Denis Derkach

Anomaly Detection: Basic Methods

2021

Definition and Examples

Outliers, Anomalies, Novelties

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

- noise;
- novelties;
- anomalies.

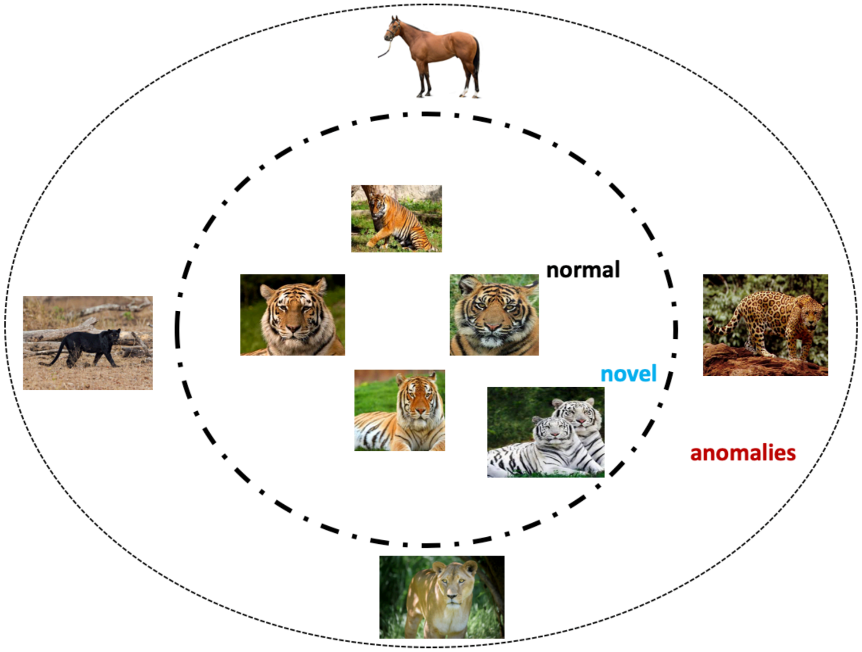
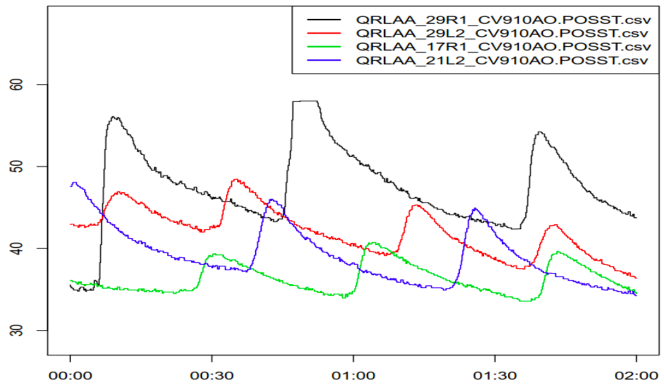


Image: R. Chalapathy and S. Chawla, Deep Learning for Anomaly Detection: A Survey

Example: LHC Cryogenic System



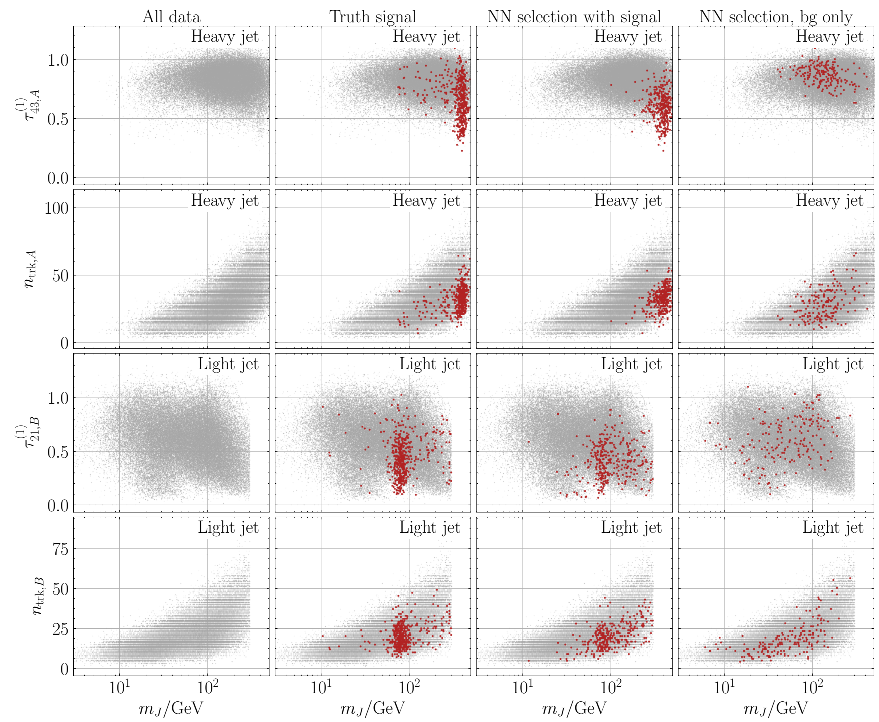
- faulty valve behaviour: range of movement if compared to the other actuators;
- immediately seen in data.

- Most obvious example of problem statement;
- anomaly points to a change in state of the system;
- anomalies can be defined as significant deviation from the data sample collected.

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

Example: New Physics as Anomaly

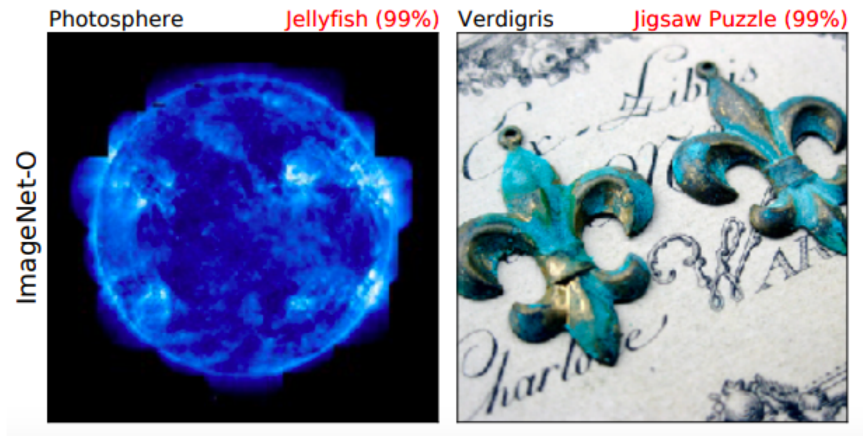
- Anomaly is our signal;
- need to analyse abundance of non-anomalous events;
- signal position is unknown.



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

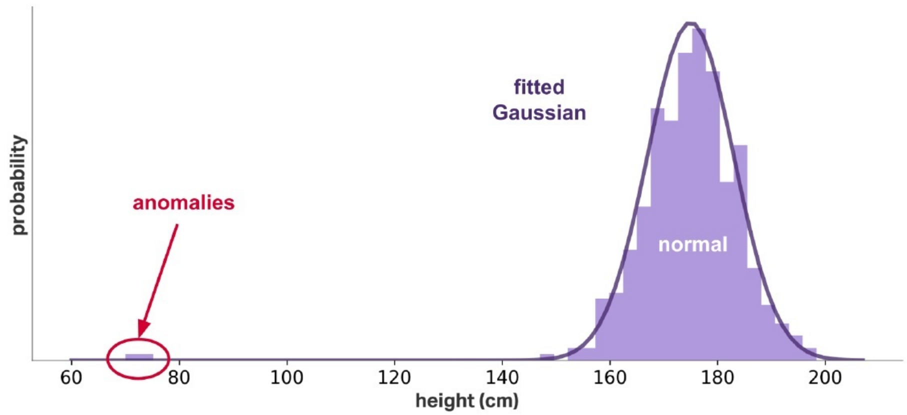
Dataset Properties

- 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 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 test instance is assigned an **anomaly score**.
    - allows outputs to be ranked
    - requires an additional threshold parameter

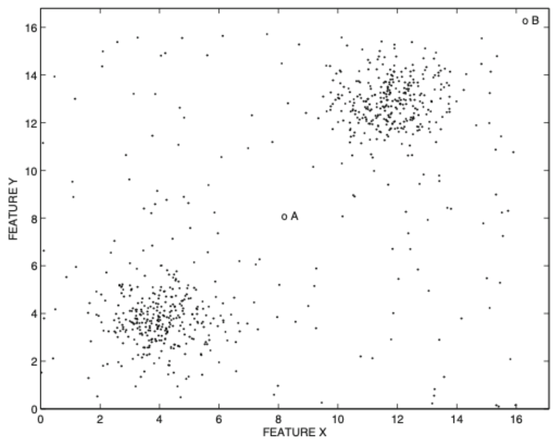
Data Model is Everything



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

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

Data Model is Everything



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

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

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

Outlier Method Evaluation

- precision at given recall;
- average precision;
- ROC AUC score;
- PR AUC score.

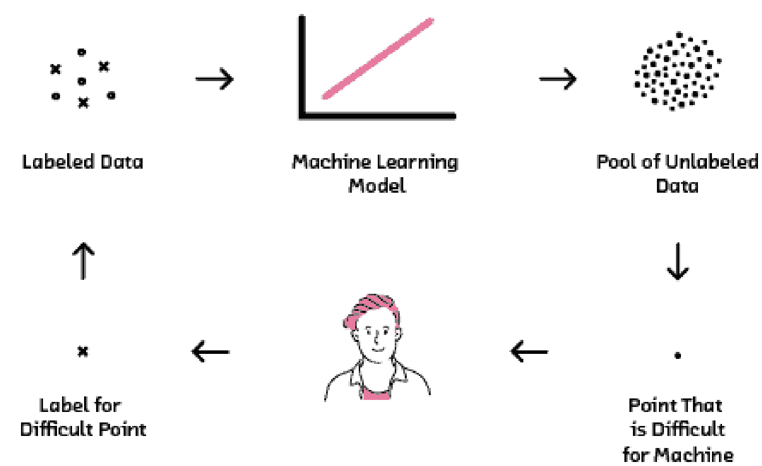
Basic methods

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?

Active learning for anomaly detection

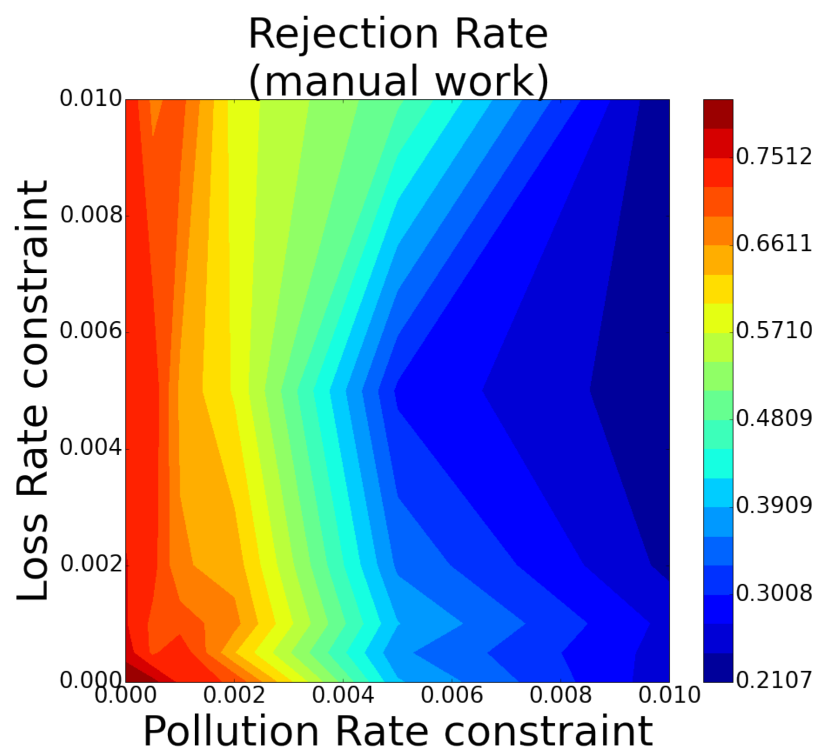
- for continuous data flow, use active learning:
  - train algorithm on existing labels;
  - check on new samples arriving;
  - ask experts to label only new examples, where classifier was not sure;
  - train new classifier.
- obtained classifier will be better in identifying anomalies.



D Pelleg, Active Learning for Anomaly and Rare-Category Detection
Figure from Cloudera blog

Example: CMS Data Certification

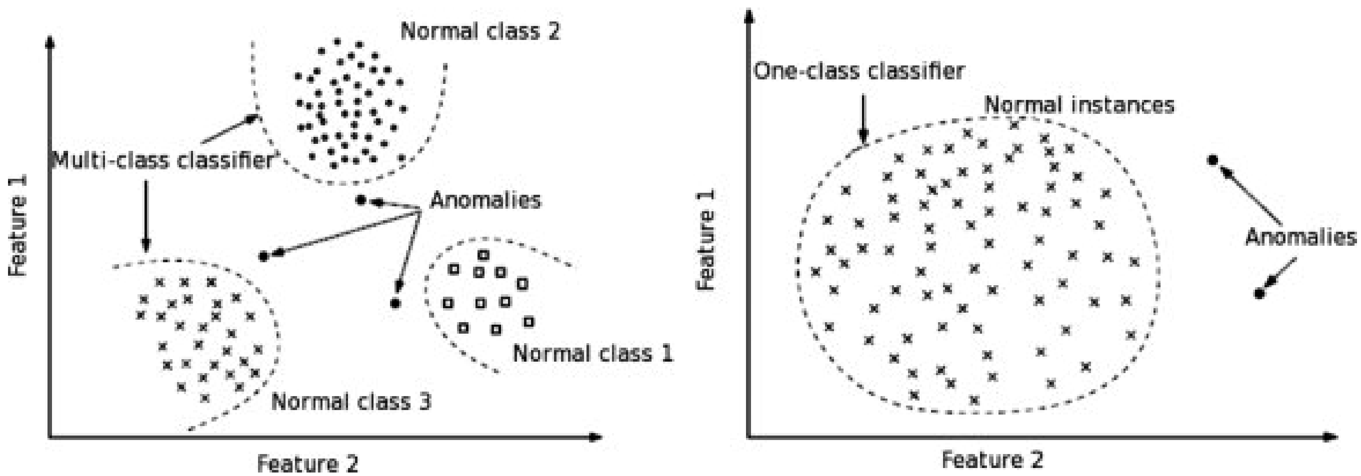
- CMS data certification problem:
  - 2010 CMS data, OpenData portal;
  - manually labeled;
- can be successfully employed in DQM settings;
- approach is able to save up to 20% manual work under tight restrictions;
- quality improves over time.



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

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.

Figure M. Chica Authentication <...>

One-class family

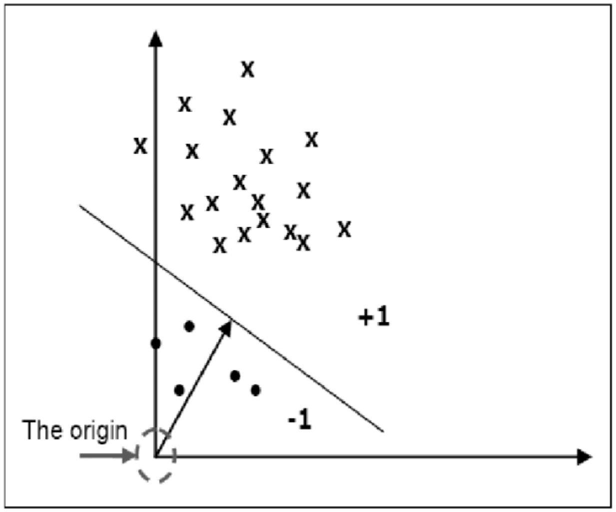
Table 1.1: Classification methods and their unsupervised analogs in outlier analysis

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

C. Aggarwal, Outlier Analysis

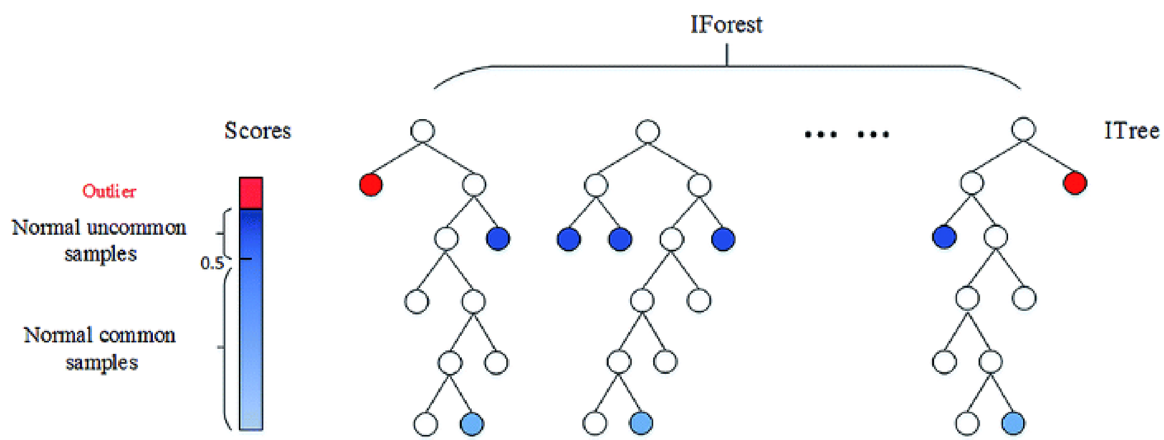
One-class Support Vector Machines

- Treat the origin as the only member of the second class.
- General idea: separate data points from origin and maximize the gap between hyperplane to the origin.
- Anomaly score: signed distance to the separating hyperplane.



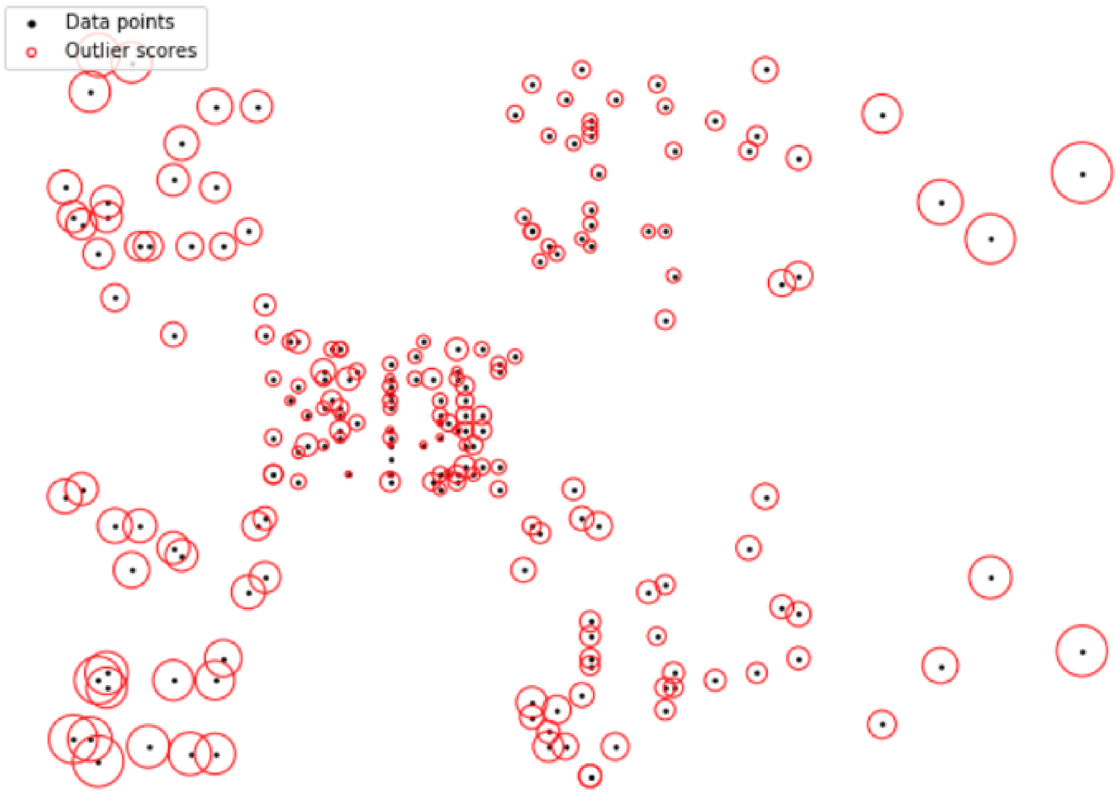
Isolation Forest

- General idea: split the sample using random projection (like in random forest case).
- Grow the tree until complete isolation of experimental points.
- Anomaly score: proporional to number of splitting needed to separate the point, averaged over a forest of such random tree.

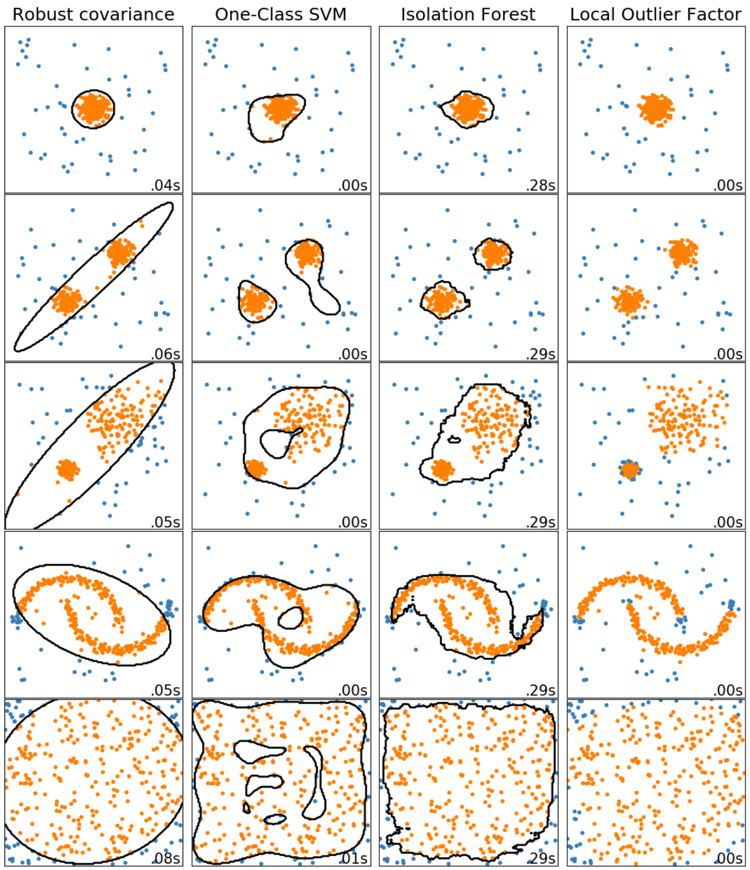


Local Outlier Factor

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



Comparison of One-class Techniques



Wrap-up

- 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.