Studio del decadimento $H \to \tau\tau$ con dati dell'esperimento CMS

Sviluppo di un algoritmo di selezione degli eventi e implementazione di un classificatore binario (S/B) con tecniche di *machine learning*

(6 cfu)

Domenico Riccardi, Viola Floris

12 marzo 2022

Abstract

Lo studio che si vuole sviluppare punta all'identificazione degli eventi di decadimento del bosone di Higgs, $m_H = 125$ GeV, in due leptoni tau. A tale scopo si ha intenzione di utilizzare il dataset realizzato dalla collaborazione CMS che raccoglie gli eventi prodotti dalle collisioni pp a 8 TeV durante il 2012 (corrispondenti ad una luminosità integrata di 11.5 fb^{-1}). I dati *reali* sono accompagnati dalle simulazioni MC descriventi diversi canali di segnale e fondo per il processo d'interesse[1].

Volendo studiare $H \to \tau_\mu \tau_h \to \mu\nu_\mu\nu_\tau + hadron$ i sample MC da analizzare sono *gluon fusion*, ggH, e *vector boson fusion*, VBF, per il segnale (canali di produzione dominanti dell'Higgs a LHC), mentre per il background la produzione di coppie $t\bar{t}$, W+jet e $Z \to ll$ (Drell-Yan).

I sample in input per il machine learning verranno prodotti da un algoritmo di selezione degli eventi in C++/ROOT sviluppato nella prima parte del progetto, integrazione 9 cfu (si rimanda al relativo abstract per i dettagli sulle modalità di selezione e sulla fisica del processo). Dunque, si scriverà un codice in puro Python che inizialmente, prendendo in input i `NameFile_selected.root` mediante tools della libreria Uproot, produca un unico Pandas dataframe successivamente elaborato nella fase di *Data Preparation*, e.g. definizione della *Features Matrix* e del *Target Vector*.

In seguito si svilupperanno due algoritmi supervised learning, una Artificial Neural Network (ANN) e un Random Forest (RF), con lo scopo di realizzare un classificatore binario segnale-fondo (S/B) che prenda in input come features un appropriato set di variabili dai sample MC processati.

[1] http://opendata.web.cern.ch/record/12350

La fase di training verrà ripetuta separatamente sulle due simulazioni per il segnale, ggF e VBF. Questo risulta necessario in quanto i due canali producono eventi fisicamente diversi e dunque l'addestramento li deve trattare, alternativamente, come un ulteriore fondo oltre quelli già elencati.

Con l'obiettivo di migliorare le prestazioni degli algoritmi, prevediamo una fase di data preprocessing e di hyperparameter tuning dei due modelli. Si confronteranno e valuteranno le performance degli algoritmi implementati attraverso la ROC curve, l'*accuracy* e il plot dello score assegnato dai classificatori per le due categorie di eventi (S/B).

Infine produrremo i plot per le distribuzioni delle variabili fisiche considerando come label sia quelle del MC, ovvero segnale-fondo, sia quelle prodotte dal classificatore definite fissando una soglia ottimale sull'output score[2]. Dunque, si genereranno anche le predizioni dei due modelli di ML sui dati *veri* per estrarre il segnale.

[2]In dettaglio, il classificatore attribuisce ad ogni evento una predizione (probabilità) tra 0 e 1. Quindi fissare una threshold significa mappare tutti gli eventi sopra soglia come segnale (S), mentre quelli sotto soglia come fondo (B), ovvero attribuire loro una label nitida. Verranno esplorate varie possibilità per la scelta dell'*optimal threshold* dalla ROC curve e dalla Precision-Recall Curve.