APPLICATIONS OF THE XGBOOST CLASSIFIER MACHINE LEARNING ALGORITHM IN PARTICLE PHYSICS

The ATLAS Project

Introduction to the XGBoost ML Model

As advancements in Artificial Intelligence and Machine Learning technology are rapidly being made, the abundant applications of these technologies are becoming more and more clear. XGboost is just one example of a powerful classification algorithm that has come out of this technological revolution. This algorithm is a gradient boosted decision tree algorithm, where an iterative, ensemble learning approach has been used, i.e after one decision tree model is made, the algorithm then identifies potential biases and trains the next decision tree model to correct mistakes made by previous models (and this is done iteratively many times) [9].

A common problem in particle physics is that, when searching for events which contain a particular particle decay, the signal events end up forming a tiny proportion of the 100,000+ events picked up by detectors in particle accelerators. Therefore, it doesn't make sense for humans to sieve through the vast quantities of data to find an elusive particle decay. Instead it may be better to use the strong classification power of the XGBoost algorithm to complete this task.

Therefore, in this project, we aim to analyse the extent to which it is possible to use the XGBoost classifier algorithm to find and classify events, in which a particular particle decay occurs, from the numerous other decays picked up by sensors in particle accelerators.

The performances of the XGBoost classifier algorithm at classifying these events will be evaluated using the Approximate Median Significance (AMS) metric [5], the derivation of which will also be covered. This metric aims to provide a standardised way of measuring performances of the model across the various scenarios by acting similar to a hypothesis test [10].

Classifying Higgs Boson Events

Prior research advancements in particle physics tells us that it is possible to represent every particle as a wave in a quantum field (the Higgs field), which suggests that there would be a particle associated with this field (the Higgs boson). The Higgs boson, which was first discovered in 2012, can be "re-discovered" by looking for the products of its decay in particle detectors at particle accelerators [6].

However, the ATLAS particle detector at the Large Hadron Collider in CERN collects around 75,000 events per second [11], and so it may be possible to utilise the classification power of the XGBoost machine learning algorithm to classify and separate the events in which Higgs Bosons were produced from the (nearly) innumerous other background events, such as the decay of the Z boson into two Taus.

We wrote a program that trained the XGBoost algorithm to classify Higgs Boson events using over 400,000 samples of data [5], some of which were signal events and others were background events, and tested it with another set of 400,000 events [5]. Out of all of the test events that the XGBoost classifier had classified as signal events, every one of them was indeed a real signal event (events where the Higgs Boson was produced). This yields an AMS score of infinity, which suggests that the XGBoost algorithm is a suitable classifier for classifying Higgs Boson events.

The idea that the XGBoost is suitable for classifying Higgs Boson can be reinforced by observing the ROC curve, which is another metric that compares the true positive rate against the false positive rate.
In this case we have a perfect model and our program show that we have successfully classified all 200,000+ events given.

As evident in Fig. 1, the TPR initially increases significantly suggesting that the XGBoost is accurate. This is because the classifier has classified every event correctly, so the TPR is always 1.

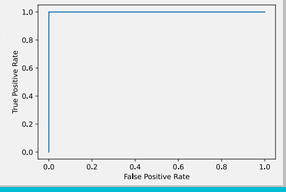


Fig.1

AMS METRIC DERIVATION

The evaluation metric used to evaluate machine learning classification models in particle physics is the approximate median significance, which is defined as:

$$AMS = \sqrt{2[(s+b)ln(1+\frac{s}{b}) - s]}$$

The formula can be derived as follows [12]:
Let the training sample be the set:

$$\{\{\mathbf{x}_i,\ y_i,\ w_i\},\ \{\mathbf{x}_i,\ y_i,\ w_i\},\ \{\mathbf{x}_i,\ y_i,\ w_i\},\ ...\}$$

Where

- $\mathbf{X}_i$ is an n length array of the features (lots data from different ATLAS sensors)
- $y_i \in$ {b, s} (class - represent whether that specific event is background or signal)
- $\mathcal{W}_i$ is the weight (more about this below)

Let S be a set of indexes of the signal events and let B be a set of indexes of the background events.

Given that the dataset can be biased (the proportions of signal and background events in the dataset is not equal to the proportions in real life). Weights are therefore used to balance this imbalance:

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

So the sum of the weights for the each class is the expected number of events that there would be in real life for that class.

Let the classifier be f(x) → {b, s} and let G be the set of indexes of the events that the classifier classifies as signals.

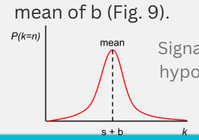
$$s = \sum_{i\in(S\cap G)} w_i \qquad b = \sum_{i\in(B\cap G)} w_i \qquad s+b = \sum_{i\in G} w_i$$

Therefore, s and b are unbiased estimators of the expected numbers of signal and background events respectively that the classifier would classify as signal or background in real life (as s and b are the sum of the weights of the events classified as being signal and background respectively). So, s+b would be an estimator of the expected number of classified signal events.

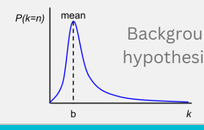
But, the actual number of classified signal events might not be equal to exactly s + b, as it's an estimate. So, let the number of actual classified signal events be equal to n.

For the hypothesis where signal and background events are abundant in the proportion that we expect (the signal background hypothesis), the distribution of the probability of values being equal to n follows a Poisson distribution with a mean of s + b (Fig. 8).

However, for the hypothesis where there are no signal events (background-only hypothesis), the distribution of the probabilities of values being equal to n follows a Poisson distribution with a mean of b (Fig. 9).



Signal background hypothesis (Fig. 8)



Background-only hypothesis (Fig. 9)

Therefore, if the classifier is effective, the probability of s + b being equal to n should be significantly greater in the signal background hypothesis (compared to the background-only hypothesis) (Fig. 10).

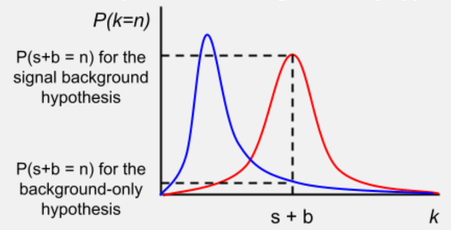


Fig.10

In order to calculate these probabilities, the Poisson distribution formula is used [8], and so for the signal background hypothesis:

$$P((s+b) = n) = \frac{(s+b)^{(s+b)}e^{-(s+b)}}{(s+b)!}$$

and for the background only hypothesis:

$$P((s+b) = n) = \frac{b^{(s+b)}e^{-b}}{(s+b)!}$$

To calculate the significance, the (modified) Wilk's theorem is used [7]: $Z = \sqrt{2[\ln(likelihood\ in\ alternative\ model) - \ln(likelihood\ for\ null\ model)]}$

$$Z = \sqrt{2[\ln(\frac{(s+b)^{(s+b)}e^{-(s+b)}}{(s+b)!}) - \ln(\frac{b^{(s+b)}e^{-b}}{(s+b)!})]}$$

$$Z = \sqrt{2[\ln((\frac{(s+b)^{(s+b)}e^{-(s+b)}}{(s+b)!})(\frac{(s+b)!}{b^{(s+b)}e^{-b}}))]}$$

$$Z = \sqrt{2[\ln(\frac{(s+b)^{(s+b)}}{b^{(s+b)}}e^{-s})]}$$

$$Z = \sqrt{2[\ln((\frac{s+b}{b})^{(s+b)}) + \ln(e^{-s})]}$$

$$Z = \sqrt{2[(s+b)\ln(1+\frac{s}{b}) - s]}$$

AMS values can take any value from 0 to infinity. The larger the AMS value the better the model is at identifying the particle

Beyond the Standard Model

The search for new particles is constantly ongoing, one example is the top quark. This quark has a coupling, that is to say connected by some force, to the standard model Higgs Boson in addition to being predicted to have a coupling with a new particle beyond the standard model (BSM) called Z' which is also a gauge boson.
To find this particle we must, as always, inspect what this particle decays into. According to models [1] this BSM Z' particle should decay into a top-antitop pair and so by bump hunting we should be able to find this particle. The cuts we made of our program were: A single electron/muon with transverse momentum above 30 GeV with ≥ 1 small-R jet close to it (i.e. ΔR(small-R jet, lepton) < 2.0), as well as one or more b-tagged small-R jets, (either the one close to the lepton or one in large-R jet). The missing transverse energy(EmissT) is to be > 20GeV and EmissT+missing transverse momentum to be > 60GeV. There is exactly 1 large-R jet, it is top-tagged with mass > 100GeV and N-subjettiness ratio less than 0.75. The particles that satisfies these requirements were plotted on a graph of transverse mass of the tt system (by summing large-R jet, small-R jet and lepton) against frequency and as you can see a particle does appear to exist (whether that particles is the Z' is a different discussion).
An XGBoost classifier was trained with >20000 data samples containing the relevant features (e.g. electron/muon number, transverse mass, R-jet data ... etc) to identify the events, which contain the decay of the Z' particle. The model was then used to classify these signal events, which were then plotted on a histogram and compared with the original histogram to see the performance of the classification as seen in Fig.2 and Fig. 3. The AMS value of this classification is infinity.

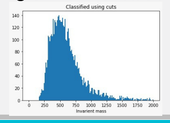


Fig. 2

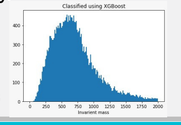


Fig.3

Kaluza Klein Gravitons

One type of particle we shall look at is a Kaluza Klein graviton hypothesised using the Randall-Sundrum model [2] (a model for gravity where gravity propagates through warped extra dimensions). In a similar way to atoms having exited states or low energy states, particles can have Kaluza Klein states where the particle has extra mass (instead of energy) in other dimensions [3].

In order to find the Kaluza Klein graviton we can search for the particles it decays into, in this case the particle decays into a gamma-gamma pair. So to find this particle we need to bump hunt gamma ray photons with transverse energies over 20 GeV [2]. To do this we made the following cuts to the ATLAS data set: the event must have two photons, it must activate the photon trigger, both photons must have a transverse energy greater than 20GeV. Once we obtained our data points we subtracted any data points that could have been formed by the Higgs -> GammaGamma decay channel and plotted a graph of transverse energy against frequency using a fitting function.

XGBoost classifiers was trained with data samples containing the relevant features (e.g. electron/muon number, transverse mass, R-jet data ... etc) to identify the events, which had good photons and the events which had photon isolation (this problem required two classifiers) . The models were then used in parallel to identify the events, which contained the decay of the Kaluza Klein graviton (both good photons and photon isolation). The events were then plotted on a histogram (Fig. 4) and compared with the original histogram (Fig. 3) to see the performance of the classification. The AMS values were 866.8 for the good photon classification and 866.7 for the photon isolation classification.

Fig.3

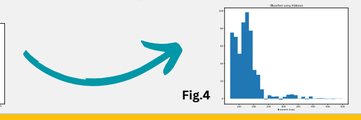
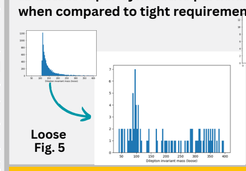


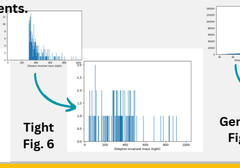
Fig.4

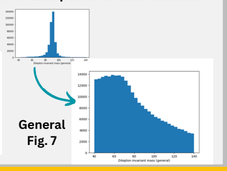
SuperSymmetry

Supersymmetry (SUSY), the idea that every fermion (particles with odd half integers spins such as protons and electrons) has a partner boson (particles that have an integer number of spins and are used to transfer a force), is a widely researched area of particle physics with many believing that some of the lightest superpartner bosons, or sparticles as they are sometimes known, may be candidates for WIMPs (dark matter explanation) [4]. In this project we will attempt to find the productions of pairs of sleptons where each slepton decays into the lightest neutrino and the corresponding lepton.

Due to the difficult nature of detecting super-symmetric particles, we looked for them with different cuts based on likelihood of false positives and negatives. The cuts needed for this are [1]: the event must have two electrons or muons of same flavour and opposite charge, transverse momentum ρT more than 20 and 25 GeV respectively, dilepton invariant mass mll larger than 40 GeV, zero b-tagged jets at 77% certainty (using MV2c10) and zero non-b-tagged jets with leptons with ρT larger than 60 GeV and (loose: stransverse mass mT2 > 100 GeV, mll > 111 GeV) or (tight: mT2 > 130 GeV, mll > 300 GeV) (The general graph, fig. 7, contains particles that have not been cut based upon their stransverse mass.). The graphs plotted by the XGBoost algorithm are below. The algorithm was trained and tested with a data ratio 1:1, producing graphs and AMS values ~1.689 (loose) and ~1.192 (tight). This may have been a result of using a smaller dataset as this would've resulted in a weaker model. the comparison of the graphs shows us that our loose ML produced graph (Fig.5) is most similar to its original result (the smaller graph), sharing a shape with the tight graph (Fig. 6) and dataset. Hence, loose requirements are the best for building an accurate model to detect super-symmetric particles although they may lead to more false positives than desirable when compared to tight requirements.


Loose
Fig. 5


Tight
Fig. 6


General
Fig. 7

Conclusion and references

From BSM particles to delving into supersymmetry we have explored a wide range of ground breaking topics in the area of particle physics and managed to apply the modern tool that is machine learning to these classification-type problems. Over the course of this poster we can see that the models generated can be a bit overly rigorous at times and result in a large proportion of signal data being cut out and dismissed as background data resulting in thin meagre graphs as seen in our exploration of SUSY and during other times can produce a near perfect result. Machine learning models are essentially a way of using pattern recognition to solve a problem, so for problem where the selection criteria are well defined ML models would not be a useful tool for researchers, as the model will only ever at best replicate the predetermined cuts. Rather, this tool may find its use in being used to identify the selection criteria for particles who's properties have yet to be defined and have very few selection criteria identified so far. In these cases models such as XGBoost may be used by researches to give them an approximate idea of the type of particles they are searching for and the properties they have.

References: To produce the XGBoost algorithm models we used a python library: https://xgboost.ai/about, [1]: https://cds.cern.ch/record/2707171/files/ANA-OTRC-2019-01-PUB-updated.pdf?version=2, the cuts required to identify kaluza klein gravitons [2]: https://indico.cern.ch/event/129980/contributions/1350939/attachments/90398/129356/evan_wulf_RS_diphoton_dpf.pdf, [3]: https://home.cern/science/physics/extra-dimensions-gravitons-and-tiny-black-holes , [4] : https://home.cern/science/physics/supersymmetry, the derivation of the ams metric [5]: http://opendata.cern.ch/record/328#, [6]: https://home.cern/science/physics/higgs-boson/what, [7]: https://handwiki.org/wiki/Wilks%27_theorem#:~:text=In%20statistics%20Wilks%27%20theorem%20offers%20an%20asymptotic%20distribution,a%20test%20statistic%20for%20performing%20the%20likelihood-ratio%20test, [8]: https://www.statology.org/poisson-distribution/, [9]: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/, [10]: https://en.wikipedia.org/wiki/Null_hypothesis, [11]: https://cds.cern.ch/record/1457044/files/ATLAS%20Fact%20sheet.pdf, [12] https://higgsml.lal.in2p3.fr/files/2014/04/documentation_v1.8.pdf

Sasan, Shizhe, Joel, Pruthvi, William

L054

KING EDWARD VI CAMP HILL SCHOOL FOR BOYS
A caring and inclusive community where everyone can do and be their best