Dimensionality Reduction for Classification using Higgs Dataset

Romero, Carlos[1], Reinhardt, Eric[2], Slivar, Ana Maria[2], Gleyzer, Sergei[2]
1. Department of Physics, Elmhurst University, Elmhurst, Illinois
2. Department of Physics, University of Alabama, Tuscaloosa, Alabama

Abstract

Machine learning algorithms can be used to reduce the dimensionality of a dataset, inferring correlations between variables, and combining them into unified (reduced) variables, thereby decreasing the number of features necessary to represent information. Throughout this research, we sought to reduce the dimensionality of a Higgs classification dataset containing signal and background events with the same signal of $WWb\bar{b}$ (semileptonic) and to evaluate our information loss using a neural network classifier.

Keras and Tensorflow were used to build a deep neural network capable of classifying between signal and background events given 7 high-level and 21 low-level features. The neural network was then used to establish benchmark ROC-AUC scores of 0.857 and 0.828 for the all-features dataset and low-level only dataset. PCA and an Autoencoder were applied to the low-level features in order to reduce their number to 20, 17, 14, and 11. Feeding them into the model used to benchmark the datasets, we achieve a best ROC-AUC of 0.828 with PCA, 0.810 with PCA, 0.780 with Autoencoder, and 0.737 with PCA, respectively.

On a different approach, some selected features, such as the missing energy magnitude and azimuthal angle, and the b-tags from different combinations of jets, were removed from the dataset in order to assess the importance of said features in the classification process. Removing both missing energy magnitude and azimuthal angle results in a ROC-AUC of 0.815, while removing the b-tags from both jets 1 and 2 results in a score of 0.798, and removing b-tags from both jets 3 and 4, results in a score of 0.814.

We establish new benchmark ROC-AUC scores for feature reduction using PCA and Autoencoders as well as selective feature removal. These benchmarks can be used to determine optimal dataset reduction methods depending on the required statistical significance for a particular data search or analysis.

Acknowledgment: This work was supported by the DOE RENEW–HEP: U.S CMS SPRINT award No. DE-SC0023680.