GNN fit tests

1 Introduction

A GNN has been used to generate predictions of the process underlying a series of ProtoDUNE events. Identification of the count of processes will be inferred by comparing the distribution of GNN scores from many events against templates. This is an alternative to event by event classification using the maximum GNN score.

These fits are performed using the iminuit python library, using a template likelihood method developed by Dembinski and Abdelmotteleb. To demonstrate this method is feasible, the robustness of the fit must be demonstrated to give confidence that sensible results will be returned when the method is applied to data.

This summary begins by describing robustness as the chosen goal, including the statistics that will be used to measure this. It then describes the methods available for testing the fit. Finally the tests to be carried out are listed.

2 Robustness

The Wikipedia introduction to Robust statistics states:

> Robust statistics seek to provide methods that emulate popular statistical methods, but are not unduly affected by outliers or other small departures from model assumptions.

It is important to note that this definition applies to "statstic"s. This refers to a statistic, N, as the result of a function applied to a set of measurements $X_1, X_2, \ldots, X_N$. I.e. T is a statistic if $N = N(X_1, X_2, \ldots, X_N)$.

2.1 Template fit robustness

The statistics for which we which to demonstrate robustness are:

$$N_1, N_2, N_3, N_4 = F(X_1, X_2, \ldots, X_N; T_1, T_2, \ldots, T_N) \tag{1}$$

The statistics of interest N_1, N_2, N_3, N_4 correspond to the number of absorption, charge exchange, single pion production, and multiple pion production events which generated the data $X_1, X_2, \ldots, X_N$. The function F represents the template fit function used. The template data $T_1, T_2, \ldots, T_N$ is shown explicitly as a separate input to the fit.

Following the definition of robustness from section 2, our tests should ensure that the fit has minimal dependence on the underlying assumptions, or outliers. In the case of a template fit, the templates themselves represent the underlying assumptions of the fit. To conclude the fit is robust, we thus require "the fit achieves the same performance over a range of modifications to the template".

The consideration above leads itself to the testing paradigm of explicitly editing the template (changing the underlying assumptions), and confirming the output statistics remain consistent. An exception to this is the outlier test, which must involve injection of outlier events into the data set.

Adjustment the statistics by considering not N_i, but rather $N_i - N_i^{\text{true}}$ allows us to make small variations in the true number of data events. This new statistic shall be called the offset statistic. According to the paradigm, which should keep the underlying data distribution the same, yet this new statistic allows sampling this data distribution whilst keeping the statistic comparable between samples. The benefit is to allow sampling the statistic many times to find a distribution of robustness estimators to improve understanding.

A final benefit of this paradigm is to explicitly separate from a method for estimating systematic uncertainties, which can be assessed by making changes to the underlying *data* distribution, without changing the template.

2.2 Robustness measurements

In the paradigm described in section 2.1, the test of robustness is that the centre of the distribution of offset statistics remains consistent as some underlying assumption is changed. This can be quantified by taking a gradient of a plot of the offset statistic values vs. the magnitude of change of the assumption. No decision has yet been made on what limit this gradient should have to be considered robust.

To improve understanding, additional measurements will be created to offer alternative view of the performance of the fit.

- A pull statistic will be tested by dividing the offset statistic with the corresponding error output by the fit.
- The χ^2 fit value between the fitted templates and the data will be returned to indicate the overall performance of the fit.
- The ratio of likelihoods from the template fit for the true data distribution over the minimum likelihood.
- The ratio of likelihoods from the template fit for the template data fractions (normalised to the number of data events) over the minimum likelihood. Note that this value *should not* be robust. If it is robust, it implies the fit is tracking the template (i.e. underlying assumptions).

3 Tools

The core tool available is a template/data generator. This is a class implemented in python which is supplied a set of GNN predictions and corresponding true regions for each of the template and data. The created instance allows for sampling and weighting of the parameters as defined by some sampling parameters. The instance begins by creating a 4D correlated histogram of true counts $T_{i,j}$. If the true counts are known, this information is kept, indexed as i. Index j runs over the (4D) with a supplied number of bins per dimension, b. Thus index j runs from 1 to b^4.

Sampling probabilities, p_i and weights, w_i may be passed for all events, or for each region individually. These represent the expected fraction/weighting of events, respectively. Both of these parameters may be distributed or fixed.

If distributed, the sampling probability is used in a binomial distribution to drawn the number of events used. The number of events $N_{i,j}$ used from process

i in bin j, given a true binned count $T_{i,j}$: $N_{i,j}$ Binom$(T_{i,j}, p_i)$. If undistributed, this is simply $T_{i,j}$: $N_{i,j} = T_{i,j} \times p_i$.

If distributed, the weights are generated by a gamma function with shape $N_{i,j} \times w_i$, noting that the sum of gamma distributions is also gamma distributed. Thus the count in a bin $C_{i,j}$, if distributed, is: $C_{i,j}$ Gamma$(w_i N_{i,j}, 1)$. If undistributed, this is simply $C_{i,j} = N_{i,j} \times w_i$.

To generate uncorrelated data, four 1D histograms are generated by summing over all by one of the 4D histogram's axis, with the axis not summed over as the axis is interest.

4 Tests

As outlined in section 2.1, the template represents the underlying assumptions. The tests should cover a range of possible discrepancies that may appear when considering real ProtoDUNE data.

The following potential discrepancies are considered:

- The template contains a different fraction of true processes than exist in data.
- The GNN score distributions in the template do not follow the distributions found in data (e.g. if the absorption scores in data are more likely to be lower than those in the template).
- Outliers do not dramatically affect the results.

The following list aims to test all the potential discrepancies:

1. Random fluctuation - randomly sample the templates, but without systematic changes.
2. Initial fit predictions - change the initial parameter estimations going into the fit.
3. Re-weighted process fractions - change the relative amount of underlying processes in the template.
4. GNN score drift - apply some weighting on the distribution of scores as a function of GNN score bin number, e.g. a linear increase in weighting from the lowest to highest bin.
5. Outliers - intentionally add extra entries into the data without similar template predictions.
6. Minimum required statistics - reducing the fraction of data in the fit until failure.