A Review of Machine Learning Techniques for Anomaly Detection

Project MIRAI - Meeting T2.2/T2.3

October 12, 2021 - Online

Pedro Santos (ISEP) – pss@isep.ipp.pt

Luís Almeida, Mário Sousa, Pedro Souto, Ricardo Morla (FEUP), Luís Lino Ferreira (ISEP)

Outline

- What is an anomaly?
- Types of Anomalies
- Data Availability & Training Mode
- One-class Classification
- A Review of Selected Anomaly Detection (AD) Techniques
- Anomaly Detection in Univariate Time-series

What is an anomaly?

R. Chalapathy and S. Chawla, 'Deep Learning for Anomaly Detection: A Survey', arXiv:1901.03407, Jan. 2019, Accessed: Aug. 04, 2021. [Online]. Available: http://arxiv.org/abs/1901.03407.

Anomalies

- Anomalies can be caused by errors in the data but can also be indicative of a new, previously unknown, underlying process.
- Hawkins [1980] defines **an outlier as an observation that deviates so significantly from other observations as to arouse suspicion that it was generated by a different mechanism.**
- Outliers are both rare and unusual:
  - **Rarity** suggests that they have a low frequency relative to non-outlier data (so-called inliers).
  - **Unusual** suggests that they do not fit neatly into the data distribution.

Novelties

- Novelties are also novel (new) or unobserved patterns in the data, that were not considered as anomalous data points.
- A novelty score may be assigned for these previously unseen data points, using a decision threshold score.
- The points which significantly deviate from this decision threshold may be considered as anomalies or outliers.
- The techniques used for anomaly detection are often used for novelty detection and vice versa.

Types of Anomaly

Point anomalies

- Point anomalies often represent an irregularity or deviation that happens randomly and may have no particular interpretation.
- For example, a big credit transaction which differs from other transactions is a point anomaly.

Contextual or Conditional Anomalies

- Some points can be normal in a certain context, while detected as anomaly in another context.
- Having a daily temperature of 35 C in summer in Germany is normal, while the same temperature in winter is regarded as an anomaly.

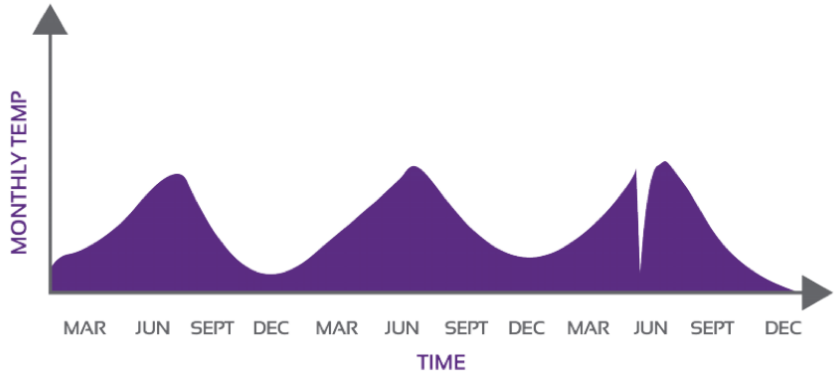
Collective or Group Anomalies

- There are cases where individual points are not anomalous, but a sequence of points are labeled as an anomaly.
- For example, a bank customer withdraws $500 from her bank account every day of a week. Although withdrawing $500 occasionally is normal for the customer, a sequence of withdrawals is an anomalous behavior.

May-22	1:14 pm	FOOD	Monaco Café	$1,127.80	→ Point Anomaly
May-22	2:14 pm	WINE	Wine Bistro	$28.00	
...					
Jun-14	2:14 pm	MISC	Mobil Mart	$75.00	
Jun-14	2:05 pm	MISC	Mobil Mart	$75.00	
Jun-15	2:06 pm	MISC	Mobil Mart	$75.00	
Jun-15	11:49 pm	MISC	Mobil Mart	$75.00	
May-28	6:14 pm	WINE	Acton shop	$31.00	→ Collective Anomaly
May-29	8:39 pm	FOOD	Crossroads	$128.00	
Jun-16	11:14 am	MISC	Mobil Mart	$75.00	
Jun-16	11:49 am	MISC	Mobil Mart	$75.00	

Point and Collective Anomalies – Card Transactions

R. Chalapathy and S. Chawla, 'Deep Learning for Anomaly Detection: A Survey', arXiv:1901.03407, Jan. 2019, Accessed: Aug. 04, 2021. [Online]. Available: http://arxiv.org/abs/1901.03407



Contextual Anomaly - Temperature data

Michael A Hayes and Miriam AM Capretz. Contextual anomaly detection framework for big sensor data. Journal of Big Data, 2(1):2, 2015.

Data Availability & Training Mode

- Anomalies are rare occurrences, so having labeled datasets is often hard to obtain their labels.
- Also, anomalous behavior may change over time, for instance, the nature of anomaly had changed so significantly and that it remained unnoticed.

Supervised anomaly detection

- Presumes existence of annotated dataset with **labels of both normal and anomalous data** instances.
- Problem maps into 'traditional' binary / multi-class classification problem; as such, well-established classification methods can be used.

Semi-supervised deep anomaly detection

- In practice, **datasets of (labelled) normal instances can be obtained** with some ease, whereas it is the anomalous data points that are harder to obtain.
- These techniques leverage existing labels of single (normally positive class) to separate outliers.

Unsupervised deep anomaly detection

- **Labelled data is hard to obtain in the first place.**
- Unsupervised deep anomaly detection techniques detect outliers solely based on intrinsic properties of the data instances.

Availability of labelled data per training mode

	Normal	Anomalous
Superv.	Y	Y
Semi-superv.	Y	N
Unsuperv.	N	N

Semi-supervised & Unsupervised

Semi-supervised

- *Operating Principle*
  - Semi-supervised or one-class classification anomaly detection (AD) techniques assume that all training instances have only one class label.
  - Semi-supervised techniques learn a discriminative boundary around the normal instances, and test instances that do not belong to the majority class are flagged as anomalous.
- *Assumptions*
  - Proximity and continuity: Points which are close to each other both in input space and learned feature space are more likely to share the same label.
  - Robust features can be learned to separating normal from outlier data points.

Unsupervised

- *Operating Principle*
  - Unsupervised anomaly detection algorithm produces an outlier score of the data instances based on intrinsic properties of the dataset, such as distances or densities.
- *Assumptions*
  - "Normal" regions in the original or latent feature space can be distinguished from "anomalous" regions in the original or latent feature space.
  - The majority of the data instances are normal compared to the remainder of the data set.

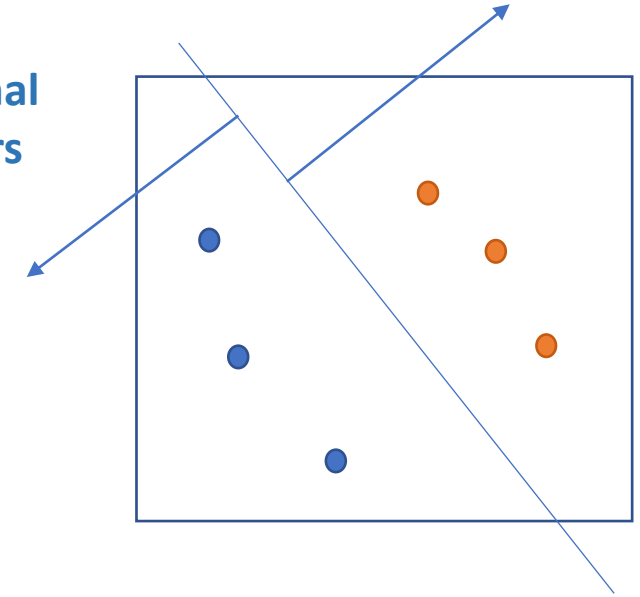
Trade-offs

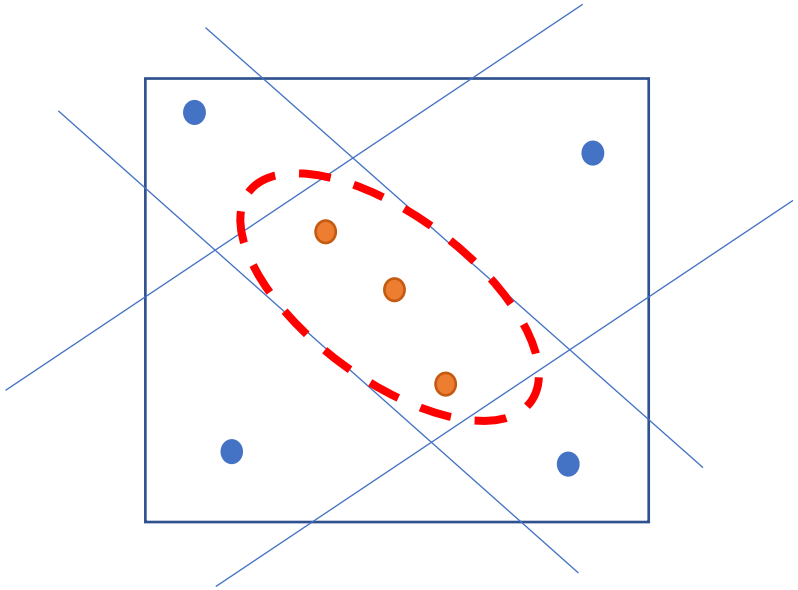
	Advantages	Disadvantages
Semi-Supervised	Use of labeled data (usually of one class) can produce **considerable performance improvement over unsupervised techniques**.	The features extracted may not be representative of fewer anomalous instances and hence **prone to the over-fitting problem.**
Unsupervised	• **Learns the inherent data characteristics** to separate normal from an anomalous data point. • **Cost-effective technique** to find the anomalies since it does not require annotated data for training the algorithms.	• Often challenging to learn commonalities within data in a complex and high dimensional space. • Unsupervised techniques are **sensitive to noise and data corruptions** • Less accurate than supervised or semi-supervised techniques.

One-class vs. N-class Classification

- One-class classification (OCC) tries to identify objects of a specific class amongst all objects, by primarily learning from a training set containing only the objects of that class [1]
- This is different from and more difficult than the traditional classification problem, which tries to distinguish between two or more classes with the training set containing objects from all the classes.

Traditional Classifiers





One-Class Classifiers

A Review of Selected Anomaly Detection (AD) Techniques

MIRAI isep Instituto Superior de Engenharia do Porto P.PORTO U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

Selected Anomaly Detection Techniques

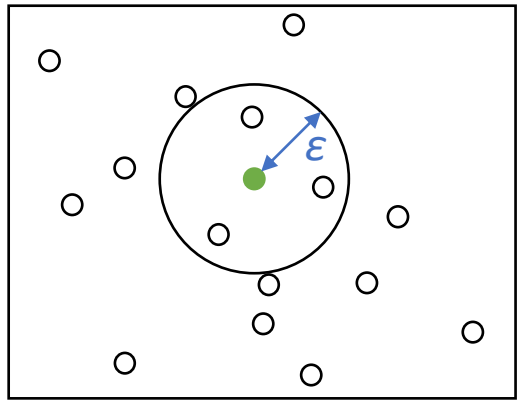
1. Clustering: **Density-based Clustering** (unsupervised)
2. Neural Network: **Autoencoders** (semi-supervised)
3. Dimensionality reduction: **Principal Component Analysis** (unsupervised)
4. Boundary-based: **One-Class Support Vector Machines** (semi-supervised)
5. Partition-based: **Isolation Forest** (unsupervised)

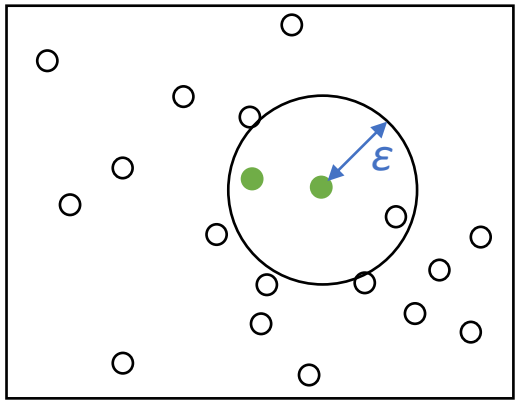
1. Density-based Clustering

J. Sander, M. Ester, H.-P. Kriegel, and X. Xu, 'Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications', p. 30.

- Unsupervised technique, often called as **DBSCAN**
- A point *p* is a *core point* if **at least *minPts* points are within distance *ε* of it**.
- Unclustered points can be interpreted as anomalies.

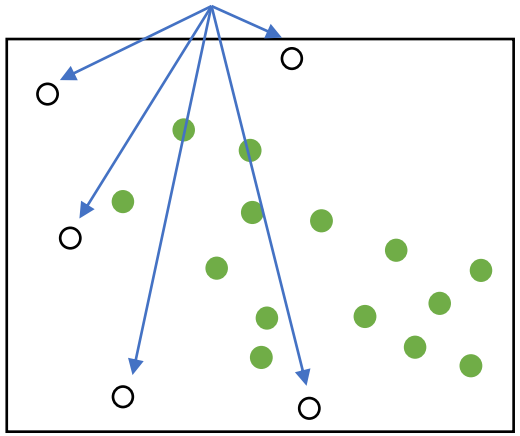
minPts = 2





...

Anomalies



2. Auto-Encoder (Neural Network)

B. Kiran, D. Thomas, and R. Parakkal, 'An Overview of Deep Learning Based Methods for Unsupervised and Semi-Supervised Anomaly Detection in Videos', J. Imaging, vol. 4, no. 2, p. 36, Feb. 2018

What are Auto-encoders?

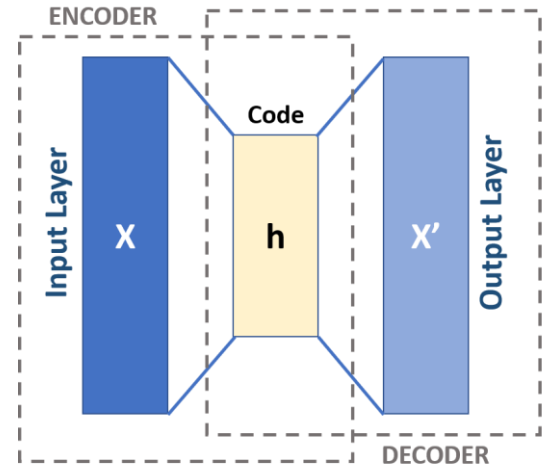
- Autoencoders represent data within multiple hidden layers by reconstructing the input data, effectively **learning an identity function**.
- The autoencoder learns a representation (encoding) for a set of data, typically for dimensionality reduction, by training the network to ignore insignificant data ("noise").

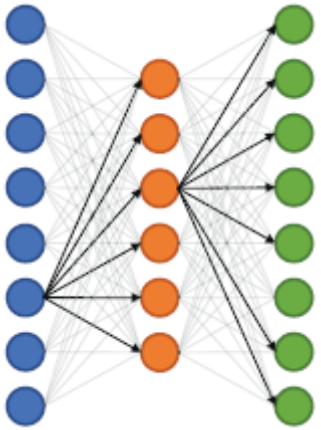
Used for Anomaly Detection

- The autoencoders, when trained solely on normal data instances (which are the majority in anomaly detection tasks), fail to reconstruct the anomalous data samples.
- **The data samples which produce high residual errors are considered outliers.**

How to use?

- The choice of autoencoder architecture depends on the nature of data. Convolution networks are preferred for image datasets. Long short-term memory (LSTM) based models tend to produce good results for sequential data.
- The choice of right degree of compression, i.e., dimensionality reduction is often a hyper-parameter that requires tuning for optimal results.



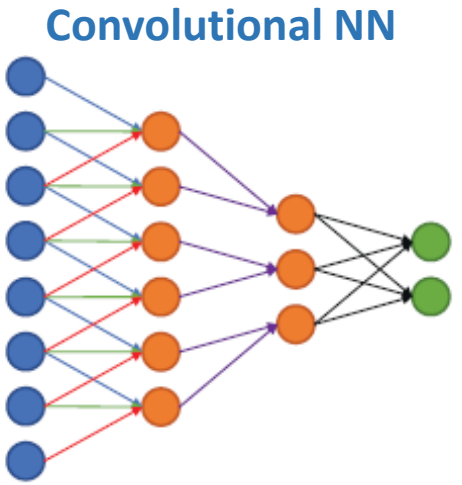


Auto-encoder

2.1 Composed (Hybrid) Architectures

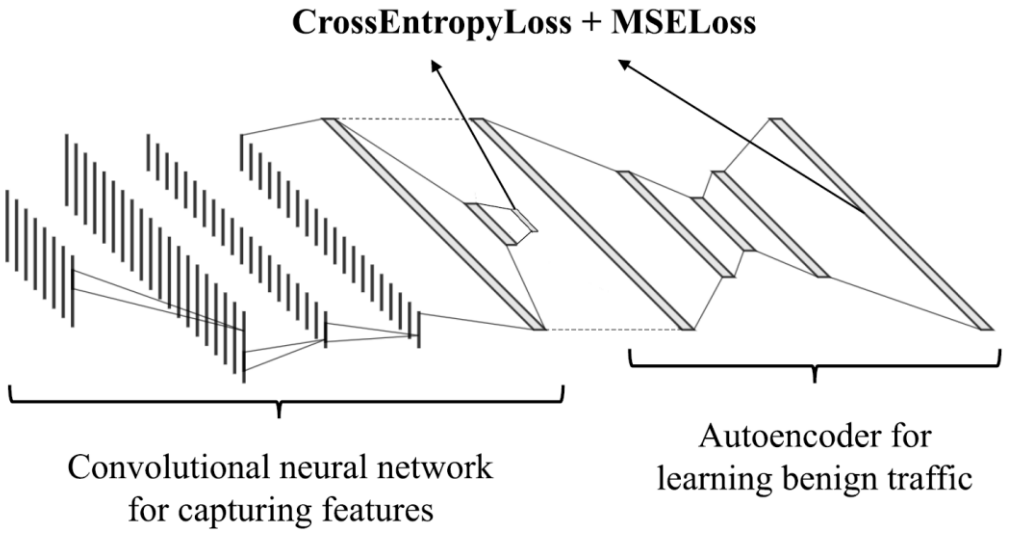
R.-H. Hwang, M.-C. Peng, C.-W. Huang, P.-C. Lin, and V.-L. Nguyen, 'An Unsupervised Deep Learning Model for Early Network Traffic Anomaly Detection', IEEE Access, vol. 8, pp. 30387–30399, 2020,.

- Sometimes, **neural networks (NN) can be concatenated.**
- The second NN is an auto-encoder, to learn a representation of the input features (for dimensionality reduction or anomaly detection)
- The first NN is dedicated to feature extraction, being selected as a function of the nature of the data. Examples:
    1. Convolution networks are preferred for image datasets.
    2. Long short-term memory (LSTM) based models tend to produce good results for sequential data.



An example:

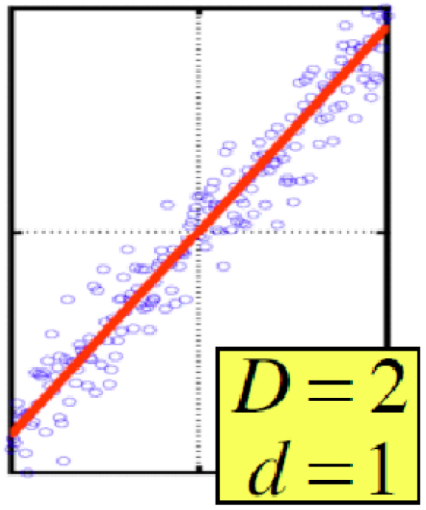
- "D-PACK (…) Consists of a Convolutional Neural Network (CNN) and an unsupervised deep learning model (e.g., Autoencoder) for auto-profiling the traffic patterns and filtering abnormal traffic."
- "In this work, we present an effective anomaly traffic detection mechanism, namely (…) notably, D-PACK inspects only the first few bytes of the first few packets in each ow for early detection."

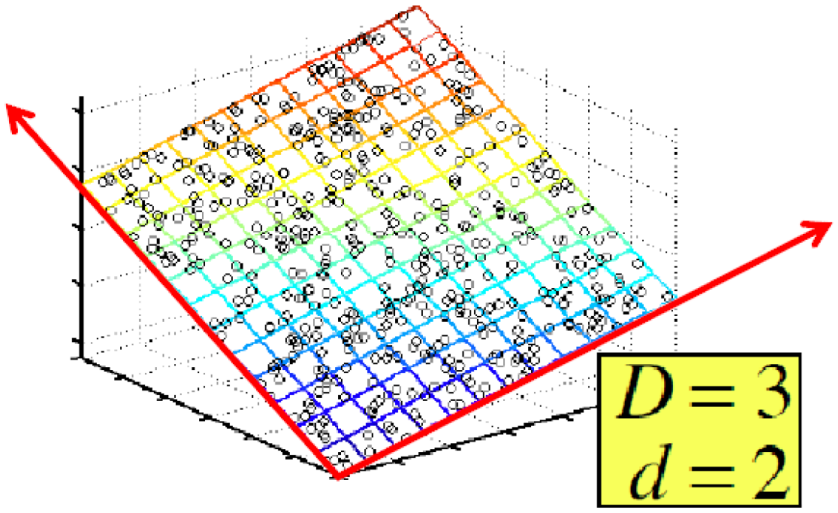


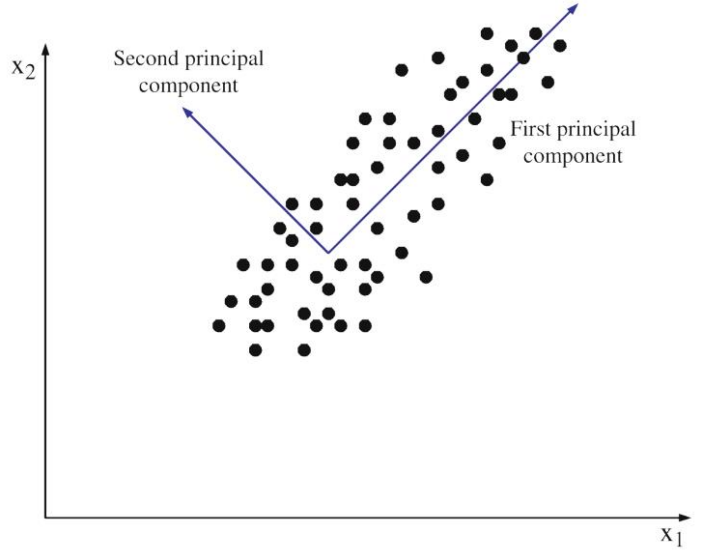
3. Principal Component Analysis (PCA)

Pearson, Karl, "LIII. On lines and planes of closest fit to systems of points in space,"*The London, Edinburgh, and Dublin philosophical magazine and journal of science,* 2.11, pp.559-572, 1901

- PCA is used most often for dimensionality reduction.
- Assumption: **data lies on or near a low d-dimensional linear subspace.**
- Axes of this subspace are an effective representation of the data.
- **Identifying those axes is known as Principal Component Analysis**, and can be obtained by Eigen or Singular Value Decomposition.
- In other words, PCA finds the directions of maximal variance in the training data.





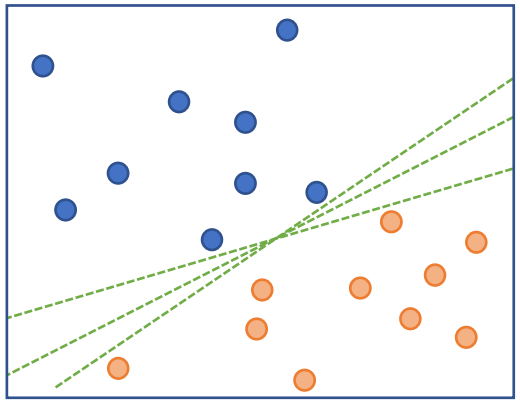


4. One-Class Support Vector Machines

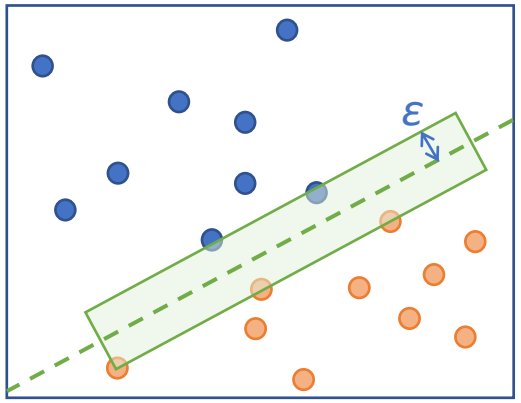
B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, 'Support Vector Method for Novelty Detection', p. 7.

Support Vector Machines (SVM)

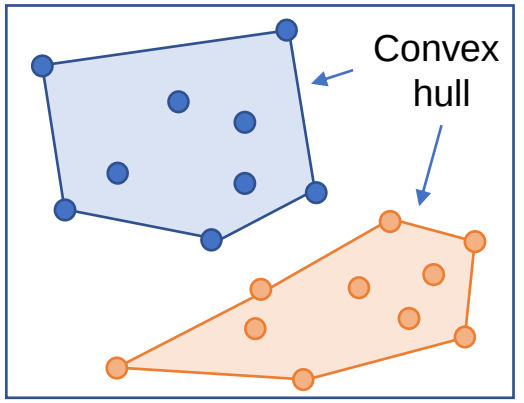
- SVMs select a decision boundary for which **the margin between data points of different classes is maximized**
- Other interpretation is that SVMs maximize the distance between the convex hulls of points belonging to each class



Other Classifiers



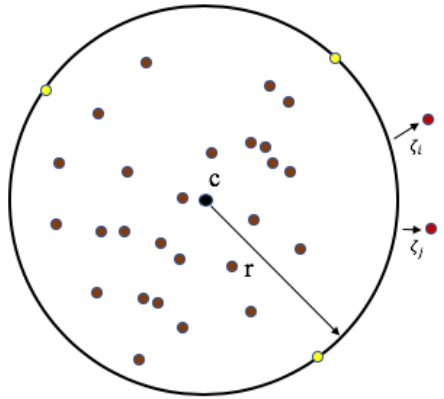
Support Vector Machines



Convex Hulls

One-Class SVM (OC-SVM)

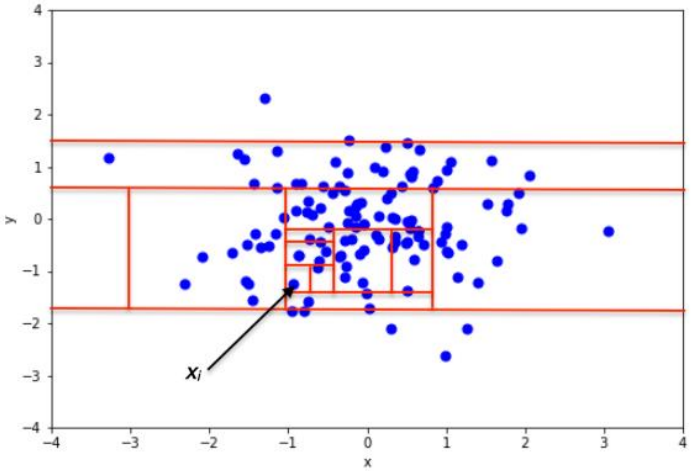
- One-Class SVM uses a hypersphere to encompass all of the instances (as opposed to using an hyperplane to separate two classes of instances).
- Data points outside the hypersphere are classified as anomalies.



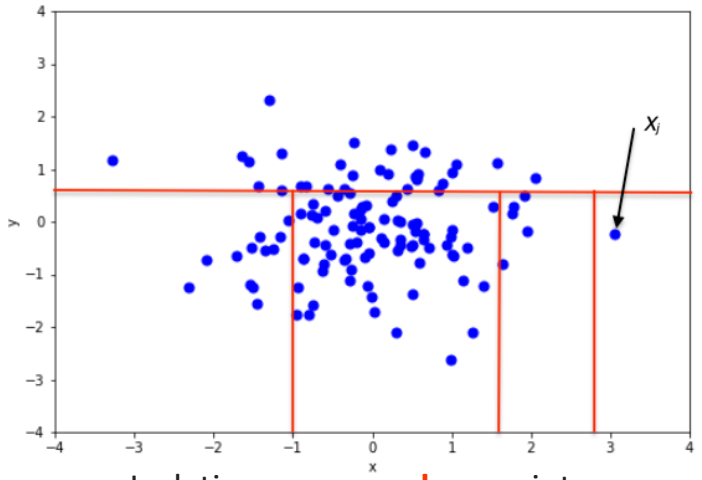
5. Isolation Forest

F. T. Liu, K. M. Ting, and Z.-H. Zhou, 'Isolation Forest', in 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, Dec. 2008, pp. 413–422. doi: 10.1109/ICDM.2008.17.

- "Most existing model-based approaches to anomaly detection construct a profile of normal instances, then identify instances that do not conform to the normal profile as anomalies."
- "This paper proposes a fundamentally different model-based method that explicitly isolates anomalies instead of profiles normal points."



Isolating a **non-anomalous** point in a 2D Gaussian distribution

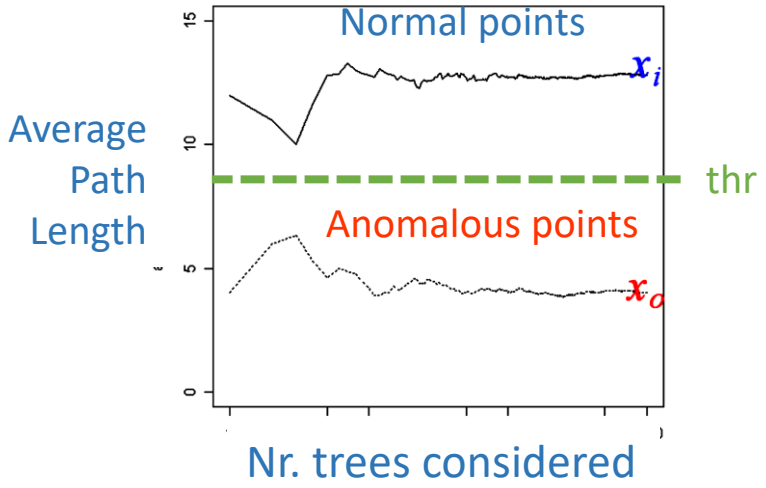


Isolating an **anomalous** point in a 2D Gaussian distribution

How it works?

- The core assumption is that **less partitioning steps are required in order to isolate an anomalous sample in a dedicated partition**.

1. Recursively generate partitions on the sample by randomly selecting an attribute
2. Randomly selecting a split value for the attribute, between the minimum and maximum values allowed for that attribute.
3. Anomalies will usually require less partitions to be isolated, i.e., have a smaller path lengths



Pros and Cons

	Type	Advantages	Disadvantages
Density-based Clustering	Unsupervised	• Low computational cost	• Selecting best parameter values (ε and minPts) for normal-class membership
Auto-encoder	Semi-supervised	• Architectures for different types of data (images, time-series, etc.)	• Computationally expensive • Dimension of neural network
PCA	Unsupervised	• Low computational cost	• Selecting best threshold value (distance to lower-d hyperplane)
OC-SVM	Semi-supervised	• Minimizes convex set of normal data	• Computationally expensive
Isolation Forests	Unsupervised	• Low computational cost	• Selecting best threshold value for path length value (used to classify as normal or anomalous)

Overview of Anomaly Detection Techniques

https://en.wikipedia.org/wiki/Anomaly_detection

- Density-based techniques: k-nearest neighbor [9][10][11], local outlier factor [12], isolation forests [13][14], and many more variations of this concept[15]
- One-class support vector machines [20]
- Replicator neural networks [21], autoencoders, variational autoencoders [22], long short-term memory neural networks [23]
- Subspace-[16], correlation-based [17] and tensor-based [18] outlier detection for high-dimensional data [19]
- Bayesian networks [21]
- Hidden Markov models (HMMs) [21]
- Cluster analysis-based outlier detection [24][25]
- Deviations from association rules and frequent item sets
- Fuzzy logic-based outlier detection.
- Ensemble techniques, using feature bagging [26][27], score normalization [28][29], and different sources of diversity [30][31]

Anomaly Detection in Univariate Time Series

MIRAI
isep Instituto Superior de Engenharia do Porto
P.PORTO
U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

Univariate Time-Series

M. Braei and S. Wagner, 'Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art', arXiv:2004.00433 [cs, stat], Apr. 2020, Accessed: Aug. 04, 2021. [Online].

Non-temporal vs. Temporal data

- The main assumption about spatial data is that the data points are independent from each other. Therefore, Anomaly detection happens by either:
    1. measuring the deviation of the abnormal points to the rest of the data
    2. clustering the whole dataset and mark all points as anomalies that lie in less dense regions.

Temporal data

- **In time-series data, it is presumed that data points are not completely independent.**
- It is assumed that the latest data points in the sequence influence their following timestamps.
- Following this, values of the sequence change smoothly or show a regular pattern. Sudden changes in the sequence can be regarded as an anomaly.

Time-series patterns

- **Trend:** if its mean is not constant, but increases or decreases over time.
- **Seasonality:** periodic recurrence of fluctuations.
- **Stationarity:** stationary time-series is a time-series having the same characteristics over every time interval. A stationary time-series will have
    1. Constant mean, thus no trend exists in the time-series.
    2. Constant variance.
    3. Constant autocorrelation over time.
    4. No seasonality, i.e., no periodic fluctuations.

Methods for Time-series Data

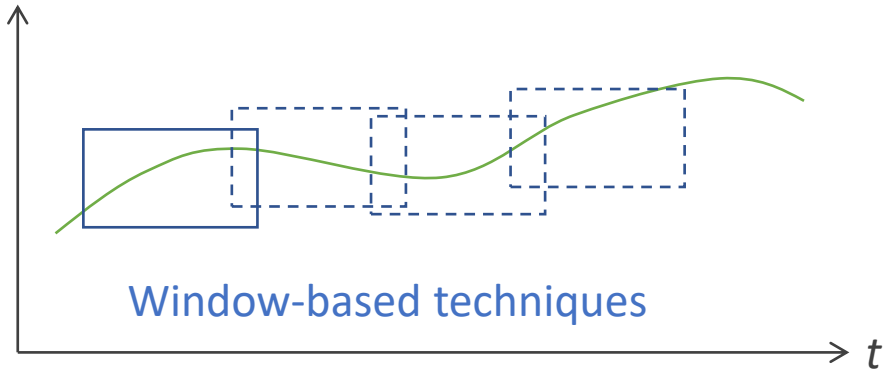
- Aggarwal [2] breaks down anomaly detection methods for time-series into two main categories:
    1. Anomaly detection based on prediction of the time series
    2. Anomaly detection based on unusual shapes of the time series

Anomaly Detection in Time-Series

M. Braei and S. Wagner, 'Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art', arXiv:2004.00433 [cs, stat], Apr. 2020, Accessed: Aug. 04, 2021. [Online].

Statistical Methods

- Anomaly detection using statistical approaches
- Autoregressive Model (AR)
- Moving Average Model (MA)
- Autoregressive Moving Average Model (ARMA)
- ARIMA Model
- Simple Exponential Smoothing (SES)
- Double and Triple Exponential Smoothing
- Time-series Outlier Detection using Prediction Confidence Interval (PCI)



Window-based techniques

Classical machine learning

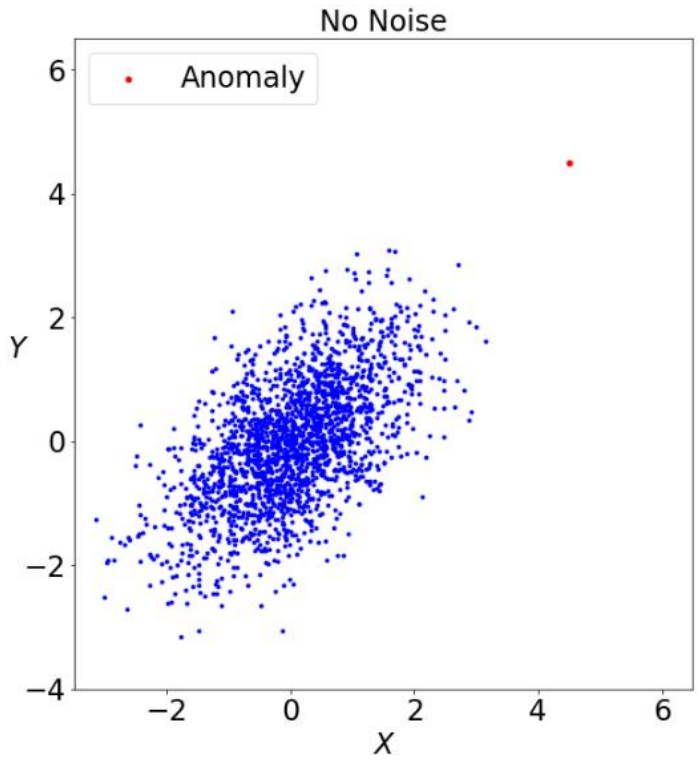
- K-Means Clustering – Subsequence Time-Series Clustering (STSC)
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- Local Outlier Factor (LOF)
- Isolation Forest
- One-Class Support Vector Machines (OC-SVM)
- Extreme Gradient boosting (XGBoost, XGB)

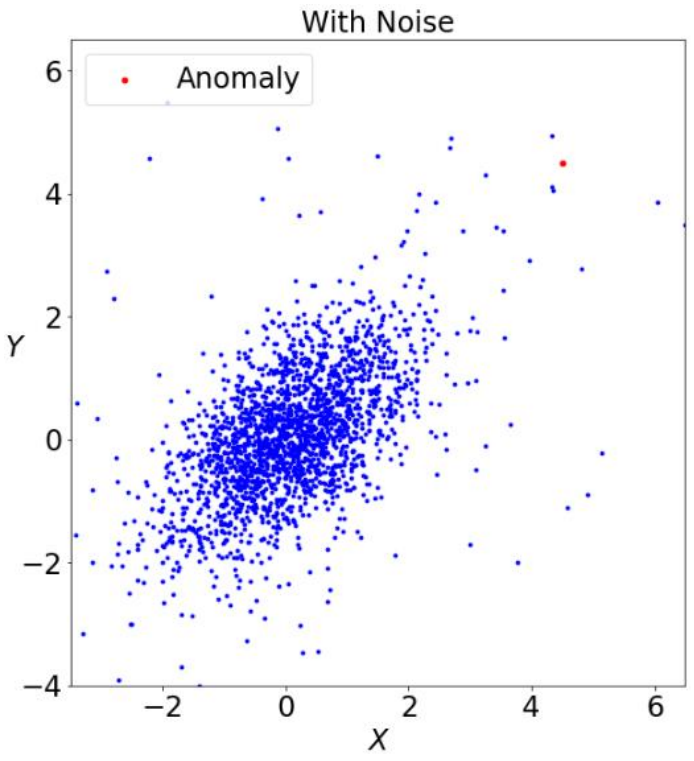
Neural networks

- Multiple Layer Perceptron (MLP)
- Convolutional Neural Networks (CNN)
- Residual Neural Network (Resnet)
- WaveNet
- Long Short Term Memory (LSTM) network
- Gated recurrent unit (GRU)
- Autoencoder

The Impact of Noise

M. Braei and S. Wagner, 'Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art', arXiv:2004.00433 [cs, stat], Apr. 2020, Accessed: Aug. 04, 2021. [Online].





Thank you

MIRAI isep Instituto Superior de Engenharia do Porto P.PORTO U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO