Taller de Aprendizaje Automático

Proyecto 1 - Bosón de Higgs

Instituto de Ingeniería Eléctrica
Facultad de Ingeniería

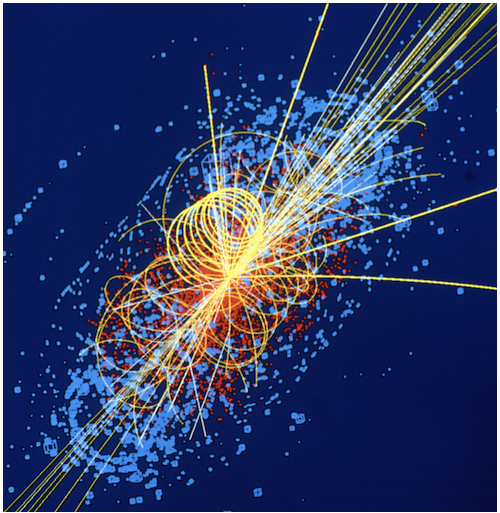
UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Montevideo, 2024

Introducción

Qué es el Bosón de Higgs ?

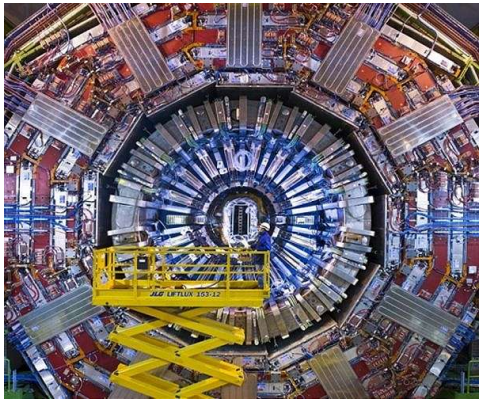
- Partícula de extrema importancia en la física de partículas.
- Propuesta por Higgs en 1964 para explicar por qué algunas partículas tienen masa.
- En 2012, la existencia de dicha particula pudo ser confirmada experimentalmente en el CERN (Centro Europeo de Investigación Nuclear)



Introducción

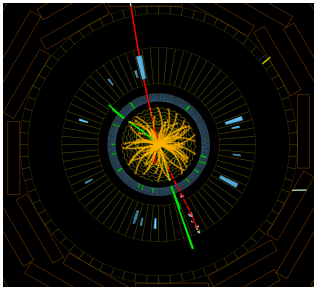
¿En qué consistió el experimento ?

- Haces de protones se aceleran en direcciones contrarias, hasta llegar casi a la velocidad de la luz.
  - Esta aceleración se realiza en el Large Hadron Collider: un anillo de 27 Km.
- Cuando dos protones chocan, se desintegran, produciendo decenas de nuevas partículas, que se disparan en todas direcciones
- El choque se produce dentro de un detector, llamado ATLAS
  - 46 metros de largo, 25 de alto y 25 de ancho, pesa 7000 toneladas y está ubicado en una caverna a 100 metros de profundidad.



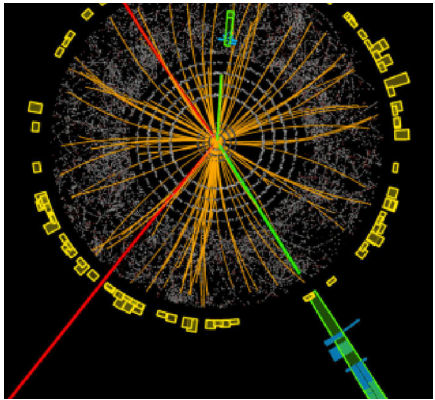
En que consistió el experimento ?

- Cada uno de estos "choques" se llaman "eventos". La cantidad de eventos que se generan es de aproximadamente 20 millones en un segundo. Cada evento contiene aproximadamente 10 partículas, que son "medidas" por el detector ATLAS en un conjunto de cientos de medidas.



Particularidades del experimento

- La mayoría de las partículas generadas por el choque NO son el Bosón de Higgs. Esto permite que se descarten la mayoría de los eventos, quedando unos 400 eventos por segundo para ser analizados.
- El Bosón de Higgs es altamente inestable: enseguida de crearse se descompone en otras partículas por lo cual su búsqueda es a partir de "rastros"'.
- En el caso del experimento ATLAS, fue la partícula *tau* la que se observó y que indirectamente confirma la presencia del Bosón de Higgs.



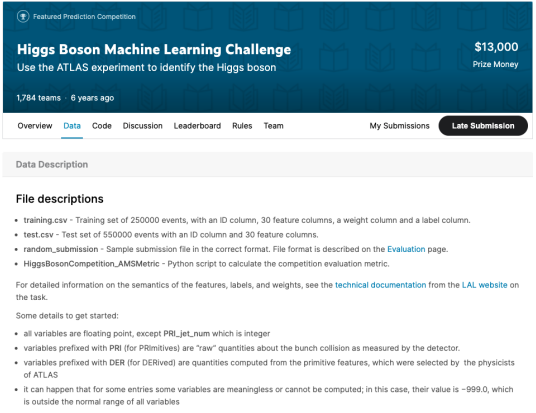
Proyecto 1 - Bosón de Higgs

Objetivo del proyecto

El objetivo del proyecto es diseñar e implementar un clasificador que permita diferenciar los eventos en los que el bosón de Higgs decayó en pares de leptones *tau* (llamados *signals*) de aquellos en que no (llamados *background*).

Particularidades del experimento

- Kaggle challenge: `https://www.kaggle.com/c/higgs-boson`
- Datos de training (training.csv): 250000 registros
- Datos de test (test.csv): 550000 registros
- Random submission: el formato para armar el envío
- HiggsBosonCompetition_AMSMetric.py: la función para calcular la métrica de evaluación



Detalles de los datos

- Cada evento está representado por un vector de 30 características.
- Algunas de estas características son obtenidas mediante mediciones directas en los sensores de ATLAS, mientras que otras características son derivadas a partir de las primeras.
- Todas las variables son en general en punto flotante.
- Todos los ángulos son en radianes entre $[-\pi, \pi]$
- El valor -999.0 indica una variable que no pudo ser medida o calculada, y es un valor por fuera del rango de todas las variables.

Detalles de los datos

- Cada evento está representado por un vector de 30 características
- Las variables que comienzan con "DER" son medidas derivadas
- Las variables que comienzan con "PRI" son medias directas

	EventId	DER_mass_MMC	DER_mass_transverse_met_lep	DER_mass_vis	DER_pt_h	DER_deltaeta_jet_jet
0	100000	138.470	51.655	97.827	27.980	0.91
1	100001	160.937	68.768	103.235	48.146	-999.00
2	100002	-999.000	162.172	125.953	35.635	-999.00
3	100003	143.905	81.417	80.943	0.414	-999.00
4	100004	175.864	16.915	134.805	16.405	-999.00

ling_eta	PRI_jet_leading_phi	PRI_jet_subleading_pt	PRI_jet_subleading_eta	PRI_jet_subleading_phi	PRI_jet_all_pt	Weight	Label
	0.444	46.062	1.24	-2.475	113.497	0.002653	s
	1.158	-999.000	-999.00	-999.000	46.226	2.233584	b
	-2.028	-999.000	-999.00	-999.000	44.251	2.347389	b
	-999.000	-999.000	-999.00	-999.000	-0.000	5.446378	b
	-999.000	-999.000	-999.00	-999.000	0.000	6.245333	b

...

Detalles de los datos

Las variables "EventId", "Weight" y "Label" no deben ser usados como datos de entrada para el entrenamiento o clasificación.

	EventId	DER_mass_MMC	DER_mass_transverse_met_lep	DER_mass_vis	DER_pt_h	DER_deltaeta_jet_jet
0	100000	138.470	51.655	97.827	27.980	0.91
1	100001	160.937	68.768	103.235	48.146	-999.00
2	100002	-999.000	162.172	125.953	35.635	-999.00
3	100003	143.905	81.417	80.943	0.414	-999.00
4	100004	175.864	16.915	134.805	16.405	-999.00

ling_eta	PRI_jet_leading_phi	PRI_jet_subleading_pt	PRI_jet_subleading_eta	PRI_jet_subleading_phi	PRI_jet_all_pt	Weight	Label
	0.444	46.062	1.24	-2.475	113.497	0.002653	s
	1.158	-999.000	-999.00	-999.000	46.226	2.233584	b
	-2.028	-999.000	-999.00	-999.000	44.251	2.347389	b
	-999.000	-999.000	-999.00	-999.000	-0.000	5.446378	b
	-999.000	-999.000	-999.00	-999.000	0.000	6.245333	b

...

Detalles de los datos

- training.csv - Training set of 250000 events, with an ID column, 30 feature columns, a weight column and a label column.
- test.csv - Test set of 550000 events with an ID column and 30 feature columns.

Links de interés

- Kaggle challenge: https://www.kaggle.com/c/higgs-boson
- Detalle del challenge: https://higgsml.lal.in2p3.fr/files/2014/04/documentation_v1.8.pdf

Una nota sobre los pesos

- El evento *signal* es mucho menos frecuente que el evento *background*: $s << b$.
- Los datos provistos son simulados, el desbalance entre las muestras de entrenamiento no es similar al de la situación real (TD: Analizar cuánto es).
- Para compensar este desbalance, se agrega una columna de pesos a todos los eventos (un factor de escala) que se usa durante el entrenamiento.
- ¿Es necesario tener en cuenta el desbalance? (¿Por qué ?)
- Cada evento tiene su peso propio, ya que los eventos no tienen todos la misma importancia.

Una nota sobre los pesos

Finalmente, la normalización tiene un significado físico. Supongamos que tenemos los siguientes datos de entrenamiento:

$$\mathcal{D} = \{(x_1, y_1, w_1), \ldots, (x_n, y_n, w_n)\}$$

definimos: $\mathcal{B} = \{i : Y_i = b\}$ y $\mathcal{S} = \{i : Y_i = s\}$.

- $\sum_{i \in \mathcal{S}} w_i = N_s$
- $\sum_{i \in \mathcal{B}} w_i = N_b$

son el valor esperado del total de eventos *signal* y *background* durante el el intervalo en que fueron adquiridos los datos.

Pregunta

¿Cuál es el valor esperado de *signal* y *background* en los datos de training ?

Medida de Evaluación

Sea g el clasificador a analizar y $\mathcal{G} = \{i : g(x_i) = s\}$. Definimos s y b como la tasa de verdaderos y falsos positivos (respectivamente). Debido a los pesos, estas tasas deben calcularse ajustadas a los mismos:

$$s = \sum_{i \in \mathcal{S} \cap \mathcal{G}} w_i$$

$$b = \sum_{i \in \mathcal{B} \cap \mathcal{G}} w_i$$

Estas dos medidas son estimadores no sesgados del número esperado de signals y backgrounds clasificados como signals por el clasificador:

$$\mu_s = N_s \int_{\mathcal{G}} p_s(x) dx$$

$$\mu_b = N_b \int_{\mathcal{G}} p_b(x) dx$$

Medida de Evaluación

La métrica de evaluación se llama *Approximate Median Significance*:

$$AMS = \sqrt{2\left((s + b + b_{reg})ln\left(1 + \frac{s}{b + b_{reg}}\right) - s\right)}$$

donde b_{reg} es un parámetro de regularización, que en nuestro caso vale 10.

Sobre la metodología de trabajo

- El objetivo didáctico del proyecto es que apliquen la metodología de trabajo vista en clase
- Recuerden revisar los conceptos vistos en clase
- En particular, revean el capítulo 2 de "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" - Aurélien Géron

Actividades y preguntas

- Exploración de los datos: entender su composición:
  - Tipo de datos
  - Datos faltantes
  - analizar correlaciones
- Entender el problema
  - Analizar el desbalance en los datos de entrenamiento
  - ¿ porqué hay que tener en cuenta el desbalance ?
  - ¿ cuál es el valor esperado de signal y background en los datos de entrenamiento ?
  - ¿ qué consideraciones hay que tener en cuenta al armar los conjuntos de validación y entrenamiento ?