A Track Identifier for OPERA and SHiP: Solution from MLHEP Summer School

Benda Xu

IPMU, U. Tokyo

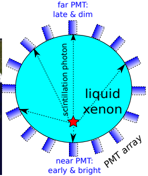
Nov. 9, 2017

Self Introduction

- Experienced in underground neutrino and dark matter experiments.

  KamLAND *Kamioka liquid scintillation anti-neutrino detector*

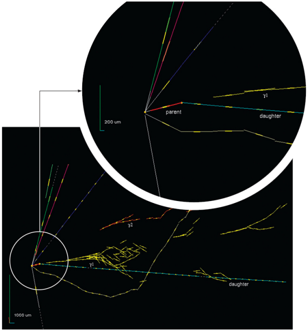
  XMASS *Xenon detector for weakly interacting massive particles*



- Also interested: statistics, machine learning, programming and computing.
  - Gentoo Linux developer on science and high perforance computing.
  - MLHEP: a learning platform I was looking for.
  - This talk grew out of the in-class contest at MLHEP.
    - ★ MLHEP:= machine learning in high energy physics summer school

Motivation: From OPERA to SHiP

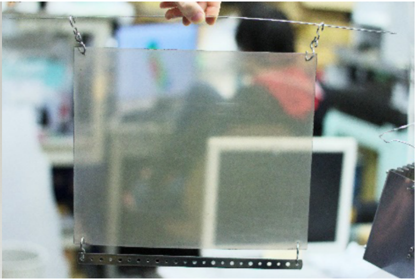
- $\nu_\mu \to \nu_\tau$ appearance observed by OPERA with excellent resolution of event topology and particle identification.
- Made possible by emulsion cloud chamber (ECC).



- SHiP is designed with a neutrino detector similar to OPERA.
  - Dark matter/hidden sector, sterile neutrino/heavy neutral lepton, ν_τ physics, lepton flavor violation. $\to$ produce energetic electrons and Electromagnetic shower by scatter or decay.
  - Recontruction of EM shower is the starting point of all the above exciting physics targets.
- Challenge: 100$\times$ more ν_τ events than OPERA.
  - Automate EM identification: explore machine learning.

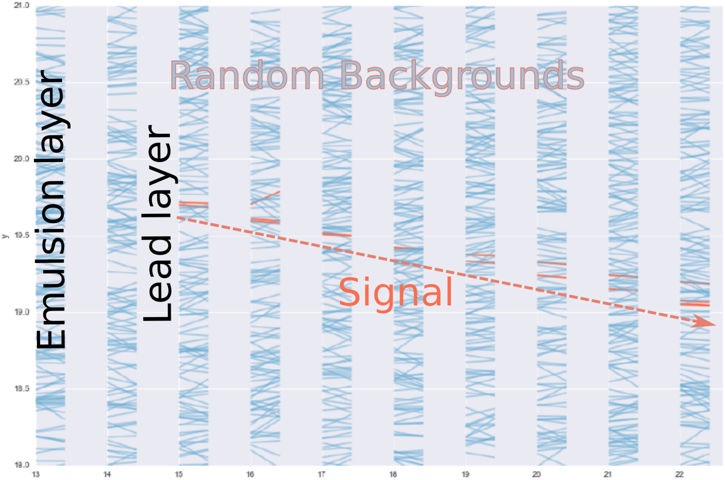
Outreach of Emulsion Technology

- Common for liquid scintillation and emulsion: legacy of industry.
- Photomultiplier tubes: vacuum tubes in the age of integrated chips.



- Emulsion: films in the age of charge-coupled device(CCD).
  - muon tomography: imaging of volcano, pyramid, etc.
- Drive them with machine learning in the age of information!

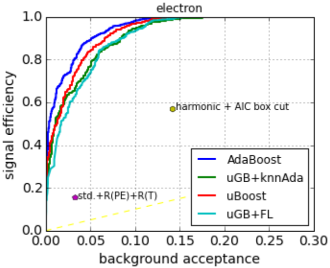
- Focusing on emulsion cloud chamber (ECC), ignoring other parts.

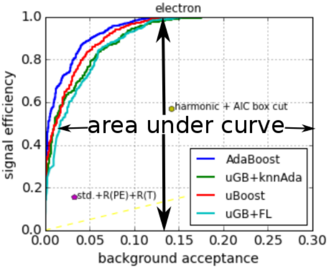


- Base track reconstructed in each emulsion layer: X, Y, Z, TX, TY, χ^2.

Terminology (II)

- ROC: receiver operating characteristic
  - In physics: signal efficiency vs. background acceptance
  - Compare parameters (curves) and cuts (points).





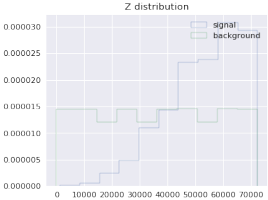
- AUC: Area under ROC curve, a performance measure.
- Contest: classification of base tracks scored by AUC.
  - Physicists also care about energy resolution, etc. Using AUC as figure-of-merit is an adaptation to Kaggle and ML community.

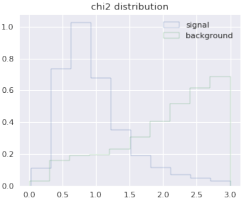
https://goo.gl/8N7BYG

https://www.kaggle.com/c/dark-matter-signal-search-episode-1

Baseline Solution – AUC 0.930

1. An eletromagnetic shower should be contained in the block to be useful: X, Y near the center, TX, TY near 0.
2. An eletromagnetic shower develops more base tracks with time: signals tend to have high-Z.



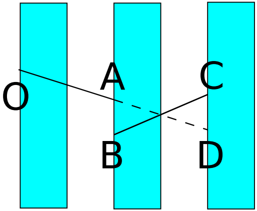


3. Signal base trackes has smaller χ^2. AUC: ~ 0.845
4. Baseline solution combining all of them: AUC: 0.930

Ideas for Exploiting Properties of Signals

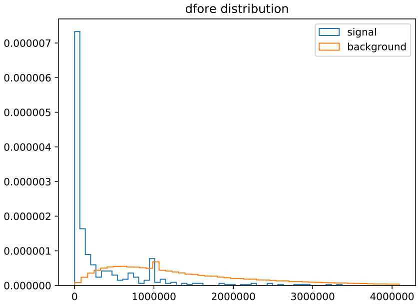
- Base tracks are connected
- EM shower base-track clusters:
    - in 3D
    - in hough space

Parameter: Base Track Connection



- For each base track OA.
  1. Select the closest downstream basetrack BC by $d^2 = \|AB\|^2 + \|CD\|^2$. Use the d^2 as a feature.
  2. Do the similar with upstream basetracks.
- Most of the signal base track have a near neighbour.

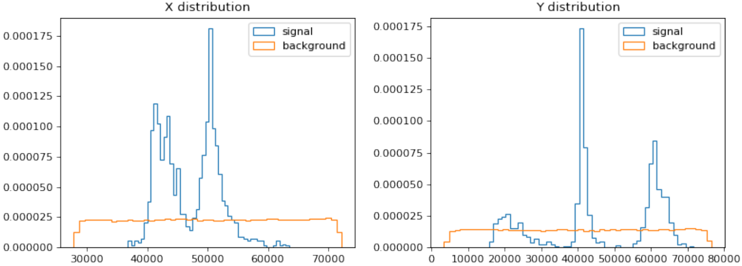
Distribution: Base Track Connection – AUC 0.995



- Signals are concentrated towards 0.
- Some layer might miss the signal base track.
  - Extension: jump 2 layers. AUC: $0.930 \to 0.993$
  - ... and up to 6 layers. AUC: $0.993 \to 0.995$

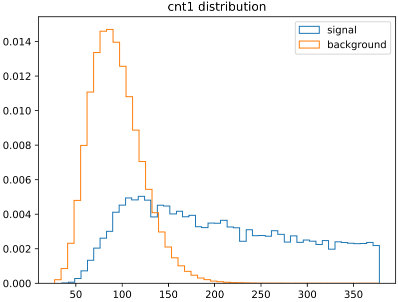
Parameter: Histogram in 3D

- Observation: EM shower base tracks tend to cluster.



- Histogram the tracks based on X, Y, Z.
- Assign each track the cell count of the histogram.

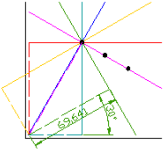
Distribution: Histogram in 3D – AUC 0.997



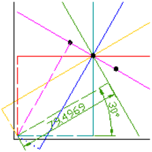
- Signals have larger bin counts.
- Convolute with a smoothing kernel $(1, 3, 1)^3$ $0 - -6$ times.
- AUC: $0.995 \to 0.997$

Hough Space

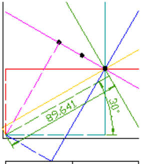
- Rational: 3D histogram ignores line nature of signals. Cure with hough transform.
- Vote (read: histogramming) for lines from a point cloud.



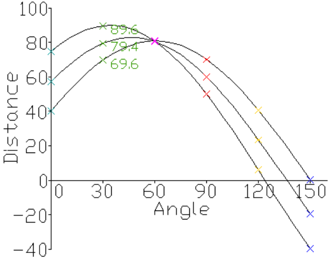
Angle	Dist.
0	40
30	69.6
60	81.2
90	70
120	40.6
150	0.4



Angle	Dist.
0	57.1
30	79.5
60	80.5
90	60
120	23.4
150	−19.5



Angle	Dist.
0	74.6
30	89.6
60	80.6
90	50
120	6.0
150	−39.6

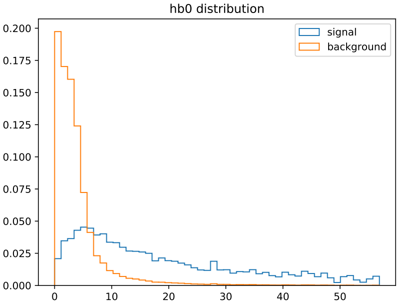


- This should work better than 3D histograms.

Parameter: Histogram in Hough Space

- A line in 3D space has degree-of-freedom 4.
  - 2 points? DOF 6.
  - A point and a unit vector? DOF 5. Degeneracy in that the point can be anywhere on the line.
  - Fix the point on Z=0 plane? Aesthetically ugly. Depending on a certain coordinate system.
- Need for a 3D parameterization suitable for ECC.
  K.S. Roberts (1988), *A new representation for a line*
- Make histograms in the Hough-Roberts space.

Distribution: Histogram in Hough Space – AUC 0.998



- Signals have larger bin counts.
- AUC: $0.997 \to 0.998$

Hyperparameter Tuning

- Tune the boost hyperparameters of extreme gradient boosting.
  - AUC: $0.99835 \to 0.99840$
  - AUC: $0.99842 \to 0.99845$ (another sample)

Concluding Remarks

- Add more parameters to develop better cuts.
- Use boosting to combine all the parameters.
- Finished in 30 hours – a very intense practice on programming and data processing.
  - Good tools: Emacs + Jupyter
  - $\to$ emacs-ipython-notebook (aka. EIN Is not only for Notebooks).
  - Jupyter Read–Eval–Print Loop(REPL) from within Emacs.
    - ★ Python, R, Julia, ROOT Cling, etc.

Prospect: 3D Convolution Neuron Network?

- We can always invent new parameters that might improve performance. Where do we stop?
  1. When the physics goal is achieved.
  2. When the researcher gets exhausted.
  3. or Let the machine invent new parameters!
- Day 4 Lecture 2, by Maxim Borisyak, Alexander Panin, Andrey Ustyuzhanin

Why bother?

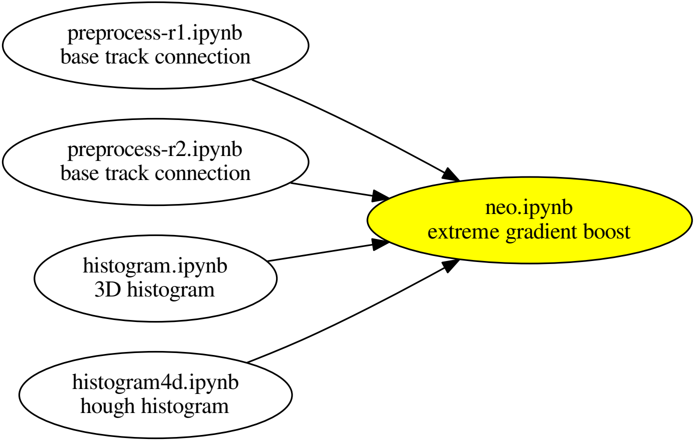
We study methods for spatial data

- 1D: Time-series, spectrograms
- 2D: Calorimeters; triggers at LHC
- 3D: Hits; tracks; events
- >3D: Spatio-temporal; descriptor manifolds

- Ways to guide the (local?) convergence of neuron network with feature engineering?

Sample Programs

https://github.com/heroxbd/mlhep2017/tree/master/episode2



Acknowledgements

- MLHEP 2017: Andrey Ustyuzhanin, Ulrik Egede, and our lecturers.
- Miha Zgubic and our classmates at MLHEP.
- OPERA collaboration for sharing data with us,
  - for pushing emulsion technology to this ultimate form of art.
  - for discovering ν_τ oscillation appearance.
- SHiP collaboration for this kind invitation.