Event Classification Using Various Machine Learning Regression Techniques

Jean Gillain
sciper no. ??????

David Desboeufs
sciper no. ??????

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.

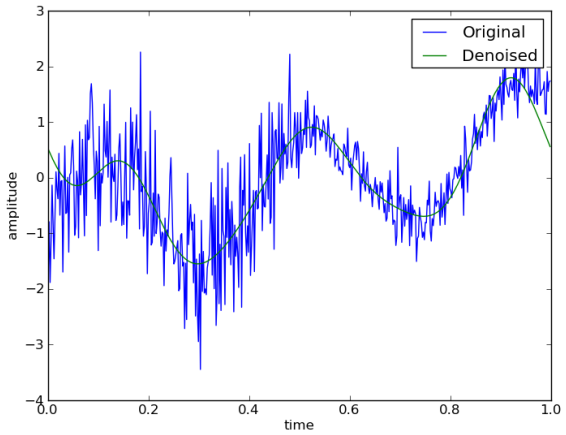

Figure 1. TODO

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 TODO (apply inverse log and poly?). 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 TODO, let us first describe a prior attempt at improving 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 order the features in increasing order of their variance. The ordered features are then subdivided into equally sized groups. We used up to 5 groups in our trials. Finally we extend the feature space by raising groups of higher variance to a higher exponent than groups of lower variance. The result are augmented feature vectors with say 5 different exponents, with features of high variance being risen to high exponents.

B. Data Processing

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.

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. 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.

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 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.

IV. Results

In this section we analyze our best performing regression and model to the models obtained using the given, required regressions.

One criterion to evaluate the performance of a gradient descent regression is the speed at which the gradient converges. In fact, the faster the gradient goes towards 0, the faster our regression will find a good model. Since many of the regressions implemented during this project are gradient descent algorithms, this comparison is relevant. Figure 2 shows gradient convergence for our implementations of various gradient descent algorithms.

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.

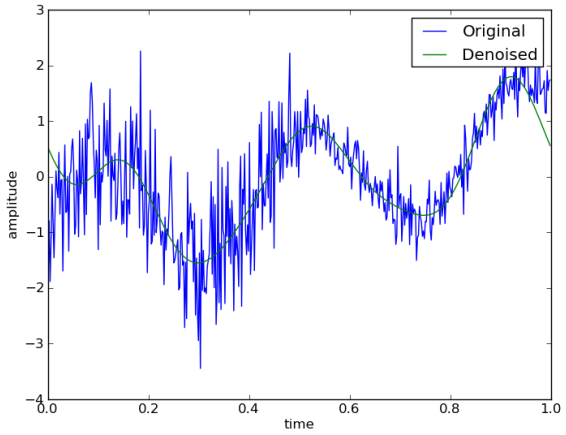

Figure 2. TODO Grid search for lin reg

Regression Technique	Obtained Model Accuracy
Least Squares GD (and SGD)	70%
Least Squares	70%
Ridge Regression	70%
Logistic Regression	70%
Regularized Logistic Regression	70%
Our Best Regression	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.