Principali informazioni sull'insegnamento		
Denominazione dell'insegnamento	***Metodi Numerici in Data Science***	
Corso di studio	*Matematica Triennale*	
Anno di corso	*III*	
Crediti formativi universitari (CFU) / European Credit Transfer and Accumulation System (ECTS):		7
SSD	*MAT/08*	
Lingua di erogazione	*Italiano*	
Periodo di erogazione	*Secondo Semestre*	
Obbligo di frequenza	*A scelta*	

Docente	
Nome e cognome	Del Buono Nicoletta/Flavia Esposito
Indirizzo mail	nicoletta.delbuono@uniba.it / flavia.esposito@uniba.it
Telefono	3288269260
Sede	*Dipartimento di Matematica, piano II, stanze 24*
Sede virtuale	*Mteams: m5denc4*
Ricevimento (giorni, orari e modalità)	Lunedì ore 11:15-12:15 per appuntamento

Syllabus	
Obiettivi formativi	Acquisizione delle tecniche numeriche di base per l'ottimizzazione di funzionali non lineari in più variabili, per la soluzione di problemi di programmazione lineare. Introduzione alle tecniche matematiche per l'analisi esplorativa dei dati e all'uso dei metodi di ottimizzazione per affrontare 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 I e della Analisi Matematica classica in una e più variabili*
Contenuti di insegnamento (Programma)	- 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* *- A. N. Langville, C. D. Meyer: Google's PageRank and beyond. Princeton Univ. Press, 2006.* *- T. Hastie, R. Tibshirani J. Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* *Second Edition, 2009*
Note ai testi di riferimento	*Alcune dispense sono disponibili alla pagina https://www.dm.uniba.it/members/delbuono*

Organizzazione della didattica			
Ore			
Totali	Didattica frontale	Pratica (laboratorio, campo, esercitazione, altro)	Studio individuale
60	*52*	*8*	*113*
CFU/ETCS			
7	*6.5*	*0.5*	

Metodi didattici	
	Lezioni in aula ed esercitazioni in laboratorio informatico

Risultati di apprendimento previsti	
Conoscenza e capacità di comprensione	○ Acquisizione delle tecniche principali per la risoluzione di problemi di ottimizzazione di tipo continuo. ○ Capacità di realizzare codici numerici efficienti che implementano le tecniche acquisite. ○ Acquisizione degli elementi di base e della terminologia essenziale utilizzata nei contesti di Data Science.
Conoscenza e capacità di comprensione applicate	○ Le conoscenze teoriche e pratiche acquisite si utilizzano in vasta parte della matematica applicata e nella risoluzione di problematiche reali.
Competenze trasversali	• *Autonomia di giudizio* ○ Capacità di individuare le giuste tecniche numeriche per affrontare e risolvere numericamente problemi di ottimizzazione derivanti da

	applicazioni reali in cui sono coinvolti dati di grandi dimensioni. • *Abilità comunicative* ○ Acquisizione del linguaggio e del formalismo matematico avanzato necessario per la consultazione e comprensione dei testi, l'esposizione delle conoscenze acquisite, la descrizione, l'analisi e la risoluzione dei problemi applicativi e di Data Science • *Capacità di apprendere in modo autonomo* ○ Acquisizione di un metodo di studio adeguato, supportato della consultazione dei testi e dalla implementazione al calcolatore delle tecniche esposte durante il corso.

Valutazione	
Modalità di verifica dell'apprendimento	*Prova Orale su programma di esame o su progetto assegnato dal docente*
Criteri di valutazione	• *Conoscenza e capacità di comprensione:* ○ Conoscenza dei contenuti • *Conoscenza e capacità di comprensione applicate:* ○ Conoscenza delle applicazioni dei concetti teorici ○ Capacità di applicare i concetti teorici • *Autonomia di giudizio*: ○ Capacità di presentare i contenuti e valutare le possibilità di applicazione dei concetti teorici • Abilità comunicative: ○ Esposizione dei contenuti ○ Capacità di analisi e sintesi • Capacità di apprendere: ○ Capacità di collegamenti interdisciplinari
Criteri di misurazione dell'apprendimento e di attribuzione del voto finale	Nella valutazione della prova orale e nell'attribuzione del voto finale si farà riferimento alla 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
Altro	