Filtración de Ruido en Eventos de Fusión de Quarks Top Usando MLP y Redes Convolucionales

Departamento de Física - Universidad de los Andes

John Mateus, Manuel Sánchez, Santiago Viteri

25 – Mayo – 2022 (Week 16)

1. Introducción

Los procesos de fusión de quarks Top que producen bosones hipotéticos llamados Z' suelen estar acompañandos de una señal de *background* que proviene de eventos descritos por el modelo estandar de partículas. Este *background* hace difícil la busqueda de fenómenos interesantes que van más allá del modelo estándar de la física de particulas. El muestreo de eventos inicia con 1200 imágenes proporcionadas para su posterior clasificación, las cuales vienen organizadas en dos grandes grupos: *Background* y *Signal*. Estas a su vez se organizan de la siguiente manera:

1. **Background:**
   - *a*) Procesos tipo $t\bar{t}h$, con 200 imágenes
   - *b*) Procesos tipo $t\bar{t}b\bar{b}/h$, con 200 imágenes
   - *c*) Procesos tipo $t\bar{t}t\bar{t}$, con 200 imágenes
2. **Signal:**
   - *a*) Procesos con $M(z') = 250$ GeV, con 200 imágenes
   - *b*) Procesos con $M(z') = 350$ GeV, con 200 imágenes
   - *c*) Procesos con $M(z') = 1000$ GeV, con 200 imágenes

donde $M(z')$ es la masa del bosón hipotético Z'. El requerimiento de los interesados[1] es poder clasificar una nueva imagen, proveniente de simulaciones o eventos provenientes de colisiones protón-protón en el LHC, en alguno de los dos grupos (*Signal* o *Background*), conociendo o postulando el valor usado para la masa del bosón Z'. Para lograr este objetivo se plantean dos métodos de análisis: *Multi-Layer Perceptron (MLP)* y *Redes Convolucionales*. En cada uno de los métodos se usarán las siguientes convenciones para clasificar los datos:

- **Clase 0:** Imágenes pertenecientes a los tres procesos tipo background.
- **Clase 1:** Imágenes pertenecientes a los tres procesos tipo signal.

El proceso entonces será:

1. Leer las imagenes y transformarlas en *arrays* para su análisis, tomando como *input* el array respectivo para cada imágen. Las imágenes vienen en formato .png y codificadas en RGB, para la cual se realiza el mapeo a escala de grises mediante la ecuación:
$$I = 0.299R + 0.587G + 0.114B\,, \tag{1}$$
de forma tal que a cada pixel de la imágen le corresponda un sólo valor de color.
2. A cada imágen se le asigna su clasificación como 0 ó 1 según corresponda, de esta manera se tendrá un array de *targets*.
3. Para MLP se toman el 75 % de los datos para entrenamiento y el 25 % restante se deja como test. Para la red convolucional se toma 85 % para entrenamiento y el 15 % restante como test.
4. Se establecen 1000 épocas de entrenamiento para MLP y 20 para la red convolucional.
5. Inicia el proceso de entrenamiento de cada método. Los resultados finales de interés serán las matrices de confusión y los valores de la exactitud (*accuracy*) de los métodos.

[1] Estudiantes del grupo de investigación en fenomenología de partículas de la Universidad de los Andes.

2. Resultados y Análisis

Se realiza la clasificación de los datos para entrenamiento y testeo (Fig. 1). El link del repositorio en GitHub se puede ver haciendo clic Acá!

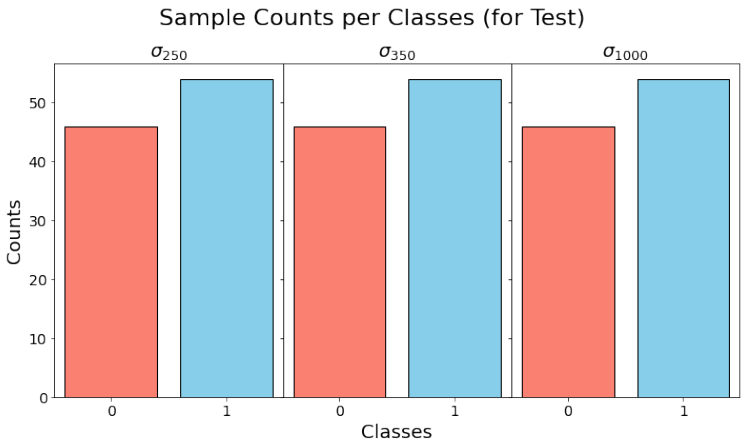

Figura 1: Clasificación de los datos de testeo para MLP. $\sigma_{250}, \sigma_{350}$ y σ_{1000} son los data sets formados a partir de los datos obtenidos de las imágenes. Los eventos de *background* para cada data set son los mismos.

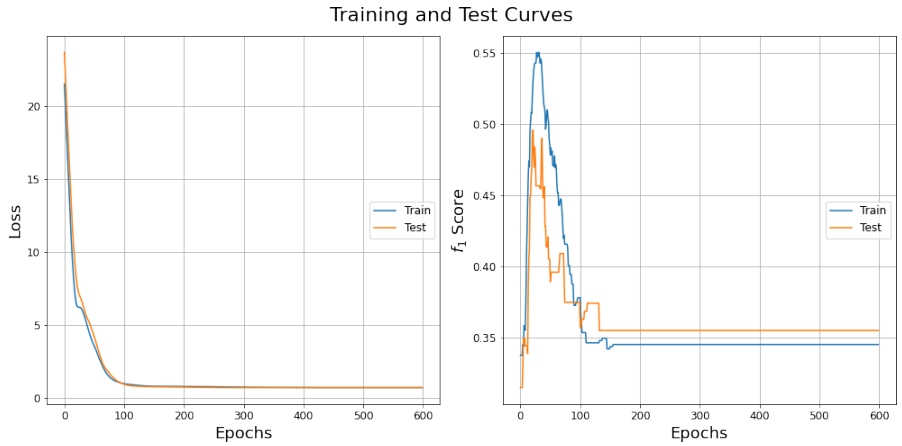

Figura 2: Comparación entre los datos de entrenamiento y testeo para el *loss* y el f_1-score tomando las imágenes en la categoria $C = 1$ con $M(z') = 250$ GeV, usando MLP como modelo de clasificación.

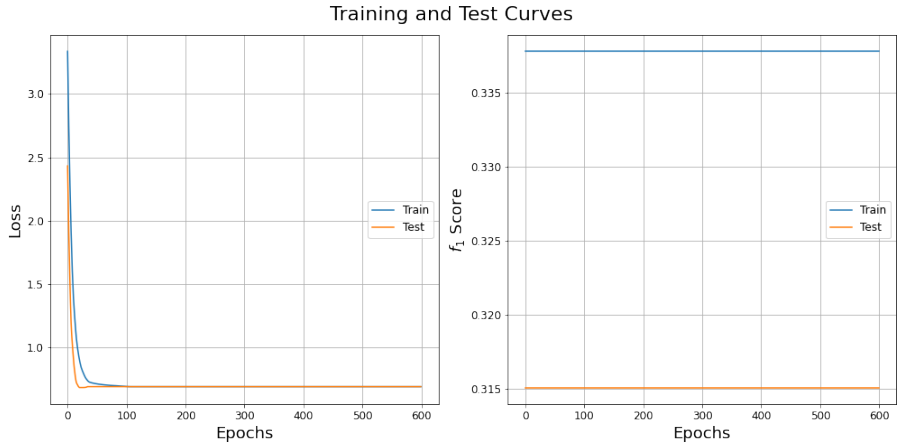

Figura 3: Comparación entre los datos de entrenamiento y testeo para el *loss* y el f_1-score tomando las imágenes en la categoria $C = 1$ con $M(z') = 350$ GeV, usando MLP como modelo de clasificación.

Las imágenes se agrupan de la siguiente manera para trabajar con MLP:

- Para cada $M(z')$ (250 GeV, 350 GeV y 1000GeV) se toman los tres grupos de background formando así tres grupos de testeo.

Se puede observar para cada gráfico en las figuras 2–4 un mínimo en los puntos para los cuales el número de épocas es 599, 21 y 599 para cada $M(z')$, con un loss-term de 0.684, 0.686 y 1.908 respectivamente.

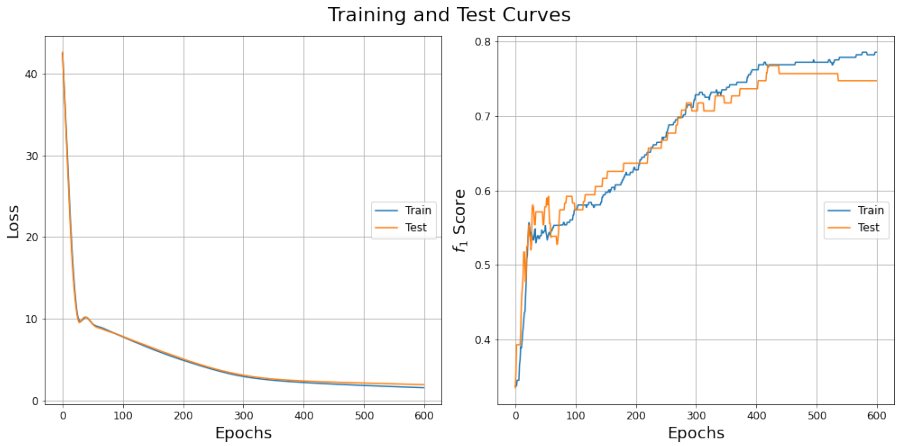

Figura 4: Comparación entre los datos de entrenamiento y testeo para el *loss* y el f_1-score tomando las imágenes en la categoria $C = 1$ con $M(z') = 1000$ GeV, usando MLP como modelo de clasificación.

Para las matrices de confusión los resultados se muestran en las figuras

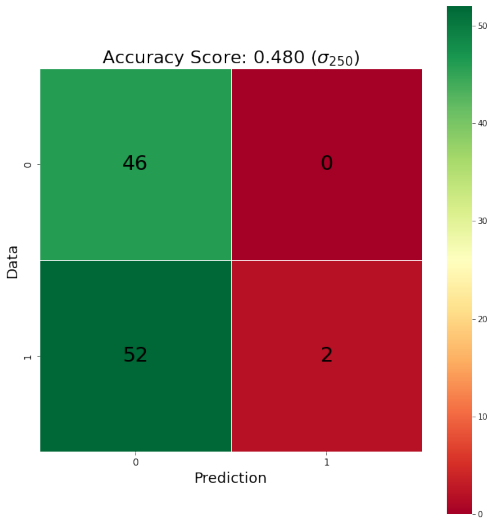

Figura 5: Matrices de confusión tomando las imágenes en la categoria $C = 1$ con $M(z') = 250$ GeV, usando MLP como modelo de clasificación.

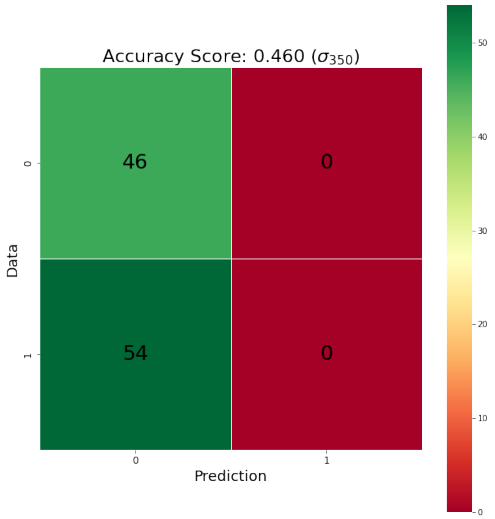

Figura 6: Matrices de confusión tomando las imágenes en la categoria $C = 1$ con $M(z') = 350$ GeV, usando MLP como modelo de clasificación.

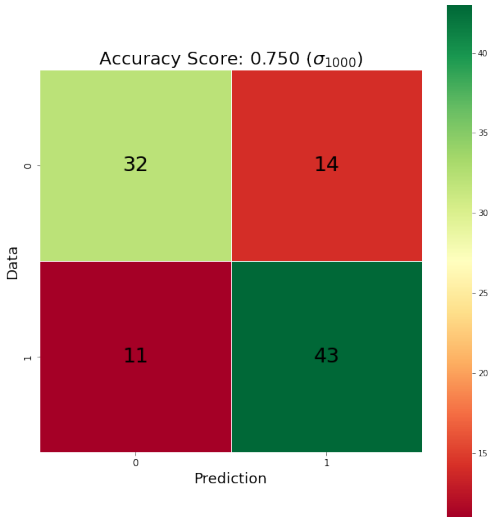

Figura 7: Matrices de confusión tomando las imágenes en la categoria $C = 1$ con $M(z') = 1000$ GeV, usando MLP como modelo de clasificación.

Los valores de *accuracy* para las distintas masas del bosón Z'van desde 0.460 hasta 0.750, mostrando que una mejor separación entre *Signal* y *Background* se da para una masa del bosón Z' de 1000 GeV.