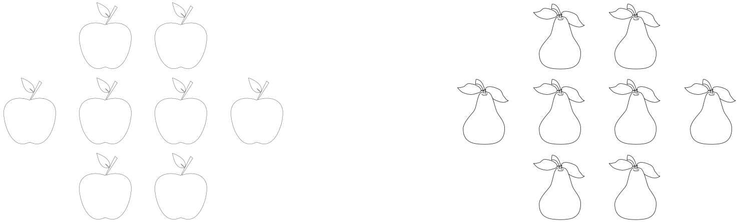
Outlier and anomaly detection

Tomáš Pevný

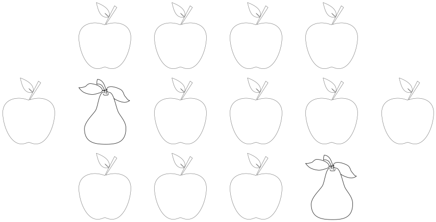
Department of Computers, Czech Technical University

December 2, 2019

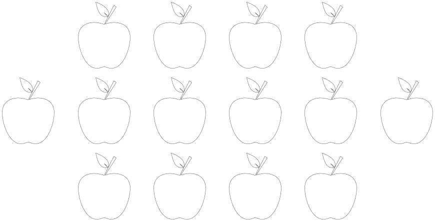
Training set for supervised binary classification



Training set for unsupervised anomaly classification



Training set for supervised anomaly classification



What are the advantages?

You can detect new types of fruit.

Definition of anomaly

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior[1].

An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism[2].

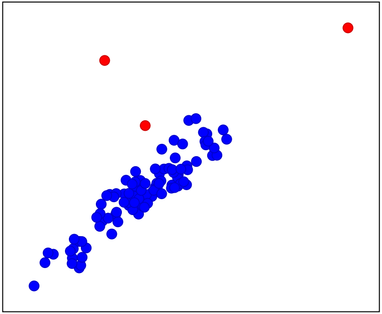
1 *V. Chandola, A. Banerjee, and V. Kumar, Anomaly detection: a survey, 2009*

2 *D. M. Hawkins, Identification of Outliers, 1980*

Formal definition of outliers / anomalies?

Outliers

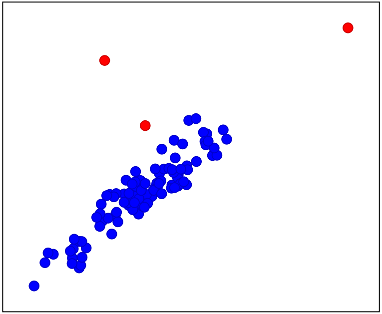
- have different statistical properties,
- or they are in low-density regions,
- or they are far from majority.


;

Formal definition of outliers / anomalies?

Outliers

- have different statistical properties,
- or they are in low-density regions,
- or they are far from majority.

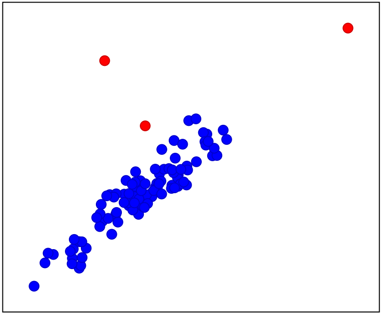

;

Definition of outliers influences the method.

Formal definition of outliers / anomalies?

Outliers

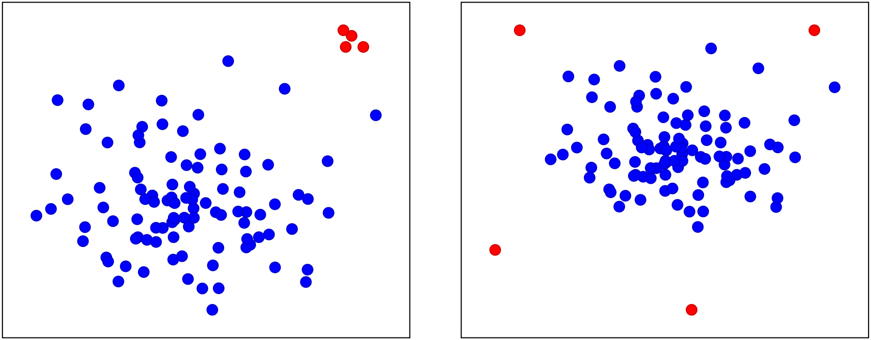
- have different statistical properties,
- or they are in low-density regions,
- or they are far from majority.



;

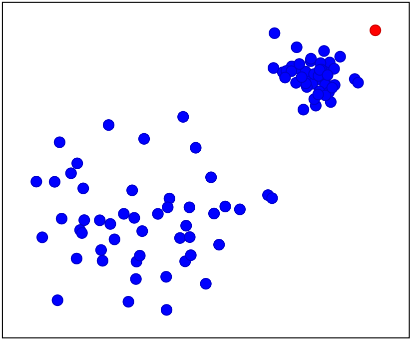
Definition of outliers is application dependent.

Types of anomalies

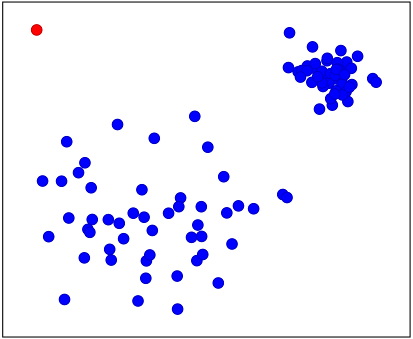


concentrated scattered

Types of anomalies



local



global

Taxonomy

- supervised vs. unsupervised
- model centric vs. data centric

Parzen window estimator — motivation

Estimate probability density and identify points in areas of low density.

E. Parzen, On Estimation of a Probability Density Function and Mode, 1962

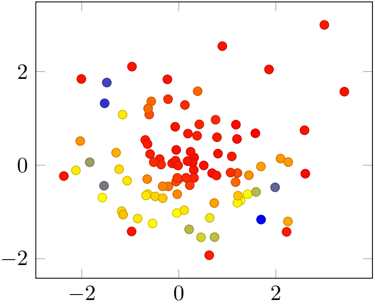
Parzen window estimator — calculation

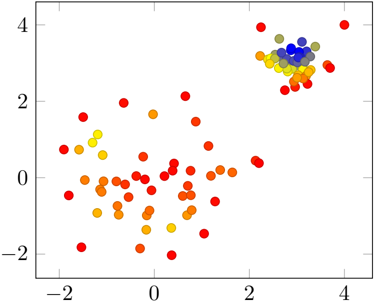
The density in point x is estimated from training points $\{x_i\}_{i=1}^N$ as

$$f(x) = \frac{1}{hN} \sum_{i=1}^{N} k\left(\frac{x - x_i}{h}\right),$$

where k is the kernel (e.g. Gaussian kernel $k(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$).

Parzen window estimator — example





Estimate probability density in each point.

E. Parzen, On Estimation of a Probability Density Function and Mode, 1962

K-nearest neighbor — motivation

Outliers are far from points / they have "empty" neighbourhood.

S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, 2000

K-nearest neighbor — calculation

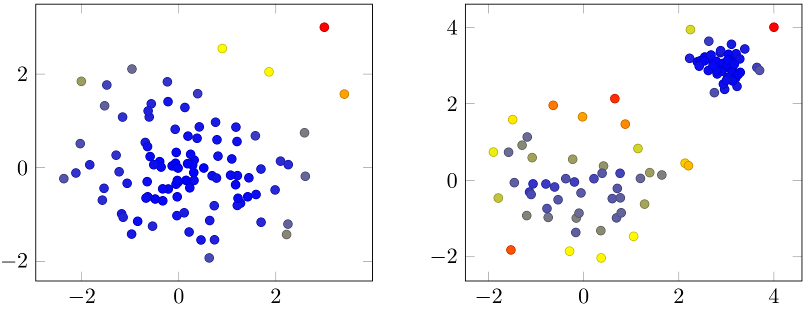
1. For sample $\{x_i\}_{i=1}^{N}$ calculate its distance to k^{th} nearest neighbor.
2. Return fraction p of samples as outliers.

Variants differs by calculating score:

- mean distance to all,
- distance to mass.

S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, 2000

K-nearest neighbor — example



S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, 2000

Local outlier factor — motivation

Outliers have low density with respect to its k neighborhood.

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, Lof: Identifying density-based local outliers, 2000.

Local outlier factor — calculation

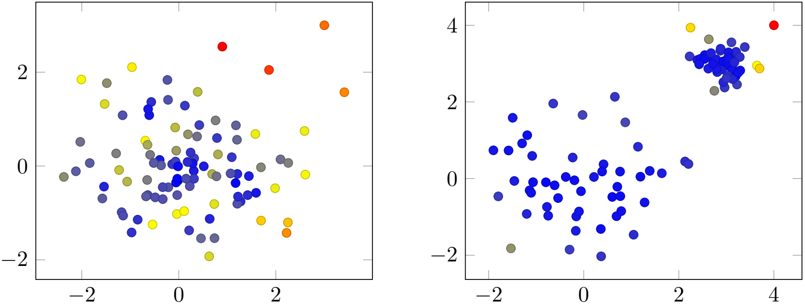
1. For every $\{x_i\}_{i=1}^{N}$ estimate the local density, $ld_k(x_i)$, as an inverse of average robust distance to k nearest neighbor.
2. Compare density of x_i with that of its k nearest neighbors, P_k,

$$lof_k(x_i) = \frac{1}{k} \sum_{x \in P_k} \frac{\mathrm{ld}_k(x)}{\mathrm{ld}_k(x_i)}.$$

3. The robust distance is calculated as

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, Lof: Identifying density-based local outliers, 2000.

Local outlier factor — example



M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, Lof: Identifying density-based local outliers, 2000.

Angle-based outlier detection — motivation

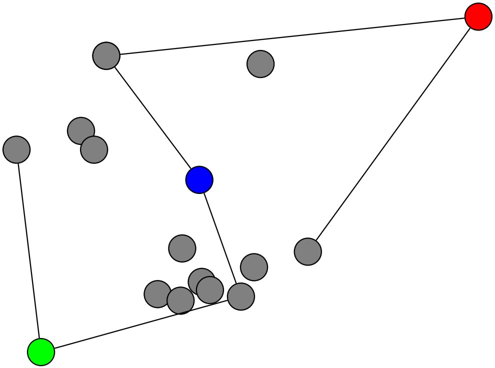
- Angles are more stable than distances in high dimensions.
- Object o is an outlier if most other objects are located in similar directions, it is on the border.
- Object o is an inlier if most other objects are located in varying directions, it is in the middle.

N. Pham, R. Pagh, A Near-linear Time Approximation Algorithm for Angle-based Outlier Detection in High-dimensional Data, 2012.

Angle-based outlier detection — motivation

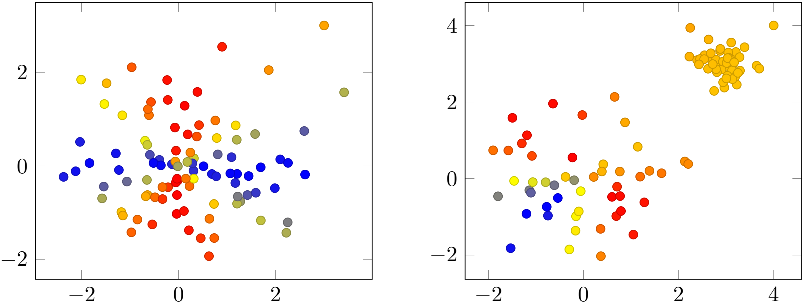
$$\mathrm{abod}(x_i) = \operatorname*{var}_{k,l \neq i} \frac{\langle x_i - x_k, x_i - x_j \rangle}{\|x_i - x_k\| \, \|x_i - x_j\|}.$$

Angle-based outlier detection



;

Angle-based outlier detection — example



Parametric anomaly detection — motivation

Robustly fit a known distribution and identify points with low probability.

Parametric anomaly detection

Multivariate Gaussian distribution

- Assumes that data follows

$$x \sim |\Sigma|^{-1}(2\pi)^{-\frac{d}{2}} e^{-(x-\mu)^{\mathrm{T}}\Sigma(x-\mu)}$$

- Related to principal component analysis and Mahalanobis distance.

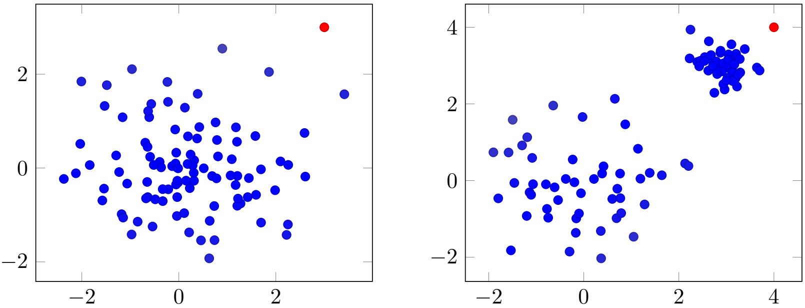
Mixture of multivariate Gaussian distributions

- Assumes that data follows

$$x \sim \sum_{j=1}^{m} w_j |\Sigma_j|^{-1}(2\pi)^{-\frac{d}{2}} e^{-(x-\mu_j)^{\mathrm{T}}\Sigma(x-\mu)}$$

- Difficult to fit.

Parametric anomaly detection — example



Density level estimation

Find the area of minimal volume, such that α fraction of probability mass is outside.

Density level estimation

$$\arg\min_{f \in \mathscr{F}, \lambda} \mathrm{Vol}(U_{f,\lambda}) = |\{x | f(x) \geq \lambda\}|$$

subject to

$$\int_{\mathscr{X}} f(x) p(x) dx \geq 1 - \alpha$$

where $\mathscr{F}$ is a class of probability density functions defined on $\mathscr{H}$.

One-class support vector machines — motivation

Estimates the support of the probability distribution allowing at most ν false positive rate.

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. Smola, R. C. Williamson, Estimating the support of a high-dimensional distribution, 2001

One-class support vector machines — calculation

training:

$$\arg\min_{w\in\mathbb{R}^d,\rho} \frac{1}{2}\|w\|^2 - \rho + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i$$

subject to

$$\begin{aligned} \langle w, x_i\rangle &\geq \rho - \xi_i \\ \xi_i &\geq 0. \end{aligned}$$

classification:

$$f(x) = \langle w, x_i\rangle - \rho > 0$$

Finds the hyper-plane separating the data from the origin with the highest margin, allowing at most ν misclassified points.

One-class support vector machines — calculation

training:

$$\arg\min_{w\in\mathbb{R}^n,\rho} \frac{1}{2}\sum_{i,j=1}^{n,n} \alpha_i\alpha_j k(x_i,x_j) - \rho + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i$$

subject to

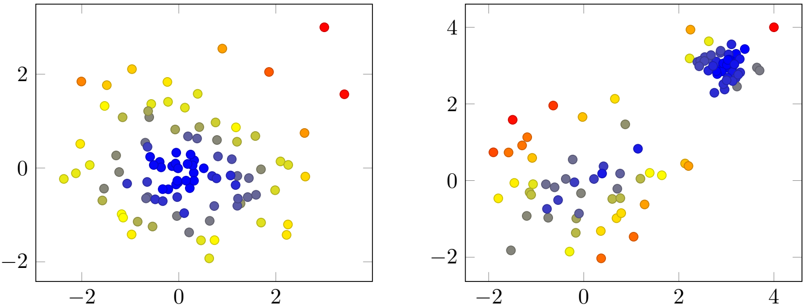
$$\begin{aligned} \sum_{j=1}^{n}\alpha_i k(x_j,x_i) &\geq \rho - \xi_i \\ \xi_i &\geq 0. \end{aligned}$$

classification:

$$f(x) = \alpha_i k(x_j,x) - \rho > 0$$

$k(x_i,.)$ is a feature map induced by the chosen kernel,
most popular choice is $k(x,x') = e^{-\gamma\|x-x'\|^2}$.

One-class support vector machines — Example

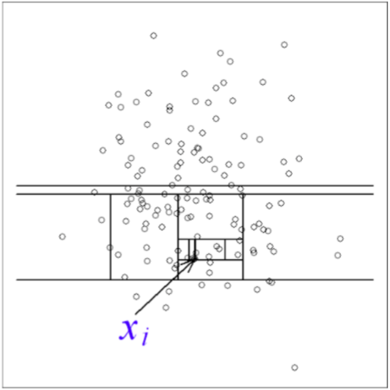


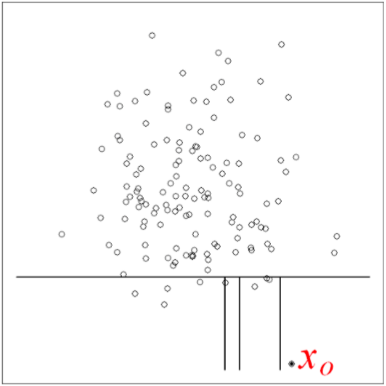
Isolation Forest — motivation

Anomalous points should be close to the root in randomly constructed tree.

F. T. Liu, K. M. Ting, Z. H. Zhou, Isolation Forest, 2008

Isolation Forest — Example





Isolation Forest — calculation

The anomaly score a sample x is defined as

$$s(x) = 2^{-\frac{E(h(x))}{c(n)}},$$

where

- $h(x)$ is depth of list containing x
- $c(n)$ is the average path length of unsuccessful search in binary search tree with n items

$$c(n) = 2H(n-1) - 2\frac{n-1}{n}$$

- $H(i) \approx ln(i) + 0.5772156649$

Frac: Supervised approach to anomaly detection — motivation

A dependency structure among features is violated for anomalies.

K. Noto, C. Brodley, D. Slonim, FRaC: Feature-modeling approach for semi-supervised and unsupervised anomaly detection, 2012

Frac: Supervised approach to anomaly detection — calculation

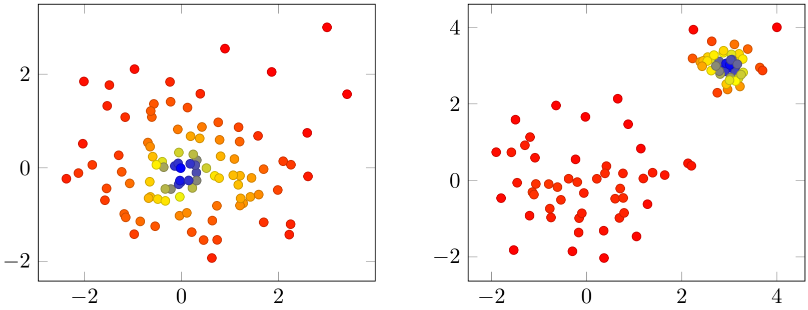
- Build a predictor of each feature x_i on rest $x_{\sim x_i}$.
- Score is proportional to the probability of estimation error

$$s(x) = \frac{1}{d}\sum_{i=1}^{d} \log p_i(x_i - o_i(x_{\sim i})),$$

where

  - $p_i(e)$ is the probability of i^{th}- estimator making error e
  - $o_i(x_{\sim i})$ output of i^{th} estimator of x_i from $x_{\sim i}$.

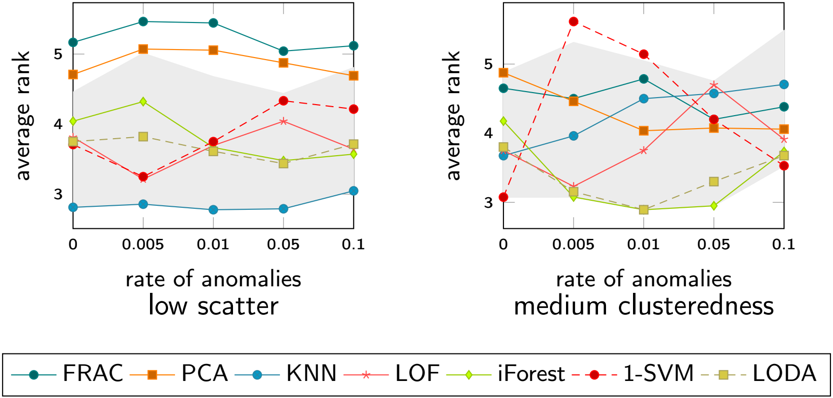
Frac: Supervised approach to anomaly detection — motivation



Experimental comparison

Comparing different methods is difficult due to lack of benchmarking problems.

Experimental comparison



Lower average rank is better.

Anomaly detection on data-streams

- Most prior art adapts batch-based algorithms by floating window or by alternating models.
- Some methods assumes continuity of data streams.

Experimental comparison

	continuous		two histograms	
dataset	AUC	time	AUC	time
covertype	0.972	4.42	0.989	3.00
http - 3	0.992	7.51	0.994	5.24
http	0.991	8.40	0.993	6.00
shuttle	0.980	0.49	0.994	0.41
smtp	0.970	1.34	0.994	1.06
smtp -3	0.871	1.35	0.886	1.11
smtp + http	0.989	9.65	0.993	7.99

Tips for successful anomaly dataction

- Understand the domain:
  - types of anomalies
  - rate of anomalies

Tips for successful anomaly dataction

- Understand the domain:
  - types of anomalies
  - rate of anomalies
- You will not get away from labelling.

Explaining the anomaly

Explaining why anomaly happened might be an invaluable information to the analyst.

The main idea

- Outliers should be separable in
  - in *few dimensions*
  - with a *large margin*.
- They should be separable by a tree of small height.
- Training multiple trees increases robustness.

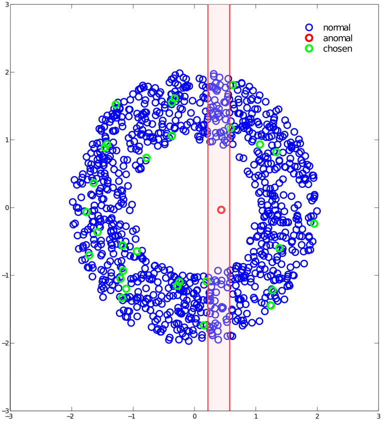
Explaining the anomaly

Summary of the Explainer algorithm

$labels \leftarrow anomalyDetector(data)$
$SRF \leftarrow \{\varnothing\}$
for all data($labels$ ==anomaly) **do**
  $T \leftarrow createTrainingSet(size)$
  $t \leftarrow trainTree(T)$
  $SRF \leftarrow t$
end for
$extractFeatures(SRF)$
$extractRules(SRF)$

Training the tree

1. Select dimension removing
    - most normal samples
    - with highest margin.
2. Repeat until sample is separated.
3. Path to leaf with anomalous sample indicates separating features.



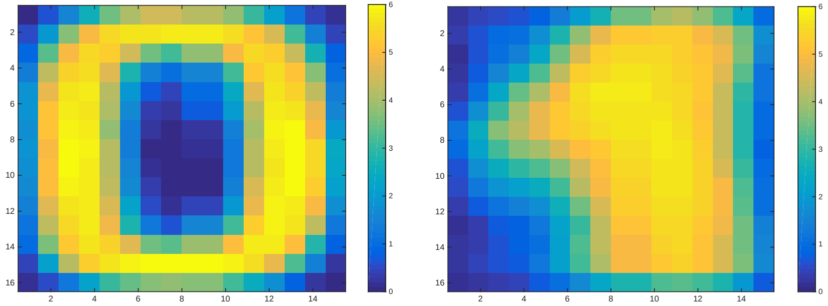
Extracting the features

- To increase robustness, train multiple trees.
- Each tree provides set of features.
- Pick the most frequent ones.

Min provides explanation using the minimal set of features.

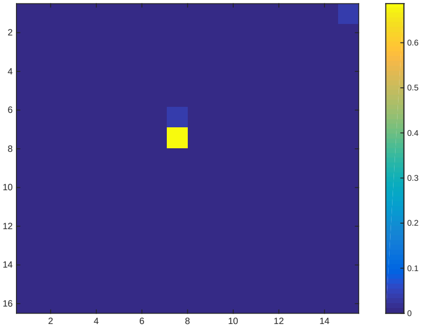
Max returns all features in which the anomaly can be detected.

Example of explanation

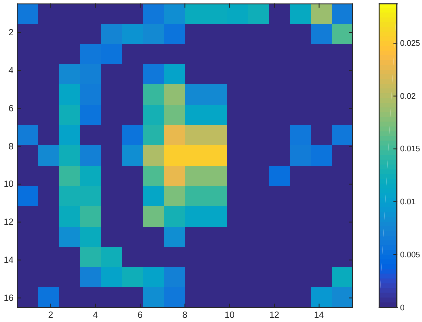


Average zero vs. average one

Features provided by the explainer



min explanation



max explanation

Summary

- Anomaly / outlier detection is not a magic bullet.
- Know strength and weaknesses of algorithm you chose.
- Learn about domain (type of anomalies).
- Anomalies might not be anomalies of interest.