HiggsML Challenge - Quantifying feature A

Vidhi Lalchand

Cavendish Laboratory, Department of Physics, J J Thomson Avenue, Cambridge. CB3 0HE

Abstract

In this short report we try to quantify the importance of feature A for Higgs classification in the Kaggle dataset. The approach used here is that of Gaussian Kernel Support Vector machines to auto-classify Higgs collisions from other background processes. Feature **A**, an analytically constructed likelihood value was proposed by Dr. Christopher Lester at the Cavendish Laboratory, High Energy Physics group. We use a set of pre-optimized parameters proposed in [1] and test the AMS performance of classifying the collisions under a model trained on a set of features which includes feature A and one that excludes feature A.

1. Feature-class correlation

Since the class variable is dichotomous, $y_i \in \{0, 1\} \forall i$, the feature vs. class correlation is computed using the point-biserial correlation. This is mathematically equivalent to the Pearson correlation for computing correlations between one continuous and one dichotomous variable. The point-biserial correlation was computed between each feature and the binary class variable. It is computed as,

$$r_{pb} = \frac{M_1 - M_0}{s_n} \sqrt{\frac{n_1 n_0}{n^2}} \tag{1}$$

where M_0 and M_1 are the mean values of the feature for the background and signal class. s_n is the inter-class standard deviation of the feature, n_0 and n_1 represent the count of samples in the background and signal class and n is the overall count. From figure 1 it is apparent that the derived features showed higher correlation with the class label than the primary features. Visually observing the distribution of features with high point biserial correlation to each class label gives some insight 2 3.

2. AMS results

There were 250,000 rows in the complete data set. We dropped rows which contain missing values (-999.0) in any of the features. The data set without any missing value tags had 68114 rows. We split this condensed data set into a training and test set in the ratio of 80:20.

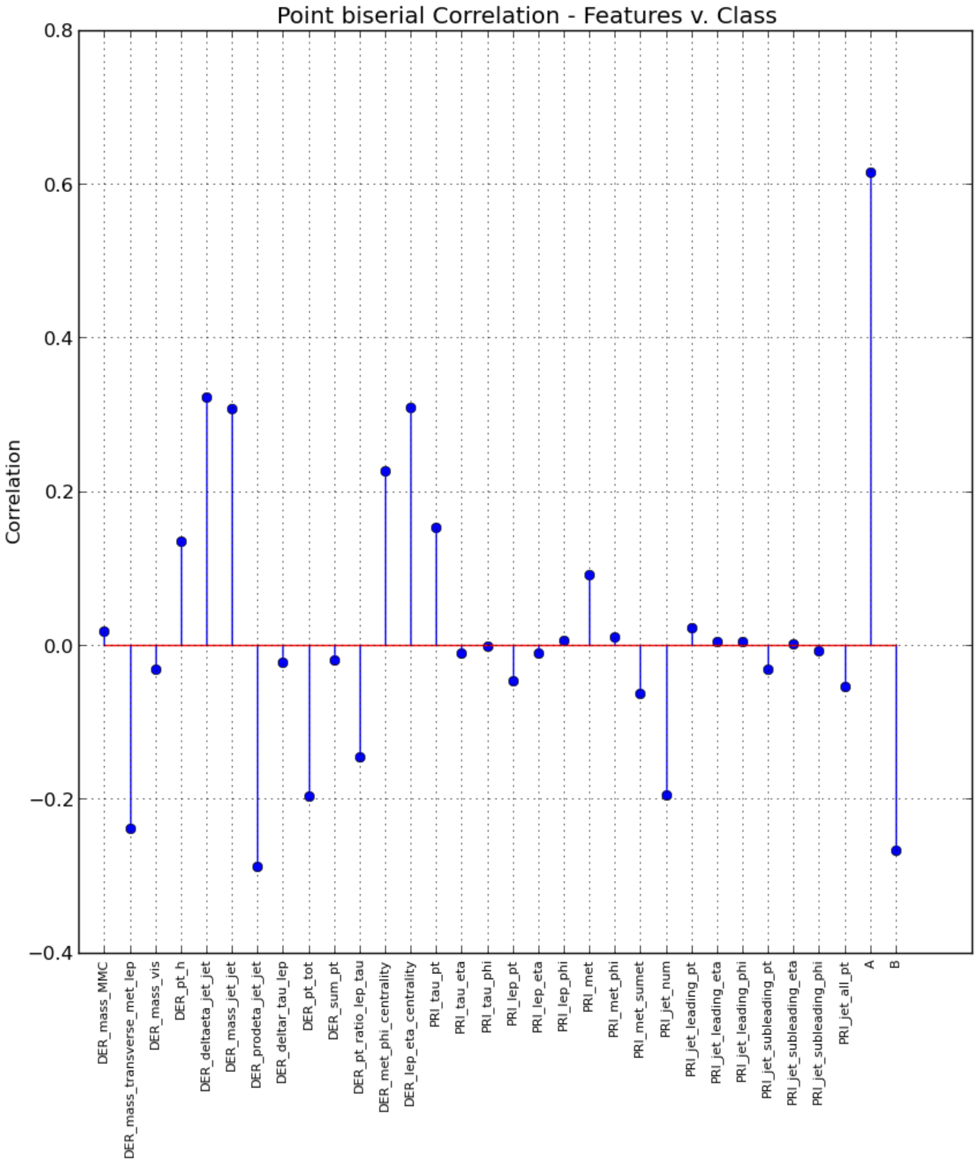

Figure 1: The figure shows the strength of correlation of derived features (DER_∗) over the primary features (PRI_∗). Derived jet features showed higher correlation and analytically derived feature A and B showed higher correlation to any of the primary features.

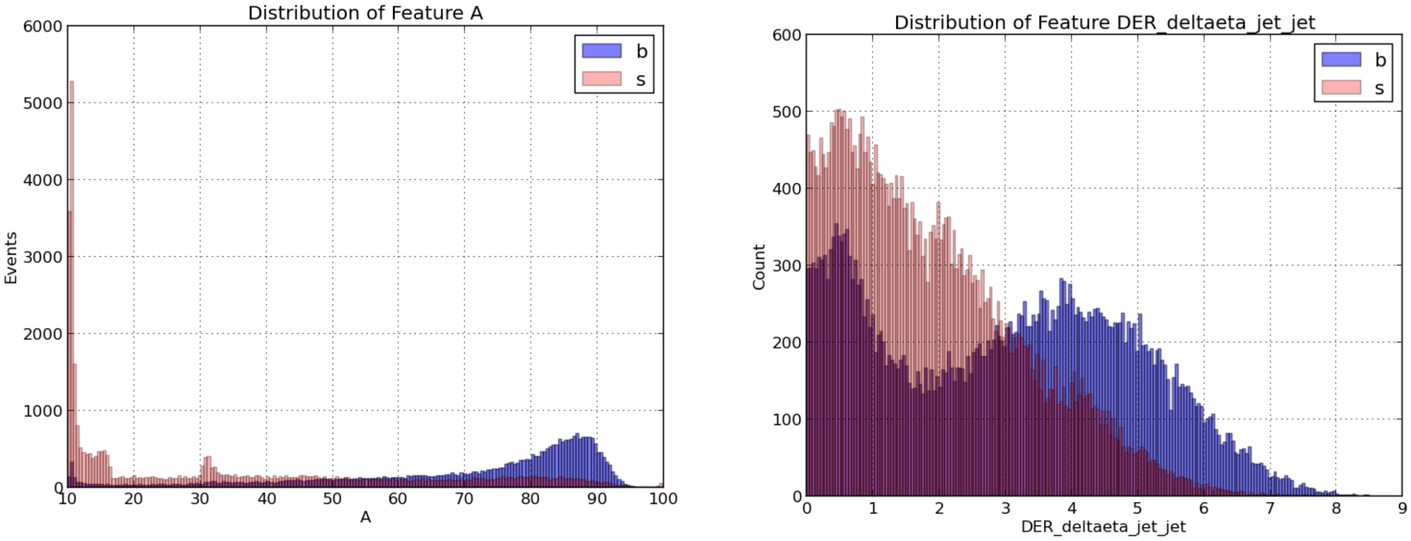

Figure 2: Distribution of Features by class breakdown

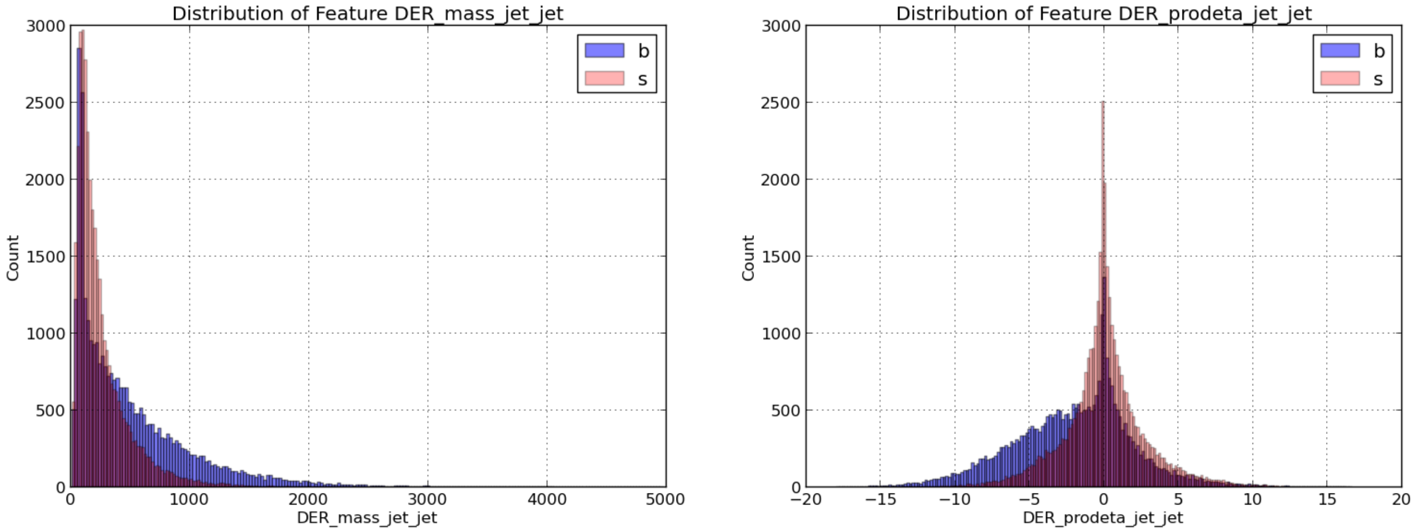

Figure 3: Distribution of Features by class breakdown

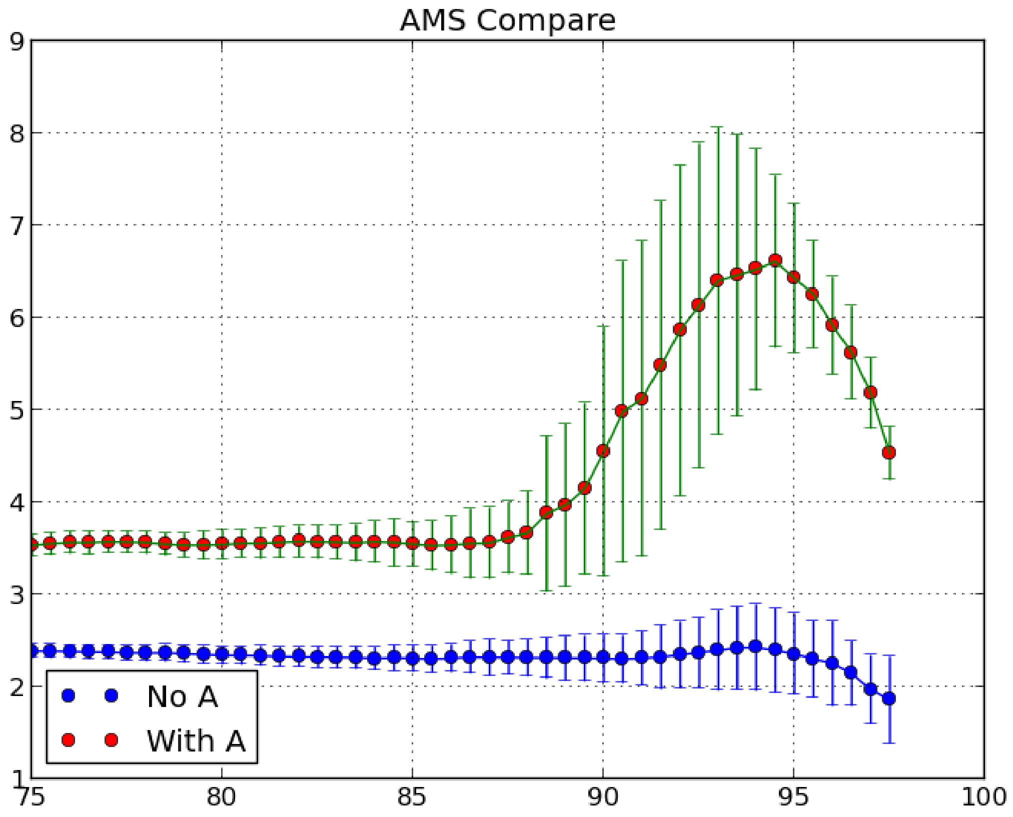

Figure 4: AMS curves with and without feature A, x-axis measures the rejection threshold (%) and y-axis shows the AMS score

3. Challenges in measuring the AMS difference

Firstly, the optimal parameters in the presence of feature A are different to the ones obtained when optimized without feature A. As a first step, we obtain two optimized models, one optimized in the presence of feature A and the other optimized without feature A. Further, the AMS value is sensitive to the choice of threshold and in order to account for this, we visualize AMS curves for both the models across a range of thresholds. The final training set which is fed to the SVM is constructed by sampling the original training set, hence repeated runs of the learning algorithm gave different AMS scores. In order to account for variational effects of sampling, we introduce error bars on the AMS score at each threshold centered on the mean AMS score computed by averaging the outcome of 100 iterations of the sampling and model fitting step. The AMS curves obtained for the two models are given in fig 4.

4. Summary

The inclusion of feature A in the dataset improves the AMS score by a significant margin. The improvement in score can be quantified by computing the difference between the mean AMS scores across a range of thresholds. For instance, the improvement in score in the threshold range of 83 to 87 is 0.98 ± 0.7. The important point to note is that we consider a relatively small fraction of the full dataset in training and testing the model. At the expense of compute time, it is possible to test how the model fares by taking into account the full training dataset of 250,000 samples by imputing the missing values either by using mean or median and testing on a larger sample of data.

References

[1] V. Lalchand, Analyzing the higgs boson using support vector machines (2016).