari e loro rappresentazione matriciale. Prodotto di matrici di dati e fattorizzazioni di matrici di dati. Fattorizzazione QR e concetti di ortogonalità tra vettori e matrici. Metodi per il calcolo di autovalori e autovettori di una matrice di dati. Decomposizione a Valori Singolari (SVD) e SVD troncata.</i></p> <p><i>-Analisi delle componenti principali (PCA): Matrice di covarianza e di correlazione dei dati. Diagonalizzazione di una matrice simmetrica. Autovalori di una forma quadratica definita positiva. Componenti principali di vettori di variabili aleatorie. Combinazioni lineari di massima varianza. Significato statistico e probabilistico delle componenti principali. Analisi delle componenti principali per insiemi di dati multivariati. Euristiche per il calcolo del numero ottimale delle componenti principali. PCA come metodo di feature selection. Interpretazione delle componenti principali: il cerchio delle correlazioni, gli Z-scores. Applicazioni reali della PCA.</i></p> <p><i>-Cenni su metodi di approssimazione low rank di matrici di dati omici. SVD, Fattorizzazioni nonnegative di matrici (NMF). Sparse Nonnegative Matrix Factorization per il Clustering.</i></p> <p><i>-Introduzione al problema di clustering, distanze utili per il clustering. Clustering Gerarchico divisivo e agglomerativo (rappresentazioni tramite dendrogramma e metodi di linkaggio: singolo, completo, medio), K-means. Metodi euristici per la scelta del numero ottimale di clusters nel K-means, Equivalenza tra NMF e K-means con vincolo rilassato di ortogonalità (Teoremi e dimostrazioni), K- medoide (cenni). Indici Interni, Indici Esterni e altre misure di bontà degli algoritmi di clustering. Algoritmi Model e Density Based (Mixture Model e DBSCAN): cenni.</i></p> <p><i>-Introduzione ai modelli di regressione Lineare Semplice e Multipla, Stima dei coefficienti con OLS (Ordinary Least Squares ed equazioni Normali). Teorema di Gauss-Markov e condizioni per il miglior stimatore lineare corretto. Interpretazione geometrica della regressione. Regressione per lo studio della collinearità e l'individuazione degli outliers in matrici di dati. Misure di valutazione dei modelli di regressione. Regressione logistica semplice e multipla. Regressione logistica e distribuzioni bernulliana e multinomiale.</i></p> <p><i>-Esercitazioni al calcolatore tramite il software statistico R https://www.r-project.org</i></p>
<p>Testi di riferimento</p>	<p><i>-Carl D. Meyer - Matrix Analysis and Applied Linear Algebra: 1st (first). Siam, 2001. http://matrixanalysis.com/</i></p> <p><i>-Gilbert Strang - Linear Algebra and Learning from Data, Wellesley-Cambridge Press, 2019 https://math.mit.edu/~gs/learningfromdata/</i></p> <p><i>-Robert Gentleman, Vince Carey, Wolfgang Huber, Rafael Irizarry, Sandrine Dudoit, Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Springer 2005. (http://www.bioconductor.org/help/publications/books/bioinformatics-and-computational-biology- solutions/)</i></p> <p><i>-Matthew He, Sergey Petoukhov. Mathematics of Bioinformatics: Theory, Practice, and Applications. Wiley&Sons 2011 https://onlinelibrary.wiley.com/doi/book/10.1002/9780470904640</i></p> <p><i>- Michael C. Whitlock, Dolph Schluter. Analisi statistica dei dati biologici. Zanichelli</i></p> <p><i>-I.T. Jolliffe, Principal Component Analysis, Second Edition, 2002, Springer https://link.springer.com/book/10.1007/b98835</i></p>
<p>Note ai testi di riferimento</p>	<p><i>Appunti e riferimenti specifici forniti dal docente</i></p>
<p>Materiali didattici</p>	<p><i>Il materiale didattico sarà reso disponibile sulla classe Teams dell'insegnamento</i></p>

Valutazione	
Modalità di verifica dell'apprendimento	<i>Prova Orale sul programma svolto nel corso delle lezioni ed esercitazioni e/o progetto assegnato dal docente</i>
Criteri di valutazione	<ul style="list-style-type: none"> • <i>Conoscenza e capacità di comprensione:</i> <ul style="list-style-type: none"> ○ <i>Lo studente deve dimostrare una adeguata conoscenza dei contenuti dell'insegnamento</i> • <i>Conoscenza e capacità di comprensione applicate:</i> <ul style="list-style-type: none"> ○ <i>Lo studente deve dimostrare una adeguata conoscenza delle possibili applicazioni dei concetti teorici e possedere una adeguata capacità di implementare tali applicazioni</i> • <i>Autonomia di giudizio:</i> <ul style="list-style-type: none"> ○ <i>Lo studente deve dimostrare una adeguata autonomia nella selezione dei concetti teorici più idonei alla risoluzione di problemi pratici</i> • <i>Abilità comunicative:</i> <ul style="list-style-type: none"> ○ <i>Lo studente deve dimostrare una adeguata capacità espositiva dei contenuti studiati e una adeguata capacità di analisi e sintesi</i> • <i>Capacità di apprendere:</i> <ul style="list-style-type: none"> ○ <i>Lo studente deve dimostrare una buona capacità di effettuare collegamenti interdisciplinari</i>
Criteri di misurazione dell'apprendimento e di attribuzione del voto finale	<p><i>Il voto finale è attribuito in trentesimi. L'esame si intende superato quando il voto è maggiore o uguale a 18. Per la formulazione del voto finali si prenderanno in considerazione i seguenti indicatori: grado di conoscenza dei contenuti e degli argomenti dell'insegnamento, capacità e correttezza nell'applicare i concetti fondamentali trattati durante le lezioni frontali e le esercitazioni, qualità della esposizione orale.</i></p> <p><i>Tutti gli argomenti del programma contribuiscono in modo uguale alla formulazione del voto finale.</i></p>
Altro	

FAC.SIMILE SCHEDA DI INSEGNAMENTO IN LINGUA INGLESE
COURSE OF STUDY: Master Course in Bioinformatics (LM-8)
ACADEMIC YEAR: 2023-2024
ACADEMIC SUBJECT: Numerical Methods in Bioinformatics

General information	
Year of the course	First
Academic calendar (starting and ending date)	First Semester (02.10.2023 - 19.01.2024)
Credits (CFU/ETCS):	6
SSD	MAT/08
Language	Italian
Mode of attendance	Teaching attendance strongly recommended

Professor/ Lecturer	
Name and Surname	<i>Del Buono Nicoletta e Flavia Esposito</i>
E-mail	<i>nicoletta.delbuono@uniba.it, flavia.esposito@uniba.it</i>
Telephone	<i>0805442711</i>
Department and address	<i>Room 24, II floor, Department of Mathematics</i>
Virtual room	<i>cuzrciw</i>
Office Hours (and modalities: e.g., by appointment, online, etc.)	<i>By appointment, to be arranged by e-mail; in-person or remote</i>

Work schedule			
Hours			
Total	Lectures	Hands-on (laboratory, workshops, working groups, seminars, field trips)	Out-of-class study hours/ Self-study hours
<i>150</i>	<i>32</i>	<i>30</i>	<i>88</i>
CFU/ETCS			
<i>6</i>	<i>4</i>	<i>2</i>	

Learning Objectives	Acquisition of classical numerical methodologies and linear algebra techniques for the analysis of structured biomedical data.
Course prerequisites	Basic knowledge of mathematics consistent with that required for admission to the degree program

Teaching strategie	- Frontal teaching conducted with the aid of slides. - Computer exercises with the support of codes in open-source languages provided by the lecturer and use of benchmark data stored in online repositories
Expected learning outcomes in terms of	
Knowledge and understanding on:	<ul style="list-style-type: none"> ○ Mathematical formalization of data analysis problems ○ Numerical approaches and techniques for data analysis ○ Understanding and use of numerical codes for analysis of biological data sets
Applying knowledge and understanding on:	<ul style="list-style-type: none"> ○ Use of basic linear algebra mechanisms for biomedical data analysis. ○ Understanding and implementation of workflows for numerical analysis of biomedical data.
Soft skills	<i>Making informed judgments and choices</i> <ul style="list-style-type: none"> ○ Understanding the structure of the proposed data and identifying the most appropriate methodology to describe it ○ Identifying the appropriate technique to answer the questions posed

	<ul style="list-style-type: none"> ○ Managing the presence of outliers, missing data and know how to handle any normalizations ○ Implementing a workflow for data analysis <p><i>Communicating knowledge and understanding</i></p> <ul style="list-style-type: none"> ○ <i>Explanation of how to identify the most appropriate numerical methods for the analysis of biological data</i> ○ <i>Identification and explanation of the most appropriate numerical technique for answering research questions related to available data</i> ○ <i>Ability to effectively communicate how to Implement a chosen workflow for the analysis of the biomedical data to be analyzed</i> <p><i>Capacities to continue learning</i></p> <ul style="list-style-type: none"> ○ Explaining the concepts acquired and developing the ability to choose the most appropriate numerical technique for analyzing the data under investigation. ○ Use the appropriate mathematical formalism to describe the data analysis process
--	---

Syllabus	
Content knowledge	<p>-<i>Linear Algebra for Bioinformatics: concepts of vector, matrix, vector and matrix operations. Concepts of matrix rules of vector spaces, spaces of matrices, generated vector spaces. Bases of vector and matrix spaces. Linear systems and their matrix representation. Product of data matrices and factorisations of data matrices. QR factorisation and concepts of orthogonality between vectors and matrices. Methods for calculating eigenvalues and eigenvectors of a data matrix. Singular Value Decomposition (SVD) and truncated SVD.</i></p> <p>-<i>Principal component analysis (PCA) for multivariate data sets: covariance and correlation matrix of data. Diagonalisation of a symmetric matrix. Eigenvalues of a positive definite quadratic form. Principal components of vectors of random variables. Linear combinations of maximum variance. Statistical and probabilistic significance of principal components. Heuristics for estimating the optimal number of principal components. PCA as feature selection method. Interpretation of principal components: the correlation circle, Z-scores. Real applications of PCA.</i></p> <p>-<i>Notes on low rank approximation methods for omics data matrices. SVD, Nonnegative matrix factorizations (NMF). Sparse Nonnegative Matrix Factorization for Clustering.</i></p> <p>-<i>Introduction to the clustering problem, useful distances for clustering. Hierarchical divisive and agglomerative clustering (dendrogram representations and linkage methods: single, complete, average), K-means. Heuristic methods for choosing the optimal number of clusters in K-means, Equivalence between NMF and K-means with relaxed orthogonality constraint (Theorems and demonstrations), K- medoid (outline). Internal Indices, External Indices and other goodness-of-fit measures of clustering algorithms. Model and Density Based Algorithms (Mixture Model and DBSCAN): outlines.</i></p> <p>-<i>Introduction to Simple and Multiple Linear Regression Models, 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. Evaluation measures of regression models. Simple and multiple logistic regression. Logistic regression and Bernullian and multinomial distributions.</i></p> <p>-<i>Computer exercises using the statistical software R https://www.r-project.org</i></p>
Texts and readings	<p><i>Carl D. Meyer - Matrix Analysis and Applied Linear Algebra: 1st (first). Siam, 2001. http://matrixanalysis.com/</i></p> <p><i>-Gilbert Strang - Linear Algebra and Learning from Data, Wellesley-Cambridge Press, 2019 https://math.mit.edu/~gs/learningfromdata/</i></p> <p><i>-Robert Gentleman, Vince Carey, Wolfgang Huber, Rafael Irizarry, Sandrine Dudoit,</i></p>

	<p><i>Bioinformatics and Computational Biology Solutions Using R and Bioconductor</i>, Springer 2005. (http://www.bioconductor.org/help/publications/books/bioinformatics-and-computational-biology-solutions/)</p> <p>-Matthew He, Sergey Petoukhov. <i>Mathematics of Bioinformatics: Theory, Practice, and Applications</i>. Wiley&Sons 2011 https://onlinelibrary.wiley.com/doi/book/10.1002/9780470904640</p> <p>- Michael C. Whitlock, Dolph Schluter .<i>Analisi statistica dei dati biologici</i>. Zanichelli</p> <p>-I.T. Jolliffe, <i>Principal Component Analysis, Second Edition, 2002, Springer</i> https://link.springer.com/book/10.1007/b98835</p>
Notes, additional materials	<i>Notes and slides provided by the instructor</i>
Repository	<i>Teaching materials will be made available on the MTeams class of the course</i>

Assessment	
Assessment methods	<i>Oral examination on the syllabus and exercises and/or project assigned by the lecturer</i>
Assessment criteria	<ul style="list-style-type: none"> • <i>Knowledge and understanding</i> <ul style="list-style-type: none"> • <i>Students must demonstrate adequate knowledge of the main topics of the course</i> • <i>Applying knowledge and understanding</i> <ul style="list-style-type: none"> • <i>Students must demonstrate adequate knowledge of the possible applications of the theoretical concepts and possess adequate ability to implement these applications</i> • <i>Autonomy of judgment</i> <ul style="list-style-type: none"> • <i>Students must demonstrate adequate autonomy in selecting the most appropriate theoretical concepts for solving practical problems.</i> • <i>Communicating knowledge and understanding</i> <ul style="list-style-type: none"> • <i>Students must demonstrate an adequate expository capacity of the studied topic</i> • <i>Communication skills</i> <ul style="list-style-type: none"> • <i>Students must demonstrate an adequate capacity in analysis and synthesis</i> • <i>Capacities to continue learning</i> <ul style="list-style-type: none"> • <i>Students must demonstrate a good ability to make interdisciplinary connections</i>
Final exam and grading criteria	The final grade is given in thirtieths. The exam is considered passed when the grade is greater than or equal to 18. The following indicators will be considered in formulating the final grade: degree of knowledge of the content and topics of the teaching, ability, and correctness in applying the fundamental concepts covered during the lectures and exercises, quality of oral exposition. All program topics contribute equally to the formulation of the final grade.
Further information	
	.