Logistic regression method for particle identification in MPD experiment

Danila A. Starikov

Peoples' Friendship University of Russia,
starikov_da@pfur.ru

GRID'2025, 7–11 July, Dubna

Goals

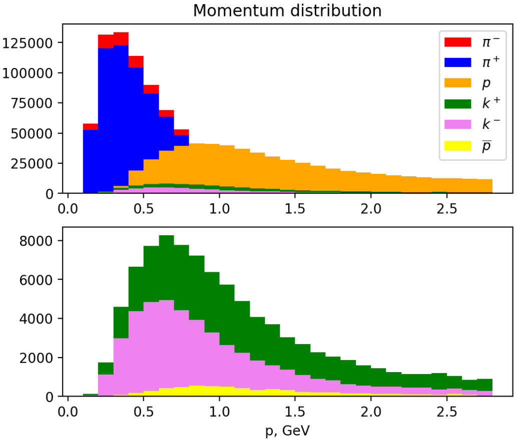
- Apply logistic regression method for particle identification problem
- Compare classification efficiency against XGBoost and N-sigma methods
- Investigate feature importance, using l_1-regularization
- Train the models on dataset with fewer features and compare results

Model data

- Dataset acquired with MPDRoot package
- 6 particle types

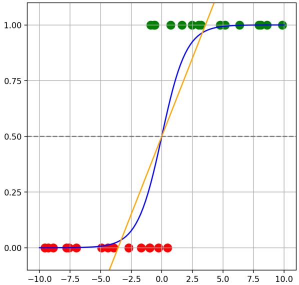
π^+	778645
π^-	851541
k^+	91423
k^-	46950
p	594156
$\bar{p}$	6357
Σ	2369072

- 14 features: **p**, **charge**, **dedx**, **m2**, **nHits**, **eta**, **dca**, **Vx**, **Vy**, **Vz**, **phi**, **theta**, **gPt**, **beta**.



Binary logistic regression

- Predicts the probability of a given data point corresponding to label 0 or 1:



$$\hat{y} = \begin{cases} 0, \hat{p} < 0.5 \\ 1, \hat{p} > 0.5 \end{cases}$$

$$\hat{p} = \sigma\left(\mathbf{x}^T \boldsymbol{\theta}\right), \sigma(t) = \frac{1}{1 + \exp(-t)},$$

- Loss function:

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Big(- y_i \ln(\hat{p}(\mathbf{x}_i)) - (1 - y_i) \ln(1 - \hat{p}(\mathbf{x}_i)) \Big) + \lambda R(\boldsymbol{\theta}),$$

где $R(\boldsymbol{\theta})$ – regularization term, λ – regularization parameter

Multinomial logistic regression

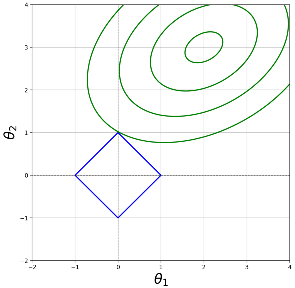
- We use softmax regression:
$$\hat{y} = \arg\max_i \hat{p}_i,$$
$$\hat{p}_i = \frac{\exp(\mathbf{x}^T \boldsymbol{\theta}_i)}{\sum_{j=1}^K \exp(\mathbf{x}^T \boldsymbol{\theta}_j)}, i = 1, 2, ..., K$$
- Each label has its own weights vector $\boldsymbol{\theta}_i$, so the model is described by weights matrix $\boldsymbol{\Theta}$
- Loss function:
$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} [y_i = k] \ln(\hat{p}_k(\mathbf{x}_i)) + \lambda R(\boldsymbol{\Theta})$$

Regularization

- l_1-regularization

$$R(\boldsymbol{\theta}) = ||\boldsymbol{\theta}||_1 = \sum_i |\theta_i|$$

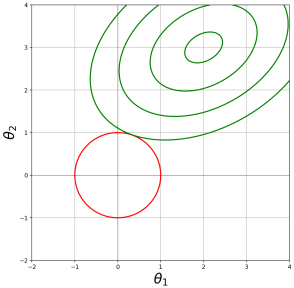
$$R(\boldsymbol{\Theta}) = ||\boldsymbol{\Theta}||_{1,1} = \sum_i \sum_j |\theta_{ij}|$$



- l_2-regularization

$$R(\boldsymbol{\theta}) = \frac{1}{2}||\boldsymbol{\theta}||_2^2 = \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta}$$

$$R(\boldsymbol{\Theta}) = \frac{1}{2}\|\boldsymbol{\Theta}\|_F^2 = \frac{1}{2}\sum_i \sum_j \theta_{ij}^2$$



Evaluation of model classification results

- $E = \dfrac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$ – efficiency,

- $C = \dfrac{\text{False Positive}}{\text{True Positive} + \text{False Positive}}$ – contamination,

Data preprocessing

- All features scaled into range $[0, 1]$
- Particles and antiparticles merged into bigger classes:

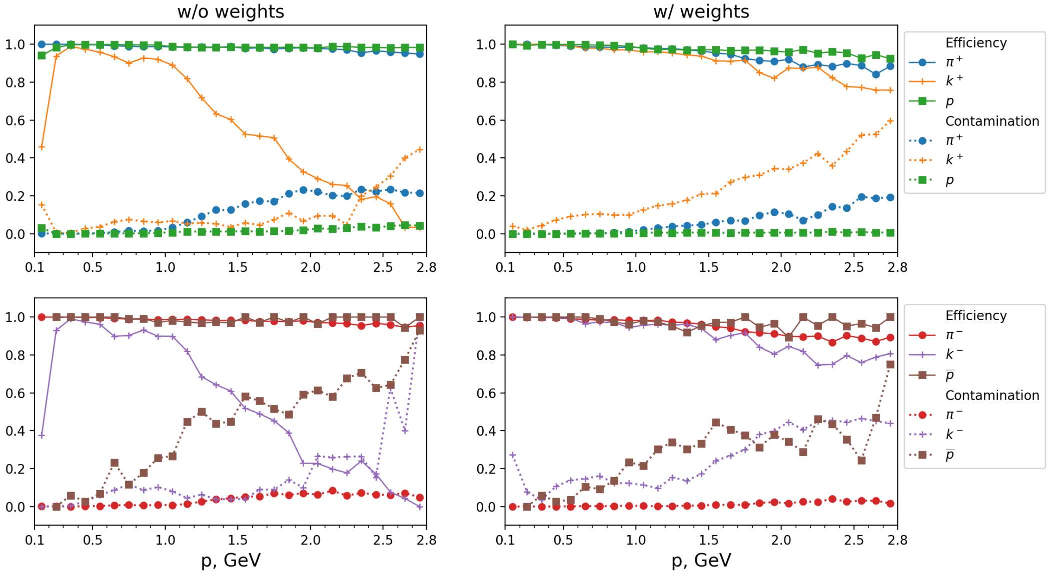
π^+	778645
π^-	851541
k^+	91423
k^-	46950
p	594156
$\bar{p}$	6357
Σ	2369072

$\Longrightarrow$

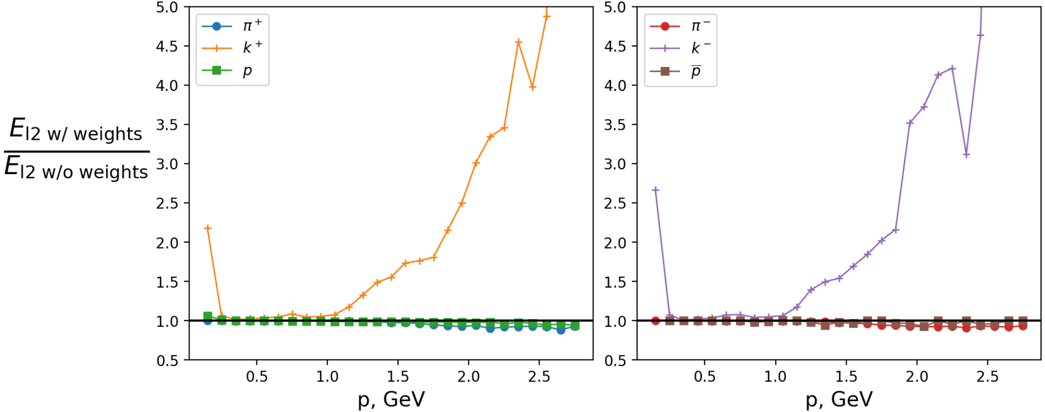
π	1630186
k	138373
p	600513
Σ	2369072

- Feature **charge** is excluded from training data, feature set is reduced from 14 to 13. Classification by **charge** is conducted separately
- Dataset is split into bins by 0.1 GeV (from 0.1 GeV to 2.8 GeV, 27 bins in total) and a separate model is trained in each bin

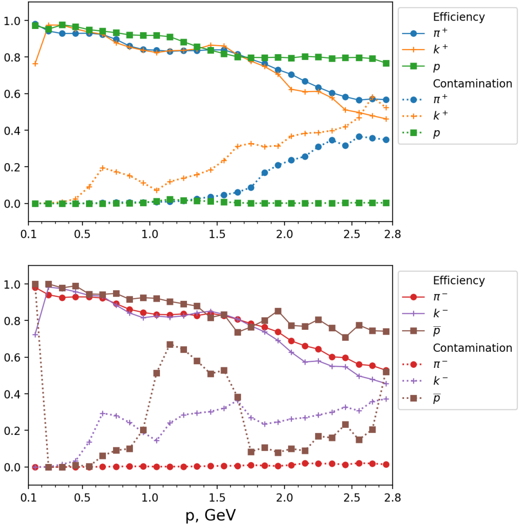
Results: l_2-regularization

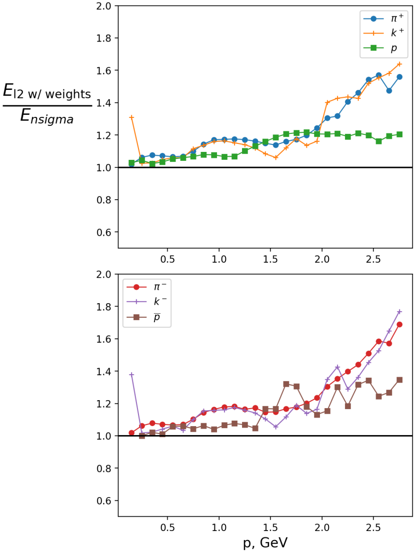


Results: l_2-regularization

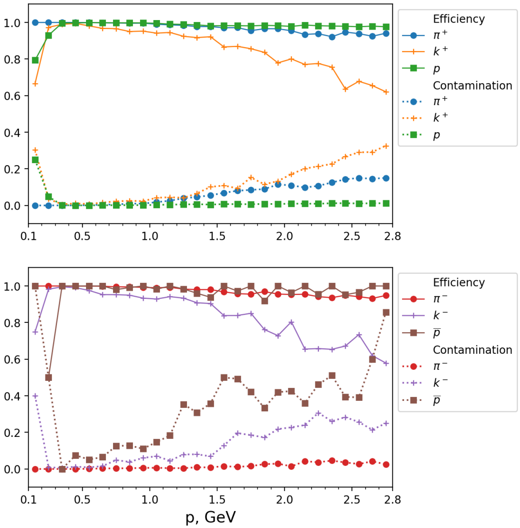


Comparison with N-sigma

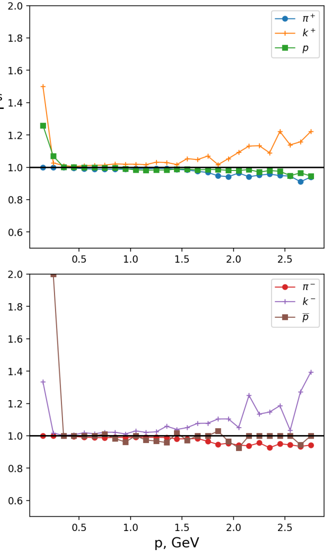




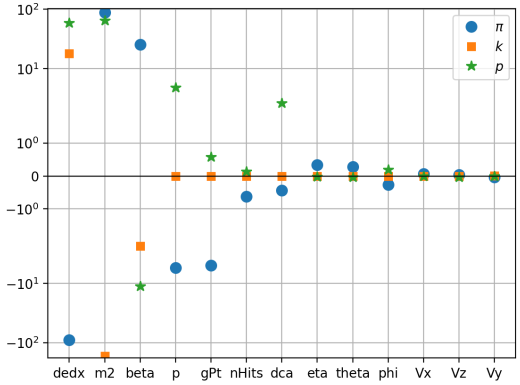
Comparison with XGBoost

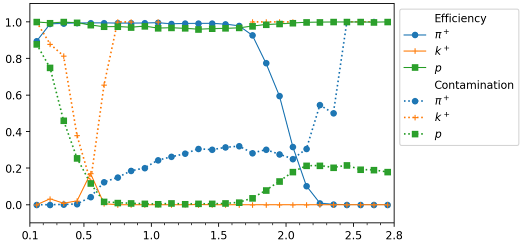


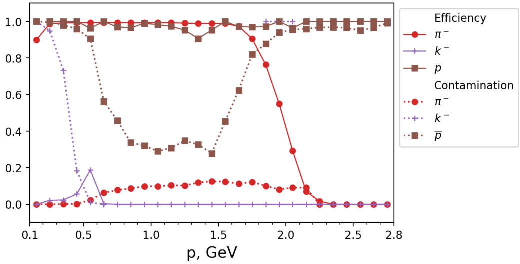
$$\frac{E_{\text{l2 w/ weights}}}{E_{XGBoost}}$$



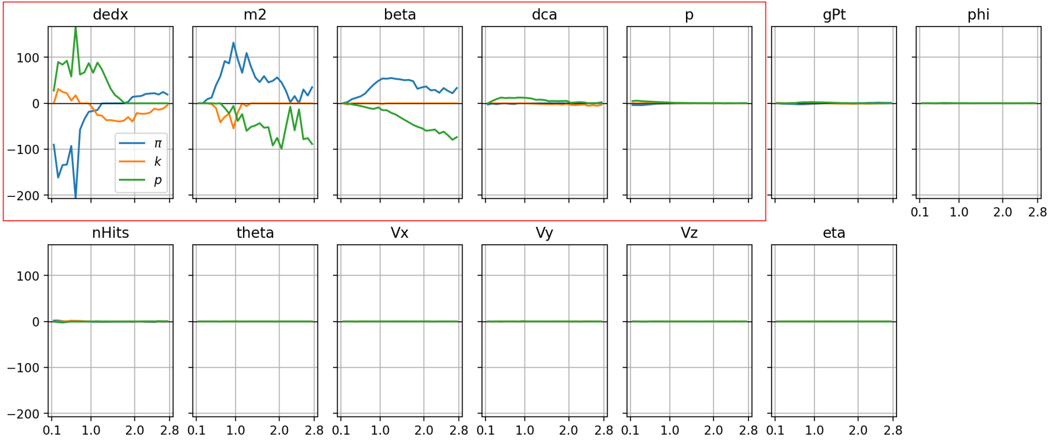
Feature importance investigation: integral case





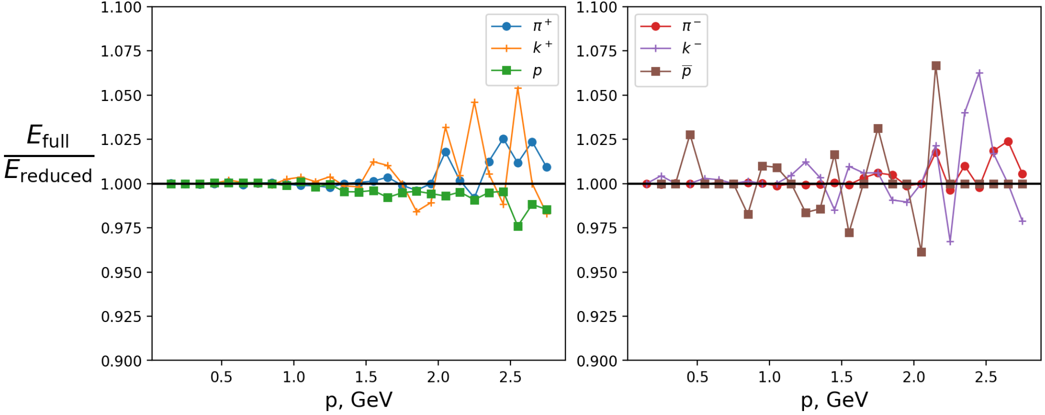


Feature importance investigation: bin-split case

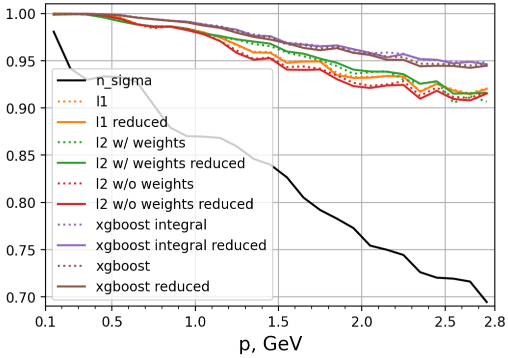


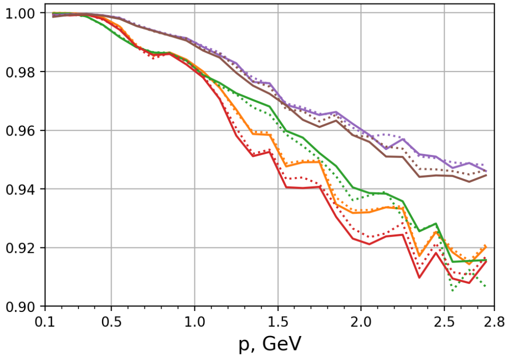
Results on reduced data: l_2-regularization

Reduced dataset contains only 6 features: **p**, **charge**, **dedx**, **m2**, **dca**, **beta**



Comparison of total efficiency





	l1	l2 w/ weights	l2 w/o weights	xgboost integral	xgboost	n-sigma
full	0.9822	0.9824	0.9804	0.9899	0.9893	0.8926
reduced	0.9821	0.9830	0.9798	0.9897	0.9888	

Conclusion

- For the first time the logistic regression method was used for the particle identification problem
- Logistic regression method compared against logistic regression method with the standard N-Sigma method of the MPDRoot package and the previously studied XGBoost model:
  - Works better than N-Sigma method
  - But loses to XGBoost model across all momentum range
- Feature importance analysis was conducted by introducing l_1-regularization:
  - Attributes **dedx**, **m2**, **beta** are significant over the entire range of moments
  - Feature weights **Vx**, **Vy**, **Vz**, **nHits**, **eta**, **phi**, **theta** were zeroed during model training
- Reducing dataset by dropping the least important features didn't impact models' prediction ability, *as expected*

Data processing, model training and results analysis were done on HybriLIT platform