Outlier Detection for Data Certification

Patomporn Payoungkhamdee

Mahidol University

patomporn.pay@gmail.com

2 August 2019

Overview

Data Quality Monitoring (DQM)

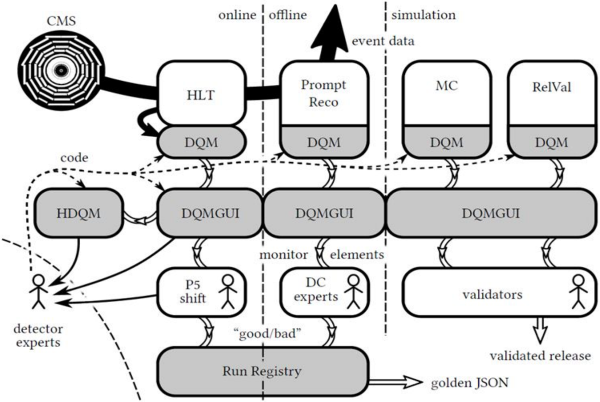


Figure: Tools and Processes of DQM, retrieved from M. Stankevicius

Data granularity in CMS (Offline)

- Reconstruct physics quantity 48 Hours after collision
- Offline shifters and detector experts check the dozens of distribution histograms to define goodness of data
- Certification is made on Run and Lumisection levels
- Lumisection(LS) is taken around 23 seconds for one interval
- Briefly criteria for bad LS
  1. Runs tagged as bad by human (whole run)
  2. DCS bits (LS levels)
  3. Shifter mark some range of LSs from other cases
- Golden JSON are the rest of them (Good LS)

[1] M. Stankevicius, Data Quality Monitoring: Offline

Objective

- **Certify data quality in lumisection granularity**
- Reduce manual work of the shifter
- Standardize data certification criteria

Expectation

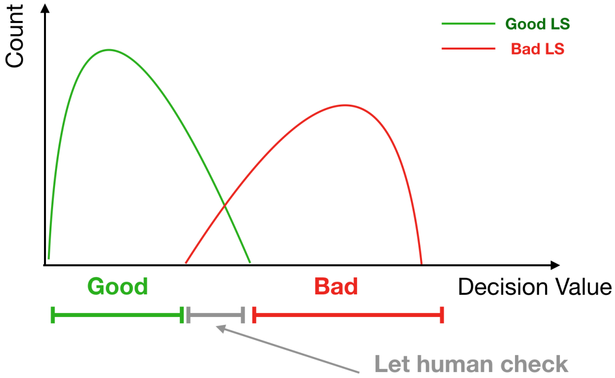


Figure: Three possible regions of prediction

Datasets

- pp collisions, 2016 data
- JetHT
- Each lumisection (datapoint) contains
  - 39 histogram of physics quantity e.g. JetPt, JetEta, JetPhi, etc.
  - Represent one histogram with 7 numbers
  - 259 Features (39×7)
- Good LS defined in Golden JSON else Bad LS
- Data splitting
  - 60% good LSs for training
  - 20% good LSs for validation
  - 20% good LSs combine with bad LS for testing

Histogram representation

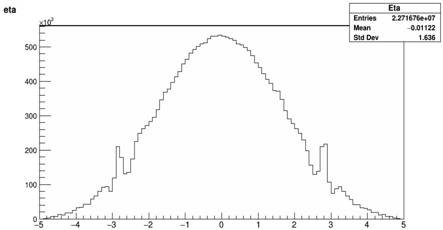


Figure: Example of Eta distribution

- Collection of physics objects e.g. photons, muons and so on
- Measurement quantity: Transverse momentum, eta, phi, etc.

1. Quantize [10%, 30%, 50%, 70%, 90%] of the histogram
2. Combine mean and rms
3. Use these **7 values to represent one histogram**

Data Preprocessing

- MinMaxScalar Transformation
- Consider Lumisection i and Feature j

$$x'_{ij} \leftarrow \frac{x_{ij} - \min_{\forall i \in S_{\text{train}}}\{x_{ij}\}}{\max_{\forall i \in S_{\text{train}}}\{x_{ij}\} - \min_{\forall i \in S_{\text{train}}}\{x_{ij}\}} \tag{1}$$

- Then our datapoint should be in range [0, 1]

Semi-supervised Learning

- Unsupervised Models
  - Schölkopf's One-Class SVM
  - Isolation Forest
  - 4 Flavours of Autoencoder
- Feed only good LS for train and validate the model
- Testing with good LS and bad LS
- Consequently, it's falling into **Semi-supervised Learning** category

Schölkopf's One-Class SVM

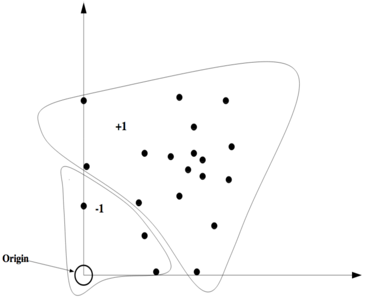


Figure: Scattering in latent space: retrieved from http://www.jmlr.org/papers/volume2/manevitz01a/manevitz01a.pdf

- Minimize (Soft Margin)

$$\frac{||w||^2}{2} + \frac{1}{\nu l} \sum_{i=1}^{l} \xi_i - \rho \qquad (2)$$

- Under

$$w \cdot \Phi(x_i) \geqslant \rho - \xi_i \ , \ \xi_i \geqslant 0 \quad (3)$$

- **Kernel**: Gaussian Base Radial function (GBF)
- Determine tangent distance from hyperplane

Isolation Forest

- Ensemble Forest from tree by subsampling (Ψ)
  - Iteratively picking up features and random value to construct the node (equivalent to step function)
  - Anomaly score evaluate from average depth of the instance over forest

$$s(x, \Psi) = \exp^{-<h(x)>/c(\Psi)} \tag{4}$$

  - where
    - $h(x)$ is the depth in tree h
    - $c(\Psi)$ normalization factor growing as $\log_2(\Psi)$ from branching

[1] https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q=isolation-forest

Vanilla Autoencoder

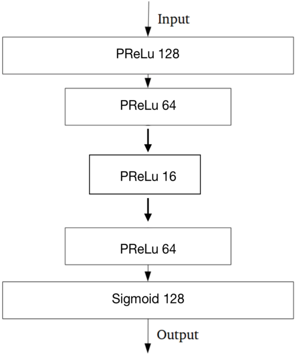


Figure: Body of Vanilla AE

- Concise the information into small latent space and reconstruct
- Loss function

$$\mathcal{L}_{\text{tot}} \equiv \frac{1}{N}\sum_{i}^{N} |x - \tilde{x}|^2 \qquad (5)$$

- Truncated normal initializer
- Adam optimizer

Sparse Autoencoder

- Similar to Vanilla AE
- Tweak by L1 Regularizaion (Prevent overfitting)
- Loss function

$$\mathcal{L}_{\text{tot}} \equiv \frac{1}{N}\sum_{i}^{N}|x-\tilde{x}|^2 + \lambda_s \sum_{j} ||w_j|| \tag{6}$$

- where $\lambda_s = 10^{-5}$

Contractive Autoencoder

- Tweak by Jacobi Matrix (Prevent variation in dataset)
- Loss function

$$\mathcal{L}_{\mathrm{tot}} \equiv \frac{1}{N}\sum_{i}^{N}|x-\tilde{x}|^2 + \lambda_{\mathrm{c}}||J_h(x)||^2 \tag{7}$$

- where $\lambda_{\mathrm{c}} = 10^{-5}$

Contractive Autoencoder

- Definition

$$||J_h(x)||^2 \equiv \sum_{ij} \left(\frac{\partial h_j}{\partial x_i} \right)^2 \tag{8}$$

- where h_j is activation function
- In our cases
  - PReLu activation function

$$||J_h(x)||^2 = \sum_j [\alpha_j H(-(w_{ji} x^i + b_j)) + H(w_{ji} x^i + b_j)] \sum_i (w_{ji})^2 \tag{9}$$

  - Sigmoid activation function

$$||J_h(x)||^2 = \sum_j [h_j * (1 - h_j)] \sum_i (w_{ji})^2 \tag{10}$$

Variational Autoencoder

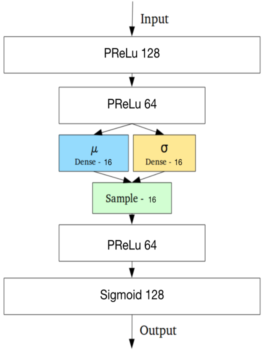


Figure: Body of Variational AE retrieved from https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

- Random "new sampling" in latent space by gaussian random generator

$$\mathcal{Z} \equiv \mathcal{N}(\mu_i, \sigma_i) \tag{11}$$

- Tweak by reduce discontinuity in latent space
- Loss function

$$\mathcal{L}_{\text{tot}} = \frac{1}{N}\sum_{i}^{N} |x - \tilde{x}|^2 + \mathcal{D}_{\text{KL}}(p|q) \tag{12}$$

Variational Autoencoder

Theorem (Kullback-Leibler Divergence)

- *"How much information is loss after represent data with function"*

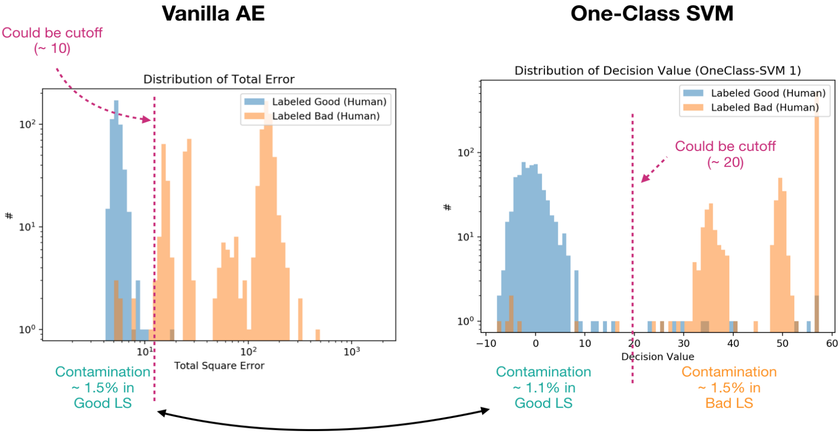
$$\mathcal{D}_{KL} \equiv < \log p - \log q > \tag{13}$$

- *Where p is observed value and q is approximaiton function*
- *Since our q is Gaussian function*

$$\mathcal{D}_{KL} = \frac{1}{2}(\mu_i^2 + \sigma_i^2 - 2\log(\sigma_i) - 1) \tag{14}$$

$$\mathcal{L}_{\text{tot}} = \frac{1}{N}\sum_i^N |x - \tilde{x}|^2 + \frac{1}{2}(\mu_i^2 + \sigma_i^2 - 2\log(\sigma_i) - 1) \tag{15}$$

Find the cutoff



Spot the same Bad => Good LS

Performance

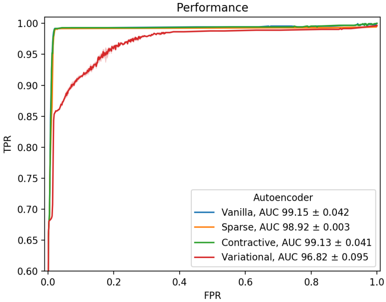


Figure: Various AE

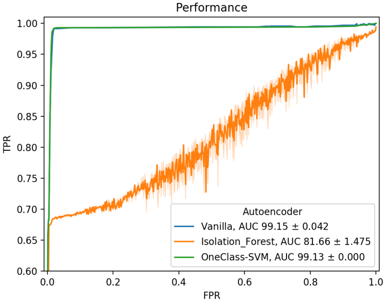


Figure: Vanilla vs SVM vs Forest

Under configuration

- **Isolation Forest** $N_{\text{tree}} = 200, \;\; \Psi = 512$
- **OneClass-SVM** $\nu = 0.1, \;\; \gamma = 0.1$(Inverse gaussian width)

Example of Reconstruction

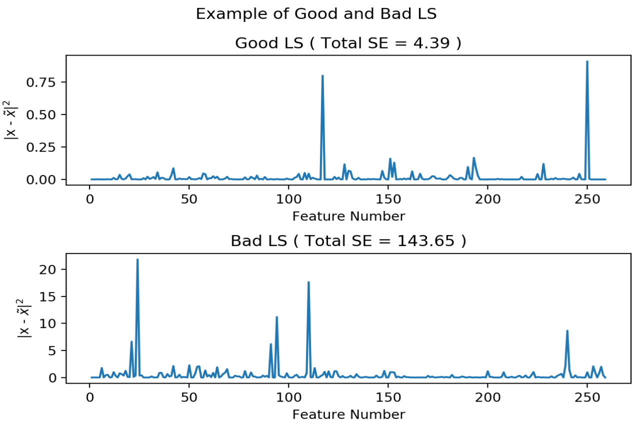
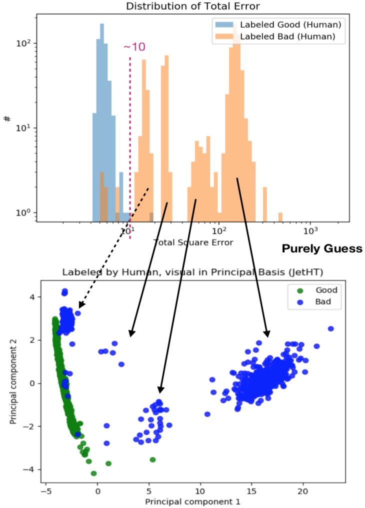
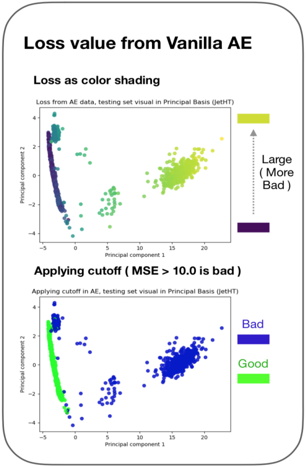


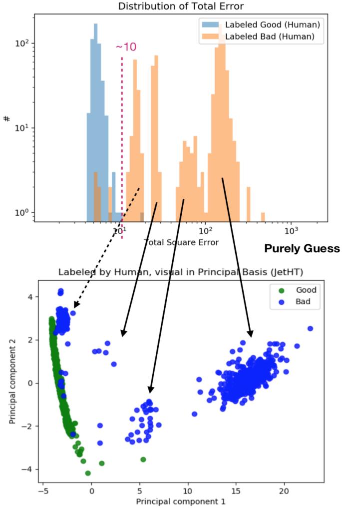
Figure: Reconstruction error from Vanilla AE

Extended Investigation



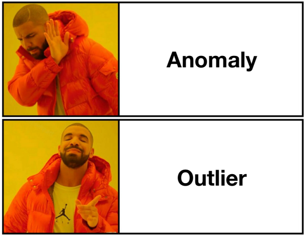
Extended Investigation





Summary

- Semi-supervised learning yield a remarkable result
- There is no grey zone from our model for this dataset
- Bad LS could be divided into two parts
    - Bad with some pattern
    - Anomaly



Future work

- Good LS from run tagged by human and DCS bits still suspicious not to be a ground truth
- Require simulation data
  - To be purely good LS for training
  - For testing the failure scenario

Acknowledgement

- CERN Summer Student program 2019
- Especially
  - Marcel Andre Schneider
  - Francesco Fiori
  - Kaori Maeshima
  - also countless DQM people :)
- GPU Resources from IBM

Thank you

Question?

Back up

ROC Curve

bla