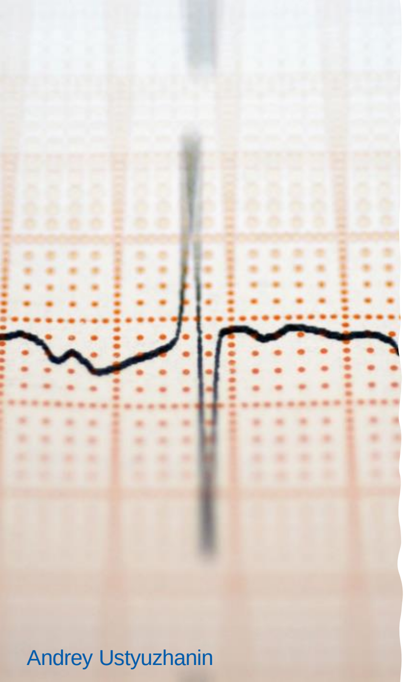


NUS National University of Singapore
Institute for Functional Intelligent Materials

MLHEP 2023

Anomaly Detection

Andrey Ustyuzhanin

INFN National Institute for Nuclear Physics

Ettore Majorana Foundation and Centre for Scientific Culture — EMFCSC

Definition and Examples

Outliers, Anomalies, Novelties

Outlier is a point that is significantly different from the the rest of the **data**:

- noise;
- novelties – differs from previous behavior;
- anomalies – differs from the bulk of data.

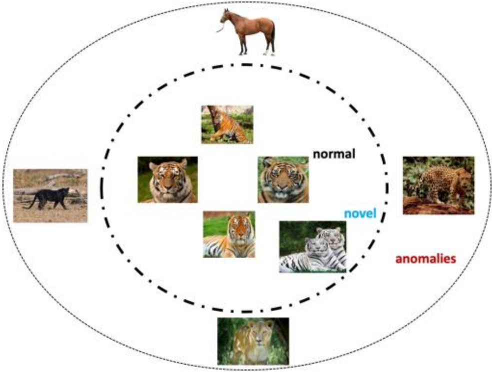
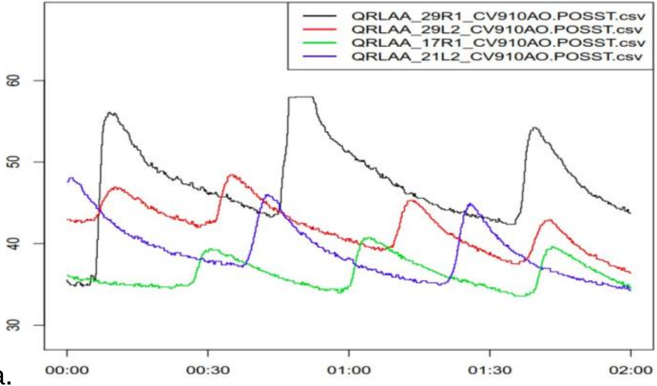


Image: R. Chalapathy and S. Chawla, Deep Learning for Anomaly Detection: A Survey

Example: LHC Cryogenic System

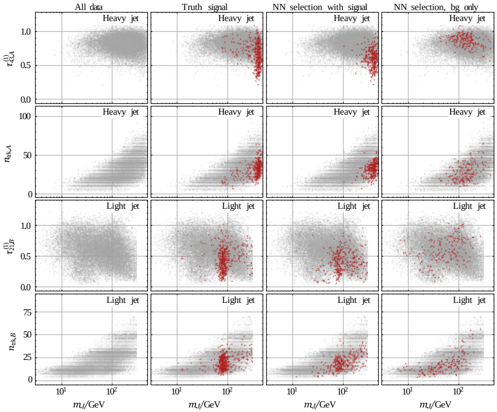
- faulty valve behaviour: one of the cryogenics valves shows an **anomalous** range of movement if compared to the other actuators;
- anomaly points indicate a change in the system's state;
- anomalies can be defined as significant deviation from the data sample collected, hence anomalies can be immediately seen in the data.



F. Tilaro et al., Model Learning Algorithms for Anomaly Detection in CERN Control Systems

Example: New Physics as Anomaly

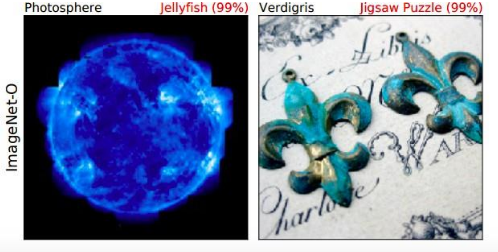
- Anomaly becomes a signal;
- Need to analyse abundance of non-anomalous events;
- Signal features/characteristics are unknown.



J. Collins et al, Extending the Bump Hunt with Machine Learning

Out-of-distribution detection

- New test set with several samples;
- test whether these samples come from distribution already seen;
- if not, the performance of ML solution might degrade (intentionally or not);
- connected to overconfidence problem for ML algorithm.



- Classes that were not previously seen by a classifier.

D. Hendrycks et al, Natural Adversarial Examples

Typical setting

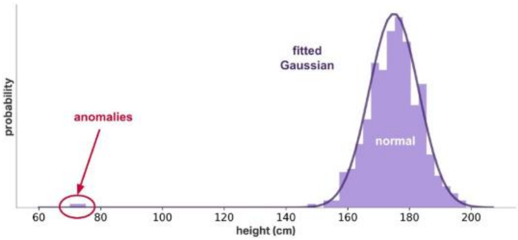
Dataset Properties

- Highly imbalanced: many data points of ”normal” class and very few, if any, of ”anomalous” class.
- Dataset can be labeled or not.
- There can be rare and unseen anomalies, that are not present in the training dataset.
- No clear separation between novelty and anomaly.
- Anomaly definition is contextual.

Output of an Anomaly Detection Algorithm

- **Label**
  - Each test instance is given a normal or anomaly label.
- **Score**
  - Each instance is assigned an **anomaly score**.
    - allows outputs to be ranked
    - May require an additional threshold parameter

Data Model is Everything

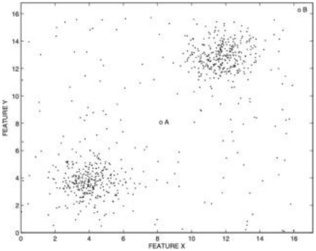


A clear candidate to detect an anomaly can be Z-score:

$$Z = \frac{x - \overline{x}}{s}$$

10/44

Data Model is Everything



A clear candidate to detect an anomaly can be Z-score:

$$Z = \frac{x - \overline{x}}{S}$$

It, however, can fail if the normal class has multimodal distribution.

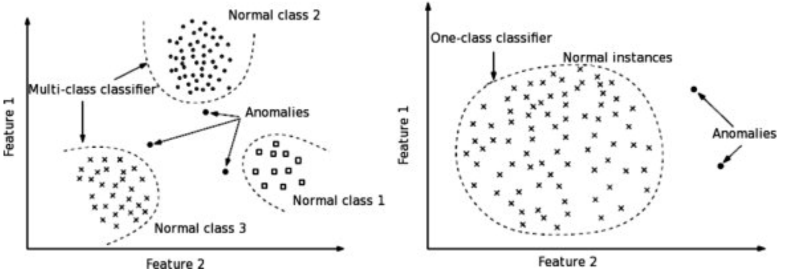
Basic methods

Usual supervised methods

- for labeled dataset;
- straightforward idea: use two- or many class classification;
- good performance if:
  - the amount of anomalous examples is big;
  - we know all types of anomalies.
- anomaly score is naturally the output of classifier;
- is it all we can do?

One-class methods

What if we say that anomaly is everything beyond the border of ”normal” class?



We only need to define how to find a border.

Uses labels of a single (normal) class only!

Figure M. Chica Authentication <...>

One-class family

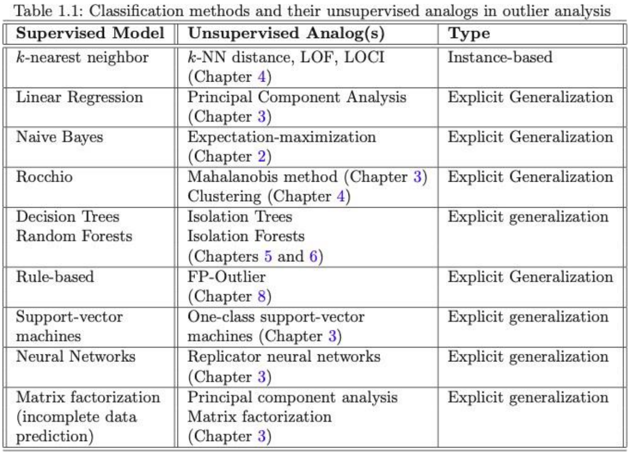


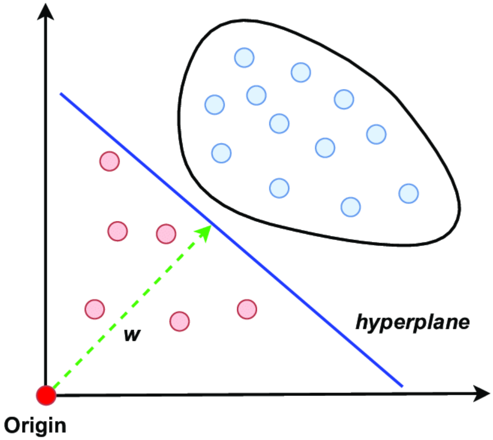
Table 1.1: Classification methods and their unsupervised analogs in outlier analysis

Supervised Model	Unsupervised Analog(s)	Type
k-nearest neighbor	*k*-NN distance, LOF, LOCI (Chapter 4)	Instance-based
Linear Regression	Principal Component Analysis (Chapter 3)	Explicit Generalization
Naive Bayes	Expectation-maximization (Chapter 2)	Explicit Generalization
Rocchio	Mahalanobis method (Chapter 3) Clustering (Chapter 4)	Explicit Generalization
Decision Trees Random Forests	Isolation Trees Isolation Forests (Chapters 5 and 6)	Explicit generalization
Rule-based	FP-Outlier (Chapter 8)	Explicit Generalization
Support-vector machines	One-class support-vector machines (Chapter 3)	Explicit generalization
Neural Networks	Replicator neural networks (Chapter 3)	Explicit generalization
Matrix factorization (incomplete data prediction)	Principal component analysis Matrix factorization (Chapter 3)	Explicit generalization

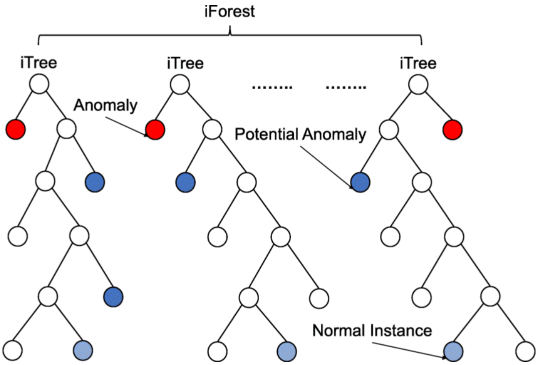
https://bit.ly/43lWLbf

One-class Support Vector Machines

- Learns a hyperplane that **encloses the data** in high-dimensional space (a) transforms the input data into a higher-dimensional space using a kernel function (e.g., Linear kernel). b) finds a hyperplane that has the largest margin to the origin while enclosing most of the data points)
- Points located outside the hyperplane are classified as **anomalies**
- Provides a **probabilistic score** of each point indicating the degree of abnormality



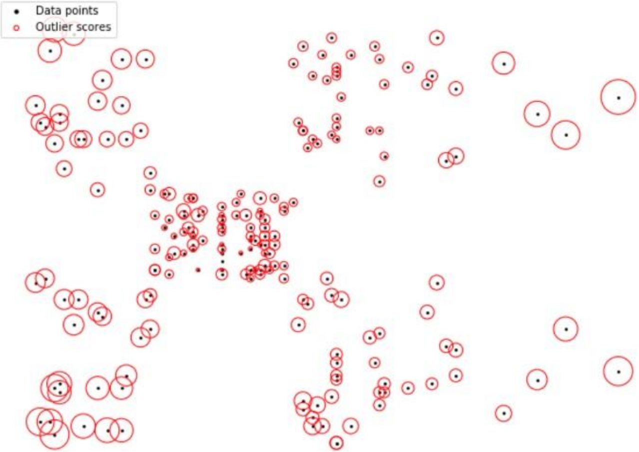
Isolation Forest



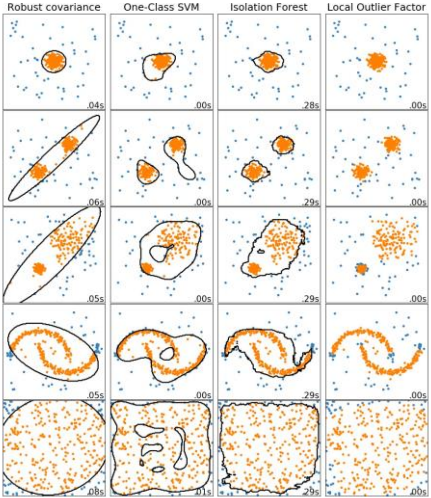
- General idea: efficiently detect anomalies in a dataset by isolating data points using an ensemble of randomized decision trees.
- The algorithm exploits the fundamental property that anomalies are few and different from the majority of the data. As a result, they can be isolated more quickly with fewer random splits in the decision trees.
- Anomalies are identified based on their average path length across all trees in the ensemble – shorter path lengths indicate a higher likelihood of being an anomaly.

Local Outlier Factor

- General idea: outliers have low density with respect to its k neighborhood.
- Anomaly score: proportional to inverse distance to k neighbours.



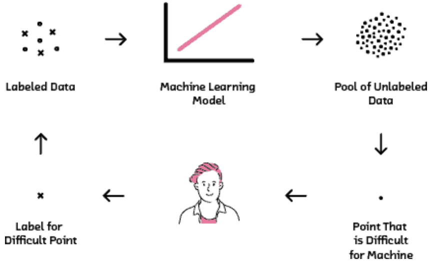
Comparison of One-class Techniques



https://bit.ly/3GBmORM

Active learning for anomaly detection

- for continuous data flow, use active learning:
  - train algorithm on existing labels;
  - check on new samples arriving;
  - ask experts to label only new examples, where classifier was not sure;
  - train new classifier.
- obtained classifier will be better in identifying anomalies.

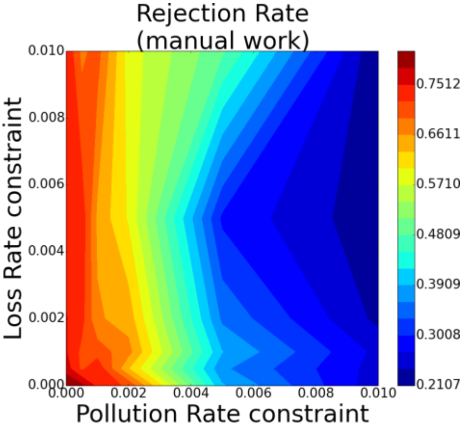

Labeled Data
Machine Learning Model
Pool of Unlabeled Data
Point That is Difficult for Machine
Label for Difficult Point

D Pelleg, Active Learning for Anomaly and Rare-Category Detection

Figure from Cloudera blog

Example: CMS Data Certification

- CMS data certification problem:
  - 2010 CMS data, OpenData portal;
  - manually labeled;
- can be successfully employed in DQM settings;
- approach is able to save up to 20% manual work under tight restrictions;
- quality improves over time.



M. Borisyak, Towards automation of data quality system for CERN CMS experiment

Pre-summary

- Anomalies are often hunted in different tasks and problem settings.
- Understanding of data is very important.
- Main evaluation scores should be used with caution due to imbalanced datasets.
- Straightforward classification might fail due to lack of ”anomalous” class.
- Once class methods provide robust outlier detection method.

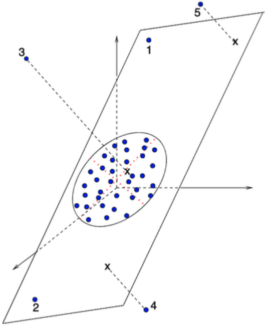
Distance scores

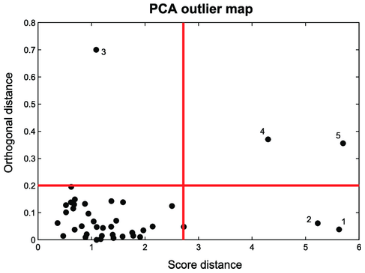
Introduction

- we know that anomalies are rare and deviate from the populous ”normal” class;
- ”normal” class is usually concentrated in some area of feature space;
- can we use this property?

Principle component analysis for anomalies

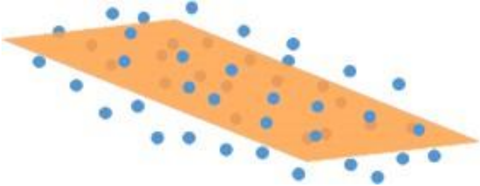
- select an r-dimensional hyperplane that minimizes the squared projection error over the remaining dimensions;
- all points X can be projected to the hyperplane (L);
- a data point, which is far away from its projection is deemed as anomalous.
- anomaly score: normalized distance of the data point to the centroid of the sample along main components.





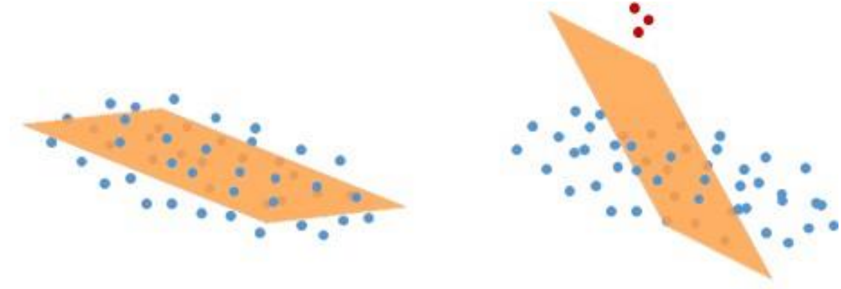
PCA: Explained

- N samples X = $\{x_1, x_2, \ldots, x_N\} \in R^{N\times n}$.
- PCA(X, r): $\min_{L:\mathrm{rank}(L)=r} = ||X-L||_2$



PCA: Outliers

- N samples $X = \{x_1, x_2, \ldots, x_N\} \in R^{N \times n}$.
- PCA(X, r): $\min_{L:\text{rank}(L)=r} = ||X - L||_2$



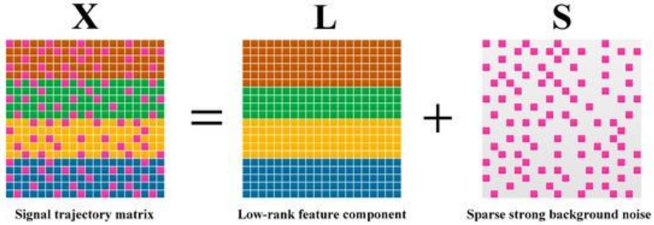
Classical PCA fails even with a few outliers

PCA for anomaly detection: issues

- sensitivity to noise
  - in presence of multiple outliers PCA can have difficulties in determining the main component.
- normalization issues
  - in case of very different feature scale, the variation of one components can eclipse other variations.
- regularization Issues
  - not really stable for small datasets.

Robust PCA

The presence of many outliers can be overcome by using Robust PCA analysis. The analysis seeks to separate low-rank trends from sparse outliers within a data matrix:



https://bit.ly/3UwAFi6

Robust PCA: some math

- We want to obtain:

$$X = L + S$$

- The basic idea:

$$\min_{L,S} \text{rank}(L) + ||S||_0$$

not convex, thus hard to optimize.

- Idea - convex relaxation:

$$\min_{L,S} ||L||_* + \lambda ||S||_1$$

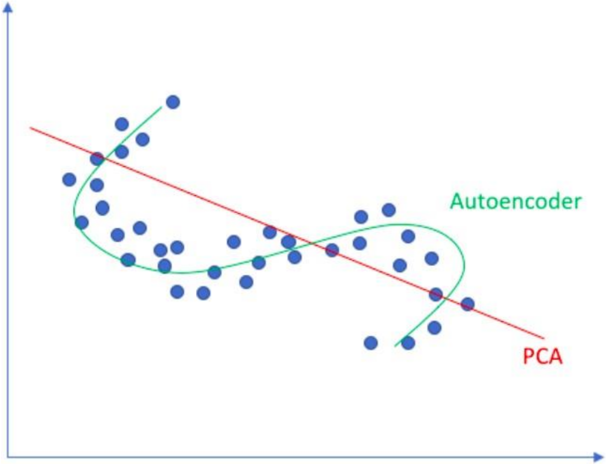
$||.||_*$ is nuclear norm, given by the sum of singular values, which is a proxy for rank

The relaxed solution converges almost always to exact one, from theory:

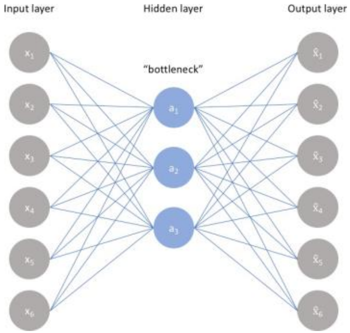
$$\lambda = {}^{1}\!/\!\sqrt{\max(n, N)}$$

Nonlinearities

What if we have a more complicated signal manifold?



Autoencoders



Two parts of the network:

- encoder $h = f(x)$;
- decoder $r = g(h)$

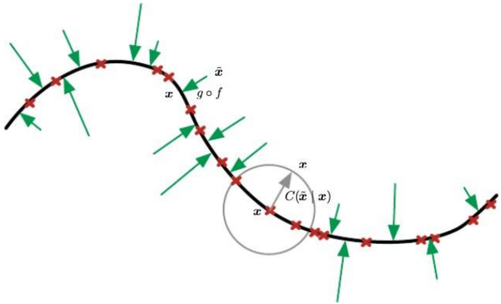
Generally, we want to find a transformation

$$g(f(x)) = x$$

The approach can be made more flexible than PCA transform.

AE: learning manifold

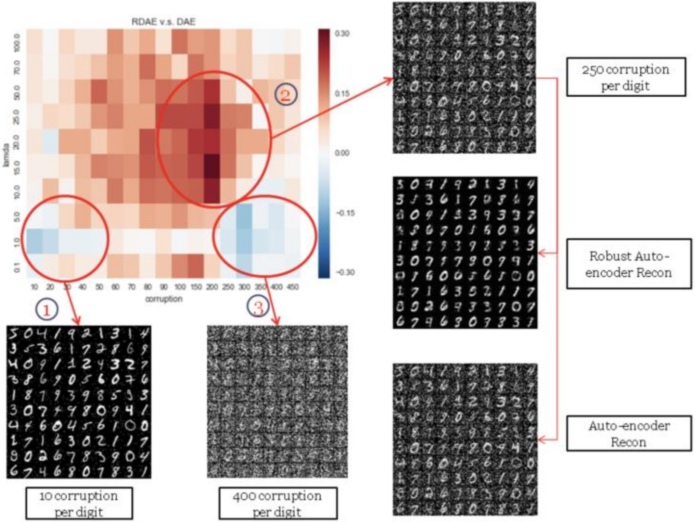
In fact, we learn a manifold, where normal class is situated:



We can keep the same anomaly score as in PCA case.

Robust Deep Autoencoder

- same problem & approach as in the Robust PCA case;
- same regularisation using sparse matrix S;
- can be learned iteratively;
- shows the difference between error rates for the features constructed by a normal autoencoder and a RDA. Red indicates where the error rates of the RDA are superior to those of the normal autoencoder, and blue indicates the opposite.



https://bit.ly/3UwAFi6

Variational Autoencoders

- ”normal” manifold can be created with probabilistic model;
- anomaly score remains distance based but we can sample from ”normal” distribution several events and average the distance.

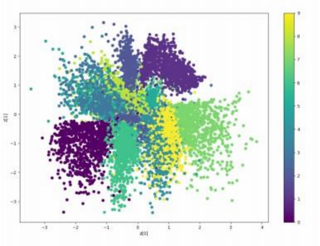


Figure 2.11: 2D plot of (variationally)autoencoded digits.

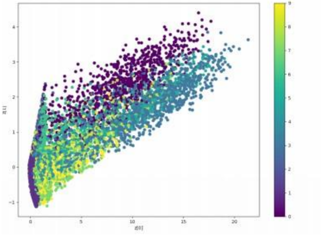


Figure 2.12: 2D plot of autoencoded digits.

Recap

- Linear methods are quite powerful for anomaly detection.
- Most of the analysis is done in the latent space.
- Issues:
  - data need to be correlated and not heavily clustered;
  - might be overfit;
  - lacks interpretability.

Probability Scores

Generative modeling

- Some generative modeling produce expilcit estimate of probability of sample:
  - Variational autoencoders.
  - Flow-based models.
- Can we use it to find anomaly?

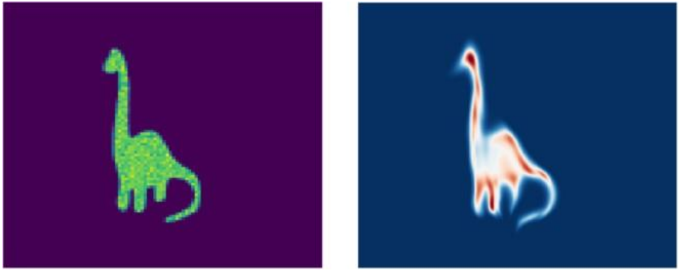


Figure: N. Schucher

Constructing Score Function

- direct probability is overly optimistic for anomalous samples (tail problem!);
- one can try to construct a different probability-based measure:
  - Watanabe-Akaike Information Criterion;
  - use in-batch dependencies.
- empirically these approaches work better.

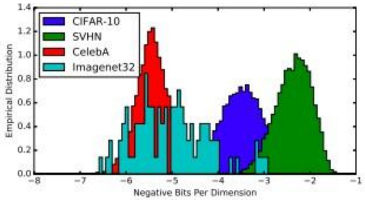


Figure 1. Density estimation models are not robust to OoD inputs. A GLOW model (Kingma & Dhariwal, 2018) trained on CIFAR-10 assigns much higher likelihoods to samples from SVHN than samples from CIFAR-10. .

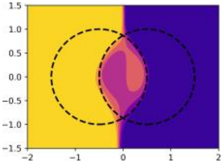
H. Choi, WAIC, but Why. Generative Ensembles for Robust Anomaly Detection

Advanced Ideas

$(1+\varepsilon)$-class classification

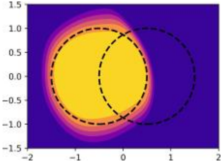
Two-class classification:

- undefined in empty regions;
- recovers proper probabilities;



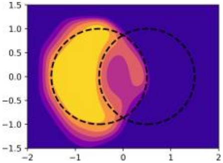
One-class classification:

- defined everywhere;
- ignores negative class;



$(1+\varepsilon)$**-class classification:**

- shifts two-class solution towards a one-class solution;



Approach Classification

Numerous approaches have been developed:

- Extreme value analysis (Z-score).
- Probabilistic and statistical models (Generative models).
- Linear models (Principle Component Analysis)/
- Proximity-based models (Clustering)
- Information theoretic models (Minimal Description Analysis).
- High-dimensional outlier detection (isolation forest).

Methods can be combined into sequential and independent ensembles.

C. Aggarwal, Outlier Analysis

LHC Olympics 2020, https://lhco2020.github.io/homepage/

Participants were offered two types of data (you could read more about the data here):

- “Monte Carlo Simulation Background” — simulated data that does not have a signal, where the physics and simulation of the detector are not entirely correct;
- “Data” — LHCO 2020 black boxes, which may contain some new signals, were revealed to the participants during this challenge.

Participants were required to report the following:

- A p-value associated with the dataset having no new particles (null hypothesis);
- As complete a description of the new physics as possible. For example, the masses and decay modes of all new particles (and uncertainties on those parameters);
- How many signal events (+uncertainty) are in the dataset (before any selection criteria).

Methods suggested:

- Deep Ensemble Anomaly Detection (combined with a mixture of neural networks with convolutional layers and Boosted Decision Trees to assign event probabilities in the signal or background categories,
- GAN-AE, VAE (see tomorrow)

Summary

- Anomaly detection problem attracts a lot attention both from researchers and practitioners communities.
- Method should be selected based on the problem to be analysed.
- Methods span wide range of families, hence always room for hyperparameter tuning.
- Many recent development in this area.

Thank you!

andrey.ustyuzhanin@constructor.org

anaderiRu

Andrey Ustyuzhanin