Noise Classification: A Feasibility Study

Aryaman Jeendgar (BITS Pilani/Princeton University)
Supervisor: Dr. Kilian Lieret (Princeton University)

February 23, 2024

Introduction

- A Learned Clustering based pipeline for charged particle tracking
- Won't go into the details of the various components of the pipeline itself — higher-level description below

Figure: The Tracking Problem

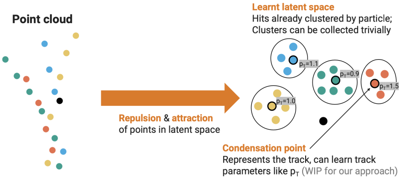
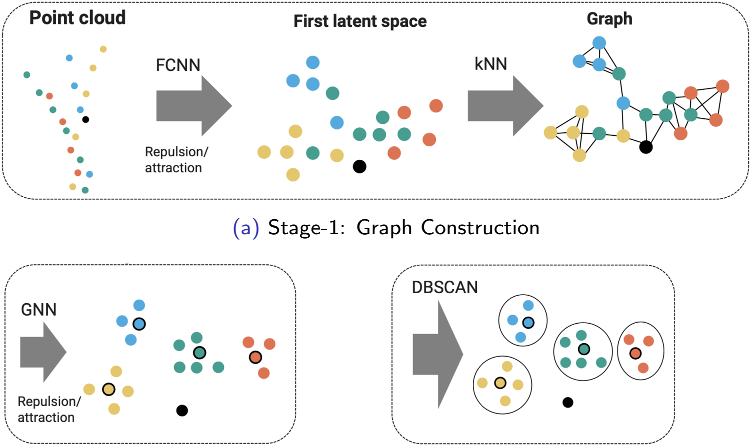


Figure: Object Condensation

The entire pipeline



(a) Stage-1: Graph Construction

(b) Stage-2: Object Condensation phase

(c) Stage-3: Collect Clusters

Figure: The entire pipeline at a glance

The Goal

- Disclaimer The entire discussion only presents results for the data recorded in the pixel detector — a deeper study for architectures for the data of the full-detector is next on my agenda
- There are noisy hits in the point cloud data i.e. detector signals that aren't due to particles from the collision
- The goal of the study was to experimentally verify the *potential* gain in performance via the pre-emptive removal of noisy hits from the dataset (before too much time was spent in the design of such a classifier)
- The results are quite promising!

The metrics

We compare the following metrics across both of the runs:

- Perfect match efficiency ($\epsilon^{\text{perfect}}$): The number of reconstructed tracks that include all hits of the matched particle and no other hits, normalized to the number of particles.
- LHC-style match efficiency (ϵ^{LHC}): The fraction of reconstructed tracks in which 75% of the hits belong to the same particle, normalized to the number of reconstructed tracks.
- Double Majority match efficiency (ϵ^{DM}): The fraction of reconstructed tracks in which at least 50% of the hits belong to one particle and this particle has less than 50% of its hits outside of the reconstructed track, normalized to the number of particles.
- Variants of each of these quantities for particles of $p_T > c\text{GeV}$ are denoted as: $\epsilon_{p_T > c}^{\{\text{DM, perfect, LHC}\}}$
- Total Validation Loss

Results

Aryaman Jeendgar (BITS Pilani/Princeton University)
Supervisor: Dr. Kilian Lieret (Princeton University)

	$\epsilon^{perfect}_{p_T>0.9}$	$\epsilon^{DM}_{p_T>0.9}$	$\epsilon^{LHC}_{p_T>0.9}$
NC	0.847145	0.9660603	0.978677
Vanilla	0.757657	0.939322	0.975234

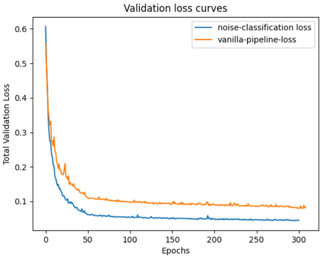


Figure: Validation loss curves

Results

Aryaman Jeendgar (BITS Pilani/Princeton University)
Supervisor: Dr. Kilian Lieret (Princeton University)

- A preliminary noise classification round produces *noticeably* better results!
- It's a binary classification problem — an appropriate choice of model still remains, but for now, it can be something simple like an `XGBoost` model or an `FCNN`
- The more important caveat is that we want our classifier to avoid false positives at *all* costs (i.e. non-noise hits being labelled as noise)
- What is the solution? Uncertainty Quantification provides a possible way out. . .

UQ: Conformal Scores

- Conformal Prediction is a straightforward way to generate prediction sets for *any* model
- Begin with a fitted model, $\hat{f}$ — generate prediction sets for this model through a small amount of data (*calibration data*)
- Conformal prediction seeks to construct a prediction set, $\mathcal{C}(X_{\text{test}}) \subset \{1, \dots K\}$ using $\hat{f}$ and the calibration data, $(X_1, Y_2), \dots (X_n, Y_n)$

$$1 - \alpha \leq \mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}})\right) \leq 1 - \alpha + \frac{1}{n+1}$$

Here, $(X_{\text{test}}, Y_{\text{test}})$ is a fresh test point from the same distribution and $\alpha \in [0, 1]$ is a user-chosen error rate

UQ: Conformal Scores

- Can be seen as a general procedure for converting a heuristic notion of uncertainty from any model and converting it to a rigorous one

Figure: Conformal Prediction

- The process:
  - Identify a heuristic notion of uncertainty using the pre-trained model
  - Define the score function $s(x, y) \in R$ (a larger score should encode a worse agreement between (x, y))
  - Compute $\hat{q}$ as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the calibration scores $(s_i = s(X_i, Y_i))$ (essentially the $(1-\alpha)$ -th quantile but with a small correction)
  - Use this quantile to form the prediction sets for new examples:

$$\mathcal{C} = \{y : s(X_{\text{test}}, y) \leq \hat{q}\}$$

UQ: Conformal Scores

- For us, the above process means that we can now quantify our 'risk-appetite' via α (and ideally perform a sweep to check for what offers best performance)
- The choice of a 'good' conformal score is a matter of design — a simple one: $s_i = 1 - \hat{f}(X_i)_{Y_i}$ (the score is large when the softmax output of the model is low, i.e. when it is *very* wrong).
- Also super straightforward to implement!

```
# 1: get conformal scores.
n=calib_Y.shape[0]
cal_smx=model(calib_X).softmax(dim=1).numpy()
# 2: get adjusted quantile
cal_scores=1-cal_smx[np.arange(n),cal_labels]
q_level=np.ceil((n+1)*(1-alpha))/n
qhat=np.quantile(cal_scores, q_level, method='higher')
val_smx=model(val_X).softmax(dim=1).numpy()
# 3: form prediction sets
prediction_sets=val_smx >= (1-qhat)
```

Other possible leads

Aryaman Jeendgar (BITS Pilani/Princeton University) Supervisor: Dr. Kilian Lieret (Princeton University)

- Quantile Regression is also another really powerful technique in the UQ toolkit — learn multiple quantiles over your model's output
- Combine the two? Conformalized Quantile Regression — ~probably overkill for our simple application
- I have some prior work in the construction of loss functions that can perform quantile regression for binary classification problems.

Summary

- Performing a noise classification preprocessing step in the pipeline produced a *marked* improvement in the evaluation metrics
- Noise classification is a simple binary classifcation problem but the context of charged particle tracking requires special focus on avoiding false positives
- Uncertainty Quantification and Conformal Prediction in particluar is a very powerful tool (and is also easy to implement) in being able to make more robust and interpretable decisions over the predictions of the model.

References

[1]: A N. Angelopoulos and S. Bates, *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification*, 2022
[2]: K. Lieret et.al. *High Pileup Particle Tracking with Object Condensation*, 2023
[3]: K. Lieret and G. DeZoort *An Object Condensation Pipeline for Charged Particle Tracking at the High Luminosity LHC*, 2023
[4]: Y. Romano et.al., *Conformalized Quantile Regression*, 2019
[5]: A. Jeendgar et.al., *LogGENE: A smooth alternative to check loss for Deep Healthcare Inference Tasks*, 2022