SYLLABUS DEL CORSO

Big Data Management and Analysis in Physics Research

2627-2-FDS02Q024-FDS02Q02402

Aims

1. Knowledge and Understanding
Upon completion of the course, students will acquire:
• knowledge of the fundamental concepts of using Big Data Analytics to interpret data in Physics
2. Applied Knowledge and Understanding
Students will acquire the following skills:
• use Big Data Analytics tools to interpret data in Physics
3. Independent Judgment
Students will be able to:
• correctly interpret strategic choices for data analysis in Physics
4. Communication Skills
Students will be able to:
• clearly explain the course topics, using appropriate scientific language.
5. Learning Skills
Students will be provided with:
• the methodological skills needed to study the course topics.

Contents

The Laboratory intends to provide detailed and updated examples of the use of Bg Data Analytics in Physics research, with a theoretical introduction to the various methodologies, examples of real data and the possibility of analyzing concrete cases in depth.

Detailed program

The course is organised as a 28-hour computer laboratory, entirely hands-on with Jupyter notebooks, structured in ten 2-hour sessions plus introductory setup and final review/project time.

Session 1 – Introduction to High Energy Physics and the scikit-hep Ecosystem
Basics of the Standard Model and particle colliders; the particle library and the PDG database; first analysis of real CMS Open Data dimuon events; the invariant mass as the key physics observable; overview of the scikit-hep toolkit.

Session 2 – HEP Data Formats: ROOT Files, uproot and awkward-array
Reading ROOT files with uproot; handling variable-length collision data with awkward-array; 4-vector arithmetic with the vector library; reconstructing the Z boson mass; efficient chunked processing of large datasets.

Session 3 – Dimuon Spectrum: Resonance Finding and Exploratory Analysis
Building the full dimuon invariant mass spectrum from real CMS data across multiple energy ranges; automated peak finding; signal-to-background estimation with the sideband method; detector resolution measurement; kinematic correlation studies.

Session 4 – Feature Engineering for HEP Machine Learning
Lorentz 4-vectors as the basis for physics features; simulating a Higgs-to-four-muons signal and background; measuring feature separation power with the Jensen-Shannon divergence; physics-motivated feature engineering; missing transverse energy; feature correlation analysis.

Session 5 – Higgs Boson Classification with BDTs and XGBoost
The ATLAS Higgs Machine Learning Challenge dataset; data preprocessing; the AMS metric for discovery significance; training an XGBoost classifier with event weights; feature importance; overtraining checks; performance across event categories.

Session 6 – Particle Identification: the MAGIC Gamma-Ray Telescope
A classic tabular machine learning workflow; exploratory analysis of Cherenkov telescope data; comparing classical cut-based selection to modern classifiers; physics-motivated threshold selection; feature importance and physical interpretation.

Session 7 – Statistical Analysis in HEP: Hypothesis Testing and CLs
Counting experiments and discovery significance; profile likelihood with nuisance parameters; the CLs method for setting exclusion limits; professional HistFactory models with pyhf; the Brazil-band sensitivity plot; goodness-of-fit testing.

Session 8 – Peak Fitting with iminuit: Measuring Particle Masses
Signal and background modelling; fitting the J/? resonance with a Gaussian-plus-exponential model; fitting the Z boson with a Voigt profile; bump hunting for unknown resonances and the danger of look-elsewhere effects.

Session 9 – Anomaly Detection for New Physics Searches
Model-agnostic searches for new physics; autoencoders and Isolation Forests for anomaly detection; latent-space visualisation; mass-sculpting and the CWoLa (weak supervision) technique; ROC comparison of methods.

Session 10 – Deep Learning for Jet Tagging
Jet physics and simulation; jet images and convolutional neural networks; Particle Flow Networks (Deep Sets) on point-cloud data; comparison with physics-motivated high-level features; graph neural networks with EdgeConv.

Each session combines a short theoretical introduction with guided Jupyter notebook exercises on real or realistically simulated CERN/ATLAS/CMS data, and closes with practice exercises to be completed independently.

Prerequisites

Basic knowledge of Python programming is recommended (variables, control flow, functions) but not mandatory, as it is briefly reviewed in the first sessions. No prior knowledge of physics is required.

Teaching form

28 hours of laboratory teaching (hands-on, computer-based), organised in ten 2-hour sessions, each combining a brief theoretical introduction with guided practice on Jupyter notebooks using real and simulated particle physics data.

Textbook and teaching resource

Slides and additional material in english will be provided to students.

Semester

Second semester.

Assessment method

Oral exam. Discussion of exercises proposed during the laboratory sessions.

Office hours

On appointment.

Sustainable Development Goals

QUALITY EDUCATION | INDUSTRY, INNOVATION AND INFRASTRUCTURE