



UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Big Data Management and Analysis in Physics Research

2627-2-FDS02Q024-FDS02Q02402

Obiettivi

1. Conoscenza e capacità di comprensione

Al termine del corso si acquisiranno:

- conoscenze dei concetti fondamentali dell'utilizzo di Big Data Analytics per l'interpretazione dei dati in Fisica

2. Conoscenza e capacità di comprensione applicate

Si acquisiranno capacità di :

- utilizzo degli strumenti di Big Data Analytics per l'interpretazione dei dati in Fisica

3. Autonomia di giudizio

Si sarà capaci di:

- Interpretare correttamente le scelte strategiche per l'analisi dei dati in Fisica

4. Abilità comunicative

Si saprà:

- esporre con chiarezza gli argomenti oggetto del corso, utilizzando correttamente il linguaggio scientifico.

5. Capacità di apprendimento

Verranno fornite:

- le competenze metodologiche per affrontare lo studio degli argomenti oggetto del corso.

Contenuti sintetici

Il Laboratorio intende fornire esempi dettagliati ed aggiornati dell'uso dei Bg Data Analytics nella ricerca in Fisica, con introduzione teorica alle varie metodologie, esempi su dati reali e possibilità di analizzarli in profondità casi concreti.

Programma esteso

Il corso è organizzato come un laboratorio informatico di 28 ore, interamente pratico su notebook Jupyter, strutturato in dieci sessioni da 2 ore più il tempo introduttivo di setup e la revisione finale/progetto.

Sessione 1 – Introduzione alla fisica delle alte energie e all'ecosistema scikit-hep

Basi del Modello Standard e degli acceleratori di particelle; la libreria `particle` e il database PDG; prima analisi di eventi reali di dimuoni dai CMS Open Data; la massa invariante come osservabile fisica chiave; panoramica del toolkit scikit-hep.

Sessione 2 – Formati dati HEP: file ROOT, uproot e awkward-array

Lettura di file ROOT con uproot; gestione di dati di collisione a lunghezza variabile con awkward-array; aritmetica dei 4-vettori con la libreria `vector`; ricostruzione della massa del bosone Z; elaborazione efficiente a blocchi (chunk) di grandi dataset.

Sessione 3 – Spettro dei dimuoni: ricerca di risonanze e analisi esplorativa

Costruzione dello spettro completo della massa invariante dei dimuoni da dati reali CMS su più intervalli di energia; ricerca automatica dei picchi; stima segnale/fondo con il metodo delle sideband; misura della risoluzione del rivelatore; studi di correlazione cinematica.

Sessione 4 – Feature engineering per il machine learning in HEP

4-vettori di Lorentz come base delle feature fisiche; simulazione di un segnale Higgs-a-quattro-muoni e del relativo fondo; misura del potere separatore delle feature con la divergenza di Jensen-Shannon; feature engineering motivata dalla fisica; energia trasversa mancante; analisi di correlazione tra feature.

Sessione 5 – Classificazione del bosone di Higgs con BDT e XGBoost

Il dataset dell'ATLAS Higgs Machine Learning Challenge; pre-elaborazione dei dati; la metrica AMS per la significatività di scoperta; addestramento di un classificatore XGBoost con pesi degli eventi; importanza delle feature; controlli di overtraining; prestazioni per categoria di evento.

Sessione 6 – Identificazione di particelle: il telescopio MAGIC per raggi gamma

Un classico flusso di lavoro di machine learning su dati tabellari; analisi esplorativa di dati da telescopio Cherenkov; confronto tra selezione classica a taglio e classificatori moderni; selezione della soglia motivata dalla fisica; importanza delle feature e interpretazione fisica.

Sessione 7 – Analisi statistica in HEP: test di ipotesi e metodo CLs

Esperimenti di conteggio e significatività di scoperta; verosimiglianza profilata con parametri di disturbo (nuisance); il metodo CLs per la determinazione di limiti di esclusione; modelli HistFactory professionali con pyhf; il grafico di sensibilità Brazil-band; test di bontà dell'adattamento.

Sessione 8 – Fit di picchi con iminuit: misura delle masse delle particelle

Modellazione di segnale e fondo; fit della risonanza J/? con un modello gaussiano più esponenziale; fit del bosone Z con un profilo di Voigt; ricerca di risonanze ignote (bump hunting) e il rischio dell'effetto look-elsewhere.

Sessione 9 – Rilevamento di anomalie per la ricerca di nuova fisica

Ricerche di nuova fisica indipendenti dal modello; autoencoder e Isolation Forest per il rilevamento di anomalie; visualizzazione dello spazio latente; mass-sculpting e la tecnica CWoLa (weak supervision); confronto tra metodi tramite curve ROC.

Sessione 10 – Deep learning per il jet tagging

Fisica e simulazione dei jet; immagini di jet e reti neurali convoluzionali; Particle Flow Network (Deep Sets) su dati a nuvola di punti; confronto con feature ad alto livello motivate dalla fisica; reti neurali su grafi con EdgeConv.

Ogni sessione unisce una breve introduzione teorica a esercitazioni guidate su notebook Jupyter con dati reali o realisticamente simulati CERN/ATLAS/CMS, e si conclude con esercizi pratici da completare in modo autonomo.

Prerequisiti

È consigliata una conoscenza di base della programmazione in Python (variabili, strutture di controllo, funzioni), ma non è obbligatoria, in quanto viene brevemente richiamata nelle prime sessioni. Non è richiesta alcuna conoscenza pregressa di fisica.

Modalità didattica

28 ore di didattica di laboratorio (pratica, al computer), organizzate in dieci sessioni da 2 ore ciascuna, ciascuna delle quali unisce una breve introduzione teorica a esercitazioni guidate su notebook Jupyter con dati reali e simulati di fisica delle particelle.

Materiale didattico

Una copia di quanto mostrato e altro materiale in inglese verrà fornito agli studenti.

Periodo di erogazione dell'insegnamento

Secondo semestre.

Modalità di verifica del profitto e valutazione

Esame orale. Discussione di esercizi assegnati nel corso del laboratorio.

Orario di ricevimento

Su appuntamento.

Sustainable Development Goals

ISTRUZIONE DI QUALITÀ | IMPRESE, INNOVAZIONE E INFRASTRUTTURE
