Project 1: Higgs Boson

Gael Moccand, Pascal Bienz
gael.moccand@gmail.com, pascal.bienz@gmail.com
Machine Learning Course, EPFL

***Abstract*—The abstract should really be written last, along with the title of the paper. The four points that should be covered:**

1) **State the problem.**
2) **Say why it is an interesting problem.**
3) **Say what your solution achieves.**
4) **Say what follows from your solution.**

I. Introduction

The Higgs boson is an elementary particle discovered at the Large Hadron Collider at CERN in 2013. In order to produce it, physicists accelerate protons and make them collide at high speeds. In some rare cases, the collision generates a Higgs boson. A major problem that arises when scientists want to observe the particle is that its life is very short. Indeed, a Higgs boson quickly decays into other particles. For that reason, it is observed indirectly by looking at the outputs of the decay. However, this process can become tricky because a Higgs boson's decay signature can be very much alike another particle's signature.

In this paper, a machine learning method that efficiently estimates the likelihood that a given measurement is due to a Higgs boson or some other particles is presented... ADD SOME DETAILS ABOUT THE METHOD HERE.

Two specific data sets are used to optimize the method. The first set S_t is called the training set and contains $N = 250000$ observations. It is used to develop the model. The second set S_v is the validation set and has 568238 events. These data are used to validate the model and make sure that we the model does not overfit the data of S_t. In both sets, the events are characterized by 30 features. Among them, 13 are "raw" quantities about the bunch collision as measured by the detector and 17 are "derived" quantities computed from the raw features, which were selected by the physicists. Finally, let's point out that in some cases the variables of some entries are not available. In order to handle the missing data, it is primordial to apply a preprocessing stage.

II. Methodology

Before digging into prediction methods, it is beneficial to handle the data. The preprocessing is key and impacts the prediction. For instance, a scatter plot matrix of each features can be useful to select or drop some features. In our case, a basic standardization by subtracting the mean and dividing by the standard deviation has been applied to the input data. Furthermore, the missing value have been replaced by the mean of the corresponding feature, which is better than a value which has no signification.

To predict the nature of the measurement, we need to find a function that best approximates the output $\mathbf{y}$ with the given inputs $\mathbf{x}$

$$y_n \approx f(\mathbf{x_n}) \ \forall n.$$

A common choice is to use a linear regression, i.e.

$$f(\mathbf{x_n}) = w_0 + \mathbf{x}_n^T \begin{pmatrix} w_1 \\ \vdots \\ w_D \end{pmatrix} =: \tilde{\mathbf{x}}_n^T \mathbf{w}$$

where $\mathbf{w} = (w_0 \dots w_D)$ are the parameters of the models. Note that we add a tilde over the input vector to indicate that it contains an offset term.

A cost function is needed to estimate how well the model does. Again, a common choice is to use the Mean Square Error (MSE):

$$\mathcal{L}_{MSE}(\mathbf{w}) := \frac{1}{N} \sum_{n=1}^{N} (y_n - f(\mathbf{x}_n))^2$$

As a starting point, it has been decided to chose a simple least squares regression to compute $\mathbf{w}$ directly. The parameters are given by the following expression:

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

which leads to a prediction for an unseen data point $\mathbf{x}_m$ given by

$$\hat{y}_m := \mathbf{x}_m^T \mathbf{w}^* = \mathbf{x}_m^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y},$$

where $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$ and $\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1D} \\ x_{21} & x_{22} & \dots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{ND} \end{bmatrix}$.

Linear models are inherently not very rich. A way to increase their representational power one can boost the input by adding a polynomial basis of arbitrary degree M:

$$\phi(\mathbf{x}_n) = \begin{bmatrix} 1 & x_{n1} & x_{n1}^2 & \dots & x_{n1}^M \\ 1 & x_{n2} & x_{n2}^2 & \dots & x_{n2}^M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{nD} & x_{nD}^2 & \dots & x_{nD}^M \end{bmatrix}.$$

We then fit a linear model to the extended feature vector $\phi(\mathbf{x}_n)$:

$$y_n \approx \phi(\mathbf{x}_n)^T \mathbf{w}.$$

Unfortunately, this tuning has a negative effect: overfitting. Regularization is a way to mitigate this undesirable behavior by penalizing the model with a parameter $\Omega(\mathbf{w})$. The optimization problem becomes

$$\underset{\mathbf{w}}{\operatorname{argmin}} \mathcal{L}(\mathbf{w}) + \Omega(\mathbf{w}).$$

The Ridge regularization $\Omega(\mathbf{w}) = \lambda ||\mathbf{w}||_2^2$ is a natural choice. It penalizes the large model weights w_i.

III. Results

In order to measure the quality of a method, the predictions obtained are sent to the predictive modelling competitions platform *Kaggle* which returns a score between 0 and 1.

As a baseline, a simple least squares together with a polynomial basis with different degrees has been used. As it can be seen on Fig.1, it is obvious that the smallest error is obtained by choosing a degree of. However not every polynomial basis could be used because some basis lead to a unsolvable solution, that's the reason why some degree point are missing on the figure.

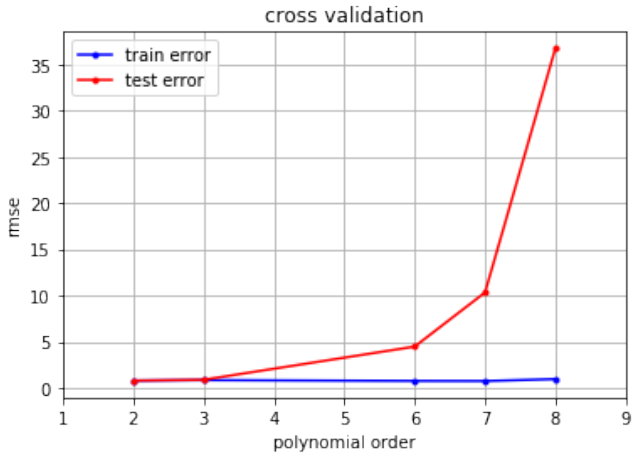

Figure 1. MSE w.r.t polynomial degree for direct least squares method

It seems that selecting raw or the derived features only gives a lower score than taking into consideration both types, even if there is information redundancy by taking both types. Moreover, there is a feature which only has integer numbers. A possible improvement would be not to feed it to the polynomial basis and use this feature directly. Unfortunately, this trick has not shown much improvement on the final score.

IV. Discussion

MSE is not a good cost function when outliers are present.

Another improvement would be also to weight the features independently because fot the moment each feature are equally weighted.

V. Summary