Data Mining - Flavours of Physics

Selene Baez Santamaria (2572529)
Andrea Jemmett (2573223)
Dimitris Alivanistos (2578740)
Vrije Universiteit Amsterdam
Amsterdam, Netherlands

***Abstract*—**

I. Introduction

The goal of this competition is to find undiscovered phenomena given data from the Large Hadron Collider (LHC). It is a classification task for the variable *signal*, which can have values 0 or 1, given the values for other 50 variables.

The competition provides a Starter Kit, training and test sets, a specific submission format and a couple check files to evaluate the submitted data. The link for these assets is here This competition was closed only five months ago, and it is still relevant in the field.

A. Objectives

- Compare the results from having domain knowledge and not.
- Objective 2.

B. Research question

C. Hypothesis

Comparable accuracy (at least as high as the winner of the Kaggle competition) can be achieved by a system that lacks domain knowledge.

D. Contribution

E. Organization

Section II explores the solutions submitted to Kaggle at the time of competition. Section III describes the dataset and the competition regulations (such as the agreement measures to be met). Section IV depicts the model we proposed to tackle the classification task. Section V refers to the experimental set up we use to test our model, while Section VI shows the results obtained. Finally, Section VII summarizes our observations.

II. Related Work

III. The Dataset

As said, the dataset was offered on Kaggle for a competition organized by institutions such as CERN and LHCb. The data have been collected directly at the LHC during experiments with high energy particles collisions. The dataset consists of a collection of collision events and their properties. The objective of the Kaggle competition was to predict whether a $\tau \to 3\mu$ decay (the one that identifies a lepton flavour) was present in the collision. From scientists this phenomenon is supposed *not* to happen, so the goal of the competition was to discover $\tau \to 3\mu$ happening more frequently than scientists currently expect.

A. Training Data

For training there is a labeled dataset ready to be used to train a classifier. The label, marked as `signal` with range in $0, 1$ where 1 identifies signal events while 0 represents background events). Signal events have been simulated while background events come from real data collected by the LHCb detectors, observing collision of accelerated particles with a specific mass range in which $\tau \to 3\mu$ can't happen.

The training dataset is given is CSV format and contains 49 features plus target label. For a detailed description of the features see Appendix A

B. Testing

For this Kaggle competition the submission procedure is different from the usual ones. The dataset comes with, besides a test set, an *agreement* and a *correlation* set. Any submission has to pass the agreement and correlation checks before being scored on the test set.

1) Agreement Test: The agreement dataset contains real and simulated events for a much more known, observed and understood decay: $Ds \to \varphi\pi$. The motivation for this check is that since the training set contains simulated data (for a phenomenon not well understood), it is possible for the classifier to reach high performances by picking up features that are not well modeled by the simulation. The check then requires the classifier not to expose a large discrepancy when applied to real-world and simulated data.

For this score, we are provided with a dataset on the control channel $Ds \to \varphi\pi$ which has the same features as the training set. This type of decay is not present in the training data. The *Kolmogorov-Smirnov* test is used to evaluate the differences between real and simulated data between the classifier distribution on each sample. The Kolmogorov-Smirnov metric has to be smaller than 0.09 to pass the agreement check.

2) Correlation Test: This test checks whether the classifier is uncorrelated with the τ mass. Because mass is a measured quantity, scientists don't trust it when building a model. Correlation with mass in an artificial signal-like peak or lead to incorrect estimations of background signals.

The `mass` column is not included in the test dataset. However, this hidden mass information is used to perform a *Cramer-von Mises* test, iteratively comparing two distributions

of a) predicted values from submission for entire dataset and b) predicted values within a certain mass region in rolling window fashion along the whole mass range. Getting similar distributions for all mass sub-regions means that the classifier is not correlated with the mass. The submission must give a Cramer-von Mises value less than 0.002 to pass the correlation test.

3) Test Set: The test set has the same columns that the training set has except for `mass`, `production`, `min_ANNmuon` and `signal`. The data contained in this dataset consists of:

1) simulated signal events for $\tau \to 3\mu$;
2) real background events for $\tau \to 3\mu$;
3) simulated events for the $Ds \to \varphi\pi$;
4) real background events for $Ds \to \varphi\pi$.

Events related to the control channel are not used for scoring, but by the agreement check (Section III-B1). One should treat all samples as coming from the same collision channel during classification.

IV. System design

The model consists of a set of neural networks, with different topologies, assembled together. The default values are $n\_models = 30$ trained for $n\_epochs = 60$ each.

A. Neural Networks design

Each neural network consists of 5 fully connected layers. The first 4 layers have a *PReLU* activation function and the last one has a *softmax* function. Furthermore, layers 2, 3, and 4 perform Dropout in order to avoid overfitting of the network. The input and output connections, as well as summary of the above characteristics, is shown in the following table:

Layer	**n of inputs**	**n of outputs**	**Dropout rate**	**Activation function**
Layer 1	N of features	75	0%	PReLU
Layer 2	75	50	11%	PReLU
Layer 3	50	30	9%	PReLU
Layer 4	30	25	7%	PReLU
Layer 5	25	2	0%	softmax

We use a *Cross entropy* loss function.

B. Model Ensemble

The models output is them combined together to produce one final prediction. So far, the aggregation of the models is done by taking the mean of the individuals.

V. Experimental set up

A. Implementation

B. Training

C. Testing/Evaluation

VI. Results

VII. Conclusion

APPENDIX

Follows a list of available features for training:

- `FlightDistance` - distance between τ and PV (primary vertex, the original protons collision point);
- `FlightDistanceError` - error on FlightDistance;
- `mass` - reconstructed τ candidate invariant mass, which is *absent in the test samples*;
- `LifeTime` - life time of tau candidate;
- `IP` - Impact Parameter of tau candidate;
- `IPSig` - Significance of Impact Parameter;
- `VertexChi2` - χ^2 of τ vertex;
- `dira` - cosine of the angle between the τ momentum and line between PV and tau vertex;
- `pt` - transverse momentum of τ;
- `DOCAone` - Distance of Closest Approach between p0 and p1;
- `DOCAtwo` - Distance of Closest Approach between p1 and p2;
- `DOCAthree` - Distance of Closest Approach between p0 and p2;
- `IP_p0p2` - Impact parameter of the p0 and p2 pair;
- `IP_p1p2` - Impact parameter of the p1 and p2 pair;
- `isolationa` - track isolation variable;
- `isolationb` - track isolation variable;
- `isolationc` - track isolation variable;
- `isolationd` - track isolation variable;
- `isolatione` - track isolation variable;
- `isolationf` - track isolation variable;
- `iso` - track isolation variable;
- `CDF1` - cone isolation variable;
- `CDF2` - cone isolation variable;
- `CDF3` - cone isolation variable;
- `production` - source of τ (*absent from test data*);
- `ISO_SumBDT` - track isolation variable;
- `p0_IsoBDT` - track isolation variable;
- `p1_IsoBDT` - track isolation variable;
- `p2_IsoBDT` - track isolation variable;
- `p0_track_Chi2Dof` - quality of p0 muon track;
- `p1_track_Chi2Dof` - quality of p1 muon track;
- `p2_track_Chi2Dof` - quality of p2 muon track;
- `p0_pt` - transverse momentum of p0 muon;
- `p0_p` - momentum of p0 muon;
- `p0_eta` - pseudorapidity of p0 muon;
- `p0_IP` - Impact parameter of p0 muon;
- `p0_IPSig` - Impact Parameter Significance of p0 muon;
- `p1_pt` - transverse momentum of p1 muon;
- `p1_p` - momentum of p1 muon;
- `p1_eta` - pseudorapidity of p1 muon;
- `p1_IP` - Impact parameter of p1 muon;
- `p1_IPSig` - Impact Parameter Significance of p1 muon;
- `p2_pt` - transverse momentum of p2 muon;
- `p2_p` - momentum of p2 muon;
- `p2_eta` - pseudorapidity of p2 muon;
- `p2_IP` - Impact parameter of p2 muon;
- `p2_IPSig` - Impact Parameter Significance of p2 muon;
- `SPDhits` - number of hits in the SPD detector;
- `min_ANNmuon` - muon identification. LHCb collaboration trains Artificial Neural Networks (ANN) from informations from RICH, ECAL, HCAL, Muon system to distinguish muons from other particles. This variable denotes the minimum of the three muons ANN. This feature should not be used for training and is *absent from the test sets*;
- `signal` - is the target variable to predict.