Machine Learning Lunch Series Problem 2

Thomas Britton (tbritton@jlab.org) David Lawrence (davidl@jlab.org)

Sept 2019

1 Introduction

1.1 Setup

One common step in particle tracking is taking the state vector of a particle at a detector plane (the particles momentum and position) and projecting where that particle would be at another point in the detector. In practice, this process requires a good understanding of the magnetic field the particle is traversing (this leads to a bending of a charged particle) and calculating the energy a particle loses as it passes through detector material. GlueX, in Hall-D, consists of a cylindrical tracking detector made up of 24 planes of detecting wires sitting inside a magnetic field.

1.2 Goal

To design and train a model that, given a set of state vectors as inputs, can predict the state vector at the next measurement plane. A state vector consists of 6 values: x,y,z,p_x,p_y,p_z where x,y,z are the particle's position and p_x,p_y,p_z are its momentum at that position.

Ultimately, the model must accept a variable number of detector planes as inputs. e.g. One track may provide 9 state vectors and the model must predict the state vector at the 10th plane while the next track may provide 22 state vectors and the model must predict the state at the 23rd plane. Inputs will always start with the state vector at the target plane followed by some number of subsequent detector planes. Each detector plane being represented by a 6-parameter state vector. The detector planes provided for a given track will always be in order and there will be no gaps. The final input parameter for each track will be the z-position of the next plane (i.e. the plane for which a prediction must be made[1]).

state vector: the particle's momentum in the x, y, and z directions (p_x, p_y, p_z) and the particle's position (x, y, z) (a second given z). In other words, the final model will need to accept a 6-parameter state vector plus a 7th parameter representing the z-location to project to. The model will the return a prediction for the remaining 5 parameters of the state vector at that z location.

2 Materials

All materials will be available **here**[2]. Independent copies can be requested by emailing tbritton@jlab.org.

Input data will be a file consisting of rows of 150 comma-separated values made of 25 blocks of 6-parameter state vectors (x, y, z, p_x, p_y, p_z). The first block (i.e. values 1-6) represents the initial state vector at the target. The next block (i.e. values 7-12) represents the state vector at the first detector plane and so on for all 24 detector planes. Participants will be required to split the data in this one training file into their own training and validation sets. The input file contains a little less than 200k tracks.

The test set will be in the form of a CSV file with 6N+1 values on each line. The value of N will be different for each line having been randomly selected to be a value between 7 and 23 inclusive. (i.e. we

[1] Note that the z-position parameter in the inputs is somewhat redundant since z-positions of each plane are constant and the index of the next plane is implied by the number of state vectors supplied on the line.

[2] https://halldweb.jlab.org/talks/ML_lunch/Sep2019/

will provide between 7 and 23 detector planes of information). An example of such a file generated from the first 10k tracks of the training file can be found in the materials directory with the name "test_in.csv".

Please note that the training data is not in the same form that will be fed into the model, although it has similarities. It is recommended participants use this opportunity to learn about dealing with data in python (e.g pandas, dataframes). There are many good tutorials for reading files line by line in python and further guidance can be given if needed.
To aid in development a small file in the format of the test set will also be given. This should ensure that all participants will be able to develop a valid testing script which is able to read in the test dataset of the above format.

3 Judging Criteria

On October 30th at noon the test set will be released. Participants will have 24 hours from this time to make a submission. A submission will consist of all scripts used by participants to both train and test. The judges reserve the right to reproduce the results of any scripts; any non-conforming submissions (eg not based on ML) will be disqualified.
Each model will have the sum of the weighted RMS values for the difference from truth for all 5 parameters computed. These weights will be the expected RMS values so as to put each parameter on equal footing. The individual or team with the lowest total score will be declared the winner.

4 Prizes

Prizes will be mainly in the form of glory and (local) fame. A couple of gift cards will also be thrown in.