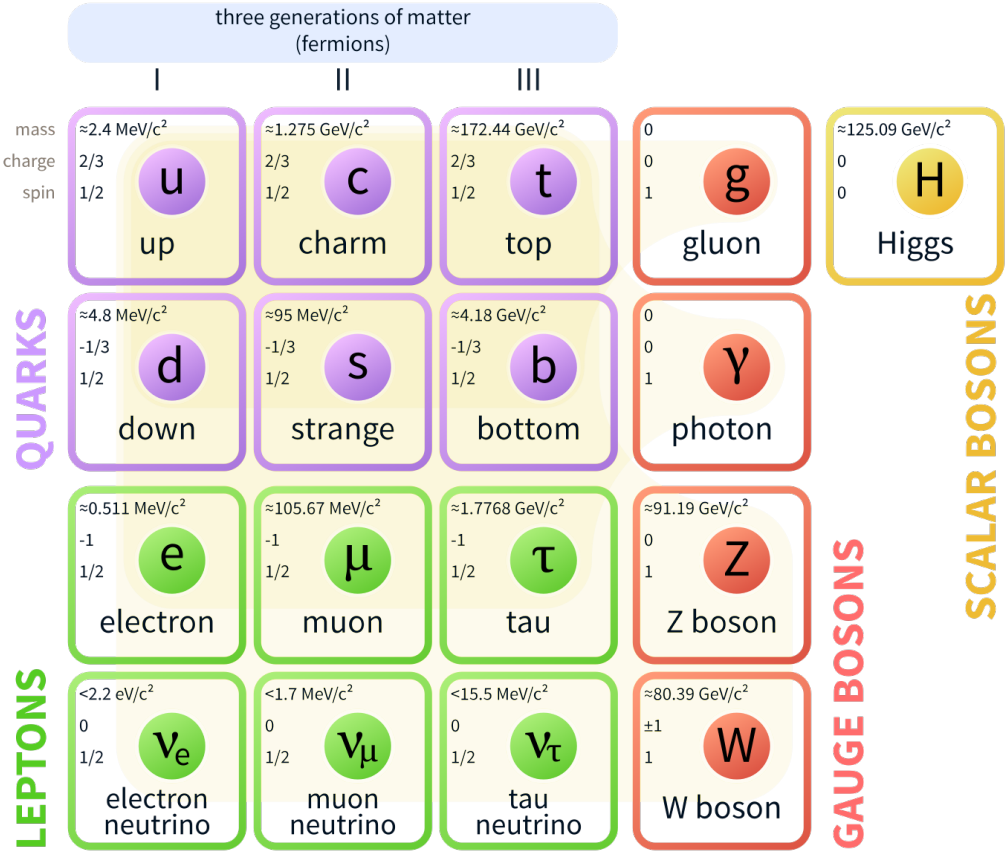

Capstone Project Proposal

Prepared by: Nora Petrova

Date: 4 June 2017

Title: Flavours of Physics

URL: https://www.kaggle.com/c/flavours-of-physics

PROPOSAL

I have chosen to complete the "Flavours of Physics: Finding $\tau \rightarrow \mu\mu\mu$" Kaggle competition as my capstone project. The competition's description and formulation can be found at its kaggle page [1].

Domain Background

The Standard Model (SM) [2] of physics is our most successful theoretical framework to date. It describes the strong, weak and electromagnetic interactions and has been consistently experimentally verified since its formulation in the 70s. Despite its success, it fails to explain some important observed phenomena, such as, matter–antimatter asymmetry in the universe or the structure of generations of elementary particles. Moreover, it does not predict the existence of dark matter or describe quantum gravity. However, SM, does not point to any new unexplored areas which we can look into in order to find answers to these questions. One such way is to find violations of symmetries that are widely believed to be conserved and are intrinsic to the structure of SM. One such symmetry around which this problem revolves is lepton flavour [3]. If significant signal for decays which do not conserve lepton flavour (for charged particles) was found in the dataset, it would challenge some of the assumptions of the SM. The implications for such a a finding would be far–reaching – it could lead to new physics and enrich our understanding on a fundamental level.

My motivation for picking this problem is two–fold: I have a deep interest in physics and am eager to bridge the gap between my physics and machine learning knowledge, and I believe this research is of great importance as it has the potential to challenge our current understanding.

Problem Statement

Given a list of collision events and their properties from the LHCb [4] experiment (mixed in with simulated data), predict whether a $\tau \rightarrow 3\mu$ (tau to three muons) decay happened in the selected collision. Because such events are believed to be rare or non–existent, the goal is to discover this particular decay ($\tau \rightarrow 3\mu$) happening more frequently than expected. The goal is to improve the discriminating power between signal events (where the decay did occur) and background events (where it did not occur). The resulting classifier cannot be too dependent on discrepancies between real and simulated data or on τ mass.

Datasets and Inputs

The datasets provided are a mixture of simulated and real data collected by the LHCb detectors observing collisions of accelerated particles with a specific mass range in which $\tau \rightarrow 3\mu$ decays cannot happen (according to the SM). For the simulated events, `signal` is set to 1 and for background events it is set to 0.

Four datasets are provided:

- `training.csv` is a labelled dataset to use for training the classifier. Background events are composed of real data and simulations.
- `check_agreement.csv` is a labelled dataset with the same features as the training dataset. It is to be used for checking the agreement between simulated and real data.
- `check_correlation.csv` is a labelled dataset (same features as the training dataset) for checking the correlation of the classifier with the tau mass.
- `test.csv` is a non-labelled (signal and background are mixed and indistinguishable) dataset which contains simulated signal events and real background data, simulated events and real data for the control channel.

The training dataset has 50 features, which include metrics, such as: flight distance, properties relating to the tau candidate (mass, momentum, life time, impact parameter and others), variables relating to final states tracks, geometric variables relating to the particles' trajectories through the detector and more.

A sample row looks like this:

```
id, LifeTime, dira, FlightDistance, FlightDistanceError, IP, IPSig, VertexChi2, pt, DOCAone,
DOCAtwo, DOCAthree, IP_p0p2, IP_p1p2, isolationa, isolationb, isolationc, isolationd,
isolatione, isolationf, iso, CDF1, CDF2, CDF3, ISO_SumBDT, p0_IsoBDT, p1_IsoBDT, p2_IsoBDT,
p0_track_Chi2Dof, p1_track_Chi2Dof, p2_track_Chi2Dof, p0_IP, p1_IP, p2_IP, p0_IPSig, p1_IPSig,
p2_IPSig, p0_pt, p1_pt, p2_pt, p0_p, p1_p, p2_p, p0_eta, p1_eta, p2_eta, SPDhits, production,
signal, mass, min_ANNmuon

18453471, 0.001578324, 0.999999344, 14.03333473, 0.681400955, 0.016039478, 0.451886147,
1.900432944, 1482.037476, 0.066666551, 0.060601622, 0.083659522, 0.208855301, 0.074342899, 8,
5, 7, 1, 0, 3, 4, 0.473951876, 0.349446565, 0.329157293, -0.57932359, -0.256309092,
-0.215444162, -0.107570335, 1.921700239, 0.866657019, 1.230708122, 0.988053977, 0.60148257,
0.277089804, 16.24318314, 4.580875397, 5.939935684, 353.8197327, 448.3694458, 1393.246826,
3842.096436, 12290.76074, 39264.39844, 3.076006174, 4.003799915, 4.031513691, 458, -99, 0,
1866.300049, 0.277559102
```

More information about the datasets and the meaning of each feature can be found in the data tab of the competition page [5] and more information about the problem, physics background, the design of the competition and evaluation procedures can be found in the research paper [6] provided in the resources section.

Solution Statement

A solution should be provided in the form of a dataset with a row for every event in the dataset and with two columns: `id` and `prediction`. The prediction should be a floating point value between 0 and 1.0, serving as the probability that the event whose ID is `id` is a $\tau \rightarrow 3\mu$ decay. The proposed solution should

first pass the pre-condition tests: agreement test [7] and correlation test [8]. I will be using the provided starter kit [9] to structure my solution as suggested by the guidelines.

As this is a supervised machine learning problem, I will explore a variety of supervised machine learning techniques and develop solutions using as many suitable classifiers as I can find and benchmark them against one another to discover the ones with most predictive power. I will look into SVMs, ensemble methods (random forests and boosting methods) [10], neural networks and others. Random forests, in particular, are quick to train, perform well, are simple to use and would therefore provide a good starting point. Because of the numerous features available, I plan on employing grid search and dimensionality reduction techniques optimising for variance.

I plan on using the information in the papers provided with the competition: Search for lepton flavour violating decay [11], New approaches for boosting uniformity [12] and the Flavours of Physics paper [13], particularly the sections "Geometric variables in LHCb detector" and "Data feature explanation" in order to understand the relationships between the variables.

Additionally, I will be performing my own independent research by looking up related papers and articles relating to similar LHC experiments, as well as experiments involving decays, in order to gain insights on the kind of information that is important for yielding accurate predictions. I will explore the existing solutions and the provided kernels [14] for additional understanding of how others approached this problem.

Benchmark Model

The best possible score that can be achieved for this competition is 1.0, or 100% accuracy in predicting the nature of collision events. I will be using the competition's leaderboard for benchmarking. There have been 3635 submissions from 673 teams to date, varying from 69.9% to 100% accuracy, with average of 97.2%.

Evaluation Metrics

The evaluation metrics are defined in the evaluation page [15] of the competition. The proposed evaluation metric is the Weighed Area Under the ROC [16] curve. The ROC curve is partitioned based on the True Positive Rate (TPR). The predictions must pass two additional checks before they are scored with weighed AUC, namely the agreement test and the correlation test (mentioned above).

Project Design

My initial approach will be to download the provided datasets and get accustomed to the variables by plotting them. My goal at this stage will be to uncover any patterns that become obvious when the information is presented visually, as well as to get more familiar with the relationship between different sets of variables. I will also explore Topological Data Analysis [17] for visualising the data using topological networks. Because the dataset has been prepared, I think minimal data cleanup will be needed, but I will explore potential ways of doing that at this stage. After that, I will scale the variables, so that they are all

treated as having equal importance to start with (they are all numeric, so no additional steps are needed). Feature selection will be the obvious next step, as there are 50 features to select from and undoubtedly some will provide more signal than others. I will look into the techniques provided by sklearn [18] and PCA for finding the principle components and reducing the dimensionality of the data. Finally, I will begin experimenting with different classifiers and measuring their performance.

References

1. https://www.kaggle.com/c/flavours-of-physics
2. https://en.wikipedia.org/wiki/Standard_Model
3. https://en.wikipedia.org/wiki/Flavour_(particle_physics)
4. http://lhcb-public.web.cern.ch/lhcb-public/
5. https://www.kaggle.com/c/flavours-of-physics/data
6. https://kaggle2.blob.core.windows.net/competitions/kaggle/4488/media/lhcb_description_official.pdf
7. https://www.kaggle.com/c/flavours-of-physics/details/agreement-test
8. https://www.kaggle.com/c/flavours-of-physics/details/correlation-test
9. https://www.kaggle.com/c/flavours-of-physics/details/starter-kit
10. http://scikit-learn.org/stable/modules/ensemble.html
11. https://arxiv.org/pdf/1409.8548.pdf
12. http://iopscience.iop.org/article/10.1088/1748-0221/10/03/T03002/pdf;jsessionid=5AAFC09D0FDA99759BB891442CE2373F.c2.iopscience.cld.iop.org
13. https://kaggle2.blob.core.windows.net/competitions/kaggle/4488/media/lhcb_description_official.pdf#page=10
14. https://www.kaggle.com/c/flavours-of-physics/kernels
15. https://www.kaggle.com/c/flavours-of-physics/details/evaluation
16. https://en.wikipedia.org/wiki/Receiver_operating_characteristic
17. https://www.ayasdi.com/blog/topology/topological-data-analysis-a-framework-for-machine-learning/
18. http://scikit-learn.org/stable/modules/feature_selection.html