Machine Learning for Data Quality Monitoring at CMS Experiment

Mridupawan Deka and Ilya Gorbunov

Joint Institute for Nuclear Research, Dubna

July 5, 2021

Introduction

- CMS detector provides a large amount of data every second that comes from millions of proton-proton collisions.
- It is necessary that both hardware and software perform optimally to ensure the high quality of acquired data.
- There are three main tasks Muon POG DQM group is responsible for:
  - Muon DQM: looking at how muons are reconstructed at CMS.
  - Data Certification: making sure that certain subsets of the data can be used for Muon analysis.
  - Release Validation: making sure that software works fine in terms of Muon reconstruction.

Motivation for Automated DQM

- Drawbacks of the manual DQM:
  - Spotting a problem latency. Human intervention requires sufficient statistics.
  - A human can analyze only a limited amount of data.
  - Unintended human errors.
  - Human decision making process depends on level of experience and understanding.
  - Changing running conditions (muons reconstruction evolves over time).
  - High demand for human resources.
- Automated DQM (ADQM)
  - Statistical tools and ML based models can help towards the automation of the present data monitoring, certification and release validation procedures.
  - DQM is basically an Anomaly detection (AD) process.

Strategy for ML-based DQM

- Collect methods used for AD and evaluate their applicability (we have contacted CMS Machine Learning Group Conveners & ML Knowledge Group Conveners (link) and ML Data Certification Group Conveners (link) for their suggestions and guidance. We have managed to collect a considerable amount of useful information). This is the scope of the current talk.
- Investigate what are the requirements for a Muon ADQM system from:
  - Analysts
  - Validators
  - Muon POG
  - Software developers
  - Missed someone???
- Choose tools to implement and build the test-bed:
  - Frontend
  - Backend
  - Algorithm and libraries that implement it
  - Anything else???
- Present the test-bed, get feedback, scale it and put into production.

Suggested Reading Materials

- ML4DQM twiki page This page collects information related to the application of ML techniques on anomaly detection of CMS data.
- ML4DQM and ML4DC differences This page describes the difference between the ML applications to online data monitoring (ML4DQM) and offline Data Certification (ML4DC).
- "Towards automation of data quality system for CERN CMS experiment", M. Borisyak *et al.*, 2017.
- "Detector monitoring with artificial neural networks at the CMS experiment at the CERN Large Hadron Collider", A. A. Pol *et al.*, 2018..
- "Deep learning for certification of the quality of the data acquired by the CMS Experiment", A. A. Pol *et al.*, 2019.
- "Anomaly detection using Deep Autoencoders for the assessment of the quality of the data acquired by the CMS experiment", A. A. Pol *et al.*, 2019.
- "The Data Quality Monitoring Software for the CMS experiment at the LHC: past, present and future", V. Azzolini *et al.*, 2019.

Limitations of ML-based DQM

- Anomalies are quite low in the CMS data. Roughly 2% of the data-set.
- Emerging and unprecedented failures/anomalies are difficult to anticipate.
- In such a situation, Supervised models are not suitable as they do not have adequate representations of anomalies to train.
- Semi-supervised or Unsupervised ML models are preferred over Supervised ones.
- Models explored: AutoEncoder (AE), Principal Component Analysis (PCA), Non-negative Matrix Factorization (NMF), etc.

Towards ML-based DQM and DC

- CMS adopts two step process towards automation of DQM: "Two step schema".
- In the first step, single histogram discriminators with single Lumi-Section time granularity are used.
- In the second and final step, results obtained by single histogram discriminator are combined to decide upon the quality of a run or lumisections.
  - Tool being developed for combining multiple histogram is ML4DQM.
- For Data Certification, semi-automated system developed for the case of the Muon Systems.
  - Tool being developed is AutoDQM. Built-in statistical tests are used to compare the "test" run to a "reference" run or runs.

Dimensional Reduction

- More layers or dimensions in data can decrease a ML model's accuracy. Known as the "Curse of Dimensionality".
- The number of dimensions can be reduced if some layers or dimensions are correlated.
- Dimensionality reduction helps in avoiding overfitting and reducing computational time.
- Two approaches to data reduction: feature selection and feature extraction.
- Both PCA and AE use feature extraction.

Motivation behind Data Reduction

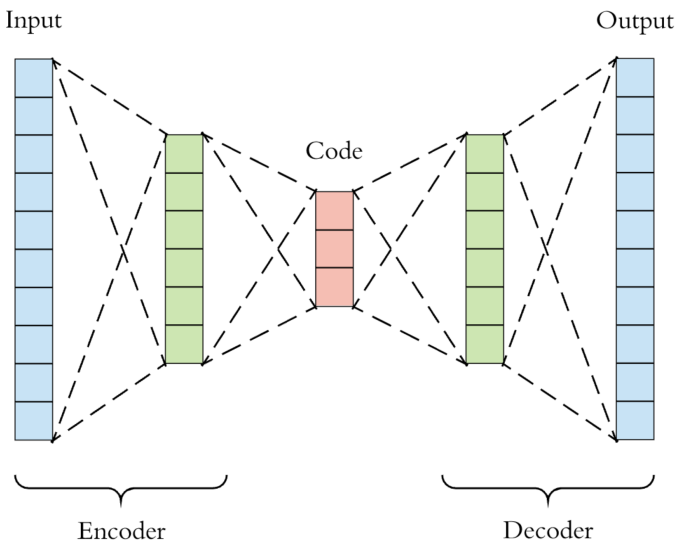
- Feature extraction approach projects high dimensional correlated data into lower dimensions.
- However, such projections result in loss of information including the outliers or anomalies.
- Then why to use?
- The answer is once the main features are identified, outliers or anomalies can easily be detected.
- In a sense, outlier or anomaly detection is a by-product of dimensional reduction.

Principal Component Analysis

- Principle Component Analysis is an unsupervised technique with linear dimensional reduction.
- Data is projected to the directions with high variance. These directions have low or zero correlations between them.
- Following are the key steps for PCA:
  1. Calculate the $n \times n$ correlation matrix from the data with n dimensions.
  2. Calculate the Eigenvectors and Eigenvalues of the correlation matrix.
  3. Take the first k-eigenvectors with the largest eigenvalues.
  4. Project the original data into these k eigenvectors resulting in k dimensional data so that $k \leq n$.
  5. Compute the Normalized Root Mean Square Error (NRMSE) between the input X and output $\hat{X}$.
  6. Higher the NRMSE, more anomalous the data is.

AutoEncoder

- Based on deep artificial neural network.
- Approximate identity mapping between the input and output layers.
- The encoder compresses the input to get to the core layer.
- The decoder reconstructs the information to produce the output.
- The decoding process mirrors the encoding process in the number of hidden layers and neurons.
- Compute the Mean Square Error (MSE).
- Higher MSE $\Rightarrow$ more anomalous data.



Some features of AutoEncoder

- The network can be trained with Keras and TensorFlow using the Adam optimizer.
- AutoEncoder can perform non-linear transformations whereas PCA uses linear algebra to transform.
- Autoencoders learn automatically from data. However, the learning is data specific.
- Autoencoders are lossy $\Rightarrow$ the output loses quality compared to the input.
- Many types of Autoencoders
  - Vanilla AutoEncoder
  - Deep AutoEncoder
  - Convolutional AutoEncoder
  - Denoising AutoEncoder
  - Variational AutoEncoder

AutoDQM

- A semi-automated system that uses statistical tests to flag anomalies.
- Developed for Data Certification for the case of the Muon Systems.
- Built-in statistical tests
  - Kolmogorov-Smirnov (KS) test for 1D histograms.
  - Pull values or χ^2 test for 2D histograms.
- Docker based deployment on CERN Openstack Virtual Machine.
- Python backend using Apache CGI and React frontend.
- Query from DQM Offline for available series, samples, and runs. ROOT files downloaded in volume of the AutoDQM container.
- Future addition of ML techniques:
  - Use of neural networks
  - Clustering algorithms (DBSCAN, k-means)
  - Autoencoders
  - Time correction in Autoencoders and PCA
  - Combining multiple runs for larger luminosity data-sets

C. Freer, I. Suarez *et al.*

Industrial Collaboration

The CMS experiment has established partnership with IBM and Yandex under the CERN Openlab framework

- IBM: to support automation of online DQM using ML.

  " Improving the use of data quality metadata via a partnership of technologies and resources between the CMS experiment at CERN and industry", Virginia Azzolini *et al.*, CHEP 2018.

- Yandex: to support automation of offline DC process using ML.

  "Towards automation of data quality system for CERN CMS experiment", M Borisyak *et al.*, Journal of Physics: Conference Series, 2017.

ML models for BSM Physics

- ML models are also being developed to search for anomalous processes as a sign of BSM Physics.
- These models are applied mostly on CERN Open Data.
- Some of the references suggested by Savannah Thais
  - "ANOmaly detection with Density Estimation (ANODE)", Nachman and Shih, 2020.

    ANODE is an unsupervised model-independent search method and is based on neural density estimation. While this work is focused on collider searches for BSM Physics, ANODE is completely general and can be applied to other areas of physics.
  - "One-Class Support Measure Machines for Group Anomaly Detection", K. Muandet and B. Schölkopf, 2013.

    One-class support measure machines (OCSMMs) aims to recognize anomalous aggregate behaviors of data points.

ML models for BSM Physics

- "Adversarially Learned Anomaly Detection on CMS Open Data: re-discovering the top quark", Oliver Knapp *et al.*, 2020.

  In this work, data driven anomaly detection using Adversarially Learned Anomaly Detection (ALAD) algorithm is applied on CMS Open Data in search for New Physics. This work shows that the data driven anomaly detection techniques, such as ALAD, can highlight the presence of rare phenomena by identifying the main features of the experimental signature at the LHC.

- "Alternate methods for anomaly detection in high-energy physics via semi-supervised learning", Mohd Adli Md Ali *et al.*, 2020.

  This work proposes an alternative classifier type called the one-class classification (OCC). OCC algorithms require only background or noise samples in its training data-set. The algorithm flags any sample that does not fit the background feature. This work also introduces two new algorithms called EHRA and C-EHRA, which use ML regression and clustering to detect anomalies in samples.

Other Methods for Anomaly Detection

- Density-based techniques (k-nearest neighbor, local outlier factor, isolation forests, etc).

  Above methods may be mostly supervised, and may suffer from the Curse of Dimensionality.

- Long short-term memory neural networks.
- Bayesian Networks.
- One-class support vector machines.
- Hidden Markov models (HMMs).
- Cluster analysis-based outlier detection.
- Fuzzy logic-based outlier detection.

Conclusion

- A number of methods have been mentioned.
- How do we choose between them?
- Next step is to collect requirements and suggestions from the people or groups in charge.

Thank you!