CS-433: Project 1

Arina Lozhkina, Bastien Aymon, and Ricardo Ferreira Ribeiro
Ecole Polytechnique Fédérale de Lausanne EPFL

(Dated: October 31, 2021)

In this project we solve a classification problem based on a CERN Higgs Boson dataset. We use data preprocessing (normalization, categorization, replacing missing values), and try basic methods (linear, Ridge and logistic regression). We split the data in 3 subsets based on the value of 'PRI_jet_num', and we manage to obtain an accuracy of 0.830 and F1 score of 0.741 on the test dataset using a Ridge regression model with trigonometric and polynomial extension of degrees 12, 11 and 13, lambdas equal to 0.00023, 0.000001, and 0.0001 respectively for the 3 data categories.

I. EXPLORATORY DATA ANALYSIS

The training data set consists of 250,000 measurements, each made of 30 features. The associated labels can take two distinct values, depending on whether the event corresponds to a Higgs boson or background noise. Notably, the data is especially ill-conditioned due to its physical nature, as some features could sometimes not be computed, in which case the corresponding value is set to -999.0. It is therefore necessary to mitigate the effect of these outliers. We also compute the correlation matrix for the whole data set for later use (Fig. 1).

II. FEATURE PROCESSING

First and foremost, the data can be split into 3 categories on a *physical* basis. Depending on the number of jets detected ('PRI_jet_num' feature), a number of features cannot be determined and will always be set to -999.0. We therefore remove the non-defined features based on the classification presented on Tab. I . This means that we train three different models individually. The benefit of this added complexity is to greatly lower the number of outliers (with value -999.0) in the datasets.

The validity of this categorization will be shown in Sect. III, as the same methods are applied to both categorized and uncategorized data sets. We also remove the highly correlated features based on the results presented in Fig. 1. Lastly, we notice the features ending in '*_phi' represent an angle. They thus have a symmetry property and can be removed, since the distribution of their labels is mostly uniform, and they are not especially informative.

TABLE I. Some features will *always* be -999.0 and can therefore be removed. The latter are indicated by a 'x' in the corresponding column.

	PRI_jet_num		
Undefined feature	*0*	*1*	*2/3*
DER_deltaeta_jet_jet	x	x	
DER_lep_eta_centrality	x	x	
DER_mass_jet_jet	x	x	
DER_prodeta_jet_jet	x	x	
PRI_jet_leading_eta	x		
PRI_jet_leading_phi	x		
PRI_jet_leading_pt	x		
PRI_jet_subleading_eta	x		
PRI_jet_subleading_phi	x	x	
PRI_jet_subleading_pt	x	x	

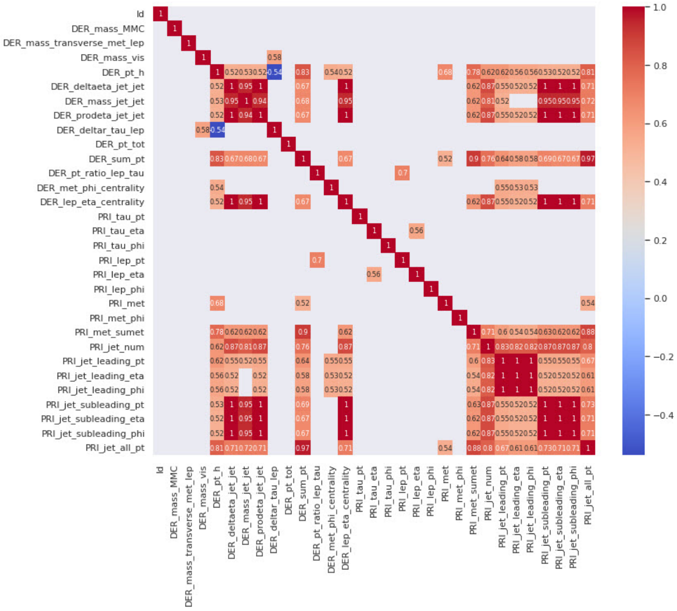

FIG. 1. Correlation matrix for the original data.

In summary, we remove features 4, 5, 6, 10, 12, 15, 18, 20, 22, 23, 24, 25, 26, 27, 28 and 29 for category 1, labels 4, 5, 6, 10, 12, 15, 18, 20, 22, 25, 26, 27 and 28 for category 2, labels 15, 18, 20, 22 and 25 for category 3 ('DER_mass_MMC' corresponds to label 0).

Despite the effort, some -999.0 values will still remain and need to be handled. A number of different techniques can be used in this case, which will be compared in the following section.

III. COMPARISON OF BASIC METHODS

We aim to determine what combination of method and feature processing technique produces the best results, in order to perform cross-validation and a detailed perfor-

TABLE II. Accuracy of the investigated feature processing methods (N = normalize, NM = normalize and set -999.0 values to the median, R = raw data, M = set the -999.0 values to the mean, Me = set the -999.0 values to the median). Numbers from 1 to 3 correspond to accuracy on the training data (80% of the available data), testing data and F1 score (both submitted on the challenge's website) respectively. Rows correspond to the least squares method, Ridge regression ($\lambda = 0.001$) and Ridge regression with a degree 4 polynomial extension ($\lambda = 0.001$) respectively.

N1	N2	N3	NM1	NM2	NM3	R1	R2	R3	M1	M2	M3	Me1	Me2	Me3
0.651	0.718	0.641	0.652	0.720	0.641	0.650	0.658	0	0.741	0.742	0.572	0.740	0.742	0.572
0.716	0.718	0.665	0.718	0.722	0.656	0.740	0.741	0.570	0.742	0.741	0.570	0.740	0.741	0.571
0.776	0.777	0.642	0.786	0.786	0.645	0.794	0.793	0.672	0.793	0.792	0.671	0.795	0.793	0.672

TABLE III. Accuracy of Ridge regression models, degree 4, $\lambda = 0.001$. Letters a to c correspond the replacing the -999.0 values by the median, the mean and doing nothing to them respectively, then normalizing in all cases. T1 to T3 are the training accuracies (80% of the data) for categories 1 to 3. The testing accuracies Te and the F1 score were obtained by submitting the prediction on the challenge website.

	T1	T2	T3	Te	F1
P4-a	0.847	0.818	0.829	0.824	0.732
P4-b	0.836	0.776	0.805	0.806	0.699
P4-c	0.835	0.766	0.787	0.799	0.685

mance evaluation on the best methods only. We begin by ignoring categorization, as shown in Tab. II. The data is then split into 3 categories as explained before, and more tests are performed. Corresponding results are presented in Tab. III. For completeness, all the methods presented in class can be found in the submission files, even if they were not used in the present analysis.

IV. DISCUSSION

From Tab. II, we determine that the Ridge regression method with polynomial extension is the best overall. We cannot yet definitely state what feature processing method is the best, but the median replacement technique seems to yield robust results. This is expected since many invalid -999.0 values are present in the data. From Tab. II, we determine that data categorization yields better results and should be used. Also, the median replacement technique seems to be the most fit, which can also be explained by the presence of some outliers, even if there are not as many in this case.

Fig. 2 presents what degree should be used in the polynomial base and what λ should be used for the Ridge regression in the 3 categories. Note that for increased accuracy, a trigonometric feature extension was implemented, which consists of adding the sine and cosine of each feature to the feature set.

Some feature processing methods were attempted but are not presented here because of their poor performance. Those include linear and quadratic interpolation, removing rows containing outliers, logarithmic and min-max normalization, etc.

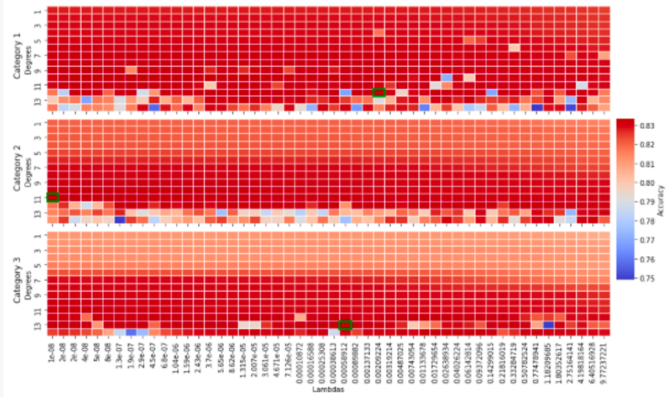

FIG. 2. Accuracy as a function of polynomial degree and lambda for all categories (the better the accuracy, the redder the color). The best results (in green) were obtained for degrees 12, 11 and 13, lambdas of 0.00023, 0.000001, and 0.0001 for categories 1 to 3 respectively.

V. CROSS-VALIDATION OF THE CHOSEN METHOD

To correctly assess the efficiency of the algorithm used to solve the problem of classifying events of the original dataset, we apply a cross-validation algorithm and calculate the accuracy metrics and loss functions. The study yielded accuracies of 84.4% for category 1, 80.6% for category 2 and 83.3% for category 3, which is satisfactory.

VI. CONCLUSION

Using the optimal values determined in the preceeding sections (see Fig. 2), we manage to obtain an accuracy of 83% on the test data after submission on the challenge's website. This accuracy is well in the range of the best results of this year. Also, tests involving more advanced methods on the raw data (*e.g.*, gradient boosting) yielded results in the same range. Provided more time, different successful methods could have been stacked for a slight accuracy improvement.