A Comparison of Machine Learning Algorithms for Identifying Higgs Boson Events From the Background

Ourania-Maria Glezakou-Elbert, Washington Representative

Introduction

In 2012, the existence of an elusive fundamental particle was confirmed [1, 2]. The Higgs boson, theorized to exist by Englert, Brout, and Higgs [3, 4], was one of the biggest missing pieces for the Standard Model, the model which describes all fundamental particles in the universe. The Higgs boson accounted for the mass of the W and Z bosons as the Higgs boson provides all elementary particles with mass via the Higgs mechanism [3, 4].

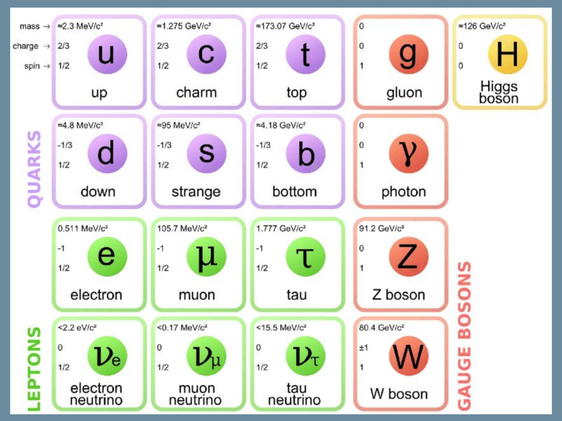
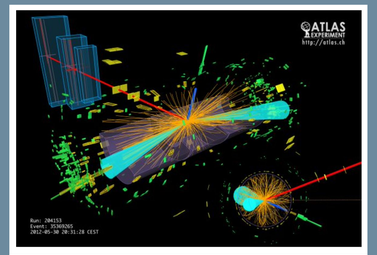


Fig.1 Table displaying the particles described by the Standard Model [5].

The Problem



In order to confirm and test the Standard Model, particle physicists employ particle accelerators, such as CERN and the ATLAS detector, where they create and detect the events they want to study.

Fig. 2 Visualization of an event in the ATLAS detector [6].

The issue is that thousands of events happen every second during these collisions [7]. CERN offloads 3200 TB of data every year from their ATLAS detector [7], not including the thousands of terabytes that occur during the collision itself. This makes the data that they receive noisy and difficult to sort through.
Being unable to accurately identify Higgs boson events in order to study their properties, such as their mass and decay channels, makes it difficult to confirm the Standard Model. Thus, machine learning (ML) is an excellent solution to the issue of finding these rare Higgs events in the noisy background [8].
Also, in order to reduce the computational cost of identifying these events, these algorithms need to be fast and computationally efficient.

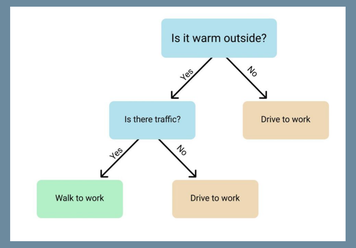
Machine Learning and Particle Physics

As long as experiments occur in particle accelerators, there will be a need for the development of off-line sorting algorithms which take in data from the detector and further process and select signal (desired) events. In 2014, ATLAS held its own Higgs Boson Machine Learning Challenge [7] and invited anyone who wanted to develop a sorting algorithm for this purpose. What remains to be formally determined is what sorting algorithm is best suited to the task.
Thus, the goal of this study is to determine which of the following algorithms, a decision tree, a support vector machine, or a neural network, is the best at identifying Higgs bosons from the background.

Goal:

To identify which ML algorithm, a decision tree, a support vector machine, or a neural network, is best suited for identifying Higgs bosons.

Decision Tree

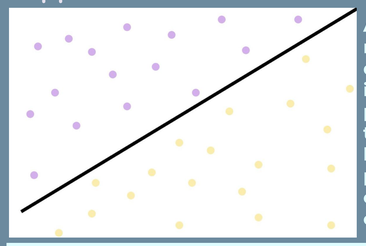


Much like the name implies, a decision tree (DT) is an algorithm which sorts given data by splitting the data into groups at nodes based on given conditions[9]. Pros of the algorithm include being able to follow and track the decisions that the DT makes.

Fig. 3: Representation of a DT determining whether someone should walk or drive to work.

However, with large amounts of data, trees can become over-complicated and fit too closely to their training data, an issue known as overfitting. Boosted DTs (BDTs) combine multiple trees and take subsets of data in order to combat overfitting. Histogram gradient BDTs bin data and then sort it through multiple trees which weight and fix complicated cuts [9].

Support Vector Machine

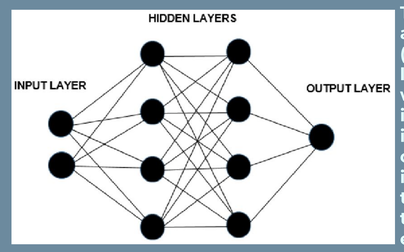


A support vector machine (SVM) plots data with *n* dimensions in an *n*-dimensional plane and then finds the hyperplane which best separates the points into distinct categories, as is seen on the right [9].

Fig. 4: Representation of a SVM model using data with two features and two possible classes: purple and yellow. The line represents the best hyperplane which classifies the elements.

The algorithm combats overfitting by maximizing the distance from the hyperplane to the nearest data point. A drawbacks is the long training time due to the algorithm having to calculate the distance from the plane to each point there is [9]. The algorithm is particularly suited for data with many features, such as the 30-feature Higgs boson dataset which will be utilized here.

Neural Network



The final algorithm is a neural network (NN). NNs combine layers of neurons which take in an input and transform it before passing it onto the next neuron if it is above a certain threshold value. By the output layer, the event is sorted [9].

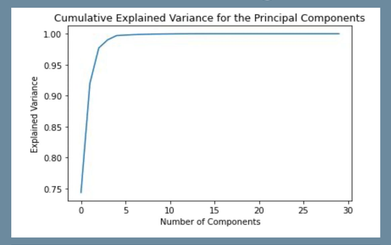
Fig. 5: Diagram representing a neural network with two hidden layers [10].

While NNs are capable of obtaining a high accuracy at times, they require large amounts of memory to function.
Additionally, they are blackboxes, meaning that it is impossible to see the decisions that the algorithm is making and it is thus impossible to fix them or see what the error is. It can also take a while for these algorithms to run as well due to their complexity [9].

Set-Up

This project makes use of the Google Co-Lab Python Notebook and the Sci-Kit Learn Machine Learning Library [11]. Data used to train the algorithm makes use of the 250,000 simulated events publicly available from the 2014 Higgs ML Challenge [7]. Each event is described by 30 features and the frequency of signal events is increased in order to more readily train the algorithm. This may have an effect on the accuracy of algorithm when implemented in a legitimate detector and must be considered as a limitation of this study.

Data Pre-Processing and Training



The dataset was split 80-10-10 to train, validate, and then test the algorithms. Validation was done in order to confirm that the hyperparameters for each algorithm were optimized in order to produce the best performance for the algorithm.

Fig. 6: Graph showing the amount of information retained after reducing the number of dimensions through PCA.

Due to issues initially attempting to run the SVM and the NN, it was decided that principal component analysis (PCA) would be done on the data in order to help the algorithm more efficiently process the data. PCA is a form of dimensionality reduction which linearly recombines the given data while still preserving the relationships between data points.

For this dataset, the graph of information retained in terms of the explained variance plateaus around 8 dimensions. This allows us to reduce the number of features of the data from 30 down to 8 while still explaining 99.9% of variance. This aided in allowing for the SVM and NN to run without crashing and helped reduce training times for them as well.

Finally, the data was also scaled for the SVM and NN in order to further streamline the process for these algorithms. For the SVM, for instance, scaling the data reduced the time needed for the algorithm to calculate the distance from each point to the dividing hyperplane by placing all values within a smaller range of one another.

Criteria

In order to compare the algorithms, the following will be compared:

- Memory usage
- Accuracy
- F1-Score
- Receiver Operating Characteristic (ROC) Curve
- Training time

As mentioned previously, when implementing this algorithm on a larger scale, it is important that it is able to run quickly and using little memory in order to lower the computational cost and make it practical.

For the evaluation of the performance of the algorithms, three metrics were used. Accuracy determines how well the algorithm sorted all events. The F1-Score is the harmonic mean of precision (how many of the positives were true positives) and recall (how many positives were identified correctly). ROC curves plot the true and false positive rates of an algorithm, with a larger area under the curve corresponding with a better performance.

Results

	Histrogram GBDT	SVM	Neural Network
Area Under the Curve	91%	84%	88%
Average Test Accuracy	83.88%	80.35%	82.48%
Average Test F1-Score	81.72%	77.12%	80.01%
Training Time (h:min:s)	00:00:8.4	00:50:32	1:34:31
Memory Usage (MB)	645.41	1167.60	1028.96

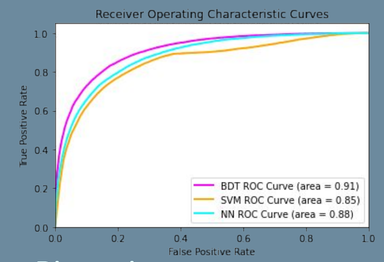


Table 1: Display of outcomes for each algorithm. Yellow shows best performance, pink/red shows worst.

Fig. :ROC curve comparing the BDT, SVM, and NN. The BDT has the largest area under the curve, showing the best performance.

Discussion

As can be seen in the table and graph above, the BDT had the best performance of all three algorithms. This is comparable to the highest accuracy achieved by a BDT in the Higgs ML challenge [12] which had an accuracy of 84%. Another study by Chen and He [13] had a similar accuracy, although the BDT from this study had a lower training time.

The SVM used here had a particularly improved performance to those in [7], but it was still not enough to establish the SVM as the best algorithm for this task due to its lower accuracy and longer training time..

It is important to note that all three algorithms had decently comparable accuracies and F1-Scores, the largest differences were with the training times.

Conclusion and Future Work

In all, the Histogram Gradient BDT was found to be the best algorithm for Higgs boson identification of the three algorithms analyzed.

In the future, it is imperative that we examine how pre-processing of the data might further improve performance as well as how the algorithm will function in more realistic and non-simulated settings. Additionally, the development of similar algorithms for researching other phenomena, such as dark matter and neutrino decays, is another area of potential future research.

Acknowledgements

Thank you so much to my mentor, Dr. Savannah Thais (Princeton) for guiding me throughout the research process.

Contact

If you wish to contact me for further discussion, please email ouraniamge@gmail.com.

Citations

[1] G. Aad, T. Abajyan, B. Abbott, J. Abdallah, S. Abdel Khalek, A. Abdelalim, O. Abdinov, R. Aben, B. Abi, M. Abolins, and et al., "Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc," Physics Letters B, vol. 716, p. 1–29, Sep 2012.
[2] S. Chatrchyan, V. Khachatryan, A. Sirunyan, A. Tumasyan, W. Adam, E. Aguilo, T. Bergauer, M. Dragicevic, J. Er ö, C. Fabjan, and et al., "Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc," Physics Letters B, vol. 716, p. 30–61, Sep 2012
[3] F. Englert and R. Brout, "Broken symmetry and the mass of gauge vector mesons," Phys. Rev. Lett., vol. 13, pp. 321–323, Aug 1964.
[4] P. W. Higgs, "Broken symmetries and the masses of gauge bosons," Phys. Rev. Lett., vol. 13, pp. 508–509, Oct 1964.
[5] A. Arbuzov, "Quantum Field Theory and the Electroweak Standard Model", 01 2018.
[6] R. M. Bianchi, "Higgs candidate decaying to 2 tau leptons in the ATLAS detector," Nov 2013. General Photo.
[7] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. K égl, and D. Rousseau, "The Higgs boson machine learning challenge," in Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning (G. Cowan, C. Germain, I. Guyon, B. K égl, and D. Rousseau, eds.), vol. 42 of Proceedings of Machine Learning Research, (Montreal, Canada), pp. 19–55, PMLR, 13 Dec 2015.
[8] D. Guest, K. Cranmer, and D. Whiteson, "Deep learning and its application to lhc physics," Annual Review of Nuclear and Particle Science, vol. 68, no. 1, pp. 161–181, 2018
[9] P. C. Sen, M. Hajra, en M. Ghosh, "Supervised Classification Algorithms in Machine Learning: A Survey and Review", in Emerging Technology in Modelling and Graphics, 2020, bll 99–111.
[10] R. Bartzatt, "Determination of Dermal Permeability Coefficient (Kp) by Utilizing Multiple Descriptors in Artificial Neural Network Analysis and Multiple Regression Analysis", Journal of Scientific Research and Reports, vol 3, bll 2884–2899, 01 2014.
[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[12] S. R. Ahmad, "Technical report of participation in higgs boson machine learning challenge," 2015.
[13] T. Chen and T. He, "Higgs boson discovery with boosted trees," in Proceedings of the 2014 International Conference on High-Energy Physics and Machine Learning - Volume 42, HEPML'14, p. 69–80, JMLR.org, 2014.