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

Sviluppo di un algoritmo di selezione degli eventi e implementazione di un classificatore binario (S/B) con tecniche di ***machine learning*** (integrazione per i 9cfu)

Domenico Riccardi

9 marzo 2022

Abstract

L'analisi che si vuole sviluppare utilizza simulazioni MC ed eventi reali resi disponibili, come Open Data[1], dalla collaborazione CMS. Il particolare canale di segnale, oggetto di studio, è $H \to \tau_\mu \tau_h$ dove l'Higgs viene prodotto mediante *gluon fusion* (ggH) oppure *vector boson fusion* (VBF), figura 1, e decade in due leptoni tau che a loro volta decadono uno in $\tau \to \mu\nu_\mu\nu_\tau$ e l'altro adronicamente.

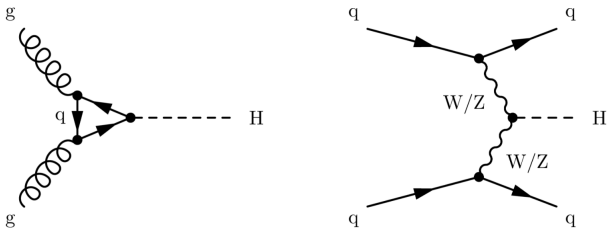

Figura 1: Diagrammi di Feynman che rappresentano, al leading order, i due modi di produzione dell'Higgs che si utilizzeranno nell'analisi: ggH (a sinistra) e VBF (a destra)

I fondi che contaminano questo canale sono diversi:

- $W + jets$: dovuto all'interazione tra due protoni che può produrre un bosone vettore W e che successivamente decade leptonicamente. Se questo viene selezionato assieme a un jet erroneamente ricostruito come un jet da tau, questo evento produce lo stesso stato finale del canale del segnale.

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

- Drell-Yan: eventi prodotti dall'annichilazione $q\bar{q}$. A seguito di quest'ultima può essere prodotto un bosone vettore Z che può decadere in una coppia $l\bar{l}$.
- $t\bar{t}$: si produce una coppia $t\bar{t}$ per *guon fusion* oppure per annichilazione $q\bar{q}$. Ciascun top decade essenzialmente in Wb e può produrre segnali simili a quelli spiegati sopra per il canale $W + jets$.
- QCD: eventi che presentano jet multipli.

I dati che si vuole utilizzare sono una parte della statistica raccolta nel 2012 da CMS e corrispondono ad una luminosità integrata $\mathcal{L} \simeq 11.5\,\mathrm{fb}^{-1}$ (Run2012B - Run2012C).

Le simulazioni MC comprendono sia i due canali di produzione dell'Higgs con successivo decadimento nel nostro canale di segnale, sia parte dei fondi elencati prima. In tabella 1 si raccolgono i processi che si intende adoperare nell'analisi assieme alla sezione d'urto e al numero di eventi che si utilizzeranno per calcolare l'opportuno fattore di scala (peso dell'evento) per confrontare i MC con i dati.

MC sample (ROOT files)	Cross section [pb]	Number of events
`GluGluToHToTauTau`	19.6	476963
`VBF_HToTauTau`	1.55	491653
`DYJetsToLL`	3503.7	30458871
`TTbar`	225.2	6423106
`W1JetsToLNu`	6381.2	29784800
`W2JetsToLNu`	2039.8	30693853
`W3JetsToLNu`	612.5	15241144

Tabella 1: Campioni MC 2012 a 8 TeV per l'analisi con annessa sezione d'urto e numero di eventi generati.

Dato l'elevato numero di eventi che è necessario analizzare, si vuole sviluppare un codice di analisi in puro linguaggio C++ che frutti funzionalità e classi del ROOT-framework (in particolare si userà la RDataFrame Class[2], con i relativi metodi, per manipolare i ROOT files). La selezione degli eventi verrà eseguita imponendo una serie di constraints sulle variabili contenute nei TTree dei ROOT files seguendo parzialmente l'analisi di questo canale svolta e pubblicata dalla collaborazione CMS[3]. L'analisi che si propone mira, dunque, alla selezione degli eventi lavorando sulle variabili cinematiche e calcolando nuove quantità che sono particolarmente rilevanti anche per la successiva fase di Machine Learning (High-level variables).

[2]Per aumentare la velocità d'esecuzione del codice, nonché per far si che il progetto tocchi il maggior numero di ambiti possibili trattati nel corso CMEPDA, si porrà enfasi anche sulla possibilità di parallelizzare l'analisi, multi-threading, offerta da questa classe.

[3]https://arxiv.org/pdf/1401.5041.pdf

Per ragioni essenzialmente di tempistica non verranno trattate: (i) la stima delle incertezze sistematiche, (ii) la determinazione del fondo di QCD, che comunque verrà ridotto al minimo richiedendo che i due leptoni nello stato finale abbiano carica opposta, (iii) l'analisi statistica e i fit alle distribuzioni.

Dopo aver prodotto i nuovi ROOT files (`FileName_selected.root`) si passerà alla rappresentazione grafica delle distribuzioni di tutte le varibili in essi contenute. Per fare questo si implementerà un nuovo script, questa volta in puro linguaggio Python[4] utilizzando alcune librerie in aggiunta a quelle standard trattate (Matplotlib, NumPy) come Uproot, Pandas e Seaborn. Il codice di plotting si interfaccerà, inoltre, con altri file Python condivisi con la parte di ML del progetto (in particolare un `infofile.py` che conterà tutte le informazioni necessarie per il plotting, e.g. numero di bins, range, titolo, labels per gli assi, etc.). Seguirà quindi la parte di Machine Learning.

NOTA: Questo abstract è da integrare con la proposta per il progetto da 6 cfu.

[4] Anche in questo caso per rendere il più completo possibile il progetto.