TÉCNICO+
FORMAÇÃO AVANÇADA

Outlier analysis

Unsupervised and supervised stances

DASH: Data Science e Análise Não Supervisionada

Rui Henriques, rmch@tecnico.ulisboa.pt
Instituto Superior Técnico, Universidade de Lisboa

Outline

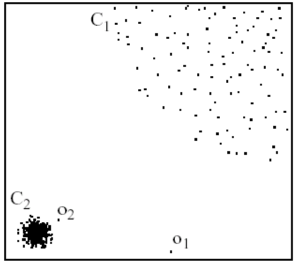
- Motivation
- Learning paradigms
- Statistical approaches
- Proximity-based approaches
- Clustering approaches
- Deep learning approaches
- Other approaches

Outlier analysis: applications

- **Fraud detection**: credit card, telecom, criminal activity in e-commerce
- **Cybersecurity** and intrusion detection (anti-viruses and network firewalls)
- **Customized marketing**: high/low income buying habits
- **Healthcare**: unusual responses to various drugs, rare diseases
- Analysis of **performance** statistics (e.g. professional athletes)
- **Adverse weather** and **seismic** prediction
- **Financial** applications: loan approval, stock tracking
- …

What is an outlier?

- Outlier ≡ anomaly ≡ exception ≈ novelty
  - outlier analysis ≡ deviant behavior analysis ≡ anomaly analysis ≡ exception analysis ≈ novelty analysis
- **Outlier**: observation that **deviates significantly** from the normal observations as if it was **generated by a different mechanism**
  - e.g. unusual credit card purchase, Michael Jordon…
  - **global outlier**: observations inconsistent with rest of the dataset
  - **local outlier**: observations inconsistent with their neighborhoods



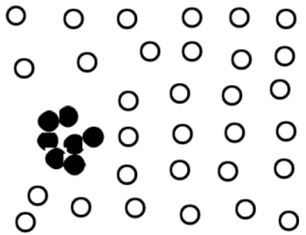
- Outliers differ from **noise data**
  - noise is random error or variance in the measured variables
  - outlier analysis should be able to discard noise

Types of outliers (1/2)

- **Global outlier** (or point anomaly)
  - observation that significantly deviates from the rest of the data (e.g. intrusion)
  - issue: find an appropriate measurement of deviation
- **Contextual outlier** (or conditional outlier)
  - observation that significantly deviates from a given context
    - e.g. 30ºC in Urbana outlier depending on whether is summer or winter?
    - variables divided into two groups
    - contextual variables: define the context (e.g. time, location)
    - behavioral variables: define the features for outlier evaluation (e.g. temperature)
  - generalization of local outliers—deviation from its local area
  - *issue*: define or formulate meaningful context

Types of outliers (2/2)

- **Collective Outliers**
  - a subset of observations that collectively significantly deviate from the whole data (even when each observation is not outlier)
    - e.g. risk groups with rare/infrequent genetic variants
  - *issues*
    - consider both individual and group behavior
    - need suitable distances



- **Final** considerations
  - a dataset may have multiple types of outlier
  - one observation may belong to more than one type of outlier

Challenges

- **Separating** normal observations from outliers
  - hard to enumerate all possible normal behaviours
  - border between normal and outlier objects is often a gray area
- **Application-specific**
  - distance metric or statistical assumptions are application-dependent (e.g. clinic data and small deviations, marketing and larger fluctuations)
- **Handling noise**
  - noise may distort the normal objects
- **Understandability**
  - explanatory detection
  - degree of outlier: likelihood of being generated by a normal mechanism

Outline

- Motivation
- **Learning paradigms**
- Statistical approaches
- Proximity-based approaches
- Clustering approaches
- Deep learning approaches
- Other approaches

Outlier analysis

- Our core premises
  - data is preprocessed (e.g., minimum artifacts/missings, numerical encodings of ordinals)
  - access to **good representations** of (complex) data ⇒ check *representation learning* class
  - controlled dimensionality of observations ⇒ check *statistics* and *dim. reduction* classes

- Two ways of categorizing outlier detection approaches:
  - whether labeled examples of outliers are given
    - **supervised**, **semi-supervised** vs. **unsupervised** methods
  - whether assumptions/knowledge w.r.t. data are available
    - **statistical**, **proximity-based**, and **clustering-based** methods vs. **deep learning**

Outlier analysis: supervised methods

Supervised outlier detection

- outlier detection as a **classification** task
  - observations validated by domain experts (e.g., fraud clearance) for training and testing
  - given an observation, the probability of the outlier class can be seen as a score
- **single-class** prediction task
  - model normal (outlier) observations and report those not matching the model

Challenges

- imbalanced classes (outliers are rare)
  - ensure that the applied learning approach can handle imbalance
    (e.g. avoid kNN, favor neural networks with weighted observations or trees/ensembles)
- catch as many outliers as possible
  - <u>sensitivity</u>/recall more important than accuracy
  - F_β-measure with higher β values

Outlier analysis: unsupervised methods

Unsupervised outlier detection

- observations “clustered” into groups, each with unique properties
  - outlier is far away from any group of normal objects
  - e.g. intrusion or virus detection, normal activities are diverse
- *Weakness*
  - unsupervised methods may have high false positive rate but still miss many real outliers (supervised often more effective)
  - cannot detect collective outliers effectively
- *How*? find clusters, then outliers are isolated observations and small clusters
  - problem 1: hard to distinguish noise from outliers
  - problem 2: costly (clustering all data when only few are outliers)

Outlier analysis: semi-supervised methods

Semi-supervised outlier detection

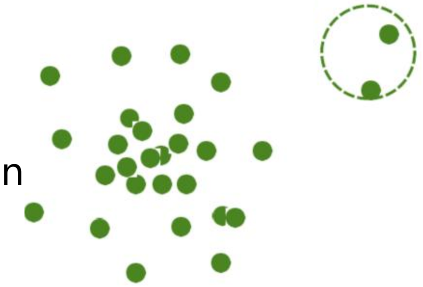
- Number of annotated observations is often small
  - annotations could be on outliers only, normal observations only, or both

- How? Semi-supervised learning
  1. if some **normal observations** are annotated
     - use labeled examples and the nearby unlabeled observations to train a model for normal observations; those not fitting the model are seen as outliers
  2. if some **outliers** are annotated (may not cover all possible outliers well)
     - get help from models for normal observations learned from unsupervised methods

Outlier analysis: statistical methods

- Statistical methods ≡ model-based methods (i.e. parametric)
  – assume normal data follows some statistical distribution (stochastic model)
  – observations not following the model seen as outliers

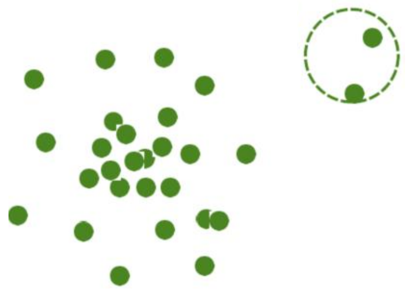
  - *example*: Gaussian distribution to model normal data
    – estimate probability of an observation fitting the distribution
    if low unlikely to be generated and thus an outlier



- **Challenge**: effectiveness depends on whether statistical assumption holds in real data

Outlier analysis: proximity methods

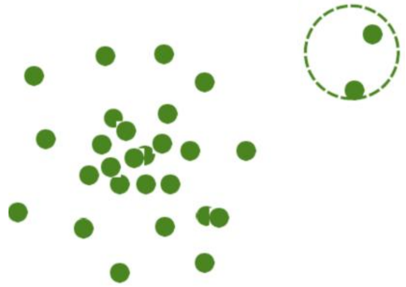
- Observation is an outlier if nearest neighbors are far away
  - proximity significantly deviates from the proximity of most observations
- Two major approaches: **distance-based** and **density-based**
  - *example*: proximity of an object using 3 nearest neighbors



- **Challenge**: effectiveness highly depends on the distance metric
  – in some applications, distances cannot be obtained easily
  – difficulty in finding collective outliers

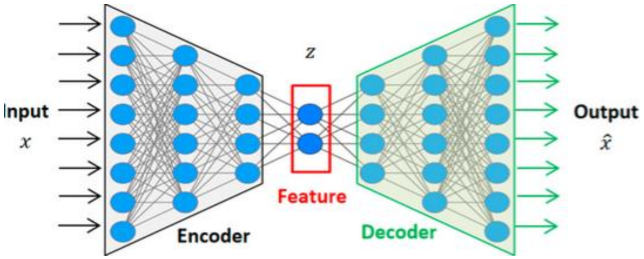
Outlier analysis: clustering methods

- Principle
  - normal data belong to large and dense clusters
  - outliers belong to small or sparse clusters
    - *example*: outliers form a tiny cluster (right)
- *How*? Clustering-based outlier detection methods
  - many forms (e.g. DBSCAN)
- **Challenge**: clustering can be expensive



Outlier analysis: deep learning methods

- Principle
  - train an autoencoder using neural networks
    - recall the aim: reconstruct an observation using a bottleneck architecture (relevant for denoising, representation… and detecting outliers!)
  - normal observations should be easy reconstructed
  - outlier observations have unexpected patterning and thus higher reconstruction error



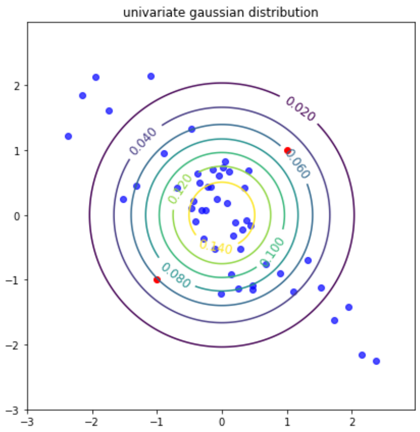
- **Challenge**: high NN capacity can expressively model deviant behavior (low reconstruction error)

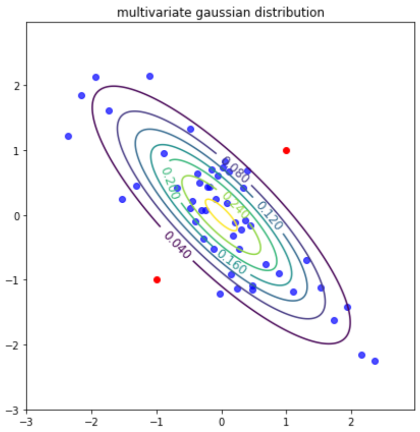
Outline

- Motivation
- Learning paradigms
- **Statistical approaches**
- Proximity-based approaches
- Clustering approaches
- Deep learning approaches
- Other approaches

Statistical-based detection
(distribution-based)

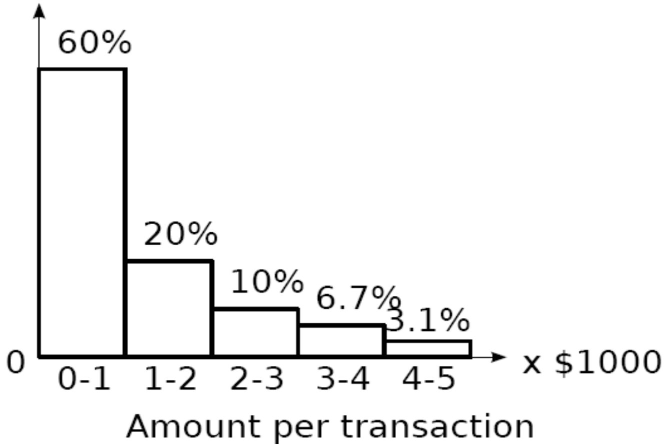
- assumes that normal data is generated by a distribution with parameters θ
- the probability density function $f(\mathbf{x}|\theta)$ gives the probability of observation $\mathbf{x}$ being generated by the distribution: the smaller, the more likely $\mathbf{x}$ is an outlier
- e.g. univariate outliers in {24.0, 28.9, 28.9, 29.0, 29.1, 29.1, 29.2, 29.2, 29.3, 29.4} assuming $\mu + 3\sigma$?
  - multivariate case?
    - multivariate Gaussian
    - naïve Bayes assumption





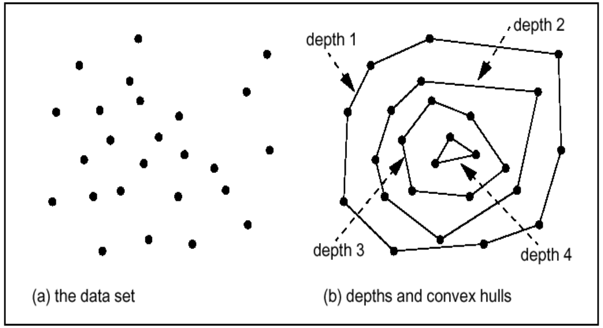
Statistical-based detection
(histogram-based)

- Model of normal data without *a priori* distribution
  - *example* (figure)
    - a transaction with amount $7,500 is an outlier, since only 0.2% transactions are >$5,000



- Challenge: fix bin size
  – *too small* → normal objects in rare bins, false positive
  – *too big* → outliers in some frequent bins, false negative

Statistical-based detection (depth-based)



- **How**
  - search for outliers at data borders
  - observations in convex hull layers
  - outliers are observations on outer layers
- Observations with **depth** ≤ ***k*** are outliers
- Basic assumption
  - outliers are located at the border of the data space
  - normal observations in the center of the data space
- Discussion
  - similar to statistical approaches (k=1 distributions) but without a priori distribution
  - convex hull computation is usually only efficient in small dimensional spaces (e.g. 3D)
  - can be extended for **outlier likelihood**: depth as scoring value

Statistical-based approaches

Learn model fitting data, and identify objects in low probability regions

- Two categories
  - **parametric**: distribution-based
  - **non-parametric**: histogram-based and depth-based

- *Strengths*
  – common and effective: many data distributions are well-approximated
- *Weakness*
  – distribution-based: assume the distribution is known
  – histogram/depth-based: overfitting to available data

Outline

- Motivation
- Learning paradigms
- Statistical approaches
- **Proximity-based approaches**
- Clustering approaches
- Deep learning approaches
- Other approaches

Proximity-based approaches

- *Intuition*: observations that are far away from the others are outliers
- *Assumption*: the proximity of an outlier deviates significantly from that of most of the others in the data set

- Two types of proximity-based outlier detection methods
  - **distance-based** outlier detection:

    an observation is an outlier if its neighborhood does not have enough other observations
  - **density-based** outlier detection:

    an observation is an outlier if its density is relatively lower than that of its neighbors

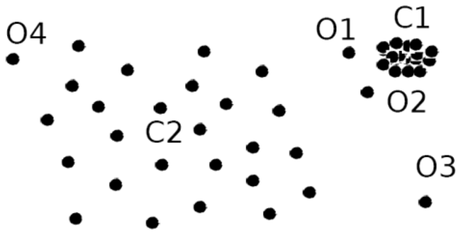
Distance-based approaches

- An observation $\mathbf{x}$ in a dataset X is a (p, d)-outlier
  if at least a fraction p of observations in X are $\geq$ distant d from $\mathbf{x}$
- An observation $\mathbf{x}$ in a dataset is a (k, d)-outlier
  if no more than k points in the dataset are at a distance d or less from $\mathbf{x}$
- Distance of the k^{th} nearest neighbor of $\mathbf{x}$ can be used as an outlier score
- Efficient computation:
  - for any observation $\mathbf{x}_i$, calculate its distance to other observations
  - if $k = pn$ observations are within r distance, terminate the inner loop: $\mathbf{x}_i$ is normal, otherwise a (p, d)-outlier

TÉCNICO+
FORMAÇÃO AVANÇADA

Density-based approaches

- Local outliers: outliers comparing to their local neighborhoods, instead of the global data distribution
  - o_1 and o_2 are local outliers to C_1
  - o_3 is a global outlier, but o_4 is not an outlier
  - distance-based clustering insufficient here

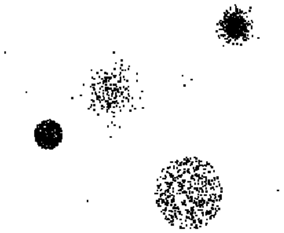


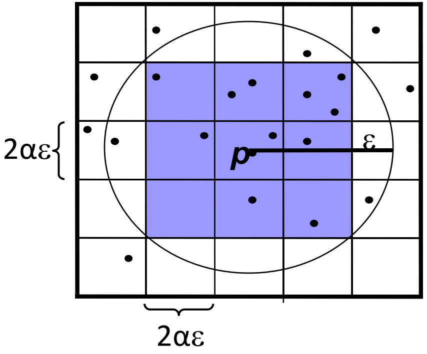
- **Intuition** (density-based outlier detection): the density around an outlier observation is significantly different from the density around its neighbors
- **Method**: use the relative density of an object against its neighbors
  - k-distance: distance between observation and its k-th NN
  - k-distance neighborhood: consider the distribution of distance to the k neighbors

Density-based approaches

Approaches essentially differ in how to estimate density

- density is simply measured by the inverse of the *k*NN distance
- examples
  - *local outlier factor* (LOF)
    - LOF ≈ 1: observation is in a cluster
      (homogeneous density around the point and its neighbors)
    - LOF ≫ 1: point is an outlier
  - *connectivity-based outlier factor* (COF)
    - motivation: in regions of low density,
      it may be hard to detect outliers
    - how: treat “low density” and “isolation” differently
      (take the ε-neighborhood instead of *k*NN)
    - test multiple resolutions (varying ε)



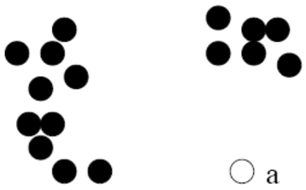


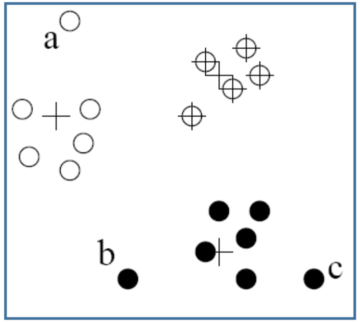
Outline

- Motivation
- Learning paradigms
- Statistical approaches
- Proximity-based approaches
- **Clustering approaches**
- Deep learning approaches
- Other approaches

Clustering approaches

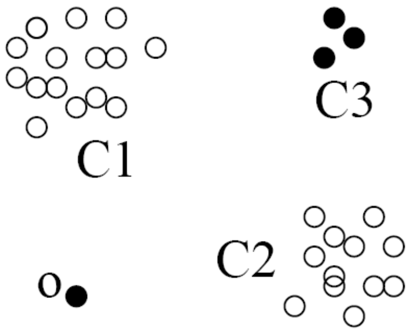
- Observation is an outlier:
  - **not belong to any cluster**
    - density-based clustering method such as DBSCAN
  - **far from its closest cluster**
    - partitioning-based clustering such as k-means
    - for each observation, assign an outlier score based on its distance from its closest center
    - if dist/averageDist is large, likely an outlier
  - **belongs to a small or sparse cluster**





Clustering approaches

- (cont.) **small or sparse clusters**
  - FindCBLOF: find clusters, sort them in decreasing size, compute statistic to detect significantly size differences
- Pros and cons of clustering-based approaches
  - *strengths*
    – work for many types of data (clusters regarded as summaries)
    – not requiring any labeled data
    – efficiency in detecting outliers once the cluster are obtained
  - *weakness*
    – effectiveness highly depends on the clustering method
    – clustering can be costly



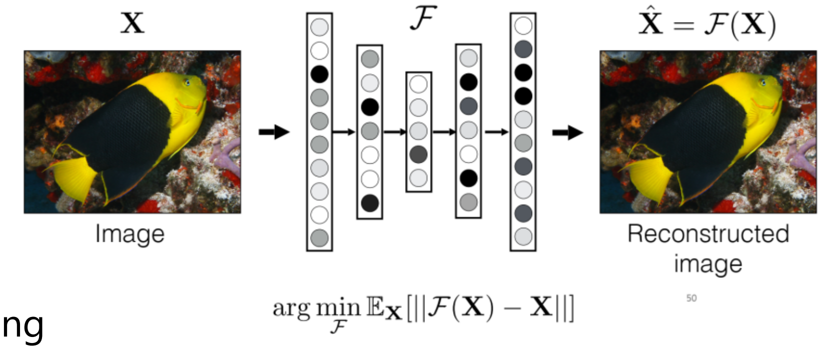
Outline

- Motivation
- Learning paradigms
- Statistical approaches
- Proximity-based approaches
- Clustering approaches
- **Deep learning approaches**
- Other approaches

Deep Learning approaches

Recall: **autoencoders** for reconstructing inputs, i.e. $\mathbf{x} \approx d(g(\mathbf{x}))$

- *goal*: learn the parameters of the encoder $g\colon X \to Z$ and decoder $d\colon Z \to X$ minimizing a reconstruction loss such as $\|\hat{\mathbf{x}} - \mathbf{x}\|^2$ where $\hat{\mathbf{x}} = d\big(g(\mathbf{x})\big)$
- *premise*: deviant behaviors are harder to reconstruct
- *principle*: reconstruction loss as a proxy for outlier likelihood

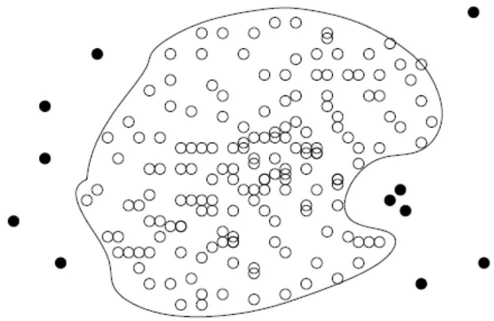


- *Pros*: inherent ability to handle complex data without an explicit data representation step
  - dense, recurrent, convolutional, transformer-based layering for multivariate, temporal, spatial, text content
- *Cons*: expressivity needs to be controlled (e.g. ensuring compact bottleneck) to avoid high capacity to reconstruct outliers

Outline

- Motivation
- Learning paradigms
- Statistical approaches
- Proximity-based approaches
- Clustering approaches
- Deep learning approaches
- **Other approaches**

Supervised approach: one-class model

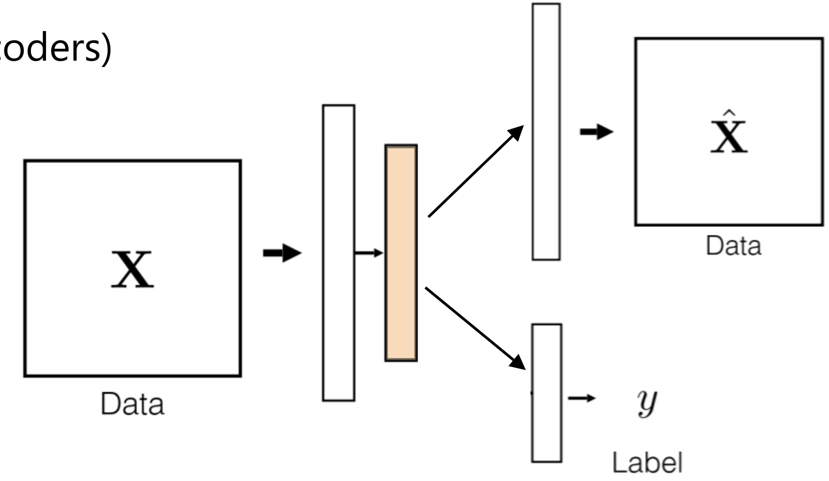


- **Idea**: in the presence of labeled data, train a classification model that can distinguish "normal" data from outliers
- **Problem**: training set is typically heavily biased (normal observations far exceeds outliers)
- Possible **solution**: one-class model
  - learn classifier to describe only the normal class.
  - learn the decision boundary of the normal class (e.g. using SVM)
  - observations that do not belong to the normal class (not within the decision boundary) declared as outliers
- **Advantage**: detect new outliers that may not appear close to any outlier in training data
- **Extension**: Normal objects may belong to multiple classes

Supervised approach: multi-task learning

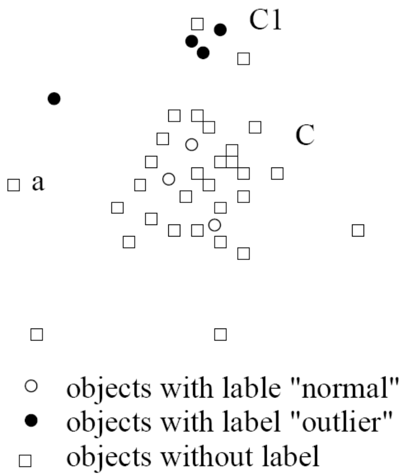
Hybrid neural networks

- combining *supervision* and *unsupervision* (autoencoders)
    - **single encoder**
    - **two decoders**
        - dedicated path for the supervised outlier detection task, other for reconstruction
        - parameters of decoders updated simultaneously or alternatively
- going beyond two dedicate paths…
  exploring synergies with other (un)supervised tasks



Semi-supervised approaches

- *How*
  - semi-supervised classification (e.g. pseudo-labeling)
  - semi-supervised clustering (e.g. membership constraints)
  - any object that does not fall into the model for C (such as a) is considered an outlier as well



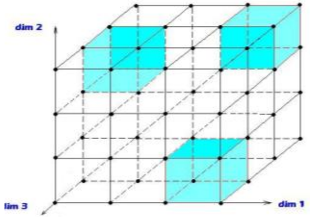
- *Pros and cons*
  - strengths: efficient, effective
  - bottlenecks:
    - quality heavily depends on the availability and quality of train data
    - difficult to obtain representative and high-quality training data

Outlier analysis in high-dimensional data

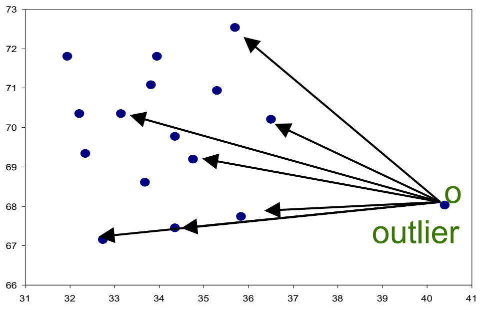
- **Challenges**
  - distance between observations becomes heavily dominated by noise
  - data in high-dimensional spaces are often sparse
  - interpretation of outliers: detecting outliers without saying why they are outliers is not very useful in high-dim data due to many involved features
- **Solutions**
  - use more robust distance functions and find full-dimensional outliers
  - find outliers in projections of the original feature space: *dimensionality reduction* works only when in lower-dimensional spaces normal instances can still be distinguished from outliers
    - PCA: components with low variance preferred since normal observations are likely close to each other and outliers often deviate from majority
    - use supervised reduction whenever possible

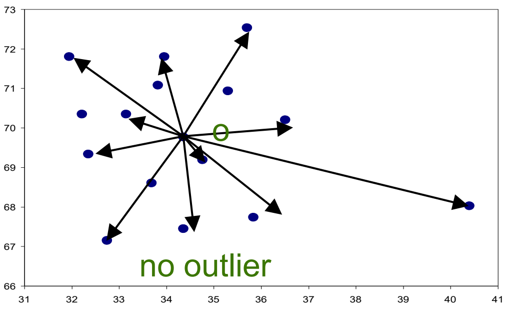
Outlier analysis in high-dimensional data

- **HilOut**: distance-based detection, yet uses ranks instead of absolute distances
- **subspaces**: find outliers in multiple lower dimensional subspaces: easy to interpret



- **ABOD**: angle-based outlier degree
  - angles are more stable than distances in high dimensional spaces
  - observation is outlier if most others are located in similar directions



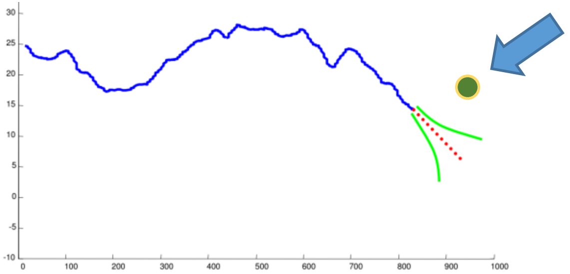


Outlier analysis on temporal data

Simple extension of traditional outlier detection techniques:

- learn good **representations** of time series data and apply previous techniques
- apply **time series clustering** and remove
  - observations untagged by density-based clustering approaches
  - observations belonging to very small clusters or clusters with loose silhouette
- model time series distributions (centroid or class-conditional "prototype" time series) and **test expectations** (how well a time series is described by the distribution)

Please note that **novelty detection** in a *single time series* (*left* image) differs from novelty detection in *time series data* (*above*)



Summary

- Outlier analysis aims to detect observations deviating from expectations
  - outliers either inconsistent with the rest data (**global**) or neighbors (**local**)
  - outliers can deviate in a given context, and appear in a group (**collective**)
- Given labeled examples, outlier analysis can be solved using **classification** with imbalanced classes (*supervised*) or guided by **semi-supervised** clustering
- **Statistical** approaches assume data is generated by a distribution to test the likelihood of an observation to be generated by the approximated distribution
- In **proximity**-based approaches, outliers have distant nearest neighbors or a density differs from the density around neighbors
- Small, compact **clusters** or unclustered observations also seen as outliers
- **Deep learning** offer expressive measures of unexpected behaviors from **reconstruction loss**
- Variants of outlier analysis for temporal data and high-dimensional data

Thank you!

Rui Henriques

rmch@tecnico.ulisboa.pt