Titolo insegnamento: Data analysis in experimental physics with machine learning
Analisi dati in fisica sperimentale con tecniche di machine learning

6 CFU, 48h

Corsi di studio
008510-101 Laurea Magistrale in Fisica ind. Fisica Nucleare e Subnucleare e Biomedica
008510-102 Laurea Magistrale in Fisica ind. Astrofisica e Fisica Teorica
008510-105 Laurea Magistrale in Fisica ind. Fisica del Sistema Meteoclimatico, Generale e delle Tecnologie Avanzate

Possibile collocamento: secondo semestre

Tipologia:
C=Affine o integrativo per il curriculum di Fisica Nucleare e Biomedica (concordato con i coordinatori didattici del curriculum)

Obiettivo formativo:
Obiettivo generale è quello di familiarizzare gli studenti con i metodi statistici più attuali per l'analisi dati in fisica sperimentale e con gli strumenti di calcolo più avanzati che si utilizzano. Si mostrerà come il machine learning possa essere applicato a problemi di analisi dati in fisica sperimentale, introducendo esempi completi tratti da applicazioni reali. Verrà data una particolare enfasi all'analisi preliminare e alla preparazione del dataset sperimentale, in particolare quando ci si confronta con grandi moli di dati ("big data"). Verranno anche fornite le basi per poter sfruttare le moderne infrastrutture di calcolo e di analisi dati distribuita nell'ambito della fisica sperimentale e le pratiche standard utilizzate per lo sviluppo software in ambito scientifico e data science.

Il corso si basa su un approccio diretto ("bottom-up") in cui i vari argomenti sono presentati a partire da problemi di analisi dati tratti dall'esperienza reale. Verranno quindi proposti esempi completi di analisi dati, che verranno seguiti da esercitazioni pratiche in aula informatica. L'obiettivo delle esercitazioni sarà quello di sviluppare le necessarie competenze per poter svolgere in completa autonomia un progetto di analisi dati e identificare le soluzioni più vantaggiose. Verranno discusse differenti classi di problemi e tecniche, al fine di poter confrontare caratteristiche e peculiarità di ciascuno strumento.

Modalità di verifica dell'apprendimento:
L'esame si svolgerà discutendo un'applicazione sviluppata dallo studente e concordata con il docente, fornendone il codice su jupyter notebook attraverso un sistema di versioning (eg github), che verrà discussa durante l'esame orale spaziando su tutto il programma.

Propedeuticità:

Nessuna propedeuticità obbligatoria, conoscenze di elementi di programmazione in python3 e di statistica

Programma:

Introduzione: statistical learning

- Analisi preliminare di un dataset: trattamento dei dati, riduzione dei dati, visualizzazione di dati, data correlation, data transformation, dataset e data frame, dimensionality reduction
- Approssimazione di funzioni tramite statistical learning: modelli lineari, tree-models, reti neurali artificiali, deep neural networks
- Discussione delle principali librerie python utilizzate in ambito scientifico: numpy, pandas, matplotlib, keras, tensorflow, pytorch, scikit-learn, xgboost.
- Good-practices per lo sviluppo software in ambito scientifico: riproducibilità dei risultati, sistemi di versioning (git)

Esempi completi di analisi dati: preparazione e valutazione del dataset, sviluppo dei modelli, valutazione e confronto dei risultati

- Fisica nucleare e subnucleare
    - Ricostruzione ottimale dell'energia di un calorimetro
    - Categorizzazione di segnale e fondo. Esempio di un'analisi in categorie di eventi, ottimizzazione della significanza statistica di un segnale. L'utilizzo di classificatori multivariati nell'ambito della scoperta del bosone di Higgs.
    - Jet tagging: CNN e RNN, trattamento di un jet come un'immagine
- Fisica biomedica
    - Trattamento delle immagini: estrazione delle features da un'immagine, image segmentation, image classification

Modelli e tecniche avanzate

- Come gestire le differenze tra dati reali e dati simulati
- Modelli parametrici: parametrized networks
- Simulazione della risposta di un calorimetro: adversarial networks, generative networks, normalizing flows
- Automatic Data Quality Monitor: variational auto-encoders

Cenni al calcolo e analisi distribuita nell'ambito della fisica sperimentale

- Infrastrutture e risorse di calcolo: da singoli computer ad una rete distribuita con farm medio/grandi
- Riduzione, replicazione, manipolazione dei dati
- Calcolo HTC distribuito e cloud

Modalità di insegnamento:
Oltre a lezioni frontali per 28h, l'insegnamento prevede 20h di sessioni hands-on presso le aule informatiche dove verranno sviluppati e discussi in modalità interattiva esempi completi di

analisi dati a partire dalla preparazione e valutazione del dataset. Per lo svolgimento degli esempi in laboratorio si utilizzerà Google colaboratory. In caso di necessità di lavorare offline per gli esercizi a casa, verranno fornite specifiche immagini docker.

Materiale didattico:
Slides e jupyter notebooks forniti dal docente

Testi consigliati:

- ML in python: Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd edition) https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/

Altre risorse possono essere trovate in rete:
https://www.deeplearningbook.org/
https://github.com/yandexdataschool/