SCHOOL OF DATA ANALYSIS

LAMBDA

Yandex Data Factory

CMS Data Quality Management

Maxim Borisyak, Fedor Ratnikov, Denis Derkach, Andrey Ustyuzhanin

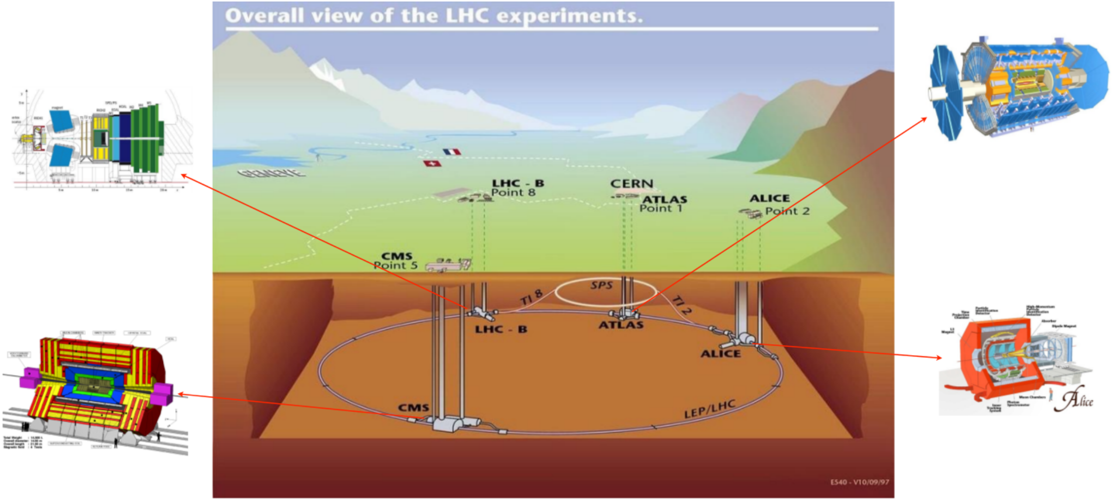
Yandex School of Data Analysis

Outline

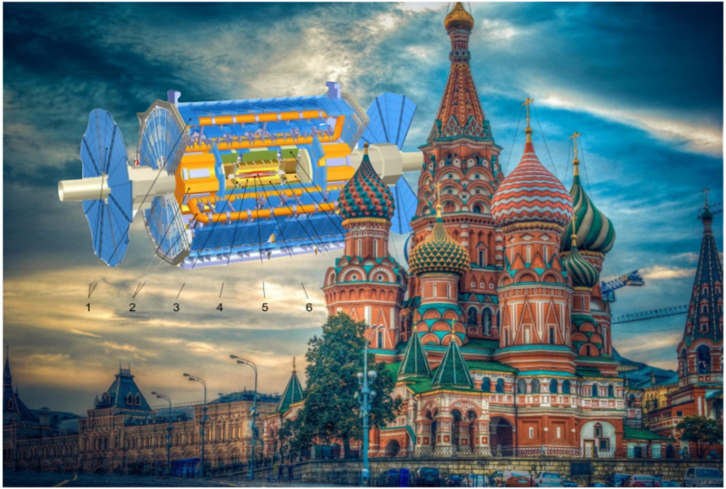
- CERN CMS brief overview;
- base solution;
- decompostion of anomalies by source.

CERN CMS

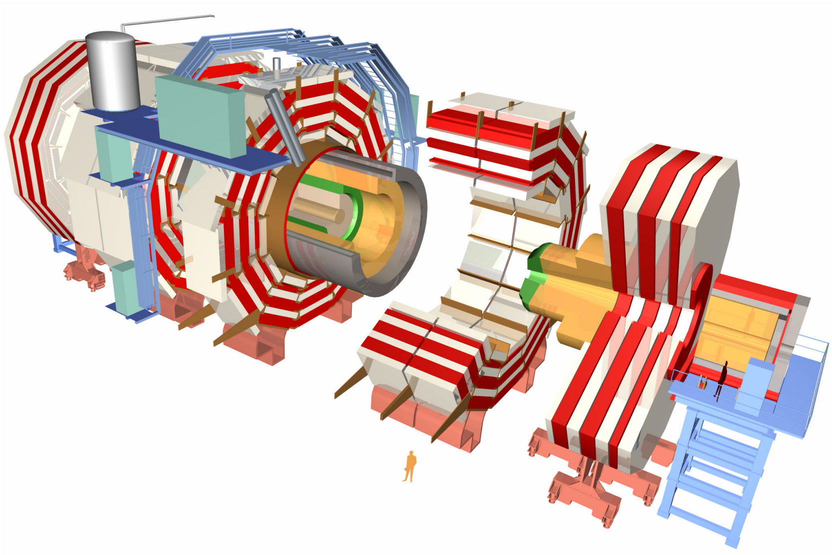
CERN LHC


Overall view of the LHC experiments.
GENEVE
LHC - B
Point 8
CERN
ATLAS
Point 1
ALICE
Point 2
CMS
Point 5
SPS
TI 8
TI 2
LHC - B
ATLAS
ALICE
CMS
LEP/LHC
E540 - V10/09/97
Alice

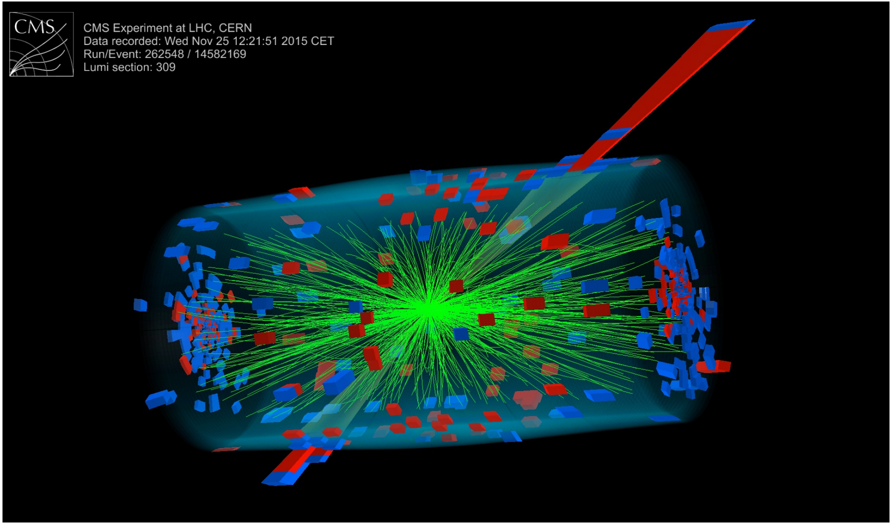
CMS detector



CMS detector



CMS detector



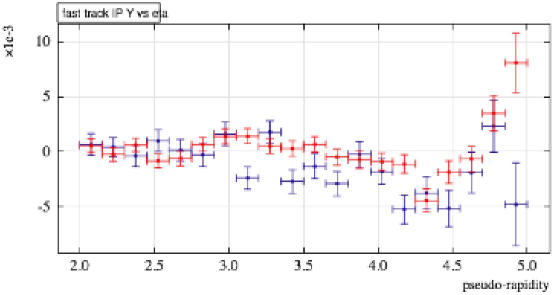
Data Quality Management

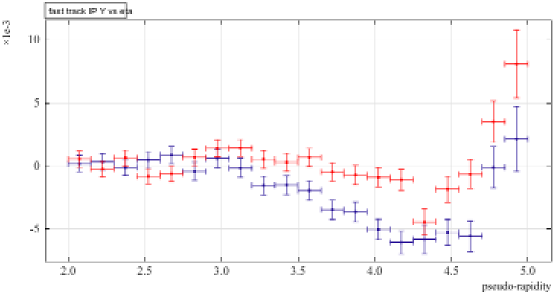
- the CMS detector is a complex system;
- data sample (lumisection) per each 20 sec.:
  - each contains a huge number of events ($10^3 - 10^4$);
  - unit of data;
- data requires validation;
  - only consistent data be used for analysis.

Current status

› experts propose high-level features;

› a number of Data Quality experts check distributions against reference ones.





Problem statement

Can Data Quality Managment be, at least partially, automated?

- assist Data Quality experts:
  - label part of the data;
  - hints for human experts (where to check).

Towards automated Data Quality

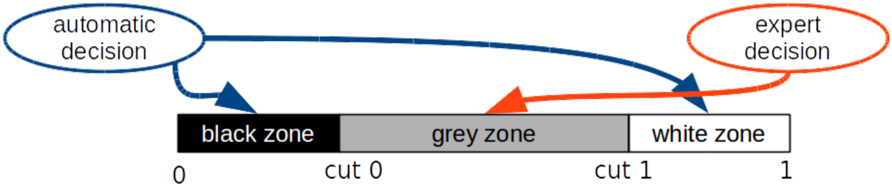
Data

- 2010 open data (`http://opendata.cern.ch/collection/CMS-Primary-Datasets`);
- data from three streams:
  - photons: events with a lot of photons,
  - muons: events with a lot of muons,
  - minibias: prescaled whole event stream;
- 4 channels ($\approx$ subsystem):
  - photons,
  - muons,
  - particle flows (proto-particles),
  - particles from calorimiter.
- normal/anomaluos samples: $2/3$ vs. $1/3$

Brute-force approach

Data Quality expert assistance:

- automatically process the most obsvious cases;
- guarantee predefined quality of automatic decisions;
- pass to human expert ambiguous cases.



Base approach

$$\text{Rejection Rate} = \frac{\mathrm{Rejected}}{\mathrm{Total}} \to \text{min}\,,$$

under constrains:

$$\text{Loss Rate} = \frac{\text{False Negative}}{\text{True Positive} + \text{False Negative}} = 1 - \mathrm{recall} \leq L_0,$$

$$\text{Pollution Rate} = \frac{\text{False Positive}}{\text{False Positive} + \text{True Positive}} = 1 - \mathrm{precision} \leq P_0,$$

Base approach

- while not out of samples:
  - train a classifier on available labeled data;
  - estimate cuts by cross-validation;
  - try to classify new sample;
  - if the score is between cuts, then:
    - pass the sample to a Data Quality expert;
    - extend dataset with the sample and the Data Quality expert label;
  - otherwise:
    - automatically label the sample.

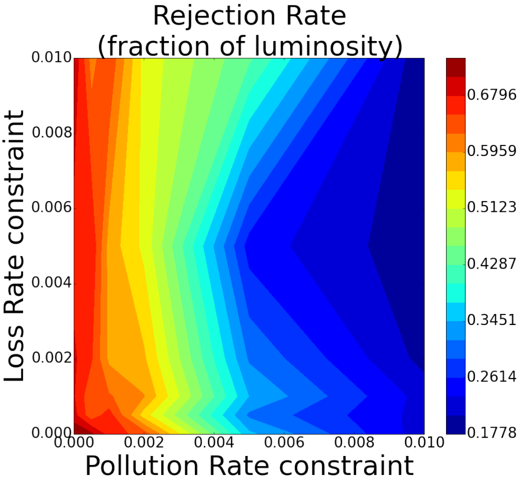
Base approach

Results:

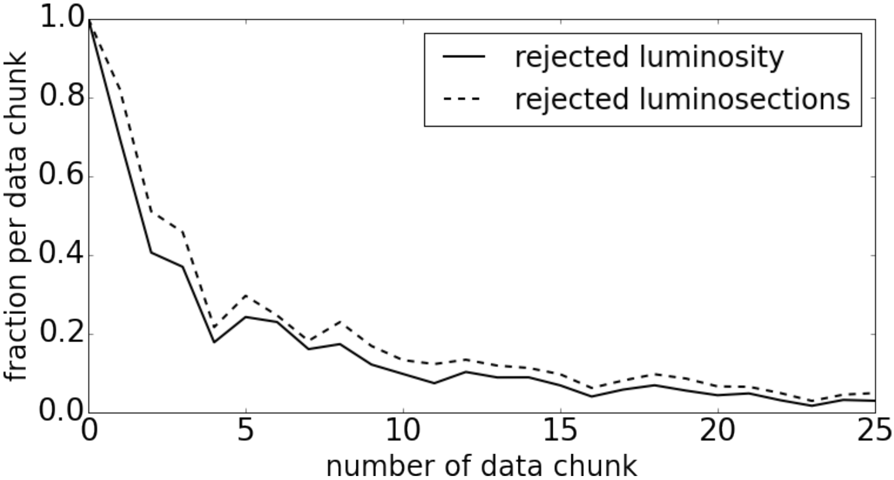
- working point:
  - Pollution Rate $< 5 \cdot 10^{-3}$;
  - Loss Rate $< 5 \cdot 10^{-3}$;
- $\approx 80\%$ manual labour saved.

Details:

- ≈ 2500 features;
- 26 chunks of data;
- ≈ 1000 samples in each chunk;
- XGBoost as underlying classifier;



Base approach



Decomposing Anomalies

Motivation

Main goal

Study how anomalies affect individual channels.

Examples.

- What channels are responsible for anomalies?
- If only photons were affected is it possible to save muonic data?
- Which plots should receive more attention from Data Quality experts?

Supervised approach

1. On features from each channel build a neural network;
2. each channel network returns a score for its channel;
3. connect networks by:
   - log. reg.
   - `min` operator (with dropout),
   - a sort of fuzzy`AND`;
4. train network to recover global labels;
5. define estimation of score for each channel as corresponding network output.

Discussion

Consider set of channels $\mathcal{C}$, an anomaly A affecting channels $C \subseteq \mathcal{C}$.

Assumption 1

Anomaly A can be detected independently from data of any channel from C.

Assumption 2

Anomaly A can not be detected from data of channels other than C.

Assumption 3

Fraction of weights of normal samples is at least $1/2$.

Discussion

Corollary 1

If an anomaly can be detected from a channel features, the channel data is anomalous.

'Theorem' 1

Under assumptions 1, 2 and 3 for all described above networks having enough degrees of freedom and data samples for training
each subnetwork has high discriminative power against anomalies affecting its channel.

Proof of the theorem

The idea of the 'proof' is to show that global minimum of loss function corresponds to the state where each subnetwork 'reacts' only on anomalies affecting its channel.

- cross-entropy loss (1 - normal lumisections, 0 - anomaly);
- outputs $f^i_{\mathrm{subnetwork}}$ of i-th subnetwork are bounded:

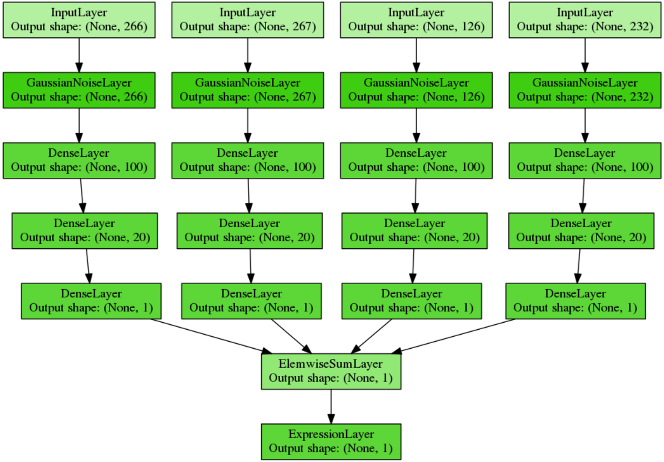
$$f^i_{\mathrm{subnetwork}} \in (0, 1)$$

- activation function for the whole network:

$$f_{\mathrm{network}} = \phi\left(\sum_{i=1}^{4} f^i_{\mathrm{subnetwork}}\right)$$

$$\phi(x) = \mathsf{exp}(x - 4)$$

Network diagram



Proof of the theorem

- consider channel $c \in \mathcal{C}$:
  - $\mathcal{A}_c$ all anomalies that does affect c;
  - $\bar{\mathcal{A}}_c$ all anomalies that does not affect c;
- relative to a subnetwork there are 3 cases:
  - no anomalies;
  - anomaly 'visible' from its channel ($\mathcal{A}_c$);
  - anomalies 'invisible' from its channel ($\bar{\mathcal{A}}_c$);
- with respect to these cases, loss of the whole network can be decomposed into:

$$\mathcal{L} = \mathcal{L}_{\mathrm{normal}} + \mathcal{L}_{\bar{\mathcal{A}}_c} + \mathcal{L}_{\mathcal{A}_c}$$

Proof of the theorem

'Lemma' 1

Under assumptions 1, 2 and 3, and the theorem's conditions in case of anomaly from $\mathcal{A}_c$ output of subnetworks corresponding to channel c is as close to 0 (anomaly) as possible.

Proof of the theorem

- $\mathcal{L}_{\mathcal{A}_c}$ and $\mathcal{L}_{\mathrm{normal}} + \mathcal{L}_{\bar{\mathcal{A}}_c}$ can be optimized independently:
  - since structure of subnetwork is sufficient to learn to separate these cases by assumptions;

$$\mathcal{L}_{\mathcal{A}_c} = -\sum_j \log\left[1 - \exp\left(\sum_{i=1}^{4} f^i_{\mathrm{subnetwork}}(X_j) - 4\right)\right]$$

- where the first sum is over samples X_j with anomalies from $\mathcal{A}_c$;
- since subnetworks are independent:
  - $\mathcal{L}_{\mathcal{A}_c}$ is minimized when output of the subnetwork built on channel c is as close to 0 as possible.
- This proves 'Lemma' 1.

Proof of the theorem

- subnetwork can not distinguish normal cases and $\bar{\mathcal{A}}_c$;
- nevertheless, since $\bar{\mathcal{A}}_c$ is still an anomaly, subnetwork receives punishment either for:
  - predicting low score for normal cases;
  - predicting large score for cases from $\bar{\mathcal{A}}_c$.
- this may result in some bias relative to the presence of anomalies from $\mathcal{A}_c$.

'Lemma' 2

Under assumptions and theorem 1 conditions, all subnetwork are unbiased, i.e. for normal cases and anomalies from $\bar{\mathcal{A}}_c$ output of subnetwork for channel c is close to 1.

Proof of the theorem

Let X be output of subnetwork for channel c under normal cases and anomalies from $\bar{\mathcal{A}}_c$, ϵ_i and ϵ'_i - sum of outputs from the rest of subnetworks, α - fraction of good lumisections, β - fraction of anomalies from $\bar{\mathcal{A}}_c$:

$$
\begin{aligned}
\mathcal{L} &= \mathcal{L}_{\text{normal}} + \mathcal{L}_{\bar{\mathcal{A}}_c} + \mathcal{L}_{\mathcal{A}_c} \\
&= -\frac{\alpha}{n_1} \sum_{i=1}^{n_1} \log \exp\left(X + \epsilon'_i - 4\right) \\
&- \frac{\beta}{n_2} \sum_{i=1}^{n_2} \log\left(1 - \exp\left(X + \epsilon_i - 4\right)\right) \\
&+ \mathcal{L}_{\mathcal{A}_c}
\end{aligned}
$$

Proof of the theorem

In the worst case scenario and by 'Lemma' 1 (at least one network reports anomaly with score $\delta \ll 1$):

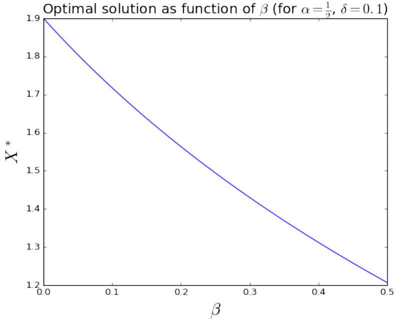
$$\epsilon < 2 + \delta$$

Solving for lower bound on optimal X:

$$\begin{aligned} \frac{\partial \mathcal{L}_{\text{worst case}}}{\partial X} &= -\alpha + \beta \frac{\mathrm{exp}(X + \delta - 2)}{1 - \mathrm{exp}(X + \delta - 2)} = 0 \\ &\Rightarrow X^* = 2 - \delta + \mathrm{log}\,\frac{\alpha}{\alpha + \beta} \end{aligned}$$

Proof of the theorem

Dataset is reweighted so that $\alpha = \frac{1}{2}$. Thus, $\beta \in [0, \frac{1}{2}]$.



X is restricted to be in range $(0, 1)$, thus minimum of $\mathcal{L}$ is achieved for X as close to 1 as possible, hence **subnetwork is unbiased**.

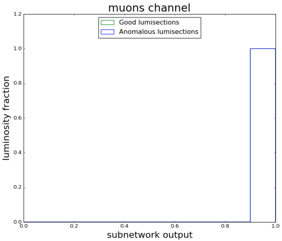
Proof of the theorem

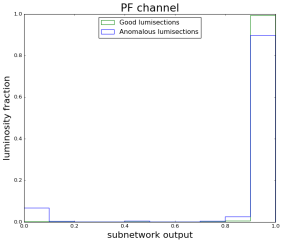
To summarize, each subnetwork return score:

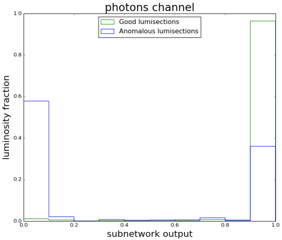
- close to 1 for normal lumisections;
- close to 1 for anomalies 'invisible' from subnetwork's channel;
- close to 0 for anomalies 'visible' from subnetwork's channel.

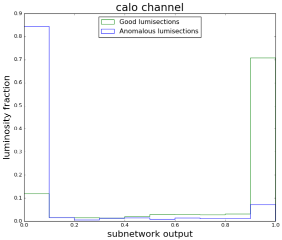
Thus, whole network 'decompose anomalies by channels'.

Results

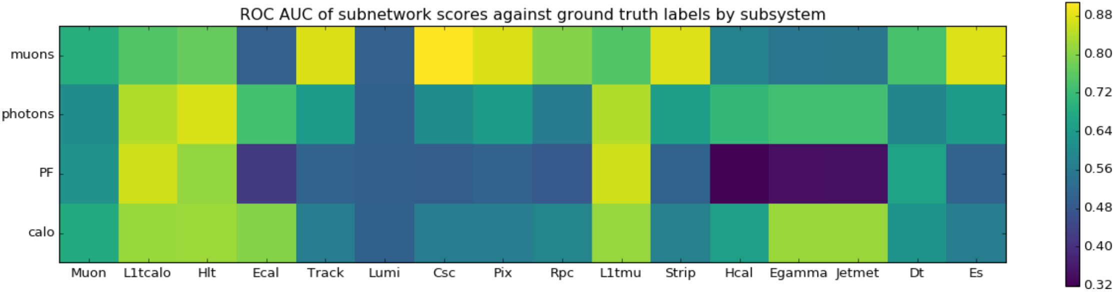

muons channel
Good lumisections
Anomalous lumisections
luminosity fraction
subnetwork output


PF channel
Good lumisections
Anomalous lumisections
luminosity fraction
subnetwork output


photons channel
Good lumisections
Anomalous lumisections
luminosity fraction
subnetwork output


calo channel
Good lumisections
Anomalous lumisections
luminosity fraction
subnetwork output

Results


ROC AUC of subnetwork scores against ground truth labels by subsystem
muons
photons
PF
calo
Muon
L1tcalo
Hlt
Ecal
Track
Lumi
Csc
Pix
Rpc
L1tmu
Strip
Hcal
Egamma
Jetmet
Dt
Es
0.88
0.80
0.72
0.64
0.56
0.48
0.40
0.32

Discussion

Results:

- scores are consistent with expert knowledge about detectors;
- scores from all methods (log.reg., min, fuzzy`AND`) are consistent with each other;

Discussion:

- assumption 1 might be hard to check in practice;
- assumption 2 is a reasonable by itself;
- assumption 3 can be artificially ensured (or even omitted).

Summary

Summary

- the CMS experiment;
- solution for Data Quality expert assistance:
  - up to 80% of saved manual labour;
- decomposition of anomalies by sources:
  - scores are consistent with expert knowledge about detectors;
- work in progress for new data (2016).

Bonus: rare anomalies

Upgraded detector

Suddenly:

- upgraded detector is much more robust;
- much less anomalies: around 1-2% of samples (against $1/3$ before);

Assumptions

Assumption 1

All normal samples are embedded into small region on a low-dimensional subspace.

Assumption 2

Every point outside this region is an anomaly.

Rare anomalies

- technically, two-class problem;
- class disbalance may lead to poor models;
- assumption 2 allows to use one-class methods.

> How can one-class methods be used in a classification problem?

One-class objective trick

Consider classification problem of a class ($\mathcal{C}$) against noise ($\mathcal{N}$):

$$\begin{aligned}
P(\mathcal{C} \mid X) &= \frac{P(X \mid \mathcal{C})P(\mathcal{C})}{P(X)} \\
&= \frac{P(X \mid \mathcal{C})P(\mathcal{C})}{P(X \mid \mathcal{C})P(\mathcal{C}) + P(X \mid \mathcal{N})P(\mathcal{N})};
\end{aligned}$$

Since, $P(X \mid \mathcal{N})$ is known ($f(X)$) and $P(\mathcal{N})$ is controled (let it be $1/2$):

$$P(\mathcal{C} \mid X) = \frac{P(X \mid \mathcal{C})}{P(X \mid \mathcal{C}) + f(X)};$$

Mixed objective

$$\mathcal{L} = -\frac{1}{2} \mathop{\mathbb{E}}_{X \sim \mathcal{C}_+} \log f(X) - \frac{1-\alpha}{2} \mathop{\mathbb{E}}_{X \sim \mathcal{C}_-} (1 - \log f(X)) - \frac{\alpha}{2} \mathop{\mathbb{E}}_{X \sim \mathcal{N}} (1 - \log f(X)) = \frac{1}{2}\mathcal{L}_+ + \frac{1-\alpha}{2}\mathcal{L}_- + \frac{\alpha}{2}\mathcal{L}_{\mathrm{noise}}$$

where:

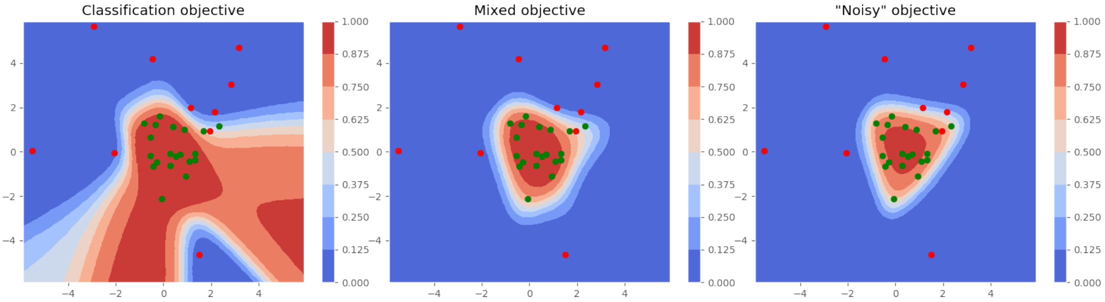
- $\mathcal{C}_+, \mathcal{C}_-, \mathcal{N}$ - positive, negative classes and noise;
- $\mathcal{L}_+, \mathcal{L}_-, \mathcal{L}_{\mathrm{noise}}$ - losses on normal, anomalous and noise examples;
- α - trade-off coefficient;

Intuition

Consider border regions:

- in presence of negative samples nearby, produce solution as in classification problem;
- otherwise, produce one-class like borders;
- $\mathcal{L}_{\text{noise}}$ can be viewed as a regularizer (restriction on possible solutions);

Toy example

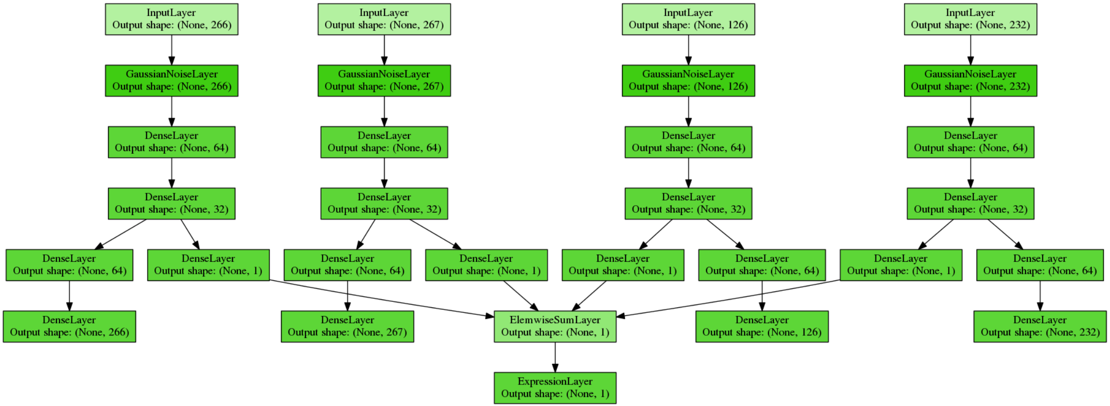


Classification objective
Mixed objective
"Noisy" objective
1.000
0.875
0.750
0.625
0.500
0.375
0.250
0.125
0.000
4
2
0
−2
−4

Discussion

- noise injection into the middle of the net;
- AutoEncoder objective to prevent collapse of the code spaces.

Net


InputLayer
Output shape: (None, 266)
InputLayer
Output shape: (None, 267)
InputLayer
Output shape: (None, 126)
InputLayer
Output shape: (None, 232)
GaussianNoiseLayer
Output shape: (None, 266)
GaussianNoiseLayer
Output shape: (None, 267)
GaussianNoiseLayer
Output shape: (None, 126)
GaussianNoiseLayer
Output shape: (None, 232)
DenseLayer
Output shape: (None, 64)
DenseLayer
Output shape: (None, 64)
DenseLayer
Output shape: (None, 64)
DenseLayer
Output shape: (None, 64)
DenseLayer
Output shape: (None, 32)
DenseLayer
Output shape: (None, 32)
DenseLayer
Output shape: (None, 32)
DenseLayer
Output shape: (None, 32)
DenseLayer
Output shape: (None, 64)
DenseLayer
Output shape: (None, 1)
DenseLayer
Output shape: (None, 64)
DenseLayer
Output shape: (None, 1)
DenseLayer
Output shape: (None, 1)
DenseLayer
Output shape: (None, 64)
DenseLayer
Output shape: (None, 1)
DenseLayer
Output shape: (None, 64)
DenseLayer
Output shape: (None, 266)
DenseLayer
Output shape: (None, 267)
ElemwiseSumLayer
Output shape: (None, 1)
DenseLayer
Output shape: (None, 126)
DenseLayer
Output shape: (None, 232)
ExpressionLayer
Output shape: (None, 1)

Results (preliminary)

Data from the previous studies:

- train/test:
    - 10k/10k positive examples; 64/6.4k negative examples;
- 800 features;

ROC AUC (32 experiments):

- 0.85 ± 0.02 test;
- 0.80 ± 0.05 control.