Data Science Challenge - Task 2

Christos Karapanagiotis

July 22, 2019

We have trained artificial neural networks to separate signal from background noise. In our first model the inputs are all the features from the data set while the output is a value in the range from 0 to 1. Values close to 1 denote Higgs signal while values close to 0 denote background noise. For the training of the neural networks we do not make use of the weights but we use them later to calculate the significance.

We have built a very simple neural network architecture with 2 hidden layers. To avoid over-fitting we split the data set into training and validation data. The validation data set corresponds to the 10% of the data set. After every epoch the validation loss as well as the validation accuracy is calculated and the best model in terms of validation loss is saved. Doing so, we keep the model which performed better on unseen data. All the parameters were selected after a short systematic study. Unfortunately the limited time did not allow further analysis.

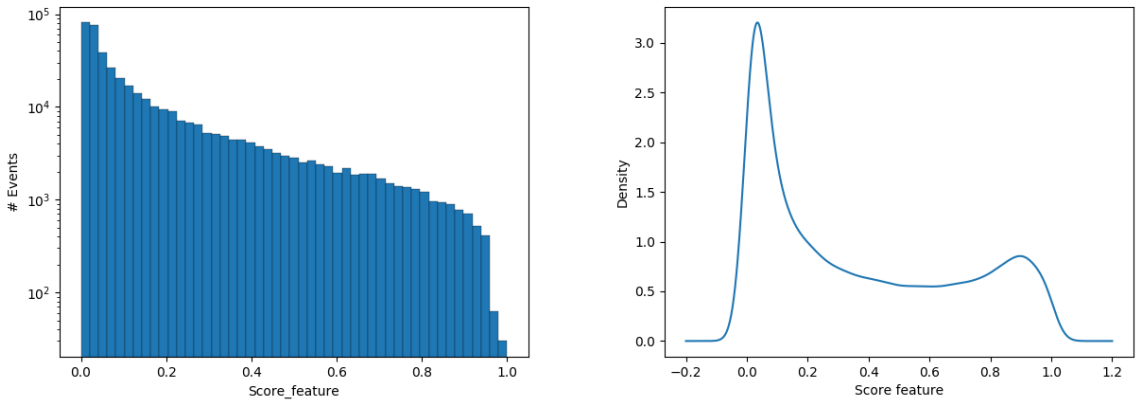

Figure 1: Histogram (left) and density plot (right) of the new score-feature using prediction model which requires all the features.

After the training of the neural network we tested our model with the test data. Fig.1 shows the histogram of the new score feature and the density plot as well. The y-scale of the histogram is in logarithmic scale for better visualization of the events close to 1. As we can see the vast majority of the events are close to 0 which agrees with the theory as we know that the events close to 0 should dominate. The other small peak in the density plot is around 0.9 and corresponds to the signal event (which are fewer). As we expected, the density of the score feature between 0.2 and 0.7 is smaller.

The metric for the performance of the model is the significance which is given by:

$$significance = \frac{s}{\sqrt{b}} \tag{1}$$

where s denotes the number of signal events that were predicted correctly (true positive) while b denotes the number of events that were predicted as signal but incorrectly (false positive) [1]. This metric is applied to signal regions that are defined by the threshold. For example, if we assume that all the events

that have signal score greater than 0.8 correspond to the Higgs signal, then the significance is calculated for the events with score grater than that.

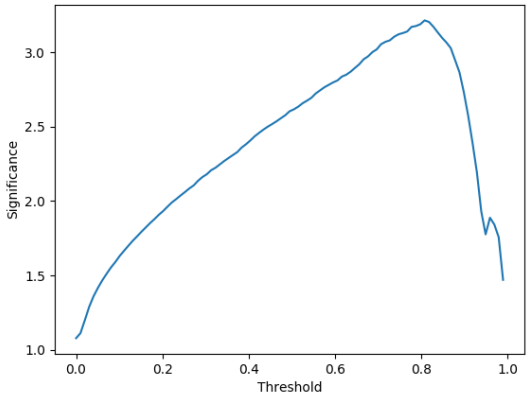

Figure 2: Significance vs. Threshold for the prediction model that requires all the features.

In Fig.2 we show the significance against the threshold. For threshold equal to 0.81 the significance takes its highest value which is equal to 3.21.

Afterwards we tried to reduce the features ignoring the five phi angle features which show a uniform distribution. The hyperparameters of the neural network as well as its architecture was the same.

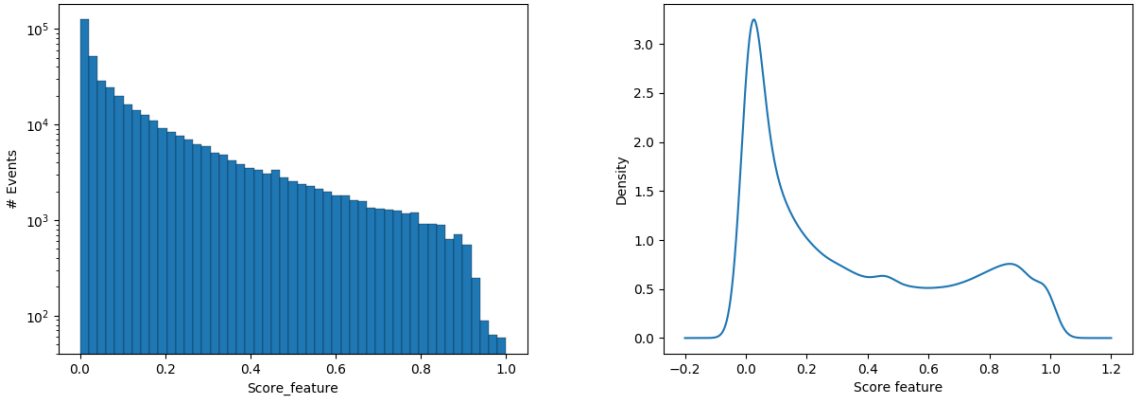

Figure 3: Histogram (left) and density plot (right) of the new score-feature using prediction model with decreased feature number.

The graphs in Fig.3 seem similar with the ones in Fig.1. A difference that one could notice is that the density plot for the first model is smoother than the one for the second model. In terms of significance the maximum value is 3.16 which corresponds to a threshold equal to 0.75.

Comments:

- The second model with the decreased number of feature does not seem to perform better on the test data set. Of course in order to have a safe conclusion we need to optimize as much as possible the networks and run them several times to test how robust they are.
- A solution for better results could be by increasing the complexity of the network or by optimizing the batch size as well as the learning rate. Additionally, a dropout layer could also be useful even though the first attempt did not improve the results considerably.

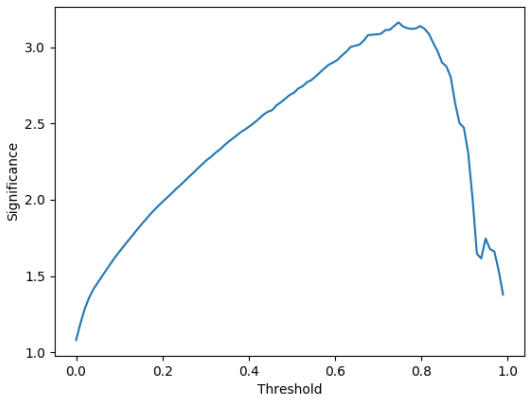

Figure 4: Significance vs. Threshold for the prediction model with decreased feature number.

- Apart from the network architecture, emphasis should be given on data. In the first task we found out that there are a lot of missing data in some columns and also some outliers. These data could be replaced with data that make more sense (e.g. with the mean). A careful analysis of these data could potentially increase the performance of the model. Apart from the missing data and the outliers, we have to emphasize more on the correlation between the features (e.g. PCA) to decide which features are good for predictions.
- Other machine learning algorithms (e.g. random forests) could be more useful for this specific problem.

References

[1] Claire Adam-Bourdarios, Glen Cowan, Cécile Germain, Isabelle Guyon, Balázs Kégl, and David Rousseau. The higgs boson machine learning challenge. In *Proceedings of the 2014 International Conference on High-Energy Physics and Machine Learning - Volume 42*, HEPML'14, pages 19–55. JMLR.org, 2014.