Lab in Analytics for Business
Seminar on: Anomaly Detection

Piercesare Secchi

Milano, 18 February 2025

Anomaly detection

Anomaly = something that deviates from what is standard, normal, or expected.

- **Anomaly detection** refers to the problem of finding patterns in data that do not conform to the expected behavior.
- Anomalies are not bad per se; indeed, an anomaly could be a novelty.
- In fact, anomaly detection, novelty detection, outliers detection are often used synonymously.

Main reference for the seminar:

Chandola V., Banerjee A., Kumar V. (2009), Anomaly detection: a survey, *ACM Computing Surveys*, 41(3),1-58

(most pictures in these slides are taken from this paper)

Anomaly detection applications

Examples:

- Fault detection
- Fraud detection for credit cards, insurance or health care
- System health monitoring
- Ecosystem perturbations and disturbances
- Intrusion detection for cyber-security
- Defect detection in image data
- Defect detection in text data
- Medical diagnosis
- Law enforcement
- Marketing

…

Challenges

- Defining what is an anomaly, implies the ability to identify expected normal behavior

- Supervised learning might be an issue (availability of a training set)

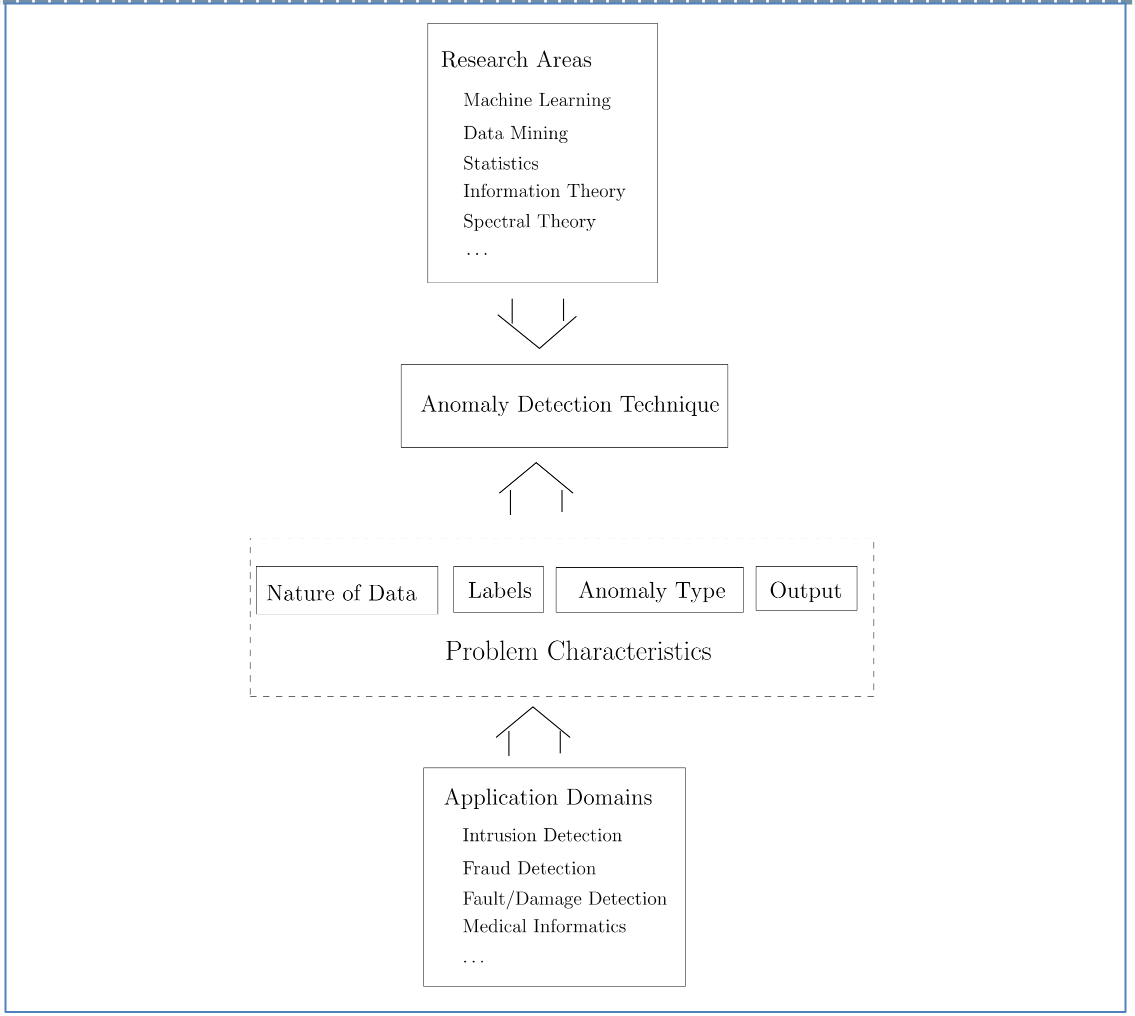
- Normal behavior evolves with time and space: what is now considered normal might not be so in the future

- What is normal and what is anomalous is domain specific

- Nature of the data and of the data space embedding

Most of the existing anomaly detection techniques solve a specific formulation of the problem.

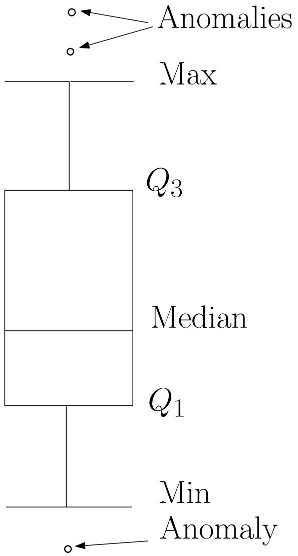
Key components of anomaly detection technique from Chandola et al. (2009)

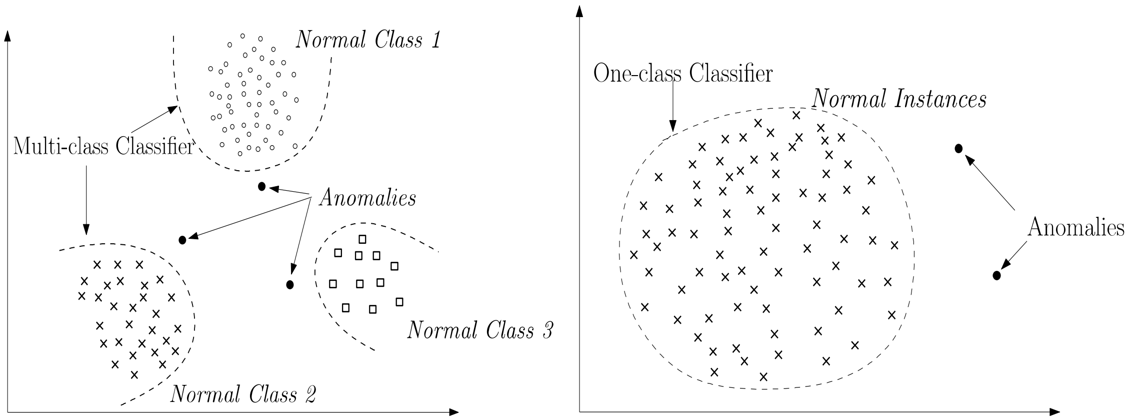


Summary

1. Type of anomaly
2. The statistical modeling approach
3. Regression for anomaly detection
4. The statistical learning approach: training sets, supervised and unsupervised anomaly detection methods
5. Nearest neighbor techniques
6. Clustering
7. Dimensional reduction and spectral methods
8. Conclusion

1. Type of anomaly: point anomaly



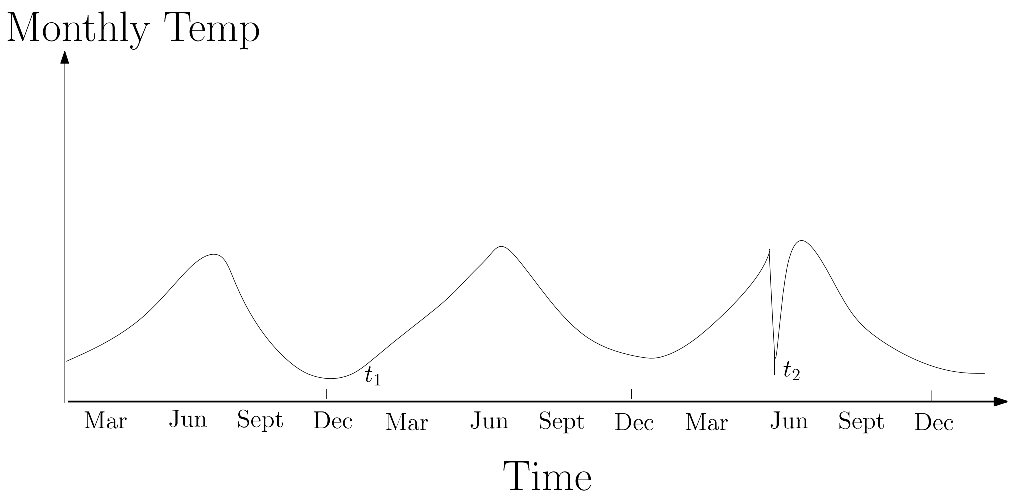


(a) Multi-class Anomaly Detection (b) One-class Anomaly Detection

Chandola et al. (2009)

1. Type of anomaly: contextual anomalies

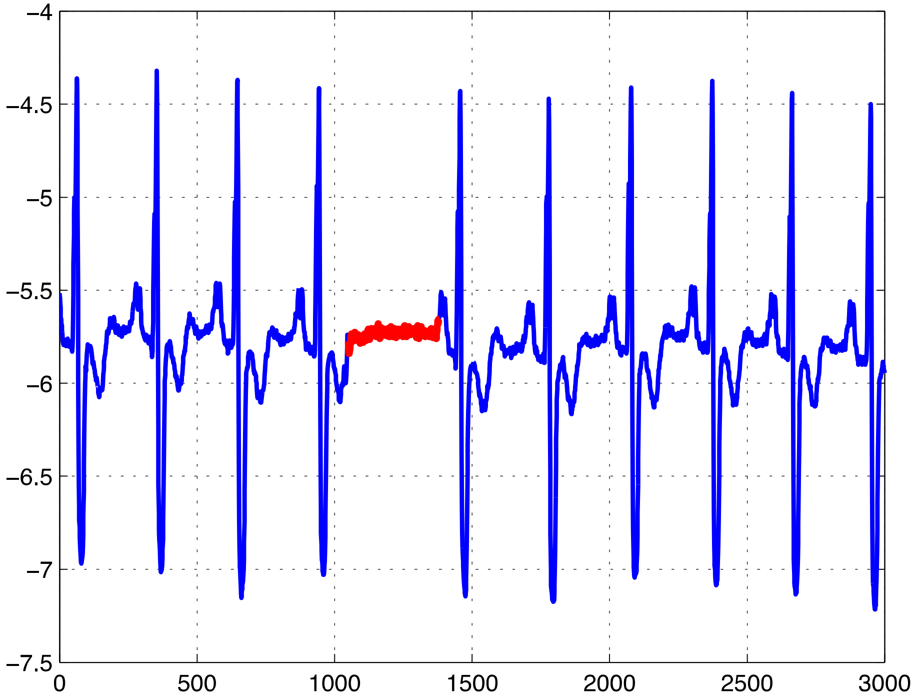
What is considered as «context» is part of the problem formulation



Temperature t_2 would not appear as an anomaly, for instance in a boxplot summarizing the marginal distribution of montlhy temperatures

Chandola et al. (2009)

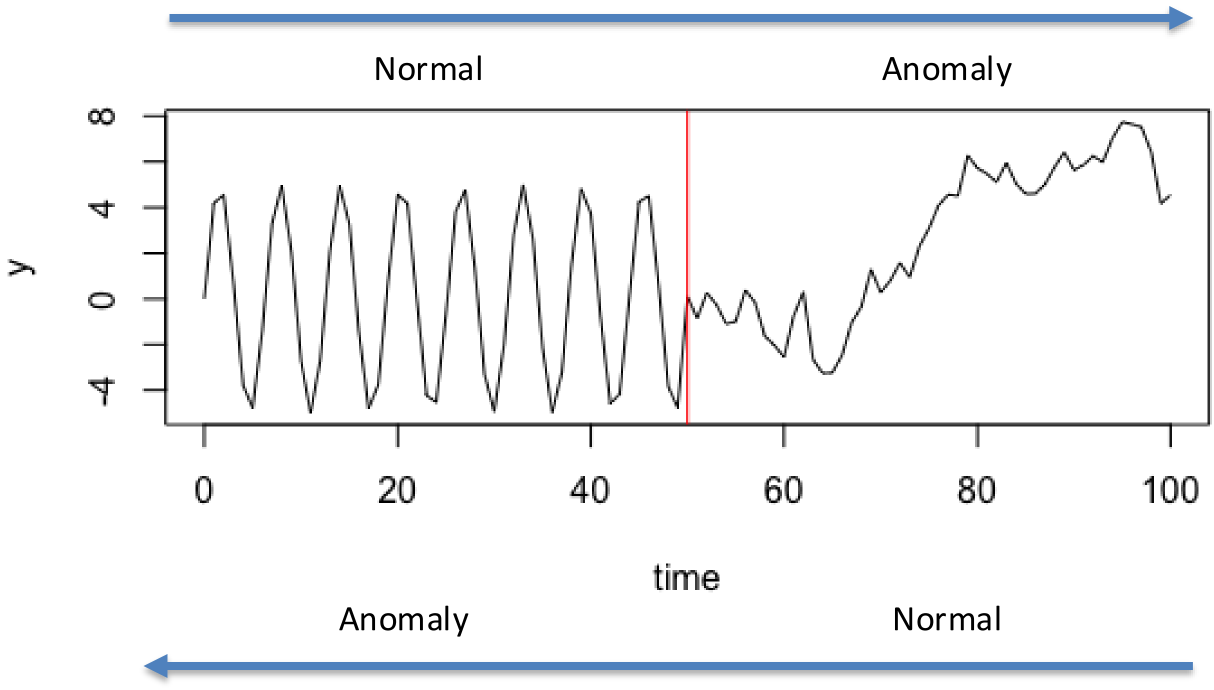
1. Type of anomaly: collective anomaly



- Signal generated by an ECG (electrocardiogram)
- Red part is an anomaly.
- Sensor failure or Atrial Premature Contraction?

Chandola et al. (2009)

1. Type of anomaly:
contextual and collective anomalies…



Normal Anomaly

Anomaly Normal

1. Type of anomaly

- In this seminar, I will mostly focus on point anomalies

- Contextual and collective anomalies are more difficult to capture (*in an automated way...)* and usually require models and methods for the statistical analysis of dependent data. E.g. Lisa maps for spatial data.

- Examples: spatial data, time-series and sequential data, segmented data, graphs,…

- General approach: learn a model from *training data*. Predict the expected behavior of *test data* based on the model. Declare the *test data* as anomalous if the observed behavior is significantly different form the expected.

- Models: regression models (see later). ARMA and ARIMA for time-series,(Hidden) Markov Models for sequential data, Space-time models, Kriging for spatial data, Markov Gaussian Random Fields for images,…

2. Anomaly detection based on the statistical model generating the data

Assumption*: normal data occur in high probability regions of a statistical model, while anomalies occur in the low probability regions of the statistical model.*

- Use data to fit a statistical model capturing the distribution which generated them; both parametric and nonparamteric models for densities (continuous or discrete) are allowed.

- Plenty of methods (parametric or semi-paramteric) for fitting densities to univariate data. Some of them have extension to multivariate data

- Once the model is fitted, the reciprocal of the density computed at a data point is the basis for building an anomaly score

- Plenty of tests based on parametric and semiparametric models, both for univariate and multivariate data, have been proposed.

2. Example: the Gaussian model

- Fit a Gaussian distribution to the data.

- Use data to estimate the mean and the (co)variance.

- Warning: sample mean and sample (co)variance estimators are not robust to outliers…

- Univariate data: declare as anomalous a data point whose distance from the mean is more than 3*sigma. Less than 0.27% of the data are this far from the mean, if the distribution generating normal data is Gaussian.

- For a general univariate distribution, Chebychev’s inequality would ask for a distance of 19*sigma to guarantee the same confidence. Not very useful for anomaly detection.

2. Example: the Gaussian model

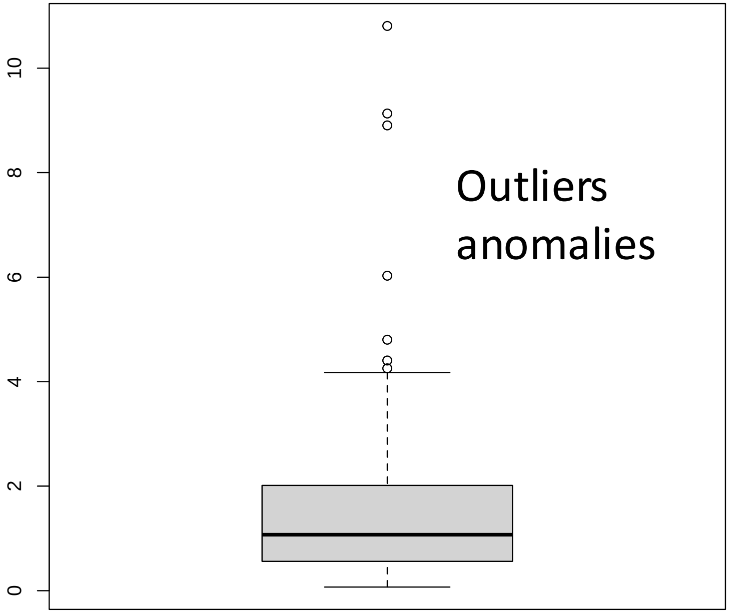
- Extend to multivariate data with Gaussian distribution, by considering as normal those data falling in the ellipsoidal contour of the density, centerd on the mean and containing a mass larger than 0.995, say

- Hotelling’s T^2 distribution as a multivariate generalization of t-Student’s distribution is often used for anomaly detection in quality control

- Fisherian approach: transform data to make their distribution Gaussian. Then use the Gaussian model to identify anomalies

2. Example: nonparametric approaches

- The boxplot rule for univariate data



Outliers
anomalies

Q1 = first quartile
Q3= third quartile
IQR= Q3-Q1

Anomalies: alla data lying outside the interval
[-1.5*IQR + Q1, Q3+1.5*IQR]

(For Gaussian data this is almost equivalent to the 3*sigma rule.)

2. Multivariate extension of boxplot: depth measures

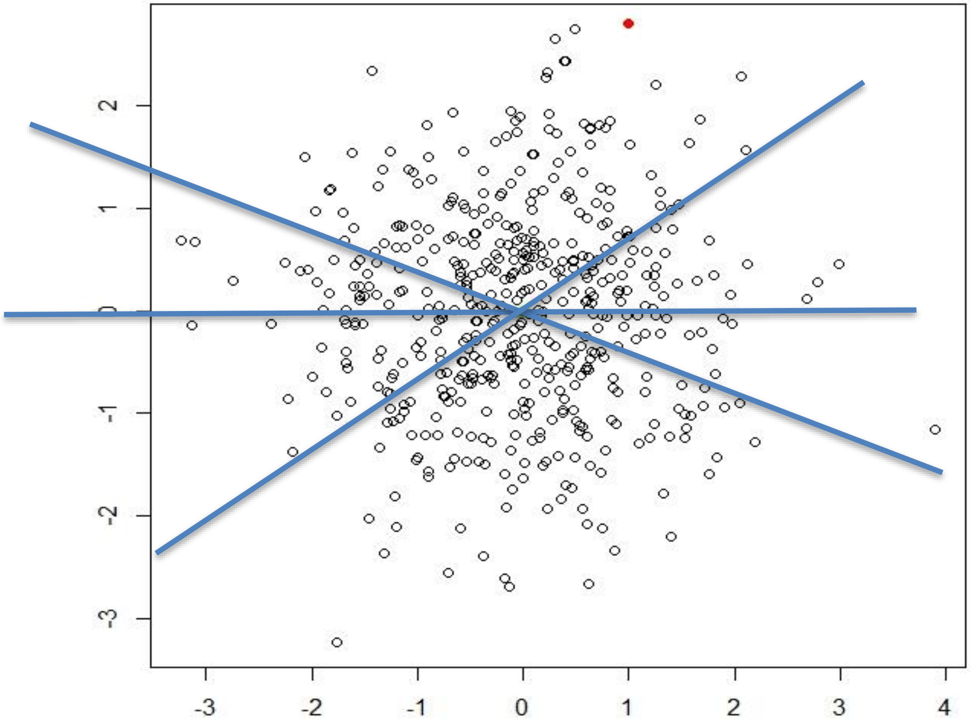
- The boxplot is based on quantiles. Extension of quantiles to the multivariate setting is not immediate, since a notion of (complete) ordering is not obvious when the space dimension is larger than or equal to 2.

- Different definitions of depth are possible, they all try to capture, in a multivariate setting, the notion of «centrality» as opposed to that of «outlyingness».

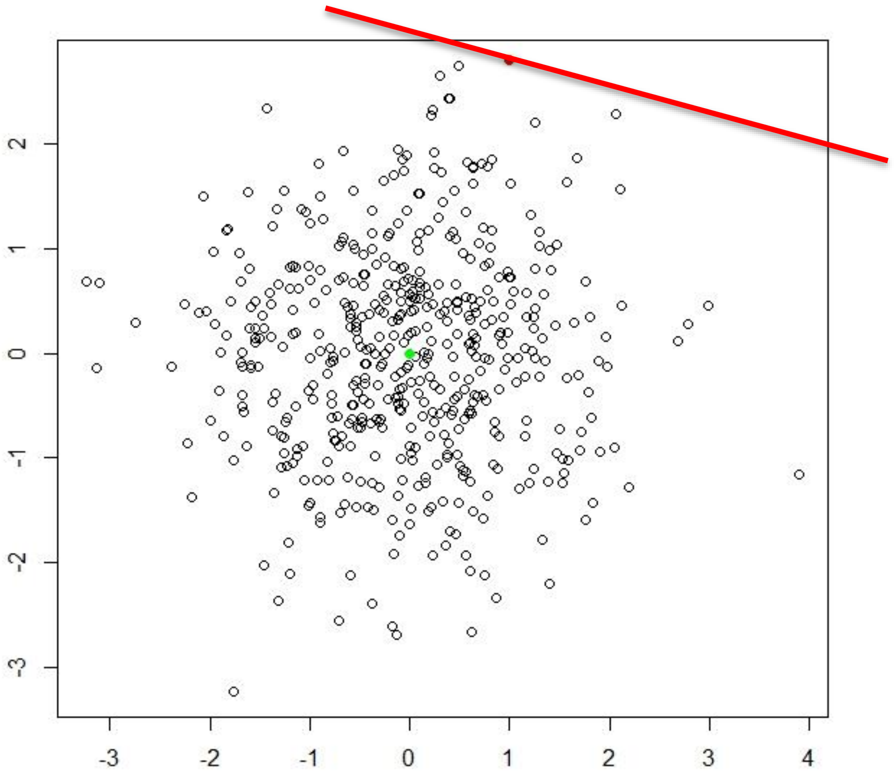
- Examples: Half Space or Tukey depth, Simplicial or Liu depth, Oja depth, Mahalanobis depth, Convex Hull Peeling depth…

- For p=2, a few pictures on the blackboard will be enough to understand…

2. Depth measures – Tukey (HD)



2. Depth measures – Tukey (HD)

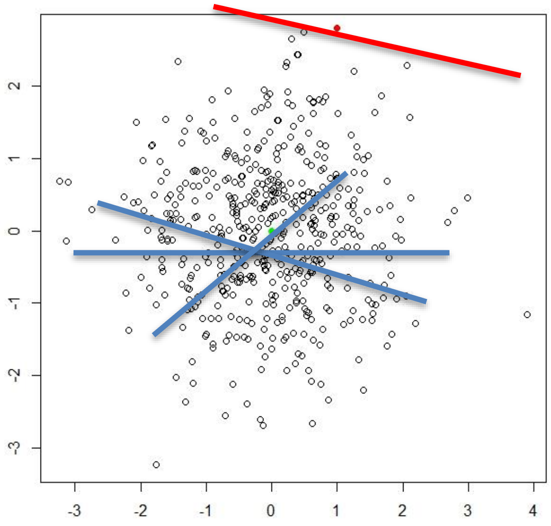


2. Depth measures – Tukey (HD)

Halfspace depth (HD) or Tukey depth

The first notion of depth has been proposed in ***1975 by John Tukey***. The halfspace depth of a point x in R^p with respect to a probability measure **P** on R^p is defined as the minimum probability mass carried by any closed halfspace containing x:

$$HD(x;\mathbf{P}) = \inf\{\mathbf{P}(\mathrm{H}) : \ \mathrm{H \ a \ closed \ halfspace, \ x \in H}\}, \ \ \mathrm{x} \in \mathbb{R}^{\mathrm{P}}$$



- ✓ Green point: maximum depth
  Median

- ✓ Red point: minimum depth
  Outlier

3. Anomaly detection using a regression model

- Given a target variable, fit a regression model to the data
- Use diagnostics on residuals to identify outliers, i.e. anomalies with respect to the target variable
- (Warning: leverage points might be anomalies wrt to the regressors)
- Common situation: the regression model is fitted on training data considered as normal, i.e. without anomalies. The magnitude of the residual of a *test case* is its anomaly score. Large residuals are indicative of anomalies, i.e. cases far from their expected values
- This approach is extensivley used in time-series analysis, where it might capture contextual anomalies

3. Anomaly detection using a regression model

- Warning: presence of anomalies (outliers, leverages, data with high Cook's distance) influence the fitting of the regression model.

- If the goal is to identify anomalies in test cases, first deal with influent cases in the training set, and then, after a satisfactory model has been fit, use it to identify anomalous test cases.

- Robust regression is less affected by the presence of anomalies in the training set

4. The statistical learning approach: training sets, supervised and unsupervised anomaly detection methods

- Supervised anomaly detection: labels (normal, anomalous) are observable on the data of the training set

- Use supervised statistical learning methods to learn how to discriminate between normal and anomalous data

- Unsupervised anomaly detection: labels are hidden

- Use unsupervised statistical learning methods (ouliers detection, cluster analysis,…) to learn how to identify normal data

- Semi-supervised anomaly detection: only normal cases are labelled

4. Supervised classification techniques

***Assumption**: using the training set, a classifier can be learnt which can discriminate between normal and anomalous units based on their features*

- Train a classifier using the training set, whose units have been labelled as normal or anomalous

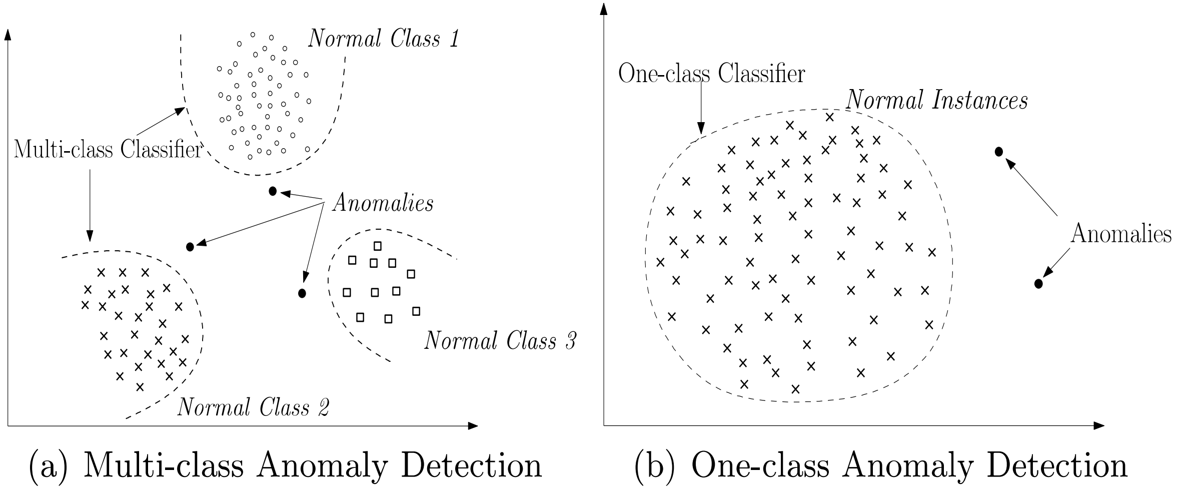
- Any classifier which is apt to the job: logistic regression, LDA, QDA, CART, Random Forests, Support Vector Machines, neural networks,…

- Use the classifier to label new units as normal or anomalous

4. Supervised Classification techniques

- One-class classifiers (e.g. one class SVM) assume that all units in the training set are normal. They learn a discriminative boundary around the normal units, represented by their features.

- A new unit whose feature falls outside the boundary is labelled as anomalous



(a) Multi-class Anomaly Detection (b) One-class Anomaly Detection

5. Nearest neighbor techniques (unsupervised)

***Assumption**: normal data occur in dense neighbors, while anomalies occur far from their closest neighbor*

- Need to specify a distance or at least a similarity (like in cluster analysis)

Two possibilities:

1. Techniques that use the distance of a data instance to its k-th nearest neighbor as the anomaly score.

2. Techniques that compute the relative density of each data instance to compute its anomaly score.

5. k-th nearest neighbor

- The anomaly score of a data instance is defined as its distance to its k-th nearest neighbor in a given data set.
- For instance: take the average distance of each unit from the k-units which are closest, and then sort the units based on this score. Use a threshold to detect anomalies.
- Other possibility, estimate the «global» density at a data point: fix a distance *d*. For each unit, count the number *n* of different units within a sphere centered in the unit and with radius *d*. Use has anomaly score *1/n*, which is proportional to the reciprocal of the density *n/Vol(sphere of radius d).*
- The dual: fix *k* and consider the radius *d* of the minimal sphere centered in the unit and containing the *k* closest units. Use *d* as anomaly score.

5. kth nearest neighbor

This global density techniques can perform poorly when data have different densities in different regions

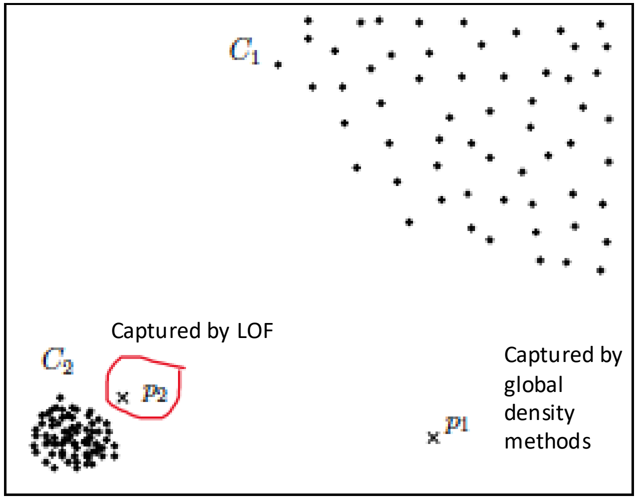


Fig. 7. Advantage of local density based techniques over global density based techniques.

Chandola et al. (2009)

5. Local outlier factor (LOF)

- For each data point *x* compute its local density: fix *k*, compute the radius of the smallest sphere containing the closest *k* neighbors and compute its volume *V*. The local density is

$$f_{loc}(x)= k/V$$

- Compute the average of the local densities for the k closest neighbors:

$$\overline{f(x)} = ave(f_{loc}(x\_j), j=1,\dots,k)$$

- LOF = $\overline{f(x)} / f_{loc}(x)$.

- If a data point is normal, LOF will be close to 1

- If a data point is anomalous, LOF will be higher than 1

6. Clustering for anomaly detection

***Assumption**: normal data lie close to their cluster centroid, anomalies lie far away from their cluster centroid*

- Use any clustering technique, and then compute the distance of any data point form its centroid. For anomalies identification, threshold those data points which are too far apart

- If anomalies are somewhat «clustered», you won't find them…

6. Clustering for anomaly detection

Assumption*: normal data belong to large and dense clusters, while anomalies belong to small or sparse clusters*

- Use clustering techniques like dendrograms and identify small or isolated clusters. Data points belonging to these small or sparse clusters are declared anomalous.
- Use cophenetic-like distances to identify anomalies

Assumption: *normal data belong to a cluster, anomalies do not belong to any cluster.*

- Use clustering techniques which do not necessarily assign every data point to a cluster: e.g. DBSCAN

5. Unsupervised methods: pros and cons

Pros:

- Decidedly nonparametric
- Mostly based on the notion of distance, they usually don't need a rich structure for the embedding space
- Easy to extend to different data types (quantitative or categorical, univariate or multivariate): just define the proper distance
- Distances can be replaced by dissimilarities without difficulties: e.g. a «distance» for which the triangle inequality does not hold

Cons:

- Curse of dimensionality
- Often computationally expensive
- Distance is domain specific (I don't consider this as a cons… Use your engineering prior knowledge!)

7. Dimensional reduction and spectral methods

***Assumption**: Data can be embedded into a lower dimensional subspace in which normal instances and anomalies appear significantly different.*

- Use Principal Component Analysis to reduce the dimensionality of the data representation

- Project data on a few PC's and proceed with your favorite method for anomaly detection

- Inspect also the projections on PCs associated to low varibiality (PCs associated to small eingenvalues). Anomalies might be visibile on those dimensions and not on the first PCs.

8. Conclusions?

- Each method for anomaly detection has strenghts and shortcomings
- There is no such thing as «the algorithm» for anomaly detection
- For any given anomaly detection problem, one or more methods will be best suited. Anomaly detection is very domain specific, use your engineering knowledge!

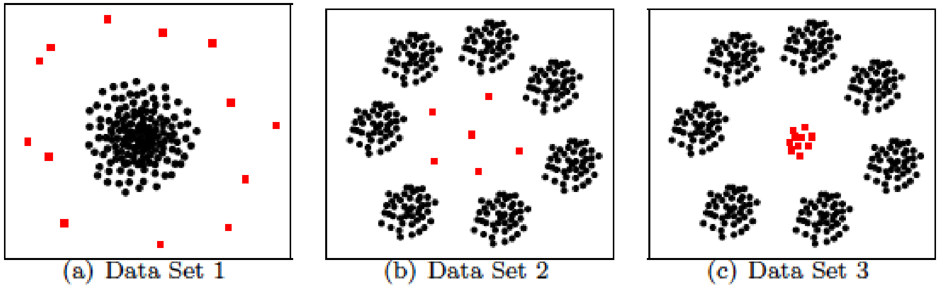


Fig. 10. 2-D data sets. Normal instances are shown as circles and anomalies are shown as squares.

Chandola et al. (2009)