Informazioni generali	**Anno accademico 2022-2023**
Denominazione dell'insegnamento	**Metodi Numerici in Data Science**
Corso di studio	Matematica Triennale (L-35)
Anno di corso	Terzo
Periodo di erogazione	Secondo semestre (28 febbraio 2023 – 31 maggio 2023)
Crediti formativi universitari (CFU)	7
Settore scientifico disciplinare (SSD)	MAT/08 – Analisi Numerica
Lingua di erogazione	Italiano
Obbligo di frequenza	No

Docenti		
Nome e cognome	Flavia Esposito (titolare)	Nicoletta Del Buono
E-mail	flavia.esposito@uniba.it	nicoletta.delbuono@uniba.it
Telefono		+39 080 544 2711
Sede	Dipartimento di Matematica Stanza 17, secondo piano	Dipartimento di Matematica Stanza 23, secondo piano
Sede virtuale	Microsoft Teams: codice 687hs39	
Pagina web	https://www.dm.uniba.it/members/esposito	https://www.dm.uniba.it/members/delbuono
Orario e modalità di ricevimento	Su appuntamento, da concordare per e-mail; in presenza o in remoto	Su appuntamento, da concordare per e-mail; in presenza o in remoto

Syllabus	
Obiettivi formativi	Acquisizione delle tecniche numeriche di base per le applicazioni di data science. Acquisizione delle tecniche numeriche di base per la Exploratory Data Analysis e dei metodi di algebra lineare numerica per il trattamento di alcuni problemi di apprendimento dai dati
Prerequisiti	Le conoscenze che in genere vengono acquisite nella laurea della classe L-35 con riferimento particolare alle discipline di Calcolo Numerico e della Analisi Matematica classica in una e più variabili
Contenuti dell'insegnamento	- Introduzione ai metodi di fattorizzazione per la Data Science. Algebra Lineare e apprendimento da dati strutturati. Rappresentazione di dati attraverso vettori e matrici. Decomposizione a valori singolari di una matrice di dati. Proprietà dei vettori singolari destri e sinistri e loro relazione con gli spazi vettoriali fondamentali generati da una matrice di dati. L'importanza della SVD nella Data Science: esempi. La trasformazione di Karhunen-Loève e la sua relazione con la SVD. Componenti principali e migliore approssimazione low-rank di una matrice di dati. Forma ridotta della SVD e teorema di Eckart-Young. Analisi di dati attraverso la fattorizzazione SVD della matrice di covarianza e di correlazione. Analisi delle componenti principali e equivalenza con il problema di massimo per la varianza dei dati. La geometria della PCA. Analisi di dati nonnegativi. Matrici positive e nonnegative. Teoremi di Perron-Frobenius. Modello Eigenface. - Il modello vettoriale dell'informazione (VSM) e il Latent Semantic Indexing. Introduzione alla formalizzazione algebrica dei problemi di information retrieval. Processo di indexing automatico, operazioni di stop-listing e stemming. Funzioni di pesatura degli index-term. Applicazione delle fattorizzazioni QR e SVD alla matrice termini-documenti e loro interpretazione geometrica. Approssimazione low-

	rank dello spazio semantico. Il processo di query-matching e misure di similarità. Confronto termine-termine e clustering di termini sinonimi. Accenno ai meccanismi di relevance feedback basati sull'utilizzo della SVD troncata della matrice termini-documenti. - Modelli basati su autovalori e autovettori per web information retrieval. La struttura a hyperlink del web e sua rappresentazione attraverso i grafi non orientati. Concetti di inlink e outlink. Modelli HITS e PageRank per il ranking di reti web. Costruzione delle matrici di Hub e Authority e loro proprietà. Costruzione della matrice di Google e sue proprietà. Riducibilità e grafi. Convergenza del modello PageRank e irriducibilità della matrice di Google. - Introduzione a problemi su reti e grafi. Esempi e problemi (problema dei ponti di Konigsberg e delle porte di una casa). Definizioni di cammini e cicli. Cammini euleriani e hamiltoniani. Grafi e rappresentazioni matriciali. Matrici di adiacenza, incidenza, matrice dei gradi e matrice laplaciana L_G di un grafo. Alcune proprietà spettrali della matrice di adiacenza e della matrice L_G. Autovalori e misure di connettività di un grafo. Cenni su grafi completi e loro proprietà. Problema del cammino di costo minimo e algoritmo di Dijkstra. - Introduzione al Machine Learning (problemi supervisionati e non supervisionati, introduzione alla classificazione e alla regressione con esempi, concetti di over e under fitting, trade-off bias-varianza), Struttura del dato (vettori, matrici e tensori), analisi esplorativa del dato (EDA), tipi di variabile e cenni di statistica descrittiva (misure di tendenza centrale, misure di variabilità, quantili) con relative rappresentazioni grafiche (box-plot, diagramma a barre, istogrammi, scatter plot). Ispezionare le relazioni tra variabili, Definizione del coefficiente di correlazione di Pearson. - Pre-processing del dato: definizione/classificazione e trattamento di missing values, definizione outliers e identificazione outliers, trasformazioni e normalizzazioni del dato. Il problema della Curse of Dimensionality e come risolverlo con tecniche di fattorizzazione lineare, Interpretazioni delle decomposizioni matriciali, Decomposizioni matriciali come sistemi di raccomandazioni (cenni). - Introduzione alla Nonnegative Matrix Factorization (NMF). Storia ed esempi. Motivazioni e interpretazione dei fattori non negativi, rappresentazione part-based e learning. Esempio eigenfaces e confronto tra VQ, PCA e NMF, Formalizzazione della NMF come problema di ottimizzazione matriciale. Analisi delle due funzioni obiettivo più utilizzate, interpretazione probabilistica della NMF. Divergenze Beta e di Bregman (definizioni ed e equivalenze). Interpretazione geometrica della NMF. Nonnegative Rank Factorization. Algoritmi numerici per il calcolo della NMF (Multiplicative Updates e Block Coordinate Descent). Teoremi di convergenza e dimostrazioni. Cenni sulla NMF regolarizzata. - 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. - Introduzione ai problemi supervisionati. Rischio Atteso e Rischio Atteso Empirico. Introduzione alla regressione, problema di under e overfitting nella regressione. 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 (errori SST, SSE, SSR, R^2, R^2 adjusted,RMSE, MAE, F statistica). Regressione polinomiale (cenni). - Introduzione ai problemi di classificazione. SVD per classificazione hand-written digits, K-Nearest Neighbors, Approcci ad albero per la classificazione (Alberi Decisionali e Random Forest). Support Vector Machines, Maximal Margin Classifier, Soft Margin Classifier, Teorema di Mercer, Misure per la bontà della Classificazione, Cross Validation (cenni). - Esercitazione in R degli argomenti trattati
Testi di riferimento	-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
Ulteriore materiale didattico	Appunti e riferimenti specifici forniti dal docente

Organizzazione della didattica				
	Totali	Didattica frontale	Pratica (esercitazioni e laboratorio)	Studio individuale
Ore	175	48	15	112
CFU	7	6	1	

Metodi didattici	
	- Lezioni frontali condotte con l'ausilio di supporti didattici (slide). - Esercitazioni al calcolatore.

Risultati di apprendimento previsti	
Conoscenza e capacità di comprensione	Acquisizione di alcune tecniche per la risoluzione di problemi di classificazione e clustering. Capacità di utilizzare codici numerici che implementano tecniche standard per l'analisi di dati reali.
Conoscenza e capacità di comprensione applicate	Le conoscenze teoriche e pratiche acquisite vengono utilizzate in campi matematici applicati e per risolvere problemi nel contesto dell'apprendimento dai dati.

Autonomia di giudizio	Capacità di individuare le tecniche numeriche adatte per affrontare e risolvere numericamente i problemi derivanti da applicazioni reali che coinvolgono i big data.
Abilità comunicative	Acquisizione del linguaggio e del formalismo matematico necessario per la consultazione e comprensione dei testi, l'esposizione delle conoscenze acquisite, la descrizione, l'analisi e la risoluzione dei problemi applicativi edi Data Science
Capacità di apprendere	Acquisizione di un metodo di studio adeguato, supportato della consultazione dei testi e dalla implementazione al calcolatore delle tecniche numeriche esposte durante il corso.

Valutazione	
Modalità di verifica dell'apprendimento	Prova Orale sul programma svolto nel corso delle lezioni ed esercitazioni o progetto assegnato dal docente
Criteri di valutazione	• *Conoscenza e capacità di comprensione*: Gli studenti devono dimostrare una adeguata conoscenza dei contenuti dell'insegnamento • *Conoscenza e capacità di comprensione applicate*: Gli studenti devono dimostrare una adeguata conoscenza delle possibili applicazioni dei concetti teorici e possedere una adeguata capacità di implementare tali applicazioni • *Autonomia di giudizio*: Gli studenti devono dimostrare una adeguata autonomia nella selezione dei concetti teorici più idonei alla risoluzione di problemi pratici • *Abilità comunicative*: Gli studenti devono dimostrare una adeguata capacità espositiva dei contenuti studiati e una adeguata capacità di analisi e sintesi • *Capacità di apprendere*: Gli studenti devono dimostrare una buona capacità di effettuare collegamenti interdisciplinari
Criteri di misurazione dell'apprendimento e di attribuzione del voto finale	Nella valutazione della prova orale e nell'attribuzione del voto finale varrà la seguente scala di valutazione dell'apprendimento: – Voto insufficiente (<18): Conoscenze frammentarie e superficiali dei contenuti, errori nell'applicare i concetti, esposizione carente – Voto 18-20: Conoscenze dei contenuti sufficienti ma generali, esposizione semplice, incertezze nell'applicazione di concetti teorici – Voto 21-23: Conoscenze dei contenuti appropriate ma non approfondite, capacità di applicare i concetti teorici, capacità di presentare i contenuti in modo semplice. – Voto 24-25: Conoscenze dei contenuti appropriate e ampie, discreta capacità di applicazione delle conoscenze, capacità di presentare i contenuti in modo articolato. – Voto 26-27: Conoscenze dei contenuti precise e complete, buona capacità di applicare le conoscenze, capacità di analisi, esposizione chiara e corretta – Voto 28-29: Conoscenze dei contenuti ampie, complete ed approfondite, buona applicazione dei contenuti, buona capacità di analisi e di sintesi, esposizione sicura e corretta – Voto 30 e 30 e lode: Conoscenze dei contenuti molto ampie, complete ed approfondite, capacità ben consolidata di applicare i

	contenuti, ottima capacità di analisi, di sintesi e di collegamenti interdisciplinari, padronanza di esposizione

Ulteriori informazioni	