Project Proposal

Domain Background and Problem Statement

At the laboratory called CERN, we are doing research on proton collision to elucidate the origins of the universe.
However, in this proton study, a large amount of data is generated and physicists have to process about 10 petabytes of data per year. I clust data processed by this physicist, thereby reducing the amount of data to be processed and clarifying the origin of the universe.

`CERN official site`

https://sites.google.com/site/trackmlparticle/

`kaggle`

https://www.kaggle.com/c/trackml-particle-identification

Datasets and Inputs

A dataset comprises multiple independent events, where each event contains simulated measurements (essentially 3D points) of particles generated in a collision between proton bunches at the Large Hadron Collider at CERN.

https://www.kaggle.com/c/trackml-particle-identification/data

These data sets contain the following information:

- Event hits
- Event truth
- Event particles
- Event hit cells

Solution Statement

For this time, we can make prediction using DBSCAN which is a type of clustering algorithm.
In addition, we can measure and verify this prediction with Custom Metric.

Benchmark Model

The model with the Public Leaderboard score of 0.2078 will be used as benchmark model.

https://www.kaggle.com/mikhailhushchyn/dbscan-benchmark

Evaluation Metrics

The evaluation metric for this competition is a custom metric.
Each hit is weighted and the score is calculated as follows.
track is uniquely matched to particles by double majority rule. Match and sum the remaining tracks and make it a total of events.

https://www.kaggle.com/c/trackml-particle-identification#evaluation

https://github.com/LAL/trackml-library

Project Design

`work flow`

- input particle physics
- deep understanding of the problem
- EDA(understanding of datasets)
- Investigation of method
- Implementation and validation

input particle physics

I do not have any knowledge of particle physics at all, so I need to do inputs.

https://www.kaggle.com/c/trackml-particle-identification/discussion/55726

deep understanding of the problem

Based on that knowledge I will understand this issue again.

EDA

Discovering a data set has a major impact on model selection and feature engineering. I can get a lot of knowledge from EDA.

Investigation of method

This different color problem is different from regular Kaggle contest. Based on knowledge of the domain, it is necessary to investigate many methods. DBScan is one of the new learning methods.

Implementation and validation

After investigation, I implement and verify various methods and build final script.