CS-433 - Project 1

Diego Dorn Sydney Hauke Ana-Maria Indreias
EPFL
diego.dorn@epfl.ch sydney.hauke@epfl.ch ana-maria.indreias@epfl.ch

1. Introduction

The aim of this project is to perform binary classification on the CERN Higgs Boson data set, in order to detect whether a Higgs boson was present during a given event. We begin with an exploratory data analysis phase, during which we discover important characteristics of the given features. Using these findings, we learn how to best pre-process our data set, and also develop a binary classification model.

2. Methodology

2.1 Data pre-processing

We start by normalizing the data. Then, we analyze the result with box plots, and notice the presence of outliers - including one with a deviation of 120. As seen in class, certain loss functions, such as the Mean Squared Error (MSE), do not perform well with outliers, so we are interested in removing these samples.

Naturally, we define a sample's **"weirdness score"** as the sum of its deviations across features. We decide to remove outliers, which we define as samples with a weirdness score ≥ 60.

We continue our data exploration by plotting the correlation matrix among features, and feature pairs against each other. We notice strong correlation ($\geq 95\%$) among three feature groups. However, according to the documentation, -999 represents an undefined value [1], and we realize that the correlated feature groups are no longer significant, as the features contain many undefined values.

After visualizing the data, we notice that some features are angles, some are undefined depending on PRI_jet_num's value, and that some distributions change favourably under a logarithmic transformation.

We decide to perform **feature expansion**, and begin by adding logarithms of selected features and 3 binary features of $PRI\_jet\_num = 0, = 1, \geq 2$.

We continue expanding features by experimenting with 2 basis functions: polynomial, as seen in class, and the truncated Fourier series, as developed in [2]. We also try pairwise combinations, such as sums, products, and trigonometric identities.

We notice that while feature expansion helps our model, it has some limits. First, the basis functions help the most when they are not used at the same time. We also notice that products and trigonometric identities have a good impact on our classifier's performance, but are too costly space-wise to add indefinitely.

To further improve our predictions, we **add a constant feature of value 1**. Surprisingly, this standalone modification yields a 75% accuracy with the MSE gradient descent regressor.

2.2 Ridge regression

A ridge regression model is based on the least squares solution. Its hyperparameter, λ, is used to penalize bigger weights, i.e. of higher l_2-norm. We also use λ as a way to obtain numerically stable matrices.

We believe ridge regression is a good choice overall because it is fast and offers a closed-form solution. One may be reluctant to use such a simple method for binary classification, perhaps out of concern for overfitting. However, in the Higgs boson dataset, the samples outnumber the features, so it would be difficult to overfit with ridge regression.

2.3 Stochastic gradient descent

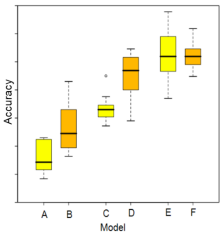
A stochastic gradient descent model iteratively updates the weights vector. At each step, it computes a partial gradient, and hence is more computationally expensive compared to ridge regression. Nevertheless, it offers more flexibility for binary classification: by clipping the predictions to [-1, 1] during training, we observe a slight increase in accuracy. For this reason, we decide to use an MSE SGD regressor as the final model.

2.4 Results

For all considered models, we transform predictions as follows: positive (> 0) predictions are mapped to 1, while the rest are mapped to -1. To measure their performance locally, we average the accuracies for five 80-20 train-test splits. The results are summarized below.

Model	Accuracy
A: Random Guess	$TODO \pm TODO$
B: Ridge Regression, no feature expansion	$TODO \pm TODO$
C: Ridge Regression ($\lambda =$ TODO	$TODO \pm TODO$
D: MSE SGD	$TODO \pm TODO$

Model	Accuracy
A: Baseline (Random guess)	0.743 ± 0.001
B: Simple logistic regression	0.782 ± 0.002
C: Handling NaN	0.797 ± 0.002
D: Polynomial Expansion (D=7)	0.817 ± 0.011
E: Regularized (D=9, λ=10^{-2})	0.821 ± 0.007
F: Magic Sauce™	0.839 ± 0.014



2.5 Techniques that were left out

We were keen to use **(regularized) logistic regression**, because it is a binary classification model "out-of-the-box". We experimented with 4-fold cross validation on the values of the regularization strength, but did not reach an accuracy as high as with simpler models. We believe this is the case because we did not choose a good gradient descent step size. Due to lack of time, we did not perform extensive cross validation on both hyperparameters.

Another technique that we do not employ is **handling the undefined values**. One can remove -999 values completely - however, 72% samples and 36.6% features have undefined values, so removing either would sharply decrease the data set size. One can also replace -999 values with zeroes, or the mean of valid values within each feature. We do not think replacement is a good technique for this data set, as -999 values are deliberately out of the normal range of the concerned features[1], and this information could be useful to our classifier.

3. Conclusions and future work

We develop a model which reaches $TODO \pm TODO$ accuracy, as evaluated using five 80-20 train-test splits. For data pre-processing, we remove outliers and make use of feature expansion techniques. Overall, we notice big improvements to regression methods when using "textbook" polynomial expansion, as well as pairwise products or trigonometric identities. Moreover, by understanding the various features' distributions, we are able to fine-tune this expansion by including logarithms of features {2, 9, 10, 13, 16, 19, 21}[1] and binary features based on the values of PRI_jet_num. For this reason, we believe that **exploratory data analysis is crucial in the development of our model.**

An interesting research direction would be to **combine four models**, each corresponding to a value of the integer feature PRI_jet_num. As detailed in the data set's documentation, certain features are undefined if $PRI\_jet\_num = 0$ or ≤ 1 [1]. One could pass new samples to the corresponding model out of the four. We did implement a simple version of this concept, but ultimately decided against using it for our final model.

[1] Feature indexing starts at 0

References

[1] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kegl, D. Rousseau. Learning to discover: the Higgs boson machine learning challenge. 2014.

[2] R. P. Adams. Features and Basis Functions. 2018.