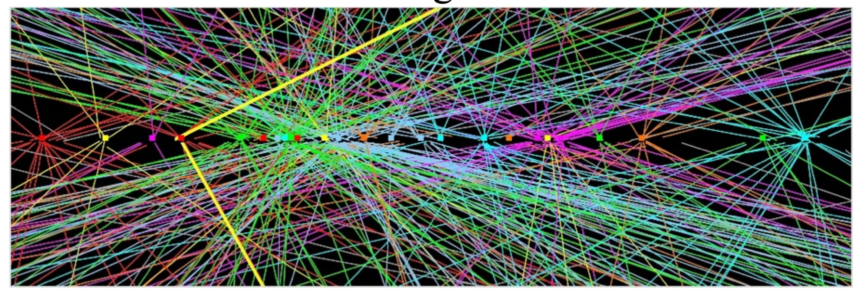
Analyzing the Putative Higgs Boson using Support Vector Machines

Mukesh Ghimire

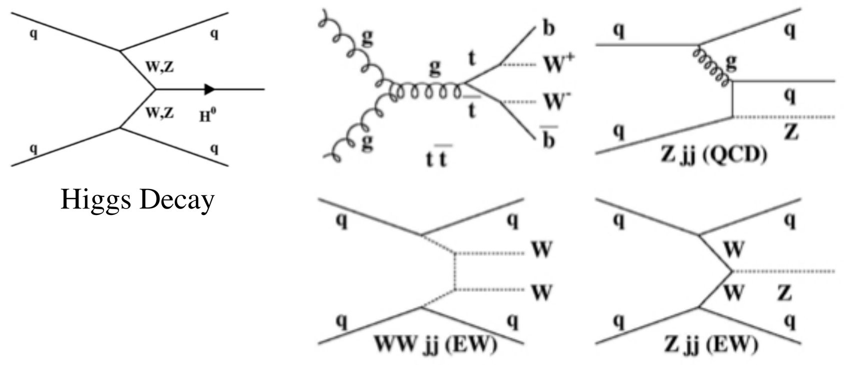
Department of Physics, Tufts University

Introduction

Event classifiers are very important tools in experimental physics because they allow us to filter out the few points of useful data from the chaotic sea of collision signals.



The Higgs Boson, in particular, leaves traces that are very well camouflaged by very similar-looking background signals, as seen in the Feynman diagrams below:



Therefore, it is important that we have an accurate event classifier in order to find Higgs events.

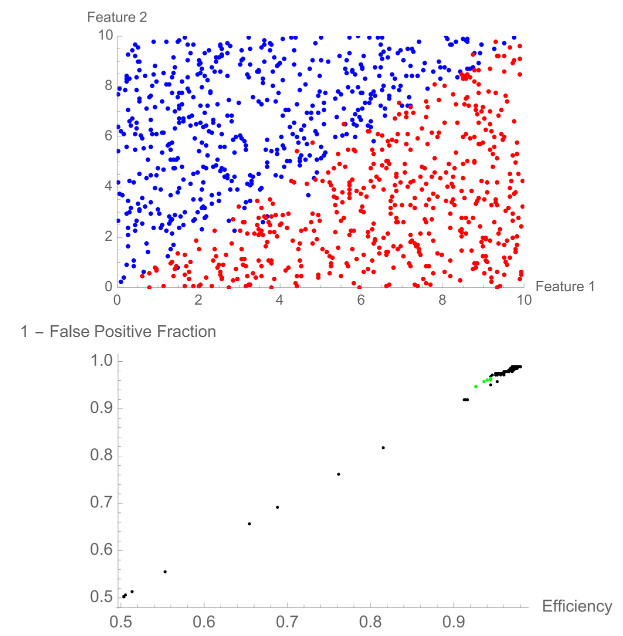
The ATLAS team at CERN used a classifier called Boosted Decision Trees (BDTs) in 2012 when they found the Higgs. We suspected that a classifier called Support Vector Machines (SVMs) may perform better than BDTs. SVMs are more modular than BDTs and have a mathematically rigorous foundation. They have various adjustable parameters and different function modes, which allows us to adjust them to suit the type of data being classified. In this project, we tested their performance across a wide range of parameters against the performance of BDTs.

Objectives

1. Test the accuracy of SVMs versus the accuracy of BDTs on various data types.
2. Test the time-efficiency of both classifiers
3. Find out which classifier is better for classifying the upcoming 13 TeV Higgs data

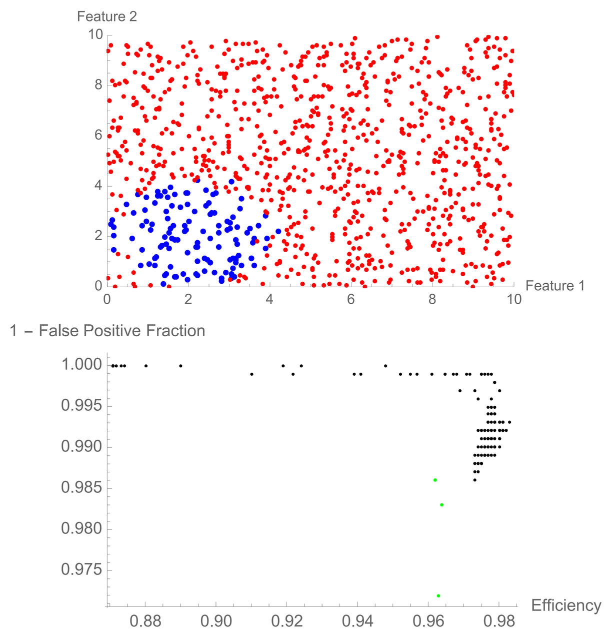
Results

Line



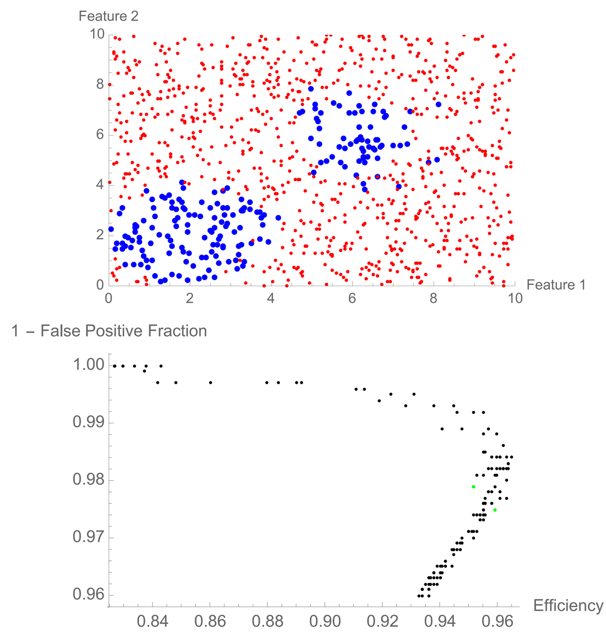
Runtimes: SVM: 32.75 s BDT: 130.37 s

Circle



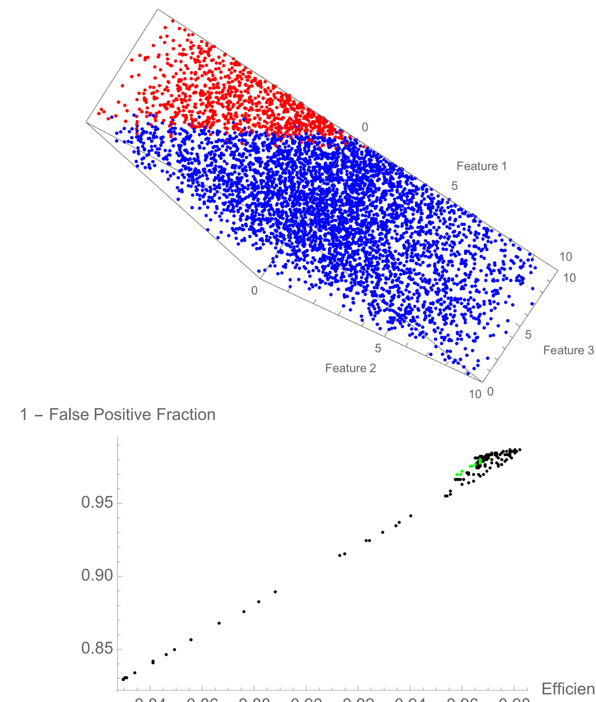
Runtimes: SVM: 25.69 s BDT: 117.63 s

Two Circles



Runtimes: SVM: 24.49 s BDT: 96.73 s

Plane

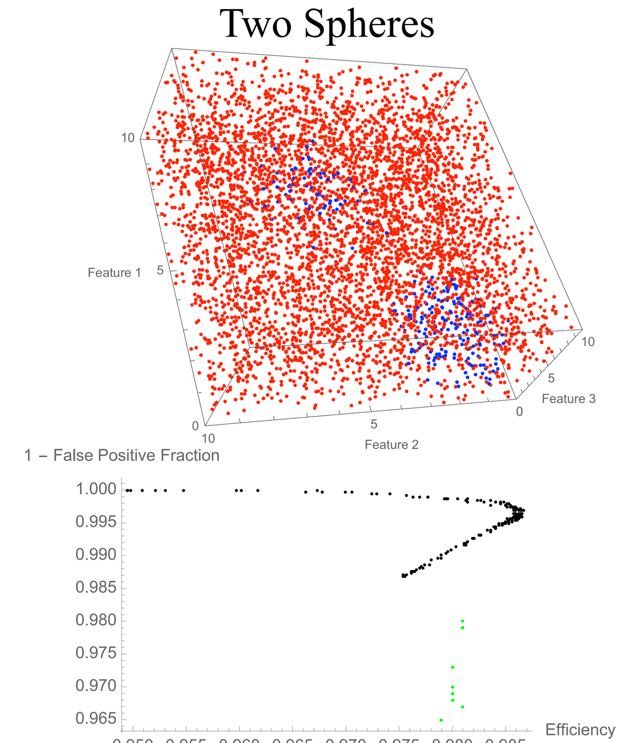


Runtimes: SVM: 520.74 s BDT: 168.44 s

Key:

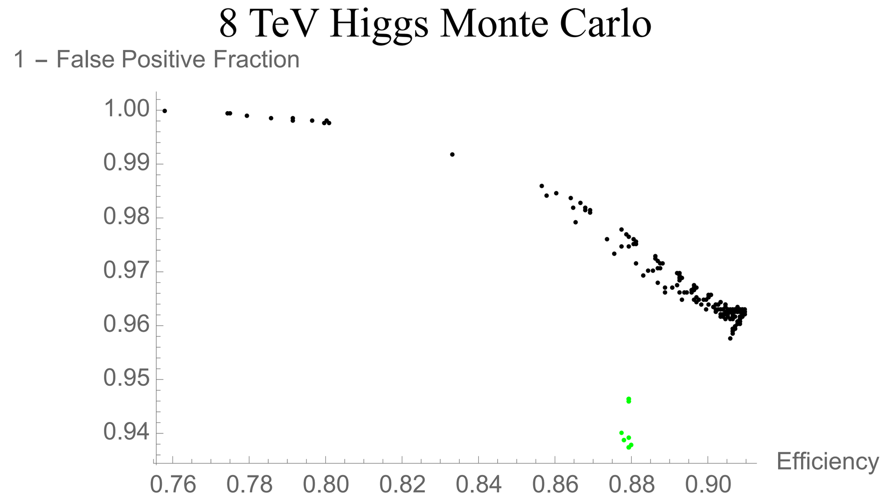
- Signal Events
- Background Events
- SVM data
- BDT data

Two Spheres



Runtimes: SVM: 349.53 s BDT: 163.88 s

8 TeV Higgs Monte Carlo



Runtimes: SVM: 3429.53 s BDT: 492.83 s

Conclusions

If we look solely at the classifying efficiencies of SVMs and BDTs, we may conclude that SVMs perform marginally better than BDTs. However, when we look at the 1 – False Positive Fractions of SVMs versus those of BDTs, we can see that SVMs are performing better in this criteria as well. This is crucial, as in physics we would rather our misclassifications be false negatives than false positives. Therefore, SVMs are statistically more powerful than BDTs.

Considering the runtimes of the two classifiers, we see that SVMs perform much faster for simpler problems, but much longer for harder problems. In particular, it is a little worrying that SVMs take so much longer than BDTs when classifying 8 features for the Higgs. It seems likely that SVMs would run much slower for more features, and data in experimental physics generally has a lot more features than 8. In fact, the 8 TeV Higgs data initially came with 31 features; luckily we could eliminate most of them because they had little bearing on the classifiers. Such may not always be the case. However, this problem can be mitigated with better computational power, which we will most likely have in the future.

In conclusion, SVMs seem to offer better data classification than BDTs across the board, but at the cost of computational power for large data.

Looking Forward

We initially wanted to test SVMs and BDTs on 13 TeV Higgs Monte Carlo data, which is relevant to the LHC runs that happened over the summer. Unfortunately we did not have access to the 13 TeV data, so we decided to test it on 8 TeV, which was relevant to the runs in 2012 when we discovered the Higgs. BDTs were used to discover the Higgs in 2012, and we have shown that SVMs would have been better.

The next order of business will be to test SVMs and BDTs on 13 TeV Monte Carlo data and to classify real LHC data.

The ultimate goal is to have an ATLAS Collaboration note published on our results.

Materials and methods

We used the LIBSVM package for our SVMs and the WEKA decision trees package for our BDTs.

We created data sets that corresponded to a linear function, a sphere, and two spheres in 2D and 3D.

We used python scripts to loop the classifiers over various parameters on the data sets we generated.

We used Mathematica to graph the results.

References

Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1—27:27, 2011.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Acknowledgments

Special thanks to my advisor Prof. Krzysztof Sliwa, my teammates Ben Nissan, Julia Rowe, and Jared Moskowitz, my donors, and Mr. Nathan Gantcher, whom my award honors.

I would also like to thank the other Summer Scholars for keeping me motivated throughout the summer.

ATLAS

Tufts UNIVERSITY