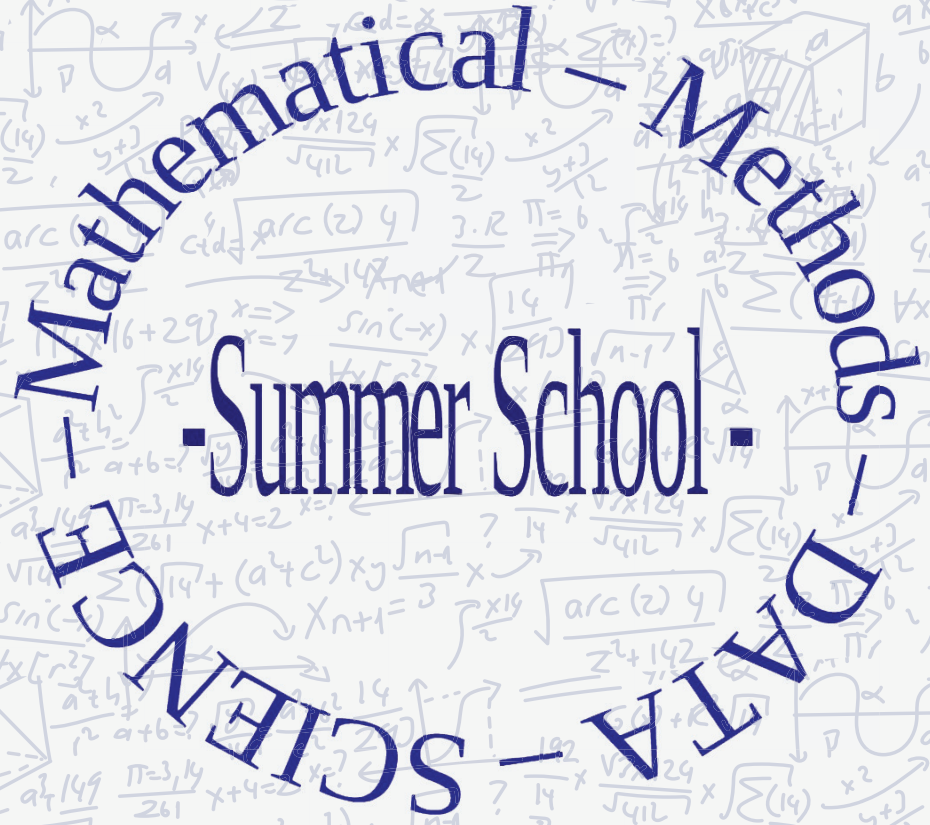
UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
DIPARTIMENTO DI MATEMATICA

INdAM
Unità di Ricerca
Università degli Studi di Bari Aldo Moro

Alumni Mathematica



Summer School Mathematical Methods in Data Science

Dipartimento di Matematica
Università degli Studi di Bari Aldo Moro
15-19 Luglio 2019

Indice

Benvenuti alla seconda edizione della Summer School MMDS 2019

La Summer School "Mathematical Methods in Data Science" è organizzata da:

- il Dipartimento di Matematica dell'Università degli Studi di Bari Aldo Moro,
- l'Associazione di Ricerca Scientifica Indipendente Alumni Mathematica,
- l'Unità di Ricerca INdAM (Istituto Nazionale di Alta Matematica) dell'Università degli Studi di Bari Aldo Moro.

L'evento propone un percorso formativo di cinque giorni incentrato sulle applicazioni della Matematica al Data Science (Scienza dei Dati).

Le giornate didattiche sono strutturate in modo da accompagnare i discenti nel graduale apprendimento di alcuni degli strumenti matematici più utilizzati nel contesto del Data Science. Gli argomenti affrontati durante la Scuola spaziano dall'analisi statistica dei dati, alla regressione lineare dei dati, ai meccanismi di riduzione di dimensionalità e estrazione di conoscenza latente da dati matriciali, fino all'introduzione dei concetti base del Deep learning. Durante ogni sessione didattica sono previste esercitazioni pratiche con i linguaggi di programmazione *R* e *Python*. Questi ambienti di lavoro open source sono introdotti durante la sessione preliminare di laboratorio del giorno martedì 16 luglio 2019.

Nella giornata di apertura della scuola sono previste specifiche relazioni tematiche tenute da esperti del settore e la *lectio magistralis* del professore Di Crescenzo, docente di Probabilità e Statistica Matematica presso il Dipartimento di Matematica dell'Università degli Studi di Salerno. Nell'arco della settimana sono proposte ai partecipanti ulteriori relazioni tematiche tenute dalle aziende sponsor della scuola. Completa il percorso di formazione la visita guidata al RECAS Data Center, una delle più potenti strutture di calcolo del Sud Italia.

Non mancheranno, inoltre, occasioni per lo scambio di idee e di opinioni: i partecipanti sono invitati ad approfittare dei momenti di incontro durante i coffee break e gli specifici "networking time".

Desideriamo, infine, ringraziare gli sponsor di questa edizione, il cui supporto è stato essenziale per l'organizzazione della scuola:

Buona scuola a tutti e non dimenticate di raccontare sui social la vostra esperienza con foto e post utilizzando l'hashtag #**MMDS19** e #**DataScienceBari**

ENJOY ☺

Il Comitato organizzatore: Anna Maria Candela, Nicoletta Del Buono, Flavia Esposito, Stefano Franco, Francesca Mazzia, Rosa Maria Mininni.

Programma

- Lunedì 15 Luglio 2019 - Aula I

08:30-09:30	**Registrazione partecipanti** *Aula XI*
09:30-10:15	**Apertura e Saluti di Benvenuto**
10:15-11:00	**Lectio Magistralis** Antonio Di Crescenzo *Modelli probabilistici e analisi statistica dei dati: alcuni casi di studio*
11:00-11:30	Coffee Break *Sala Riunioni*
11:30-12:00	**Seminario Tematico** Giuseppe Pirlo (CINI) *Stability Analysis in Data Sequences*
12:00-12:30	**Seminario Tematico** Valeria Ruggiero (GNCS) *Data Science: un punto di vista matematico*
12:30-13:00	**Seminario Tematico** Matteo Longoni e Gianpaolo Francesco Trotta (Moxoff) *Matematica e trasferimento tecnologico come asset di innovazione*
13:00-14:30	Lunch Time *Sala Riunioni*
14:30-16:10	**Sessione tecnica** Gianluigi Riccio *Data Sharing*
16:10-16:30	Coffee Break *Sala Riunioni*
16:30-19:00	**Sessione tecnica** Gianluigi Riccio *Data Sharing*

- Martedì 16 Luglio 2019 - Aula XI

09:00-10:40	**Sessione 1** Flavia Esposito *Laboratorio di Introduzione al linguaggio R*
10:40-11:00	Coffee Break *Sala Riunioni*
11:00-13:30	**Sessione 1** Flavia Esposito *Laboratorio di Introduzione al linguaggio Python*
13:30-14:30	Lunch Time *Sala Riunioni*
14:30-16:10	**Sessione 2** Nicoletta Del Buono *Analisi delle Componenti Principali: teoria e applicazioni*
16:10-16:30	Coffee Break *Sala Riunioni*
16:30-19:00	**Sessione 2: Laboratorio** Nicoletta Del Buono *Analisi delle Componenti Principali: teoria e applicazioni*

- Mercoledì 17 Luglio 2019 - Aula XI

09:00-10:40	**Sessione 2** Nicoletta Del Buono *Analisi delle Componenti Principali: teoria e applicazioni*
10:40-11:00	Coffee Break *Sala Riunioni*
11:00-12:40	**Sessione 2: Laboratorio** Nicoletta Del Buono *Analisi delle Componenti Principali: teoria e applicazioni*
12:40-13:30	**Seminario Aziendale** Alfredo Abrescia Finconsgroup
13:30-14:30	Lunch Time *Sala Riunioni*
14:30-16:10	**Sessione 3** Claudia Angelini *Clustering per Applicazioni Biomediche*
16:10-16:30	Coffee Break *Sala Riunioni*
16:30-19:00	**Sessione 3: Laboratorio** Claudia Angelini *Clustering per Applicazioni Biomediche*

- Giovedì 18 Luglio 2019 - Aula XI

09:00-10:40	**Sessione 4** Claudia Angelini *Regressione Lineare, approfondimenti e Applicazioni mediante R*
10:40-11:00	Coffee Break *Sala Riunioni*
11:00-12:40	**Sessione 4** Claudia Angelini *Regressione Lineare, approfondimenti e Applicazioni mediante R*
12:40-13:30	**Le aziende sponsor incontrano i partecipanti della MMDS Summer school**
13:30-14:30	Lunch Time *Sala Riunioni*
14:30-16:10	**Sessione 4: Laboratorio** Claudia Angelini *Regressione Lineare, approfondimenti e Applicazioni mediante R*
16:10-16:30	Coffee Break *Sala Riunioni*
16:30-19:00	**Sessione 4: Laboratorio** Claudia Angelini *Regressione Lineare, approfondimenti e Applicazioni mediante R*

- Venerdì 19 Luglio 2019 - Aula XI

09:00-10:40	**Sessione 5** Roberto Bellotti e Nicola Amoroso *Deep Learning: introduzione e un caso di studio*
10:40-11:00	Coffee Break *Sala Riunioni*
11:00-12:40	**Sessione 5** Roberto Bellotti e Nicola Amoroso *Deep Learning: introduzione e un caso di studio*
12:40-13:30	**Visita guidata al** **Recas Data Center**
13:30-14:30	Lunch Time *Sala Riunioni*
14:30-16:10	**Sessione 5: Laboratorio** Roberto Bellotti e Nicola Amoroso *Deep Learning: introduzione e un caso di studio*
16:10-16:30	Coffee Break *Sala Riunioni*
16:30-18:10	**Sessione 5: Laboratorio** Roberto Bellotti e Nicola Amoroso *Deep Learning: introduzione e un caso di studio*
18:10-19:00	**Chiusura dei lavori e consegna degli attestati**

Lectio Magistralis

Modelli probabilistici e analisi statistica dei dati: alcuni casi di studio

Antonio Di Crescenzo
Dipartimento di Matematica, Università degli Studi di Salerno,
Via Giovanni Paolo II n. 132; 84084 Fisciano (SA), Italy
Università degli Studi di Salerno
adicrescenzo@unisa.it

La costruzione e la validazione di modelli matematici per la descrizione di fenomeni reali si sta proponendo con crescente vigore nel panorama delle sfide poste agli studiosi moderni, per via delle grandi moli di dati che le nuove tecnologie rendono accessibili e più facilmente trattabili. Matematici, informatici, statistici e ingegneri sono sempre più frequentemente chiamati ad utilizzare in modo sinergico le loro competenze scientifiche per risolvere con metodo problemi attuali che richiedono l'estrazione d'informazione dai dati.

Ruolo essenziale è svolto dalla messa a punto di modelli probabilistici idonei alla formalizzazione dell'evoluzione temporale di sistemi soggetti a casualità, tra cui ricadono larga parte dei fenomeni correntemente sotto indagine. La costruzione di modelli matematici adeguati va poi accompagnata dalla necessaria verifica della loro bontà mediante l'analisi statistica dei dati, spesso basata su confronto tra previsioni teoriche e sintesi di natura statistica delle informazioni disponibili.

Prenderemo in esame alcuni recenti studi illustrativi basati su modelli evolutivi di natura probabilistica.

- La regione vulcanica dei Campi Flegrei è caratterizzata dal bradisismo, un fenomeno vulcanico consistente in moti verticali di tipo alternante in cui si susseguono fasi discendenti e ascendenti. Un modello stocastico idoneo a descrivere tali movimenti è basato su un processo di moto browniano guidato da un processo del telegrafo generalizzato. La conoscenza della legge di probabilità del processo stocastico in questione consente di effettuare l'analisi quantitativa di alcuni parametri rilevanti che regolano i processi di inflazione/deflazione, come velocità e costanti di tempo, e di realizzare previsioni sulla tendenza futura degli spostamenti. L'analisi statistica relativa si basa sulla regressione lineare con vincoli e su test statistici atti a confermare la bontà del modello.

- Nell'ambito dei processi di conteggio ampio interesse è riposto sul processo di Poisson e sue generalizzazioni, come i processi di Poisson composti. In particolare, tra questi rientra il cosiddetto processo di conteggio geometrico, caratterizzato dall'avere distribuzione marginale geometrica, tempi di interarrivo con distribuzione di Pareto, incrementi correlati positivamente, proprietà di sovradispersione, comportamento asintotico differente dal processo di Poisson. Ciò rende il processo geometrico idoneo a descrivere la dinamica temporale di certi fenomeni d'interesse in sismologia ed in affidabilità del software. Ad esempio, tra i casi di studio esaminati rientrano le sequenze di terremoti significativi in Italia negli anni dal 1900 al 1999, e l'occorrenza di errori in sistemi elettronici per la telefonia.

Bibliografia

1. F. Travaglino, A. Di Crescenzo, B. Martinucci, R. Scarpa (2018) A new model of Campi Flegrei inflation and deflation episodes based on Brownian motion driven by the telegraph process. *Math. Geosci.* 50:961–975.

2. A. Di Crescenzo, F. Pellerey (2019) Some results and applications of geometric counting processes. *Methodol. Comput. Appl. Probab.* 21:203–233.

CV breve: Antonio Di Crescenzo è professore ordinario di Probabilità e Statistica Matematica presso il Dipartimento di Matematica dell'Università degli Studi di Salerno, e membro del collegio dei docenti del Dottorato di Ricerca in Matematica, Fisica ed Applicazioni. I suoi interessi di ricerca includono la teoria e la simulazione dei processi stocastici con applicazioni alla modellistica in biomatematica e ai sistemi di file d'attesa. Si dedica inoltre a problemi nell'ambito della teoria dell'affidabilità anche con l'intento di fornirne applicazioni in altri settori disciplinari, particolarmente in biocibernetica e modellistica stocastica. È autore di numerose pubblicazioni su riviste internazionali, ed è membro del comitato editoriale di alcune riviste scientifiche internazionali. È stato tutor per diverse tesi di dottorato di ricerca. Ha partecipato all'organizzazione di vari congressi internazionali e workshop. È inserito nell'albo "Register of Expert Peer Reviewers for Italian Scientific Evaluation".

Seminari Tematici

Stability Analysis in Data Sequences

Giuseppe Pirlo
Consorzio Interuniversitario Nazionale per l'Informatica
Dipartimento di Informatica, Università degli Studi di Bari Aldo Moro
giuseppe.pirlo@uniba.it

Questa presentazione illustra alcune recenti strategie per l'individuazione di regioni di stabilità e di variabilità in serie temporali. Tali strategie, che utilizzano tecniche innovative di Intelligenza Artificiale basate su algoritmi multipli di matching, consentono di sviluppare soluzioni originali per problemi in diversi domini applicativi. In particolare saranno mostrate alcune applicazioni nel dominio del "Failure Prediction" per dispositivi meccanici e di "Personalized Verification" nel dominio dei sistemi Biometrici per la verifica dell'identità personale attraverso l'analisi della firma manoscritta.

CV breve: Giuseppe Pirlo, laureato nel 1986 in Scienze dell'Informazione, è attualmente Professore Ordinario di Sistemi di Elaborazione delle Informazioni. Già Pro Rettore Vicario dell'Università degli Studi di Bari Aldo Moro è il Referente dell'Università di Bari per l'Agenda Digitale e le Smart City, delegato alla Digitalizzazione dell'Università degli Studi di Bari ed alla Sperimentazione di Reti 5G. Giuseppe Pirlo è il Presidente di AICA (Associazione Italiana per l'Informatica ed il Calcolo Automatico) della Sezione Puglia.
L'attività di ricerca di Giuseppe Pirlo riguarda diversi campi tra i quali Intelligenza Artificiale, Pattern Recognition, Sistemi Intelligenti, Document Processing, Sistemi Biometrici e architetture di calcolo ad elevate prestazioni. Su questi temi è autore/coautore di oltre 300 articoli su riviste internazionali e/o presentati a conferenze internazionali. Curatore di numerosi libri e guest editor di diversi special issue è associate editor e revisore di prestigiose riviste scientifiche internazionali. Si occupa da molti anni di problematiche legate all'innovazione, al trasferimento tecnologico ed alle competenze digitali.

Data Science: un punto di vista matematico

Valeria Ruggiero
Gruppo Nazionale per il Calcolo Scientifico-INDAM
Dipartimento di Matematica e Informatica, Università di Ferrara
valeria.ruggiero@unife.it

L'interesse per le tematiche del Data Science è aumentato esponenzialmente negli ultimi anni, portando alla nascita di una nuova figura professionale - quella del Data Scientist - definita dalla Harvard Business Review come "la professione più sexy del XXI secolo". Si intende fornire qualche spunto sul ruolo della matematica nel Machine Learning, ossia nella capacità di estrarre le connessioni di un insieme di dati, allo scopo di acquisire conoscenza per generare previsioni e prendere decisioni.

CV breve: Valeria Ruggiero si è laureata in Matematica all'Università di Ferrara nel 1978. Nel 1981 è diventata ricercatore presso l'Università di Ferrara e nel 1992 professore associato prima all'Università di Modena e poi a Ferrara. Dal 2000 è professore ordinario di Analisi Numerica all'Università di Ferrara. Dal 2013 è Direttore del Gruppo Nazionale di Calcolo Scientifico dell'Istituto Nazionale di Alta Matematica (INdAM). E' stata coordinatore nazionale di progetti di ricerca (PRIN 97 "Numerical Analysis: Methods and Mathematical Software", FIRB 2001 "Parallel algorithms and Nonlinear Numerical Optimization", PRIN 2008 "Optimization methods and software for inverse problems") finanziati dal MIUR. La sua attività di ricerca riguarda lo sviluppo e l'analisi di metodi numerici per sistemi di grandi dimensioni, il calcolo parallelo, l'ottimizzazione non lineare e le relative applicazioni. I suoi più recenti interessi riguardano i metodi variazionali per problemi inversi e, più specificatamente, i metodi del primo ordine a metrica variabile di tipo forward–backward e primali-duali perla ricostruzione di immagini. I contributi della sua attività di ricercasono contenuti in più di 50 pubblicazioni in riviste scientifiche internazionali, in atti di convegno e capitoli di libro, in alcuni pacchetti software e nella recente monografia "Inverse Imaging with Poisson data" di cui è coautore. E' membro dell'Editorial Board di Computational Optimization and Applications. (https://sites.google.com/a/unife.it/valeria-ruggiero).

Matematica e trasferimento tecnologico come asset di innovazione.

Matteo Longoni e Gianpaolo Francesco Trotta
MOXOFF S.p.A.
Spinoff del Politecnico di Milano
Via Simone Schiaffino 11/19 - 20159 Milano
www.moxoff.com
matteo.longoni@moxoff.com, gianpaolo.trotta@moxoff.com

Come i processi di digitalizzazione, le nuove frontiere di capacità di acquisire segnali e trasmetterli e i sempre più efficaci ed efficienti sistemi di storage, di calcolo e di visualizzazione delle informazioni alla base dell'Industry 4.0 favoriscono la più crescente richiesta di Data Scientist. Apprendere, estrarre, trasformare e trasferire conoscenza con il nostro lavoro. Dal mondo puramente accademico all'impresa nata e legata a quel mondo: una testimonianza di analogie e opportunità.

CV breve: Matteo Longoni, laurea in Ingegneria Aerospaziale – Master Aerodinamica, e tesi su meshing avanzato per applicazioni biomediche. Dopo un'esperienza di 2 anni in università su modellazione matematica e simulazioni e in un'azienda di progettazione e produzione impianti idraulici per aeronautica, è entrato nel 2010 nella allora neo-nata Moxoff. Qui ha iniziato con l'operatività di progetti in campi industriali e settori aziendali molto diversificati, ma tutti accomunati da un approccio basato su tecniche avanzate di modellistica matematica e statistica, simulazione e analisi di dati. Oggi ricopre il ruolo di sviluppo commerciale e responsabile delle attività di marketing e comunicazione, ed entra in contatto tutti i giorni con sfidanti problemi industriali, sempre nuovi, da affrontare e risolvere con soluzioni di innovazione industriale.

CV breve: Gianpaolo Francesco Trotta, laurea in Ingegneria Informatica e dottorato in Ingegneria Meccanica. Dopo una tesi sperimentale in Ingegneria Informatica durante la quale ha progettato e sviluppato sistemi di Interazione Uomo-Macchina basati su realtà Virtuale in ambito Biomedico, ha continuato approfondendo queste tematiche durante il dottorato progettando e sviluppando innovativi sistemi di supporto alle decisioni e interfacce immersive di interazione su tematiche Industry 4.0. Concluso il dottorato ha deciso di approfondire le dinamiche alla base di sistemi di data-leak, di estrazione dell'informazione e di previsione usando tecniche avanzate di modellistica matematica e statistica, simulazione e analisi, lavorando come Analyst and Software Developer in Moxoff.

Sessioni Didattiche

- **Sessione tecnica**: Data Sharing, Giuseppe Riccio e Lucio Fiamingo (pag. 15)
- **Sessione 1**: Laboratorio di Introduzione ai linguaggi Python e R, Flavia Esposito (pag. 15)
- **Sessione 2**: Analisi delle Componenti Principali: teoria e applicazioni, Nicoletta Del Buono (pag. 17)
- **Sessione 3**: Clustering per applicazioni Biomediche, Claudia Angelini (pag. 18)
- **Sessione 4**: Regressione lineare, approfondimenti a applicazioni mediante R, Claudia Angelini (pag. 18)
- **Sessione 5**: Deep Learning: introduzione e caso di studio, Roberto Bellotti e Nicola Amoroso (pag. 19)

Sessione tecnica: *Data Sharing*

Gianluigi Riccio e Lucio Fiamingo
AIS system, Datonix
gianluigi.riccio@gmail.com, lucio.fiamingo@gmail.com

In un mondo ideale, i dati potrebbero arrivare in data set puliti e facili da scaricare e tutto sarebbe fantastico.
Ma il mondo è crudele, quindi una volta che si entra in possesso di un data set da analizzare, bisogna pulirlo e metterlo in un formato usabile.
Una volta che i dati sono pronti per essere preparati, bisognerebbe esplorarli, ricavarne il profilo, trattare i valori mancanti e le anomalie, valutare le cardinalità e le statistiche di primo e secondo ordine.
Questo lavoro dovrebbe essere strutturato e possibilmente automatizzato in modo tale da non sprecare il tempo e le energie dei data Scientist, che in tutto questo si sta focalizzando sempre di più nell'intelligenza artificiale perché le tecniche di Deep Learning e Reinforcement Learning, stanno trasformando il modo di usare i dati e stanno migliorando molte applicazioni data-driven (diagnosi medica, sistemi di guida autonoma, robotica, etc.).

Durante la sessione gli ingegneri Fiamingo e Riccio presenteranno:

- la metodologie più diffuse ed agili di data engineering
- i requisiti delle infrastrutture di data sharing
- le caratteristiche e le possibilità delle Data Management Solution for Analytics di mercato
- le tecniche di AI più diffuse e più recenti,
- le problematiche relative all'incorporazione di AI nei prodotti di nuova generazione.

CV breve: Lucio Fiamingo è co-fondatore di AISystem, una start up innovativa che realizza soluzioni di intelligenza artificiale. Nato e cresciuto a Napoli si laurea in ingegneria elettronica nel 1986. Ha collaborato con molte aziende italiane e internazionali come : Italtel Telematica (IT); Ascom Authophon (CH); Ravisent Technologies (USA); IPM Group (IT) ricoprendo diversi ruoli e responsabilità. Tra i più significativi progetti realizzati da Lucio va segnalato il primo telefono senza fili italiano. Lucio detiene vari brevetti di soluzioni hardware, firmware, e software.

CV breve: Gianluigi Riccio è co-fondatore di AISystem. Prima di AI System, Gianluigi ha realizzato datonix, un Data Base per Soluzioni Analitiche basato sulla matematica frattale. Durante gli anni '90, per 8 anni, è stato ricercatore nel principale servizio di ricerca della società Meta Group negli Stati Uniti. Prima di ciò Gianluigi ha avuto ruoli esecutivi in società come QueryObject Inc. Aeritalia Saipa, Olivetti OPE, Gartner Group Italia, and Eureka Consortium. Gianluigi è ingegnere e Professore di Elettronica e Sistemi.

Sessione 1: *Laboratorio di Introduzione ai linguaggi Python e R*

Flavia Esposito
Dipartimento di Ingegneria Elettrica e dell'Informazione,
Politecnico di Bari
flavia.esposito@poliba.it

In questa sessione verranno introdotti i concetti preliminari dei linguaggi di programmazione R e Python. Si farà uso del software Anaconda e dei programmi preinstallati in questo per la programmazione in R e Python. Per entrambi i linguaggi, tramite esercitazioni dal vivo con i propri calcolatori, verranno illustrate le modalità di trattamento dei vari tipi di dati, di liste, vettori, matrici e dataframes. Verranno forniti gli strumenti per ispezionare un dataframe facendo uso dei dataset precaricati negli ambienti di lavoro. Infine verranno illustrati alcuni esempi di importazione e esportazione di dataset e risultati.

CV breve: Flavia Esposito è assegnista di ricerca presso il Dipartimento di Elettronica e Informatica del Politecnico di Bari. Ha conseguito il titolo di dottore di ricerca in Matematica nel 2019 presso l'Università degli Studi di Bari Aldo Moro discutendo una tesi dal titolo "Nonnegative Matrix Factorization for Knowledge Extraction from Biomedical and other real world data". Nello specifico lavora nel campo dell'analisi numerica e dei metodi di ottimizzazione di tipo low-rank per la fattorizzazione di grandi matrici di dati. Ha effettuato un periodo di ricerca in Belgio presso l'Universitè de Mons sotto la supervisione del Prof. Nicolas Gillis, uno dei massimi esperti mondiali nell'utlizzo delle tecniche di fattorizzazioni nonnegative. E' impegnata nel campo della divulgazione delle scienze matHematiche con l'associazione di promozione sociale Alumni Mathematica ed è fermamente convinta che la matematica sia alla base di gran parte dei processi applicativi moderni e che la collaborazione e lo scambio di idee con esperti di altri settori possa aiutare a mettere questa disciplina al servizio di tutti.

Sessione 2: *Analisi delle Componenti Principali: teoria e applicazioni*

Nicoletta Del Buono
Dipartimento di Matematica,
Università degli Studi di Bari Aldo Moro
nicoletta.delbuono@uniba.it

L'Analisi delle Componenti Principali (Principal Component Analysis - PCA) è un metodo matematico utilizzato per estrarre informazioni da un dataset rappresentandole tramite un numero di variabili ridotto rispetto a quello di partenza. Questo metodo trasforma le variabili correlate presenti nel dataset in studio in un nuovo insieme di variabili artificiali chiamate "componenti principali (PCs)" che rappresentano la maggior parte della varianza nei dati. Nella lezione si introduranno i concetti generali legati alla PCA, le nozioni di matrice di covarianza e correlazione dei dati, di assi principali e di rotazione dello spazio degli osservabili e i meccanismi matematici per calcolare le PCs. Si discuteranno alcune euristiche per il calcolo del numero di PCs da conservare e i meccanismi empirici per interpretare le PCs.
La fattibilità della PCA sarà illustrata mediante lo studio di alcuni dataset di benchmark elaborati mediante i linguaggi open source Python e R.

CV breve: Nicoletta Del Buono è professore associato di Analisi Numerica presso il Dipartimento di Matematica, dell'Università degli Studi di Bari Aldo Moro. L'attività di ricerca è stata principalmente rivolta allo studio di metodi numerici per equazioni differenziali ordinarie, calcolo di funzioni di matrici, fattorizzazioni di matrici e loro applicazioni. Più recentemente si è occupata di metodi di approssimazione low rank per matrici di dati, concentrandosi in particolare sulle fattorizzazioni di matrici con vincoli di non-negatività e su problematiche applicative quali a esempio, l'analisi di microarray e di dati biomedici, il clustering di collezioni di documenti testuali. E' autore di diversi lavori scientifici pubblicati su riviste internazionali e ha partecipato a numerosi convegni nazionali e internazionali nonché a diversi progetti di ricerca finanziati dall'Università degli Studi di Bari, dal gruppo INDAM-GNCS, dal MIUR e dalla Fondazione Cassa di Risparmio di Puglia. Completano l'attività scientifica alcuni periodi di ricerca all'estero svolti presso la Bath University (U.K.) e la North Carolina State University, Raleigh (USA).

Sessione 3: *Clustering per applicazioni Biomediche*

Claudia Angelini
Laboratorio di Statistica e Strumenti di Calcolo per la Bioinformatica
Istituto per le Applicazioni del Calcolo "Mauro Picone" - CNR, Napoli
c.angelini@na.iac.cnr.it

In questa lezione verrà introdotto il problema del clustering, saranno presentati i principali metodi per il clustering gerachico (di tipo aggregativo e divisivo) e per il clustering partizionale (K-means e PAM). Quindi, saranno illustrati i metodi per la scelta del numero di clusters e per la validazione di una clusterizzazione. Se il tempo lo consentirà, saranno forniti cenni relativamente al clustering mediante l'utilizzo di misture.

Le metodologie presentate saranno illustrate attraverso alcuni esempi svolti mediante R.

Sessione 4: *Regressione lineare, approfondimenti e applicazioni mediante R*

Claudia Angelini
Laboratorio di Statistica e Strumenti di Calcolo per la Bioinformatica
Istituto per le Applicazioni del Calcolo "Mauro Picone" - CNR, Napoli
c.angelini@na.iac.cnr.it

In questa lezione verrà presentato il modello di regressione lineare (semplice e multipla) e la tecnica dei minimi quadrati per la stima dei coefficienti di regressione. Saranno presentate le condizioni di Gauss-Markov e le principali proprietà del modello lineare con in minimi quadrati, sia da un punto di vista teorico che pratico. Quindi, saranno discussi i problemi connessi alla presenza degli outliers, alla multi-collinearità tra le variabili, e alla deviazione rispetto alle ipotesi del modello.
Successivamente saranno presentati i concetti alla base della selezione dei modelli, incluso la best subset selection. Inoltre, verranno presentate le principali tecniche di regressione penalizzata: Ridge regression, Lasso regression ed Elastic net regression e sarà introdotto il criterio della cross-validation per la scelta del parametro di regolarizzazione.

Le metodologie presentate saranno illustrate attraverso alcuni esempi svolti mediante R.

CV breve: La dott.ssa Claudia Angelini è ricercatrice presso l'Istituto per le Applicazioni del Calcolo, CNR. La sua attività principale è dedicata allo sviluppo di metodi statistici per l'analisi di dati biologici. È stata coordinatrice scientifica di diversi progetti, tra cui i progetto CNR-Bioinformatica "Metodi matematici e statistici per la genetica e la proteomica", progetti CNR-RSTL "Metodi di selezione delle variabili bayesiane con applicazione alla genomica" , il progetto LAGSHIP "InterOmics" e nel progetto FLAGSHIP "Epigenomica". È coautrice di oltre 50 articoli apparsi su riviste internazionali peer-reviewed e anche di numerose altre pubblicazioni in atti di conferenze e in capitolo di libri.

Sessione 5: *Deep Learning: introduzione e un caso di studio*

Roberto Bellotti e Nicola Amoroso
Dipartimento Interateneo di Fisica "M. Merlin"
Università degli studi di Bari "Aldo Moro"
Istituto Nazionale di Fisica Nucleare (Bari)
roberto.bellotti@ba.infn.it
nicola.amoroso@ba.infn.it

In una prima lezione affronteremo le tematiche generali del machine learning e specificatamente del deep learning. Dopo aver offerto un sintetico richiamo sulle problematiche che differenziano in modo sostanziale apprendimento supervisionato e non-supervisionato, si porrà l'accento sulle possibilità offerte dalla tecniche di deep learning di costruire un ponte tra questi due ambiti e offrire nuove soluzioni a problemi di ricerca concreti.

Nella seconda lezione affronteremo un caso studio specifico riguardante l'applicazione di queste tecniche alle neuroimmagini. Si discuterà in particolare l'applicazione del deep learning a (i) problemi di classificazione e (ii) regressione in ambito medicale, come per esempio la diagnosi precoce di malattie neurodegenerative e l'invecchiamento cerebrale.

Nelle esercitazioni utilizzeremo il linguaggio R per mostrare fattivamente come applicare le tecniche di deep learning ai due casi studio affrontati nella seconda lezione.

CV breve: Nicola Amoroso si è laureato con lode in Fisica presso l'Università degli studi di Bari "A Moro" ove ha anche ottenuto un dottorato in Fisica Applicata. Attualmente è un ricercatore a tempo determinato dell'Università di Bari e si occupa dello studio, dello sviluppo e dell'analisi di soluzioni basate sul paradigma delle reti complesse e del cloud computing per analisi di dati biomedicali provenienti da neuroimmagini. La sua attività di ricerca è principalmente dedicata alla progettazione e allo sviluppo di sistemi di supporto alle decisioni nell'ambito delle malattie neurodegenerative, specificatamente attraverso l'applicazione di tecniche di deep learning e, più in generale, di tecniche di analisi per big data.

CV breve: Roberto Bellotti è professore Ordinario di Fisica Applicata dal 2017. Co-autore di oltre 200 pubblicazioni scientifiche su riviste internazionali sul tema dell'estrazione dell'informazione da dati eterogenei: dalla fisica delle alte energie alla fisica medica, in particolare ha incentrato le sue più recenti attività di ricerca sui temi dell'intelligenza artificiale. Coordinatore del gruppo di Fisica Medica e responsabile sin dal 2008 di programmi di Alta Formazione concernenti il calcolo scientifico e le sua applicazioni, la sua attività di ricerca è tra l'altro dedicata ai temi di Industria 4.0 e l'utilizzo di tecniche di analisi di Big Data.

Seminario Aziendale

Il lato oscuro del Machine Learning e Deep Learning: il meccanismo di apprendimento dell'Intelligenza Artificiale

Alfredo Abrescia
Finconsgroup srl

Oggi le fonti di dati digitali crescono senza sosta, con dati di tutti i tipi che provengono da molteplici sorgenti. Ma se le analisi di questi dati sono necessarie quasi ovunque, gli approcci attuali per svilupparle sono lenti e costosi. II moderno mondo degli analytics si è allargato oltre il tradizionale mondo del data warehouse e si sono sviluppati metodi di analisi dati che automatizzano la costruzione di modelli analitici. Durante il seminario saranno affontate le principali tecnologie di Artificial Intelligence, Advanced Machine Learning e Deep Learning in grado di supportare il business aziendale grazie alla maturità tecnologica raggiunta sia nel calcolo computazionale sia nella capacità di analisi in real-time di enormi quantità di dati eterogenei.