Physics at the Large Hadron Collider: A Data Analytic Approach

Krish Vijayan (2019PH10634)
Ishaan Watts (2019PH10629)

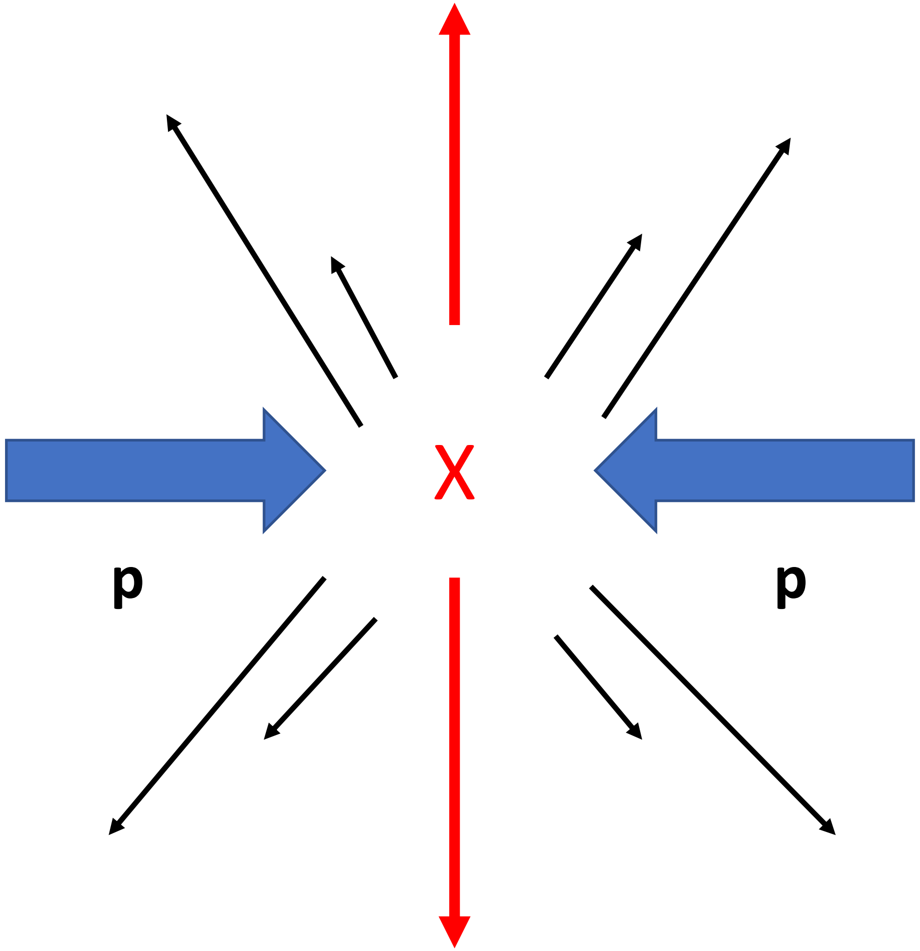
Advisor: Prof. Abhishek Muralidhar Iyer

Contents

Large Hadron Collider

The Large Hadron Collider at CERN, which is currently the world's largest and most powerful particle accelerator, has the potential to unlock interactions beyond the Standard Model (SM) of physics.

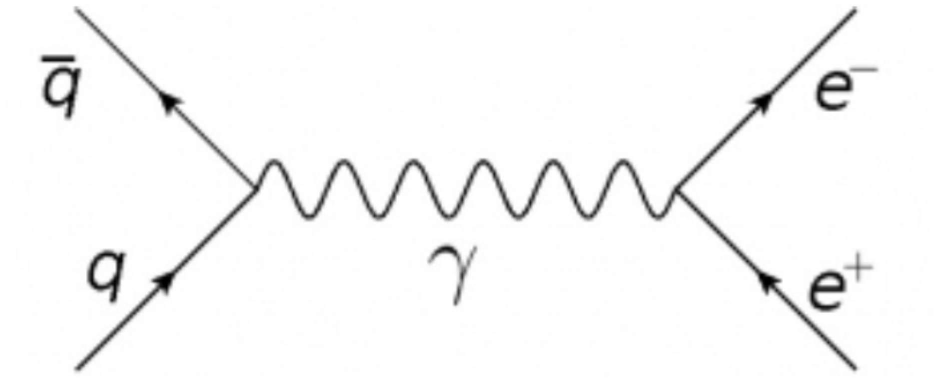
Four vectors are identified in terms of traditional objects depending on their imprint on the detector



Final states are four-vectors

Typical Process

A typical process which contributes to the dataset used in this project is the collision of a quark (q) and an anti-quark (q’) to become a photon (γ) which further decays into an electron (e-) and a positron (e+) in the final state.

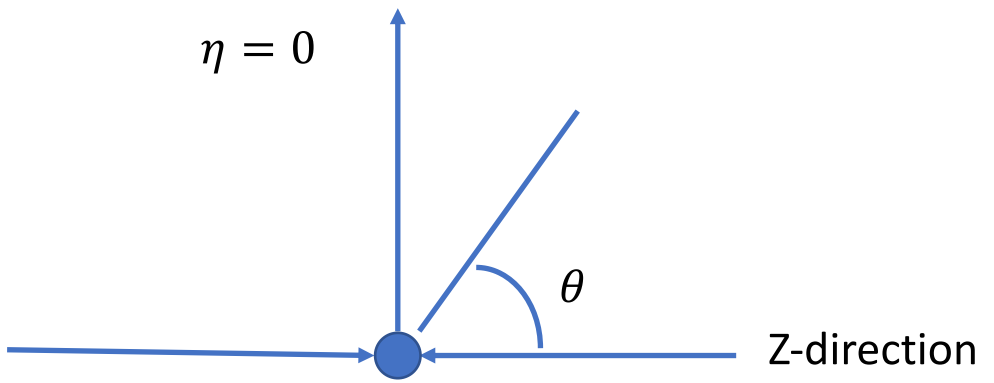


Typical Process

The final state of this process is a set of four-vectors from which different features can be constructed, which include:

$$p_T = \sqrt{p_x^2 + p_y^2},$$
$$\eta = -ln(tan(\frac{\theta}{2})).$$

Here p_T is the transverse momentum, η is the pseudorapidity and θ is the angle of deflection measured from the z direction.



Problem Statement

- Basic idea of LHC is to compare 2 datasets. How can one reject one dataset in favor of another.
- The discovery of an expected signal, such as the Higgs Boson, is associated with a p-value. The trained model should ideally predict whether a given event is an anomaly with a high probability based on features such as the energy and momenta of different kinds of particles.

Generation of Dataset

In order to reject one dataset in favour of the other, the difference in pseudorapidities of the final state particles is considered to be a distinguishing feature:

$$\Delta\eta = \eta_1 - \eta_1'$$

Another important metric which is relevant to the classification is the Signal Discovery Significance (Z), given by:

$$Z = \sqrt{\sum_{i=1}^{N}(2(s_i + b_i)log(1 + \frac{s_i}{b_i}) - 2s_i)}$$

(summed over all the bins, si and bi are the expected numbers for signal and background events in the i^{th} bin)

Two datasets are used for different tasks, which are:

1. A dataset containing a background and a signal (bump) with 46,612 and 4,601 values of $\Delta\eta$ respectively is used to calculate the signal discovery significance.

2. The dataset used to train the Neural Network models and classify signals is a signal and a background, which contain 122,854 and 133,619 values of $\Delta\eta$ respectively.

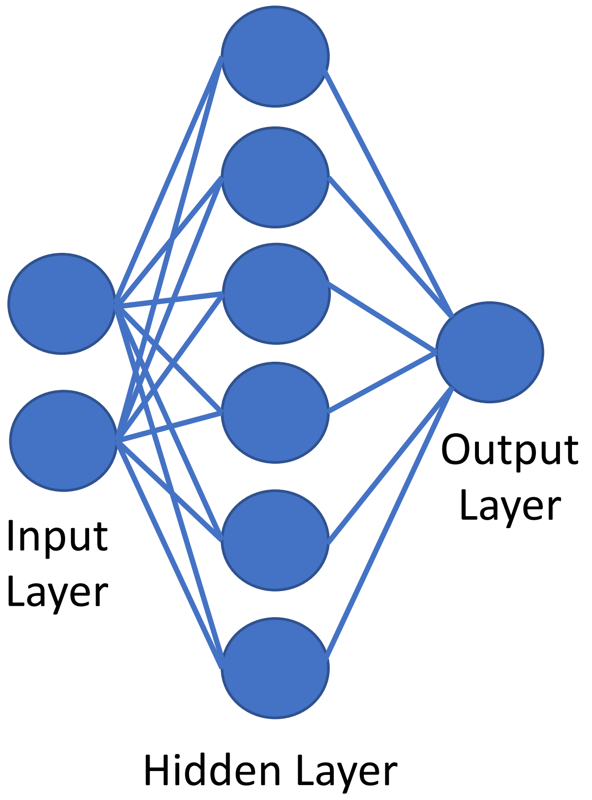
https://inspirehep.net/literature/1607893

Model Architecture

- An Artificial Neural Network (ANN), with one input layer, one hidden layer and an output layer is constructed.
- The input features used are delta-eta values and the bin density for each delta-eat values.
- We have used a binary cross entropy loss function, given by:

$$L(y,p) = -\sum_{i=1}^{m} \left(y_i \log(p_i) + (1-y_i)\log(1-p_i)\right)$$

- where $y_i \in \{0, 1\}$ is the actual class the i^{th} example belongs to and p_i is the probability with which the model predicts it to belong to the 1st class. This loss function is summed over all the m samples to get the total loss, which is minimized over 500 epochs.



Model Architecture

- The activation functions used between these layers are a Rectified Linear Unit (ReLU) and Sigmoid respectively. Mathematically, these functions are given by:

$$\sigma(z) = \frac{1}{1 + e^{-z}},$$

$$ReLU(z) = max(0, z).$$

- The choices for the loss optimizer are the Adam Optimizer, RMSProp and Stochastic Gradient Descent (SGD), which are the most common choices for training.

Results and Discussions

Plot of values from 1st dataset.

The value of Z for the given distribution is calculated to be 2183.79. A high value of Z indicates that for a sufficiently large background we can differentiate between signal and background curves.

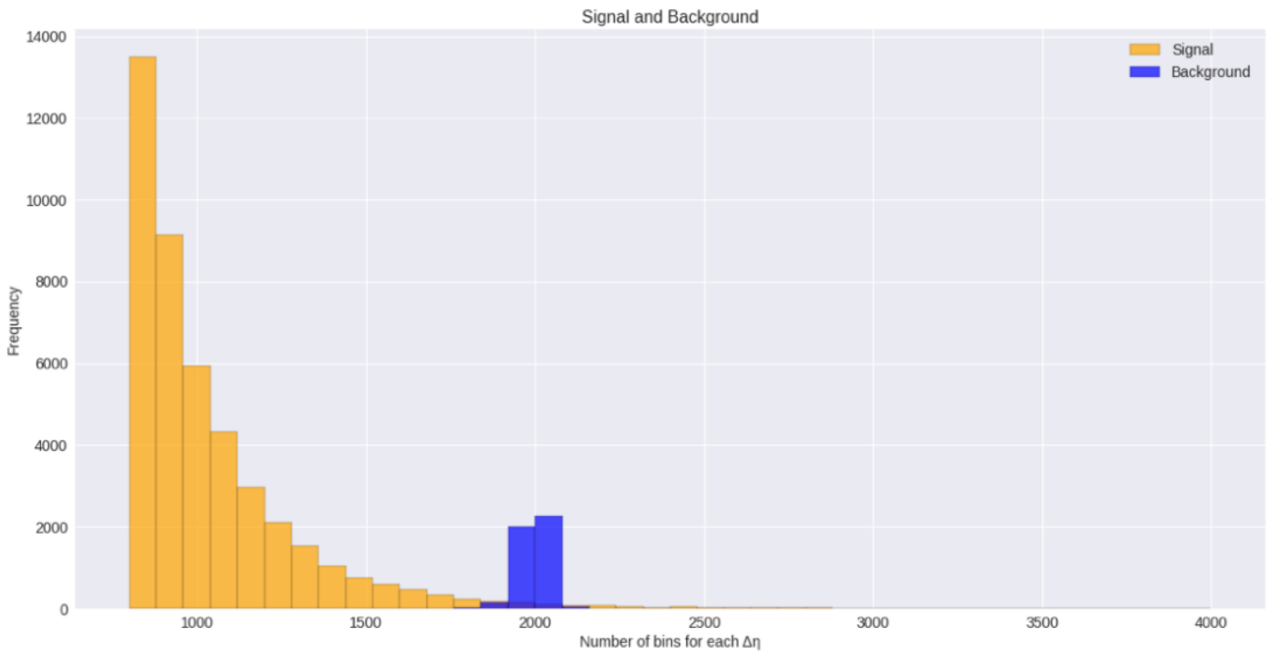


Figure 3: Histogram of the range of values of $\Delta\eta$ vs the frequency, after normalization

Results and Discussions

Subsequently, values from the classification dataset (Data 2) are grouped into 50 equal buckets based on their frequency, which yields the following plot of the dataset in order to get a clear visual representation of the distributions:

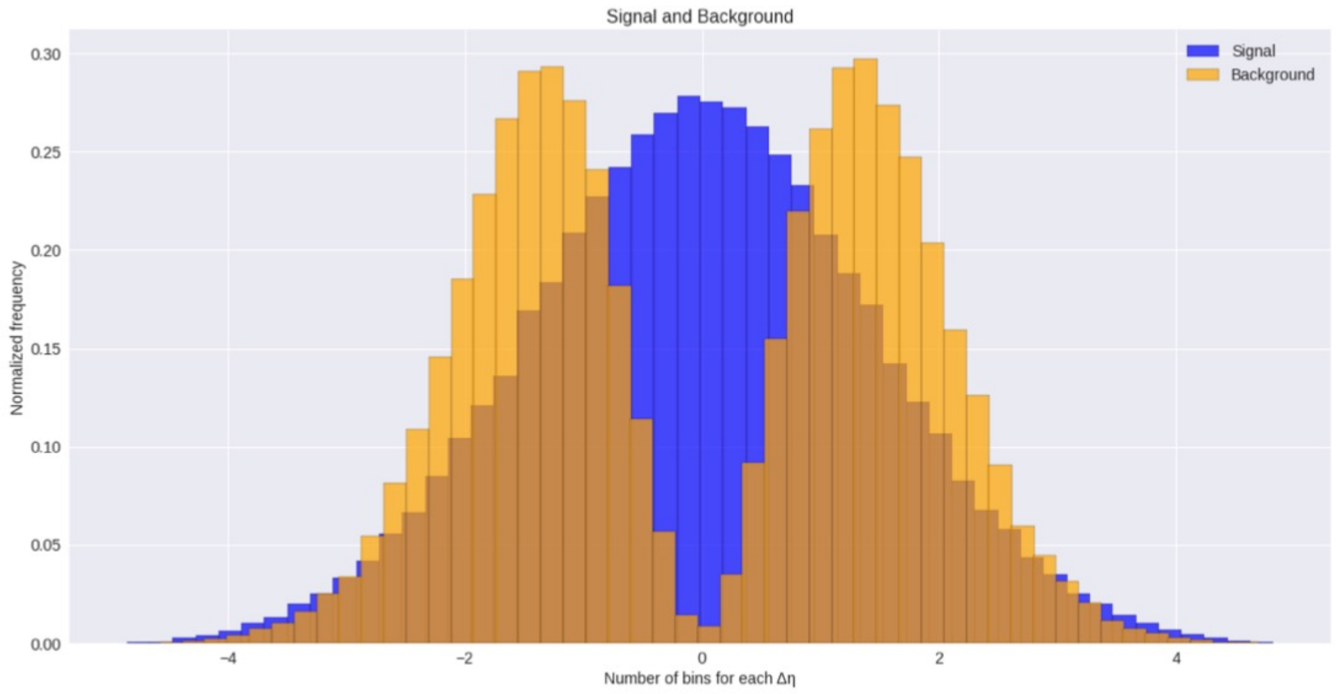


Figure 4: Histogram of the range of values of $\Delta\eta$ vs the frequency, after normalization

Results and Discussions

Best values for hyperparameters (such as the number of hidden units, learning rate and choice of optimizer) found using grid-search framework Optuna.

Best model ---> 112 hidden units, Adam optimizer and a learning rate of 0.036 performs the best on the training data, having an accuracy of 96.83% on the test data as well.

It is also observed that the overall accuracy increases as the number of units increase, along with the added advantage of the Adam Optimizer. On the other hand, SGD performs poorly.

Accuracy	Learning Rate	Number of Hidden Units	Optimizer
0.9697437577005260	0.03598887220487900	112	Adam
0.9685272696080720	0.03751239226618550	117	Adam
0.9676382975405110	0.1901150721313650	106	Adam
0.9666401534646520	0.13782620155439200	107	Adam
0.9666401534646520	0.037699042076936100	116	Adam
0.9664841934527990	0.03070600731157850	127	Adam
0.9657823733994600	0.06902679124302690	96	Adam
0.9655952213852370	0.3594409050775390	110	Adam
0.9638484692524840	0.1805661584264290	89	Adam
0.9629127091813660	0.04336196818313000	104	Adam
0.9628035371730690	0.014614354553899900	111	Adam
0.9625851931564750	0.01639707423541860	84	Adam
0.9611503610474280	0.3290987715142460	128	Adam
0.9611347650462420	0.27865346408328700	127	Adam
0.9607916530201660	0.0709261027865521	118	Adam
0.9590137088850420	0.06169820318734170	29	Adam
0.9564247726882830	0.05011494874138600	44	Adam
0.884963895257256	0.14003501850013000	86	Adam
0.8834666791434680	0.024268874389818100	119	RMSprop
0.8728613983374660	0.09534304874862510	5	Adam
0.8602754253809320	0.004446329002890750	15	RMSprop
0.8381602957001830	0.007755953330658100	54	RMSprop
0.8376144356586970	0.008956919381978440	45	RMSprop
0.8318907032236930	0.026818144099999000	69	RMSprop
0.658993434083501	0.12827425933783600	76	SGD
0.6570283379341540	0.001006797400324050	28	Adam
0.6483881532775000	0.1972605559928290	99	SGD
0.6378920444797950	0.2974787338618750	98	SGD
0.6377984684726840	0.0024160698288270000	97	SGD
0.563389946817636	0.004130082798527250	52	SGD

Results and Discussions

For this model, the BCE loss varies with 500 iterations as shown in Figure 5. This shows that the model decreases the loss without any significant oscillations around local minima:

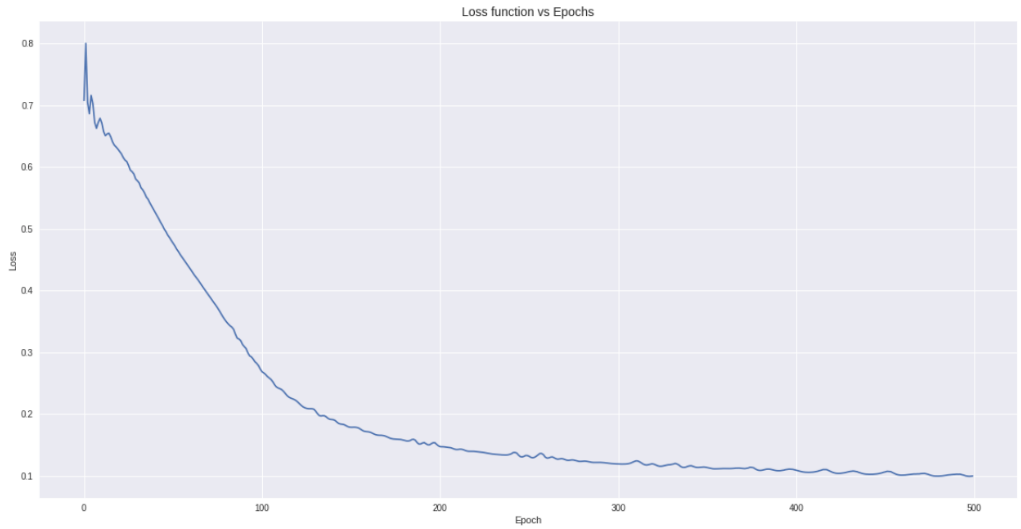


Figure 5: Loss function vs Epochs for the ANN

An ROC curve was plotted for this model's predictions, which yields the following graph. The area under the curve is found to be **0.9973**:

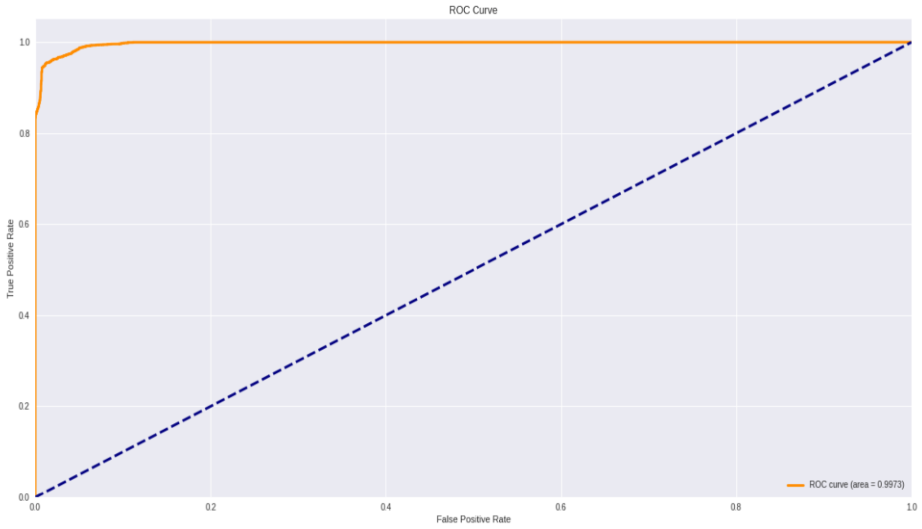


Figure 6: Plot of the *True Positive Rate (TPR)* vs the *False Positive Rate (FPR)*

Conclusion and Future Work

We are able to distinguish between distributions with high accuracy in a Supervised Learning task using machine learning models.

The current models presented are all Supervised Learning approaches to solve the classification problem. In the future, the use of Unsupervised Learning algorithms can be used for clustering of the points and predict the distribution to which they belong.

Further, more processes can be studied to derive complex features to obtain more meaningful and interpretable results. Given an unseen signal we want to distinguish it from background.

Thank You

QnA