Machine Learning Project Proposal

Michael Z Reynolds — Spring 2021

March 29, 2021

1 Topic and Motivation

I will do a clustering, classification, and optimization study of the SUSY data found in the course reference article. More specifically, I will perform some clustering on the data's features to do some regression analysis. There are a total of eighteen features, but I won't need all of them, so part of the approach will be dimension reduction. I am interested in looking at the transverse momentum, pseudo-rapitidy, azimuthal angle, and mass of the particles. This data set is very close to the type of data I'll be working with, and analysis I'll be doing in my dissertation research. This is the primary motivating factor, as I believe that this project topic will give me a good foundation and initial exposure to working with this type of data.

The physics question will be to identify dilepton decays from heavier quarks such as the J/ψ and Υ mesons. These heavy particles provide interesting and vital information about the nature of hot, dense, nuclear matter such as the quark-gluon plasma. For example, in my dissertation research I will be analysing the elliptic flow of the J/ψ meson in proton-proton collisions at the CMS detector. The J/ψ meson is formed from a charm-anticharm quark pair, and the Υ is from a bottom-antibottom quark pair. Particles that we wish to analyze cannot be measured directly because they are produced immediately after the collision and have extremely short lifetimes before decaying into smaller, more stable particles. These final state particles are what actually hit the detectors, the signals from which we can reconstruct what happened in the collision and identify interesting phenomena. When doing an analysis, the particles of interest are the signal and everything else is background. So my analysis of this data will serve to identify and tag the dileptons decays of J/ψ and Υ mesons for further physics studies.

2 Data and Approach

This data set was produced from Monte Carlo simulations, by Baldi et al, specifically to apply machine learning classification to collision events. These events are identified through dileptions; that is to identify two leptons (electrons or muons) as originating from the same vertex, as opposed to background processes. As mentioned before, the data has 18 features; the first 8 of which are direct measurements pertaining to the dileptons themselves, and the rest are functions of the first eight; that is they are higher level functions that are convenient for conveying important physics information. However, I will not need these higher level features as they do not pertain to my particular physics question. In total, there are 5 million events. I will have to determine how many to use for training and testing. Since it's such a large filesize (2GB), I will have to pull small samples from the data to begin my analysis with a more manageable data size.

The important features are p_T (transverse momentum), m_{inv} (invariant mass), $|\eta|$ (pseudo-rapidity), and ϕ (azimuthal angle). The identify the dileptonic decays from the heavy mesons of interest, I will set thresholds on these parameters that qualify as potential matches. For example, the J/ψ has a mass of around 3 GeV, and the Υ as a mass of about 9 GeV. These masses inform the constraints I can put on dilepton pairs in their mass feature.

First I will have to visualize the data to decide which clustering method will be best: kmeans, dbscan, Gaussian mixture, or another method not from the class assignments. Then I will determine the best method of dimension reduction: variance thresholding (probably not), multidimensional scaling, principle component analysis, or stocastic neighbor embedding (tsne). It may be prudent to look at the way the data is clustering after dimension reduction, so I will do another visual pass before committing to a method. During this state of the analysis, I can also look at the entropy of the clusters as well as the normalized mutual information. During the clustering process, there will be tuning parameters (depending on which method I use), so I will have to optimize those parameters. This is done by running over a range of values and plotting.

Further analysis could include performing regression analysis of the cross-entropy loss functions, and MCMC methods. It will be hard to determine which methods will be best until I start looking at the data. Once I hone in on the best methods, I will bring in the entire data set, or at least a significant portion of it, for final analysis.

3 Timeline and Difficulty

I am requesting to give my presentation in the second round, April 26^{th}. I am taking my qualifying exam on Friday, April 23^{rd}. I would like my presentation for this project to be after my qualifying exam. In this case, I will have 4 weeks to complete my project. It will be a challenging endeavor given my level of python and data science knowledge, and it's possible that I may not be able to complete the debugging process to present a clean, working code. However, this is not my expectation and I will work hard to see the analysis to completion. In order to keep progress going smoothly, I will aim to complete the code for one of each technique (clustering, classification, optimization) each week, with the last week for debugging and final tweaking. This will be guided by weekly meetings with Prof Holmes to answer and specific questions I have about approach or structure.

4 Summary

This project will essentially be an application of many of the techniques learned in class onto the same data set. This will put all of these methods into the same context in which I will be able to explore and understand how these data analysis techniques work together. Understanding them separately through isolated homework assignments is good, but putting it all together into a cohesive framework will be much better for me to build a strong foundation in machine learning. I will address the physics question of classifying a handful of features to identify events that came from heavy quark meson decays, which then sets the data up for a variety of important physics analyses. This will take some time trying different methods until I find the best that work for my data. But in the end I should be able to present on the number and nature of heavy quark meson decays in this data set.