UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Introduction To Machine Learning for Physics

2627-3-E3001Q093

Obiettivi

Questo corso introduce i principi fondamentali del Machine Learning attraverso la lente della fisica. Superando l'approccio "scatola nera" (black-box), esploreremo come i diversi algoritmi di Deep Learning vengano progettati e implementati nella ricerca. Gli studenti acquisiranno esperienza pratica applicando questi strumenti a svariati set di dati — inclusi gli eventi registrati dai rivelatori di fisica delle particelle, le onde gravitationali e le immagini astronomiche, mediche o ambientali — per poi concludere con un progetto finale basato su un test di ipotesi statistica. L'obiettivo è padroneggiare sia il "come" funzionano gli algoritmi, sia il "perché" della loro applicazione nelle scienze fisiche.

Contenuti sintetici

1. La cassetta degli attrezzi del fisico e fondamenti statistici
Dalla verosimiglianza alla perdita (Loss): Il Lemma di Neyman-Pearson; Strutture dati e Pre-elaborazione dei dati.

2. Shallow Learning e alberi di decisione
Boosted Decision Trees (BDT); Validazione in un regime a statistica limitata; Metriche per la scoperta.

3. Deep Learning: simmetrie e architetture
Deep Neural Networks (DNN); Convolutional Neural Networks (CNN); Graph Neural Networks (GNN).

4. Dati sequenziali e modelli generativi
Serie temporali (RNN e Attention); Modelli generativi: VAE (Autoencoder Variazionali), Modelli di diffusione e Normalizing Flows.

5. Inferenza statistica e il "risultato fisico"
Anomaly Detection (Rilevamento delle anomalie); Limit Setting (Definizione dei limiti); Teorema di Wilks.

Programma esteso

1. La cassetta degli attrezzi del fisico e fondamenti statistici

- Dalla verosimiglianza alla perdita (Loss): Il Lemma di Neyman-Pearson come fondamento del Machine Learning, dimostrando che la Binary Cross-Entropy (entropia incrociata binaria) è la log-verosimiglianza negativa di un processo di Bernoulli.
- Strutture dati: Andare oltre le tabelle (es. array di NumPy o dataframe di Pandas). Introduzione agli Awkward Arrays per dati di fisica "irregolari" (ragged, con numero variabile di particelle per evento) e a uproot per l'integrazione dei file di ROOT.
- Pre-elaborazione dei dati: Standardizzazione, gestione del mascheramento dei rivelatori/valori anomali (outliers) e importanza delle unità di misura fisiche nella riscalatura (scaling).

2. Shallow Learning e alberi di decisione

- Boosted Decision Trees (BDT): Perché rimangono la "baseline" (linea di riferimento) in diverse applicazioni scientifiche. Introduzione a XGBoost e LightGBM.
- Validazione in un regime a statistica limitata: K-folding, verifiche di overtraining (sovrallenamento) e il bilanciamento tra Bias e Variance (compromesso bias-varianza).
- Metriche per la scoperta: Passare dall'accuratezza (Accuracy) alle curve ROC, alla purezza (Purity), all'efficienza (Efficiency) e alla significatività (Significance).

3. Deep Learning: simmetrie e architetture

- Deep Neural Networks (DNN): Ottimizzazione tramite discesa del gradiente, utilizzando la Backpropagation (retropropagazione dell'errore) e la regola della catena (Chain Rule, interpretata come calcolo differenziale).
- Convolutional Neural Networks (CNN): Elaborazione delle immagini nei telescopi, calorimetri, satelliti e nella diagnostica medica.
- Graph Neural Networks (GNN): Risoluzione di problemi di riconoscimento di pattern (pattern recognition) nella fisica delle particelle.

4. Dati sequenziali e modelli generativi

- Serie temporali (RNN e Attention): Analisi degli impulsi elettronici nei rivelatori di particelle (es. LArTPC) o dei dati di deformazione (strain) delle onde gravitationali (es. LIGO). Introduzione al meccanismo di attenzione (Attention).
- Modelli generativi:
- VAE (Autoencoder Variazionali): Riduzione della dimensionalità e fisica dello spazio latente.
- Modelli di diffusione e Normalizing Flows: Stima della densità condizionale e loro applicazione per la simulazione veloce dei rivelatori.

5. Inferenza statistica e il "risultato fisico"

- Anomaly Detection (Rilevamento delle anomalie): Apprendimento non supervisionato (Autoencoder) per la ricerca di "Nuova Fisica" in assenza di un modello teorico.
- Limit Setting (Definizione dei limiti): Utilizzo dei punteggi (score) del Machine Learning come input per i fit di verosimiglianza profilo (Profile Likelihood Fits).
- Teorema di Wilks: Traduzione delle prestazioni del Machine Learning in significatività espressa in "Sigma" (?).

Prerequisiti

Laboratorio di Calcolo e Statistica. Le competenze e le conoscenze acquisite durante questo corso, sia sulla programmazione in Python che sui metodi statistici, sono considerate fondamentali e saranno date per assunte.

Modalità didattica

Lezioni: 50% Teoria / 50% Laboratorio pratico di programmazione (Hands-on).

Materiale didattico

- “Hands-On Machine Learning with Scikit-Learn and PyTorch” - Aurélien Géron (2025) - N.B. In biblioteca è disponibile una vecchia edizione dello stesso autore che copre i medesimi argomenti, intitolata “Hands-On Machine Learning with Scikit-Learn, Keras and Tensorflow”
- “Understanding Deep Learning” - Simon Prince
- Goodfellow, Bengio e Courville, “Deep Learning”, MIT Press
- “Pen and Paper Exercises in Machine Learning” - arXiv:2206.13446

Periodo di erogazione dell'insegnamento

Secondo semestre

Modalità di verifica del profitto e valutazione

Agli studenti verrà fornito un set di dati "alla cieca" (blind dataset) e dovranno completare uno specifico compito di analisi ideato dai docenti. Il progetto richiede il soddisfacimento di una serie di obiettivi tecnici e la documentazione del processo in una relazione scritta formale. Ogni relazione sarà valutata in base alla sua chiarezza, all'intuizione fisica e alla completezza. In seguito alla valutazione positiva della relazione, gli studenti accederanno a una prova orale. Questa sessione inizierà con una discussione tecnica sul progetto dello studente, per poi ampliarsi in un esame approfondito su qualsiasi argomento trattato durante il corso.

Orario di ricevimento

Il ricevimento si tiene su appuntamento, da concordare preventivamente via e-mail con il docente.

Sustainable Development Goals

ISTRUZIONE DI QUALITÁ | PARITÁ DI GENERE | IMPRESE, INNOVAZIONE E INFRASTRUTTURE