Semester Project Update #1

Describe in some detail the dataset(s) you intend to work with. What data elements exist, how are they structured, what features you hope to extract, etc. This is also the place to explain where the data came from, and any limitations on your use/sharing of the data or your work on the data.

This section includes your findings from initial data cleansing, exploratory statistics and visualizations, and additional information on reduced features selected through dimensionality reduction (if appropriate).

The dataset that we intend to work with is the Particle Tracking dataset from the Large Hadron Collider at CERN. Since this data was made publicly available by CERN for the TrackML Particle Tracking Challenge in 2018, there are no restrictions on our use or sharing of the data.

We are given a test dataset with 125 collision events upon which to evaluate our model; a training dataset, 8850 events split into five different files, as well as a sample set (the first 100 events from the training dataset); and some informational files about the geometry of the detectors. These are all provided by CERN in zipped .csv files, that can be downloaded and then read into a Pandas DataFrame for ease of access and manipulability. Additionally, CERN provides a trackml python library in GitHub to simplify some of the data handling, which will most likely be used extensively to aid in data visualization and processing.

Each event has four associated files, containing the hits, the hit cells, the particles, and the ground truth for the event. The hit files contain the identification numbers for the hit itself and the detector group/layer/module location of the hit, as well as the x-y-z coordinates for the hit. The truth files have the hit identification number, the particle identification number, the true hit location and particle momentum, and the weight of the hit (for the scoring metric). The particle files contain the particle ID number, the particle type, initial position and momentum, charge of the particle, and number of hits from this particle. The hit cell files contain the hit ID number, how much charge a particle has deposited on the cell, and the cell coordinates. The cells are the smallest positional identifier on the detectors, and can be used to more accurately track association between hits and particles.

The geometry of the detectors is important information to know because the detector is built from concentric silicon slabs that have been subdivided several times. The largest groups are volumes, subdivided into layers, which are then divided into modules, which are made up of cells. Each of those have ID numbers except the cells, which have a gridded identification system. Each module has a different local position and orientation, so a transformation must be made between the local coordinates of the hit and the global coordinates of the hit to get the actual path of the particle.