Machine Learning system mimicking student's choice in Particle Data Analysis laboratory activity

V Wachirapusitanand, N Suwonjandee, B Asavapibhop and N Srimanobhas

Department of Physics, Faculty of Science, Chulalongkorn University 254 Phyathai Rd Wangmai Pathumwan Bangkok 10330

E-mail: vichayanun@hotmail.com, snarumon@gmail.com

Abstract. In Particle Data Analysis laboratory activity, aimed at undergraduate and high school students, the student is tasked with classifying collision events which contain two muons decaying from J/ψ meson. The activity provides 2000 collision events from the CMS detector, selected by CMS outreach community. However, classifying 2000 collision events by hand can be a tedious task for any human, so a smaller subset of collision events are usually used in the activity to save time. We built a machine learning classifier which mimic the student's classification based on a subset of collision events handed to the student, using some information from data in corresponding collision event. The information used in this system is parts of muon trajectory, extracted from files suited for CMS event viewer on the internet, as well as the four-momentum of both muons, available from the same source. With this system, students can input a subset of graded events into the system, and the system will be able to illustrate the results if the student worked on all 2000 collision events using his/her logic. Users can download the code from our repository and follow easy instructions to replicate this activity.

1. Introduction

The discoveries of new particles based on the Standard Model are made through rigorous experiments involving particle accelerators, particle detectors, huge amount of data outputted from the detectors, and thorough statistical analysis on the outputted data. The Large Hadron Collider (LHC), located at European Organisation for Nuclear Research (CERN) is the most famous particle accelerator, surrounded by particle detectors. The Compact Muon Solenoid (CMS) is one of them, detecting particles created by particle collisions within its chamber. Numerous particles are discovered in part by experiment data from CMS, most notably the Higgs particle, discovered in 2012.

To illustrate this rigorous process of particle discovery to high school and undergraduate students, CMS collaboration has approved data on 2000 collision events to be used for the purpose of education and outreach, now available in [1].The dataset contains events with two detected muons having an invariant mass in the range of 2 - 5 GeV. These collision events may contain a J/ψ meson, a neutral hadron containing a charm quark and an anticharm quark, with its mass of 3.1 GeV [2]. One of the possible decay channels for J/ψ mesons is into two oppositely charged muons.

An activity is set up to identify J/ψ meson in the dataset. A student working on this activity will be assigned to classify these collision events by looking at muon trajectories within the

detector one by one, grading collision events into 4 classes. These graded classes will determine whether or not the event contains the J/ψ meson candidates. However, since the goal of this activity is not to rediscover the J/ψ meson, but rather inform the student the process of data analysis used in high-energy physics, a smaller subset of the dataset, 500 events in this case, can be used instead.

In this project, we have devised a system based on machine learning to mimic the student's classification based on a subset of events graded by the student. The system is accompanied by some important information derived from each collision event. With this system, the student can input grades according to each event, and the system will be able to illustrate the results as if the student worked on the complete set of 2000 events using the same grading logic on the event subset.

2. Data preparation and extraction

In any collision event, two muons are visualised inside the CMS detector, with silicon trackers being the innermost component, detecting charged particles, and muon chambers being the outermost component, solely detecting muons. In the activity, students will look through each of the collision events in an event display, grading them into four classes: 0 being an event with two muons with the same charge, 1 being an event with two oppositely charged muons with their trajectory end points in the silicon tracker, 2 being an event with at least one out of two oppositely charged muons has its trajectory end point in the muon chamber, and 3 being an event with both oppositely charged muons have their trajectory end points in the muon chamber. From these conditions, we can expect that collision events given grade 3 have the highest chance to contain the J/ψ meson.

The data required in our system can be separated into two parts: the muon information in each event and the student's classification result. The muon information part consists of important parameters that can be extracted from an event data file used to visualise through an event display, and a corresponding entry in a spreadsheet containing four-vector information of each muon in each event. The student's classification result contains grades given by the student on each unique event.

Our system will condense the data from two parts and give an output text file containing the same number of events in the student's classification result, with corresponding parameters associated with each event. The parameters are event identification number, grade given by the student, four-vector of both muons, start and end points of the trajectory of each muon, and inverse of trajectory curvature at the start and end points of each muon.

The system will also have separated dataset extracted from the full set of 2000 collision events, containing all important parameters except the grades given by the student. To make mimicking student's decision more accurate, a cut-based discriminator is also built based on the inverse of curvature and the trajectory end points from both muons.

3. Features

Our system is based on Python source code and Jupyter Notebook, which can be used easily by the student. It can illustrate several exhibits based on each student's grading logic, giving different results for each different input from the students. The program, along with instructions, can be found at `github.com/vicha-w/jpsi-ml-classifier`.

3.1. Student confusion matrix

The first exhibit which can be generated by the program is the student confusion matrix, as shown in figure 1. The student confusion matrix shows the comparison of number of graded events separated by "truth grades" obtained from a cut-based classifier. Vertical axis represent grades given by student ("Graded") and horizontal axis represent grades given by cut-based

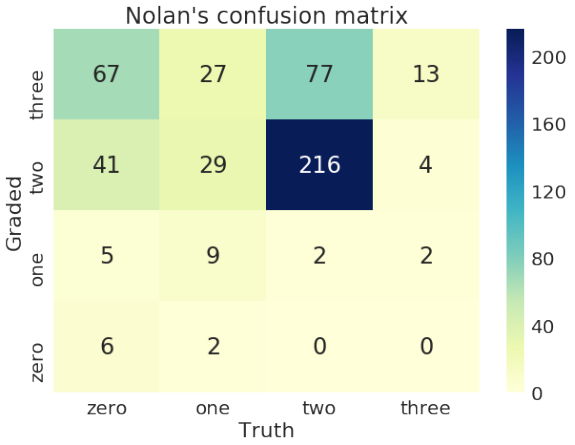


Figure 1. An example of student confusion matrix. The horizontal axis represents "Truth grades" obtained by the cut-based classifier, and the vertical axis represents grades obtained by the student.

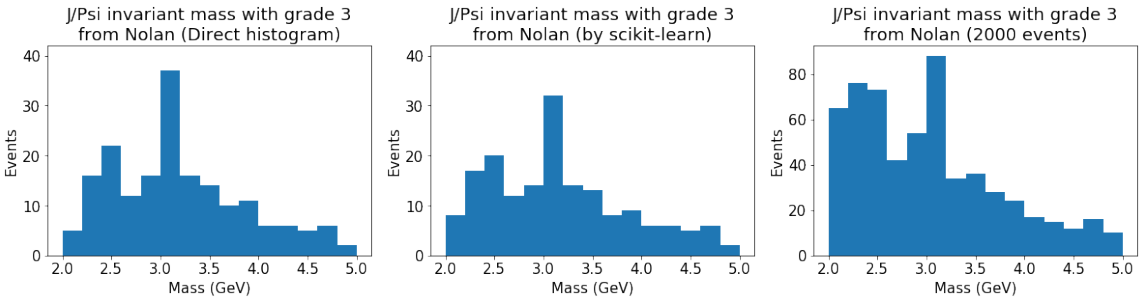
classifier ("Truth"). If the student classifies the events perfectly, the matrix will have only diagonal elements, showing all events have their grades in agreement with the cut-based classifier.

3.2. Machine Learning prediction on datasets

Our system contains four different random forest classifiers [3], a subset of machine learning classifiers, which are trained from events separated by their truth grades. Each classifier will receive and learn the relationship between each event and the decisions made by the student, and output into four different grades. This configuration gives us more prediction accuracy than by using one classifier alone. An important aspect is these classifiers are trying to mimic the student, not the cut-based discriminator, so their output can disagree with the cut-based discriminator itself, just like the student's grades. Based on this set of machine learning classifiers, we can derive a histogram showing J/ψ invariant mass calculated from events with grade 3 given by the machine learning classifiers, as shown in figure 2b.

3.3. J/ψ invariant mass histogram prediction

Following the training of machine learning classifiers described in section 3.2, we can let the classifiers predict the student's grading on the complete dataset of 2000 collision events, giving a histogram of J/ψ invariant mass with more statistics, as shown in figure 2c. The histogram, however, is mainly based on the student's grading decision, so if the student performed poorly on the classification task, the program will simply give worse histogram with less concludable information on the J/ψ meson invariant mass.



(a) Student-based, 500 events **(b)** ML-based, 500 events **(c)** ML-based, 2000 events

Figure 2. Invariant mass histograms based on different scenarios.

Table 1. Accuracy of trained classifier, separated by dataset.

	Test dataset				Train dataset				Test + Train dataset			
Student	0	1	2	3	0	1	2	3	0	1	2	3
A	0.88	0.96	0.97	1.00	0.99	1.00	1.00	1.00	0.94	0.98	0.99	1.00
B	1.00	0.98	0.99	1.00	1.00	0.99	1.00	1.00	1.00	0.99	1.00	1.00
C	0.80	0.93	0.95	1.00	0.99	1.00	1.00	1.00	0.92	0.97	0.98	1.00
D	0.83	0.89	0.97	1.00	0.99	0.97	0.99	1.00	0.92	0.94	0.98	1.00
E	0.81	0.81	0.89	0.88	0.96	1.00	0.99	1.00	0.90	0.93	0.95	0.95
F	0.88	0.90	0.92	1.00	1.00	0.99	1.00	1.00	0.95	0.95	0.97	1.00
G	0.95	1.00	0.97	1.00	0.99	1.00	1.00	1.00	0.98	1.00	0.99	1.00
H	0.81	1.00	0.96	1.00	0.99	1.00	0.99	1.00	0.92	1.00	0.98	1.00

4. Classifier training results

We have tested our system with grade inputs from 8 students. During the training phase, it is possible that there is only one event in which a certain grade given is different from cut-based grade, e.g. there is only one event in which the student gives grade 3 and the cut-based discriminator gives grade 0. In this extreme case, separating the data into test and training set is not allowed. To alleviate this, we doubled the dataset events to guarantee that the data can be separated into test and training dataset. After the dataset has been doubled, it will be separated into test and training dataset, with 40% of events retained (200 events) as test dataset. The accuracy for each classifier, tested on test and train dataset, from each of the student's input is shown in table 1, indicating that our system can mimic each individual student.

One important aspect is the fact that our system's goal is to mimic the decisions made by an individual student, not a collective of students, hence only the data from the student is relevant to the training of machine learning system each time the student uses it. Also, machine learning models such as this one will try its best to find the relationship hidden within the data it has been inputted as a training dataset. We can check whether the relationship is also applicable to the data it has not been trained before by determining the accuracy of a test dataset.

5. Conclusion

We have created a system which can mimic the grading decision by students while classifying collision events provided by CMS collaboration. With the use of machine learning techniques, the system can illustrate the results of discovery of J/ψ meson while allowing the student to work on a subset of collision events. The results illustrated by the system will mainly based on the grading quality from each unique student.

Acknowledgments

This research is funded by Chulalongkorn University, Government Budget, the Special Task Force for Activating Research (STAR), and "CUniverse" research promotion project by Chulalongkorn University (grant reference CUAASC).

References

[1] McCauley T 2014 Dimuon events with invariant mass rangge 2-5 GeV for public education and outreach *CERN Open Data Portal* DOI:10.7483/OPENDATA.CMS.SW96.PFX3

[2] Tanabashi M *et al* (Particle Data Group) 2018 *Phys. Rev.* D **98** 030001

[3] Pedregosa F *et al* 2016 Scikit-learn: Machine Learning in Python *J. Mach. Learn. Res.* **12** 2825