Event Classification Using Various Machine Learning Regression Techniques

Jean Gillain
sciper no. 331411

David Desboeufs
sciper no. 287441

Mathieu Caboche
sciper no. 282182

***Abstract*—This project aims at classifying events recorded at CERN. This is achieved by implementing various machine learning regression techniques and eventually using them to create a good model. The project also required us to process the given data in order to achieve the best possible results.**

I. Introduction

In this report we will give a brief but detailed explanation of the thought and development process we went through in order to train a model enabling classification of a physical event. The assigned goal was to create a model capable of classifying data records from CERN as occurrences or non-occurrences of the Higgs boson.

Section II describes the given data set in more detail. Over the course of this project we used several different regression techniques learned during the course. In order to find the best possible model we then used one of the implemented regressions with pre-processed data and tuned their parameters as needed. An important part of creating a good model is to have a good training data set. That is why data analysis and cleansing was a major part of this project as well. The methods used to analyze and clean our data sets, as well as the steps we went through to find the best regression method are described in section III. Finally, we will compare the models obtained by various regression techniques to our best performing implementation in section IV.

II. The Higgs boson Data Set

The given data set for model training consists of some 250000 records which classify an event as a Higgs boson occurrence or not. The events are recordings of 30 features. The testing data set is identical except that classification is left up to our model.

Let us describe the training data in more detail so as to highlight the importance of analyzing and subsequently cleaning the latter. A look at the records quickly reveals that the value -999 appears very frequently and for a number of different features and records. This value could thus be interpreted as an error of measurement. Furthermore we can notice that some features take on only very few different values while others almost never repeat the same value twice. This raises the question of feature importance, i.e. which features impact the classification a lot and which do not (or even lead to misclassification). Section III describes how we decided to clean the data set with respect to the above mentioned concerns.

III. Models and Methods

As required by the project guidelines we implemented 6 regression techniques. We will not go into the details of their implementation. Rather we will describe how we tried to obtain a best performing model and with which solution we eventually did get the best results. We shall also cover how we cleansed the data sets and for what reasons.

A. Finding the Best Regression Technique

The regression which we decided to use in order to get the best model is the ridge regression. This decision is based on the fact that ridge regression performed best among all of our implemented regressions, as shown in table I.

To further improve on the performance of ridge regression we cleaned the input data and used feature extension. Section III-B describes the steps we took to clean the data and why we decided to clean the input data as we do. Before we dive into the details of our final feature extension, let us first describe a prior attempt at improving ridge regression.

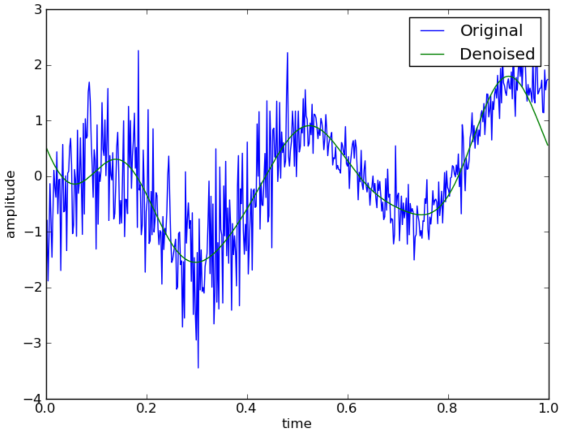

Figure 1. Grid search for our implementation of ridge regression

Our initial intuition was that features with large variance should impact the resulting model more than features with low variance. Thus we decided to use augmented feature vectors on the input data according to feature variance. We first sort the features in increasing order of their variance. The ordered features are then subdivided into equally sized groups. The number (and size) of groups is a parameter we are free to adjust. Finally we extend the feature space by building polynomials of higher exponent for groups of higher variance. The result are augmented feature vectors with say 5 different polynomials (if we have 5 groups).

As it turns out, the previously described implementation did not provide an actual improvement over classical ridge regression on the raw (non cleaned) data. That is why we implemented a different feature extension. We first apply the inverse logarithm function (1) and then a fixed polynomial expansion of degree 9. Note that this introduces a constant term.

$$f(x) = \log \frac{1}{1+x} \tag{1}$$

Finally, we need to adjust the parameters of our regression. This is the very last step, after data cleaning and feature extension. In order to find the best parameters ($lambda$ and $gamma$), we simply use a grid search algorithm on possible parameter values. Figure 1 shows the obtained results for our best performing regression (data cleaning, feature extension and ridge regression). The best values found are thus: $lambda =$ and $gamma =$

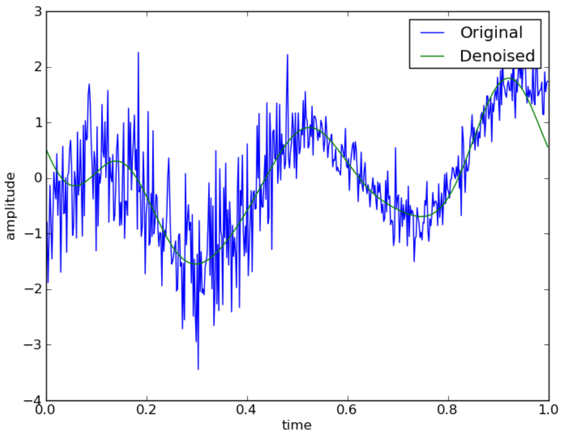

Figure 2. Correlation matrix of all features

B. Data Processing

Let us briefly go over some data cleaning that did not prove effective in improving model performance.

- Removing features that contain over a given percentage of error values (value -999).
- After applying the previous point, removing lines (measurements) that still contain error values.
- Remove features that contain few different values (frequency analysis).

Here are the data cleaning steps that dis prove useful and which we actually use in our best performance regression.

- Replace all error values by the mean of the feature in which they appear. We do not consider error values for mean calculation.
- Compute the correlation matrix between all features and keep only one feature in cases where multiple features are correlated. The correlation matrix is shown in figure 2.
- Standardize all features. This step is the last step of the data cleaning.

C. Model Validation

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est

IV. Results

In this section we analyze our best performing regression (respectively the model obtained by the latter) to the models obtained using our implementations of other regression techniques.

To compare the performance of the models resulting from our various regression implementations we can simply compare the model accuracy. I.e. we check what percentage of a known classified set is correctly classified by our model, as described in section III-C. The model accuracies obtained by our various regressions are listed in table I.

Note that all regressions are run on the raw, uncleaned and unprocessed data except for *Our Best Regression*.

Regression Technique	**Accuracy (Avg.) - STD**
Least Squares GD (and SGD)	70% - 70%
Least Squares	70% - 70%
Ridge Regression	70% - 70%
Logistic Regression	70% - 70%
Regularized Logistic Regression	70% - 70%
Our Best Regression	70% - 70%

Table I
Accuracy of Various Models Obtained by Our Regressions

V. Summary

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.