Anomalies in Data

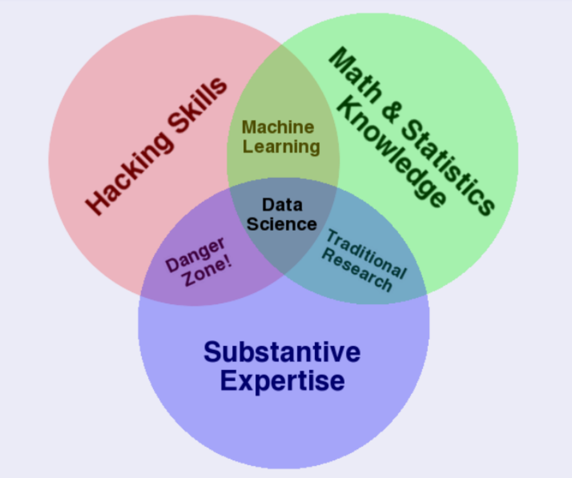
Maximilian Toller

KDDM2

SCIENCE
PASSION
TECHNOLOGY

Anomalies in Data

Recall from earlier



What are *Outliers*?

A recap from KDDM1

What are *Outliers*?

Definitions

- *An observation that appears to* ***deviate markedly*** *from other members of the sample in which it occurs*.
  (Grubbs, 1969)
- *An observation (or subset of observations) which appears to be* ***inconsistent*** *with the remainder of that set of data.*
  (Barnett and Lewis, 1974)
- *An observation, which deviates so much from other observations as to arouse suspicions that it was generated by a* ***different mechanism****.*
  (Hawkins, 1980)

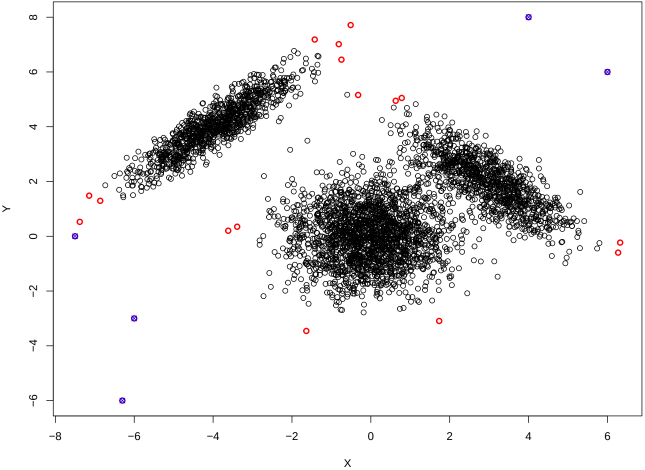
What are *Outliers*?

Examples (easy)

Inliers

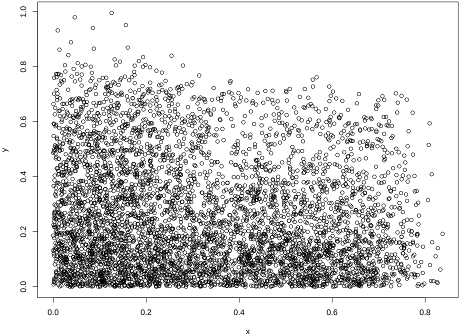
Outliers
(Grubb, Barnett)

Outliers
(Grubb, Barnett, Hawkins)



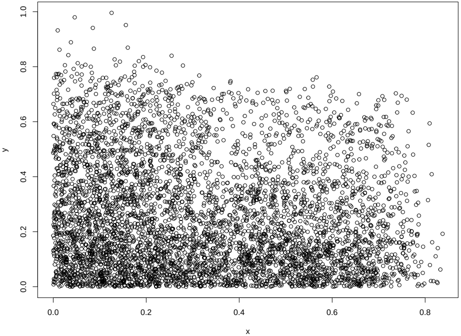
What are *Outliers*?

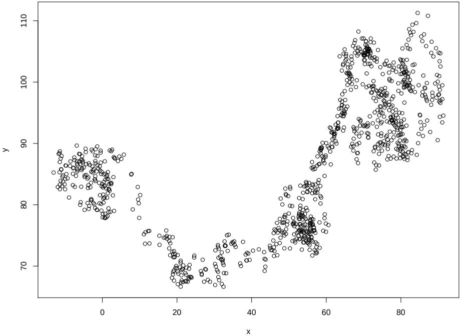
Examples (more difficult)



What are *Outliers*?

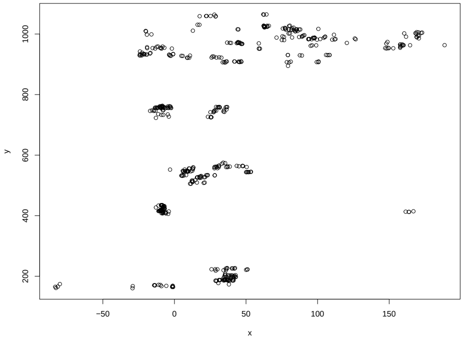
Examples (more difficult)





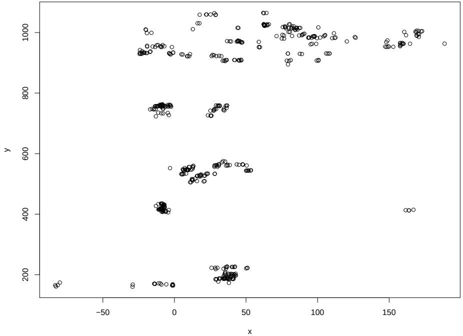
What are *Outliers*?

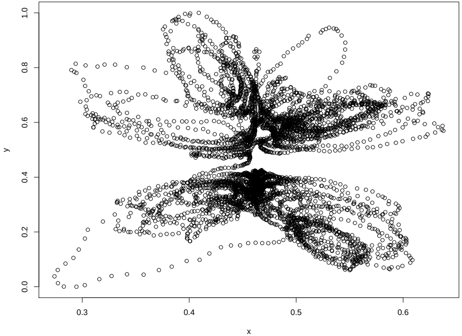
Examples (more difficult)



What are *Outliers*?

Examples (more difficult)





Methods: Preview

- There are many outlier detection methods:
  - Local outlier factor
  - Angle-based outlier degree
  - Artificial neural networks
  - . . .
- Why are there so many?

What are *Anomalies*?

What are *Anomalies*?

Difference from Outliers

- In literature, *outlier* and *anomaly* are used interchangeably
- For both, only vague definitions exist that are very similar
- However, the terms have different origins and different typical use:

Outliers typically...	**Anomalies** typically...
...are motivated by statistics.	...require context.
...are unusual data.	...are abnormal events.
...are investigated by traditional researches and statisticians.	...are investigated by data analysts and data scientists.

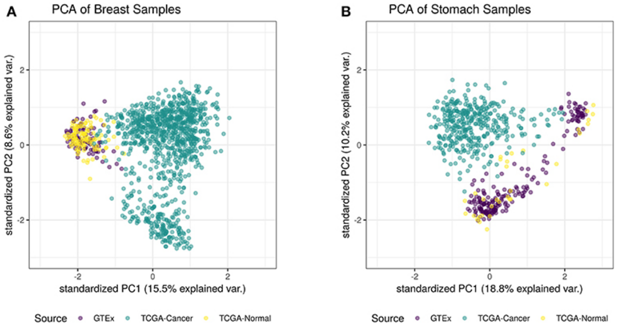
Example: Credit card fraud

- Billions of dollars lost every year
- Fraudulent transactions often significantly different
- Difficult to disguise fraud s.t. it is not visible on any scale

What are *Anomalies*?

Example: Cancer

- One of the most common causes of human death
- Disease with abnormal cell growth
- Cancer has abnormal gene expression signature



The role of context

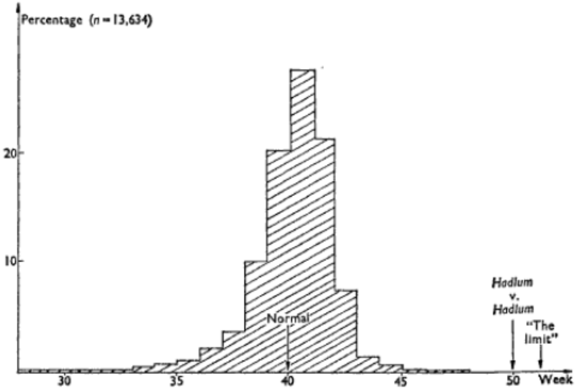
- Abnormality is context-dependent
  - Discordant data problem (credit card fraud example)
    - Many normal observations
    - Rare outlying data
  - Anomaly class problem (cancer example)
    - Normal data class
    - Anomaly classes
- Can data define abnormality?

Unlikely, Discordant and Contaminated Data

How to interpret suspicious data

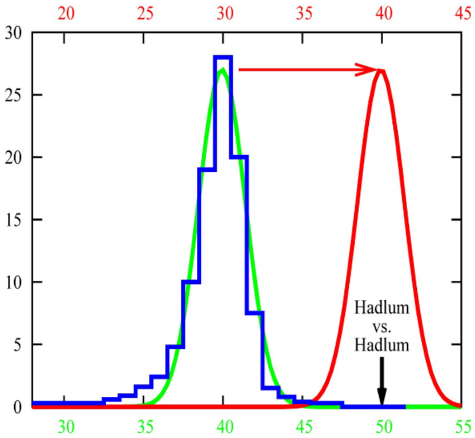
The Case of Hadlum vs Hadlum

- Mr Hadlum accuses Mrs Hadlum of adultery
- Sole evidence: Birth of child 349 days after Mr Hadlum left the country
- Average human gestation period: 280 days



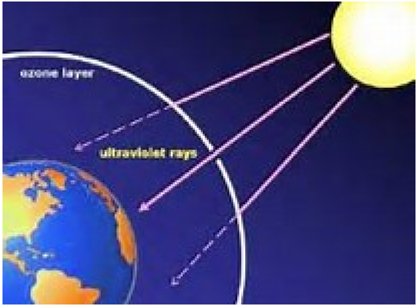
The Case of Hadlum vs Hadlum

- Mr Hadlum conjectured different distribution (red)
- Judges did not find Mrs Hadlum guilty, since 349 days unlikely, but not impossible (blue)
- (Modern research showed that more than 340 days is impossible)



The Antarctic Ozone Hole

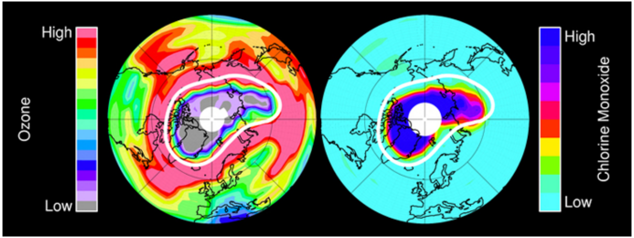
- Ozone layer protects Earth from solar radiation
- Damaged by human emissions of chlorofuorocarbons
- High depletion (hole) above poles



https://de.wikipedia.org/wiki/Datei:Ozone_layer.jpg

The (Ant)Arctic Ozone Hole

- Farman et al. (1985) discover hole in field study
- Authors hesitant to publish
- Nimbus satellite data showed no drop
- Problem: Largely deviating values discarded as measurement errors



NASA/JPL-Caltech

Definition

Unlikely data

- Position of judges
- "Random drop of ozone not caused by humans"
- Data unlikely but still normal
- No correction
- Action: none

Discordant data

- Position of Mr Hadlum
- Ozone field study by Farman et al. (1985)
- Data too unlikely to be normal
- Correction of model
- Action: investigate

Contamination

- "Wrong day of birth?”
- Satellite measurement error
- Data incorrect or misleading
- Correction of data
- Action: remove

Implications

- It is hard to classify data as *unlikely*, *discordant* or *contaminated*
- No universal decision criterion
- Domain knowledge as remedy
- Ultimately subjective

Strategies

1. Try to ignore anomalies (Not interesting)

2. Find anomalies for investigation or removal (Interesting)

Robust Statistics

Data Analysis in Presence of Anomalies

Introduction I

- Setting
  - Potentially contaminated dataset
  - Majority uncontaminated
  - Cannot find or remove contamination, e.g. inserted by attacker
- Task: Analyze data in spite of contamination, understand what is normal

Introduction II

- Challenges
  - No prior information about data
  - Contamination may be arbitrarily “bad” (adversarial)

- Question: Which methods are suitable?

Example: Mean and variance

- Two common estimators
  - Sample mean $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$
  - Sample variance $\hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2$
- Mean and variance are influenced by contamination
  - Original $x = [1, 3, 2, 1, 9, 2, 3, 2, 3, 2, 2, 1]$ $\quad \bar{x} \approx 2.58 \quad \hat{\sigma}_x^2 \approx 4.63$
  - Clean $\quad y = [1, 3, 2, 1, \quad 2, 3, 2, 3, 2, 2, 1]$ $\quad \bar{y} = 2 \quad \hat{\sigma}_y^2 = 0.6$

Example: Mean and variance

- What happens when attacker corrupts data unfavorably?

Example: Mean and variance

- What happens when attacker corrupts data unfavorably?
  - Attack #1 $\quad a_1 = [1, 3, 2, 1, 900, 2, 3, 2, 3, 2, 2, 1]$
    $\bar{a}_1 \approx 76.83 \quad \hat{\sigma}^2_{a_1} \approx 67200.88$

Example: Mean and variance

- What happens when attacker corrupts data unfavorably?
  - Attack #1 $a_1 = [1, 3, 2, 1, 900, 2, 3, 2, 3, 2, 2, 1]$
    $\bar{a}_1 \approx 76.83 \quad \hat{\sigma}^2_{a_1} \approx 67200.88$
  - Attack #2 $a_2 = [1, 3, 2, 1, 900000000, 2, 3, 2, 3, 2, 2, 1]$
    $\bar{a}_2 \approx 7.5 \times 10^7 \quad \hat{\sigma}^2_{a_2} \approx 6.75 \times 10^{16}$

Example: Mean and variance

- What happens when attacker corrupts data unfavorably?
  - Attack #1 $a_1 = [1, 3, 2, 1, 900, 2, 3, 2, 3, 2, 2, 1]$
    $\bar{a}_1 \approx 76.83$ $\hat{\sigma}^2_{a_1} \approx 67200.88$
  - Attack #2 $a_2 = [1, 3, 2, 1, 900000000, 2, 3, 2, 3, 2, 2, 1]$
    $\bar{a}_2 \approx 7.5 \times 10^7$ $\hat{\sigma}^2_{a_2} \approx 6.75 \times 10^{16}$
  - Attack #3 $a_3 = [1, 3, 2, 1, \infty, 2, 3, 2, 3, 2, 2, 1]$
    $\bar{a}_3 = \infty$ $\hat{\sigma}^2_{a_3} = \infty$

Example: Mean and variance

- What happens when attacker corrupts data unfavorably?
  - Attack #1 $a_1 = [1, 3, 2, 1, 900, 2, 3, 2, 3, 2, 2, 1]$
    $\bar{a}_1 \approx 76.83 \quad \hat{\sigma}^2_{a_1} \approx 67200.88$
  - Attack #2 $a_2 = [1, 3, 2, 1, 900000000, 2, 3, 2, 3, 2, 2, 1]$
    $\bar{a}_2 \approx 7.5 \times 10^7 \quad \hat{\sigma}^2_{a_2} \approx 6.75 \times 10^{16}$
  - Attack #3 $a_3 = [1, 3, 2, 1, \infty, 2, 3, 2, 3, 2, 2, 1]$
    $\bar{a}_3 = \infty \quad \hat{\sigma}^2_{a_3} = \infty$
- $\rightarrow$ Mean and variance are not *robust*.

Example: Median and MAD

- Two different estimators
  - Median $m(X)$
    - Any real number satisfying $P(X \leq m(X)) \geq 0.5$ and $P(X \geq m(X)) \geq 0.5$
    - For finite data $\boldsymbol{x} = [x_1, \ldots, x_n]$: $m(x) = \frac{x_{\lfloor (n+1)/2 \rfloor} + x_{\lceil (n+1)/2 \rceil}}{2}$ (middle value)
  - Median Absolute Deviation (MAD) $\zeta(x) = m(|\boldsymbol{x} - m(\boldsymbol{x})|)$

- Median and MAD are less influenced by contamination

- Median and MAD are less influenced by contamination
  - $a_1 = [1, 3, 2, 1, 900, 2, 3, 2, 3, 2, 2, 1]$
    $m(a_1) = 2 \quad \zeta(a_1) = 1$

Median and MAD are less influenced by contamination

- $a_1 = [1, 3, 2, 1, 900, 2, 3, 2, 3, 2, 2, 1]$
  $m(a_1) = 2 \quad \zeta(a_1) = 1$
- $a_2 = [1, 3, 2, 1, \infty, 2, 3, 2, 3, 2, 2, 1]$
  $m(a_2) = 2 \quad \zeta(a_2) = 1$

- Median and MAD are less influenced by contamination
  - $a_1 = [1, 3, 2, 1, 900, 2, 3, 2, 3, 2, 2, 1]$
    $m(a_1) = 2 \quad \zeta(a_1) = 1$
  - $a_2 = [1, 3, 2, 1, \infty, 2, 3, 2, 3, 2, 2, 1]$
    $m(a_2) = 2 \quad \zeta(a_2) = 1$
  - $a_3 = [\infty, 3, 2, \infty, \infty, 2, \infty, 2, 3, 2, 2, \infty]$
    $m(a_3) = 3 \quad \zeta(a_3) = 1$

- Median and MAD are less influenced by contamination
  - $a_1 = [1, 3, 2, 1, 900, 2, 3, 2, 3, 2, 2, 1]$
    $m(a_1) = 2 \quad \zeta(a_1) = 1$
  - $a_2 = [1, 3, 2, 1, \infty, 2, 3, 2, 3, 2, 2, 1]$
    $m(a_2) = 2 \quad \zeta(a_2) = 1$
  - $a_3 = [\infty, 3, 2, \infty, \infty, 2, \infty, 2, 3, 2, 2, \infty]$
    $m(a_3) = 3 \quad \zeta(a_3) = 1$
  - $a_4 = [\infty, \infty, 2, \infty, \infty, 2, \infty, 2, \infty, 2, 2, \infty]$
    $m(a_4) = \infty \quad \zeta(a_4) = \infty$

- Median and MAD are less influenced by contamination
  - $a_1 = [1, 3, 2, 1, 900, 2, 3, 2, 3, 2, 2, 1]$
    $m(a_1) = 2 \quad \zeta(a_1) = 1$
  - $a_2 = [1, 3, 2, 1, \infty, 2, 3, 2, 3, 2, 2, 1]$
    $m(a_2) = 2 \quad \zeta(a_2) = 1$
  - $a_3 = [\infty, 3, 2, \infty, \infty, 2, \infty, 2, 3, 2, 2, \infty]$
    $m(a_3) = 3 \quad \zeta(a_3) = 1$
  - $a_4 = [\infty, \infty, 2, \infty, \infty, 2, \infty, 2, \infty, 2, 2, \infty]$
    $m(a_4) = \infty \quad \zeta(a_4) = \infty$
- $\rightarrow$ Median and MAD are robust estimators of central tendency and dispersion

Definition

- A *statistic* $T(\cdot)$ maps data to single value, i.e. $T : \mathbb{R}^n \to \mathbb{R}$
- Examples: *mean*, *minimum*, χ^2 *tests*, ...
- Robust Statistics = Robust + $T(\cdot)$

Definition

A statistic $T(\cdot)$ is robust if it behaves favorably as the data it is computed on increasingly deviates from the assumptions made by $T(\cdot)$.

About mean and variance I

- What is estimated by
  - sample mean $\bar{x} = \hat{\mu}_X = \frac{1}{n}\sum_{j=1}^{n} x_i$?
  - sample variance $\hat{\sigma}_X = \frac{1}{n-1}\sum_{j=1}^{n}(x_i - \bar{x})^2$?
- By the strong law of large numbers (L.L.N.)
  - $\bar{x} \overset{\text{a.s.}}{\to} \mu_X = \mathbb{E}[X] \qquad (n \to \infty)$
  - $\hat{\sigma}_X \to \sigma_X \qquad (n \to \infty)$

About mean and variance II

- The strong L.L.N. *assumes* $x \overset{\text{iid}}{\sim} \mathcal{D}(\cdot)$.
- Anomalies typically follow a different distribution
  - A single anomaly might break iid assumption
  - $\bar{x}$ and $\hat{\sigma}_X$ become *biased* towards anomaly

Bias

- Mean $\bar{x}$ and median $m(x)$ are affected differently by contamination
- $\rightarrow$ Different amount of contamination needed to *bias* them
  - Single corrupted observation will add bias to $\bar{x}$
  - At least $\frac{n}{2}$ corrupted observations needed to bias $m(x)$
- Question: How do we measure the impact of contamination on bias?

Breakdown point I

Definition

Let $T_n(\cdot)$ be an estimator of θ and let $T_n(\boldsymbol{x}_n) = \hat{\theta}$. Further, let $0 < k < n$ observations in x_n be contamination to an arbitrary value. Then the breakdown point $\beta^\star$ of T_n is given by

$$\beta^\star_T(n) = \min\left\{\frac{k}{n}\middle| |\mathbb{E}[\hat{\theta}] - \theta| = \sup b(T_n, \theta)\right\}$$

Breakdown point II

- In simple terms
  - The smallest fraction of corrupted observations that T_n cannot handle
  - Assess robustness with
    1. Corrupt observation
    2. Check bias
    3. Repeat until worst possible output reached

Breakdown Point: Example

- Some breakdown points
  - Mean $\beta^{\star}_{\bar{x}}(n) = \frac{1}{n}$
  - IQR $\beta^{\star}_{I}(n) = \frac{n}{4}$
  - Median $\beta^{\star}_{m}(n) = \frac{n}{2}$
  - Perceptron $\beta^{\star}_{p}(n) = \frac{1}{n}$
- Easy to test on small dataset
  1. Contaminate a few observations
  2. See how statistic/algorithm behaves

Recap of last few slides I

- Robustness is about deviations from assumptions
  - Every meaningful statistic/algorithm $T(\cdot)$ assumes *something*
    (no-free lunch theorems)
  - Robust methods are consistent and become slowly biased towards contamination
- Robustness can be measured with the (asymptotic) breakdown point

Recap of last few slides II

- Want to test if $T(\cdot)$ is robust?

  1. Find dataset X where assumptions of $T(\cdot)$ hold
  2. Compute $T(X)$
  3. Contaminate X to X' so that assumptions of $T(\cdot)$ are violated
  4. Compute $T(X')$

Final Remark: Efficiency

- Robust methods are needed when anomalies in data
- Robustness alone is not enough
- $T(\cdot)$ also needs to be good at estimating θ
- Statistical efficiency

Anomaly Detection

Introduction

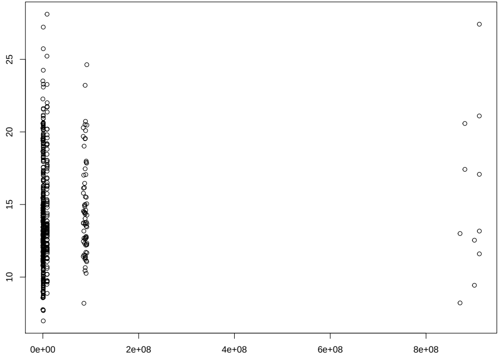
- There are many “anomaly detection” methods
  - Density-based techniques
  - One-class support-vector machines
  - Artificial neural networks
  - ...
- Why are there so many?
  - Performance depends largely on dataset (Why?)
  - There are many types of anomalies
  - Different settings require different methods

Objective

- Apparent goal: Detect when something unexpected/abnormal happens
- What data is available?
  - Given data might contain very many anomalies . . .
  - . . . or none.
- $\rightarrow$ True goal: Need to learn what is normal
- Normality is typically defined by the problem context, not by data

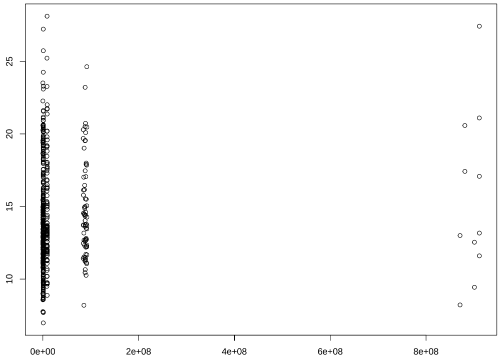
Anomaly Detection

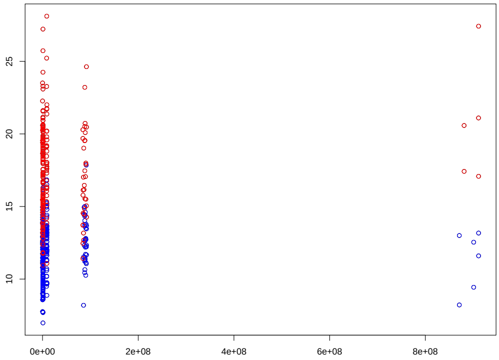
A classical pitfall



Anomaly Detection

A classical pitfall





How can we learn what is normal?

- Expert-based (traditional)
  1. Acquire domain expertise
  2. Analyze data and formulate rules
  3. Test rules

- Model-driven (traditional statistics)
  1. Understand problem
  2. Make assumptions and model
  3. Compare model with data

- Data-driven (data science)
  1. Analyze data
  2. Derive model from data & problem understanding
  3. Search deviations from model in data

How can we learn from data what is normal? I

1. Labeled data with normal and anomalous records
   - Goal: Learn to detect labeled anomalies
   - Reduction to classification problem
   - \+ Super easy compared to other settings!
   - – What about new anomalies?

How can we learn from data what is normal? II

2. Labeled data with only normal records (and maybe unlabeled data)

- Goal: Learn boundaries of what is normal
- No assumptions made about anomalies
- \+ Best setting for successful anomaly detection!
- – Setting very rare

How can we learn from data what is normal? III

3. Unlabeled data

- Goal: Find deviating data
- Hard to learn what is normal
- + Most common practical setting
- – Impossible to truly solve (needs strong assumptions)

Overview: Settings and Methods

1. Fully labeled data	Supervised anomaly detection
2. Labeled normal data	Unsupervised anomaly detection
3. Unlabeled data	Method-based anomaly detection

Setting: Fully Labeled Data

Overview I

- Setting
  - Labeled training set
  - Learn to classify normal and abnormal data
  - $\rightarrow$ Classification problem
- Examples
  - Distinguish between normal cell growth and cancer
  - Recognize attack signatures in normal web traffic

Setting: Fully Labeled Data

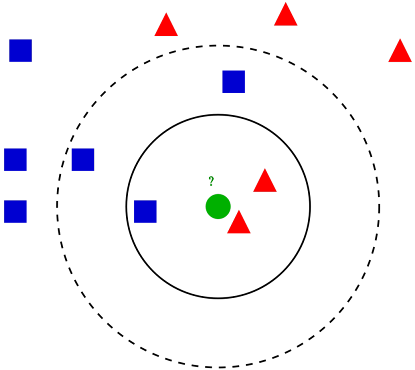
Overview II

- Suggested approach: Supervised learning
  - Statistical regression methods
  - Support vector machines
  - Classical neural networks
  - Deep neural networks
  - ...

Setting: Fully Labeled Data

Method 1.1: K-nearest neighbor classification

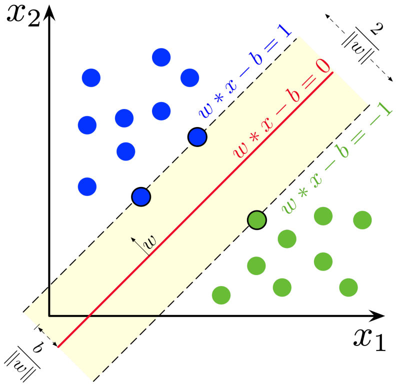
- Class of query is class of k^{th} nearest neighbor
- $\rightarrow$ Anomalies are close to each other
- Critical component: Distance function
  - Euclidean distance
  - Mahalanobis distance
  - ...



Setting: Fully Labeled Data

Method 1.2: Support Vector Machines

- Construct hyperplane that separates classes
- To solve nonlinear problems, needs extension
- Kernels
  - Polynomial
  - Radial basis function
  - Hyperbolic tangent



By Larhmam - Own work, CC BY-SA 3.0
https://commons.wikimedia.org/wiki/File:SVM_margin.png

Problems I

- While supervised methods learn to classify data as normal or anomalous. . .
- . . . they do not learn what is normal
  - Only boarder between seen anomalies and normal learned
  - Unseen anomalies not considered

Problems II

- Only applicable when all possible types of anomalies are known
- Examples:
  - Detect cheating at simple gambling $\rightarrow$ Always unusually high winnings
  - Classical (naive) anti-virus approaches $\rightarrow$ Learn attack signatures

Setting: Labeled Normal Data

Overview I

- Setting
  - Dataset with only normal data
  - Learn what is normal
  - Decide how likely unlabeled data are normal

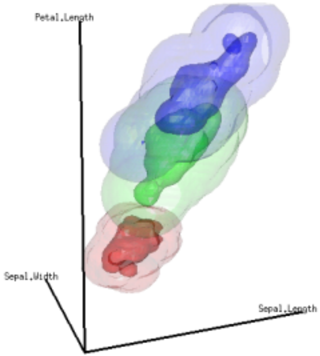
Overview II

- This is the most promising setting!
  - Not restricted to certain anomaly types
  - Ideal for handling new anomalies
- Labeled normal data rare in practice
- Suggested Approach: Unsupervised Learning

Setting: Labeled Normal Data

Method 2.1: Multivariate kernel density estimation

- Estimate probability density functions
- Assigns probabilities to entire space
- Assumption: Unlikely = Anomalous
- *Needs good kernel function*



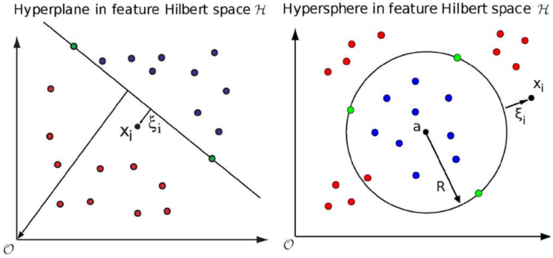
Duong, Tarn. "ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R." Journal of Statistical Software 21.7 (2007): 1-16.

Setting: Labeled Normal Data

Method 2.2: One-class support vector machines

- Planar approach
  - Hyperplane between data and origin
  - Maximize distance
- Spherical approach

  (support vector data descriptors)
  - Hypersphere around data
  - Minimize volume
- *Needs good kernel function*

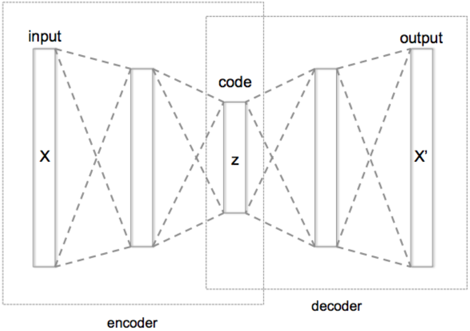


Muñoz-Marí, Jordi, et al. "Semisupervised one-class support vector machines for classification of remote sensing data." IEEE transactions on geoscience and remote sensing 48.8 (2010): 3188-3197.

Setting: Labeled Normal Data

Method 2.3: Autoencoders

- Learn to replicate data
- Collect reconstruction error for unlabeled queries
  - Low error: normal
  - High error: anomaly
- Important: Needs large training data set!



Setting: Unlabeled Data

Overview I

- Setting
  - Unlabeled dataset
  - No context information available
  - Limited domain expertise
- Worst scenario
  - How distinguish between normal and anomalous?
  - No method for learning normality
  - How can detection results be evaluated?

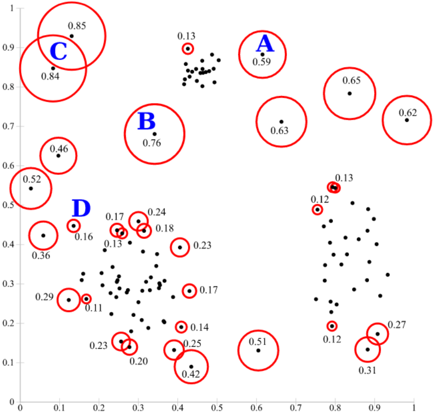
Overview II

- Solution: Make assumptions
  - No learning without assumptions (no free lunch theorems)
  - Assume that outliers according to method Y are anomalies
- Important: Use simple detection methods!

Setting: Unlabeled Data

Method 3.1: Local outlier probability

- Local Outlier Factor
  - Estimate local density
  - Low local density → anomaly
  - How to interpret deviation?
- Local Outlier Probability
  - Estimate local density
  - Estimate outlier probability

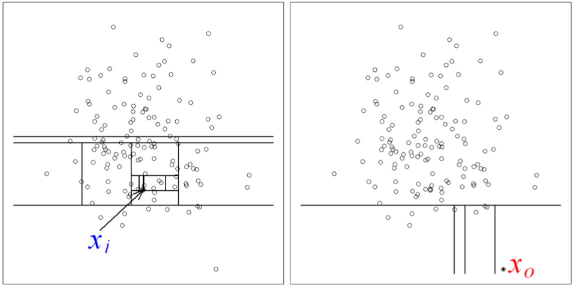


Kriegel, Hans-Peter, et al. "LoOP: local outlier probabilities." Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009.

Setting: Unlabeled Data

Method 3.2: Isolation forest

- Isolation tree
  1. Randomly split data with hyperplane
  2. Repeat until every point isolated
  3. Evaluate number of partitions
     - Few partitions to isolate → anomaly
     - Many partitions to isolate → inlier



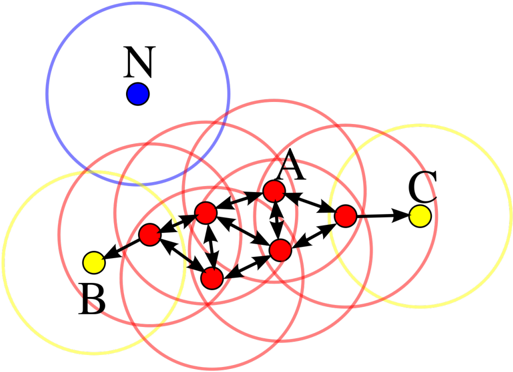
- Isolation Forest
  1. Grow many isolation trees
  2. Compare trees

Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest." 2008 Eighth IEEE International Conference on Data Mining. IEEE, 2008.

Setting: Unlabeled Data

Method 3.3: DBSCAN

- Cluster data according to density
  1. Compute k-NN distances
  2. Check which data have many neighbors
  3. Connect "dense" data
  4. Points in no cluster are anomalies
- Returns clustering and anomalies



LBy Chire - Own work, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=17045963

Final Remarks

Tools

- Robust statistics
  - https://cran.r-project.org/web/views/Robust.html
  - https://www.iumsp.ch/en/software/robust-statistics
  - AstroPy
- Anomaly detection
  - DDoutlier
  - ELKI
  - anomaly (R package)
  - scikit-learn
  - Tensorflow, Keras

Further Reading

Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." ACM computing surveys (CSUR) 41.3 (2009): 15.

Zimek, Arthur, and Peter Filzmoser. "There and back again: Outlier detection between statistical reasoning and data mining algorithms." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8.6 (2018): e1280.

Campos, Guilherme O., et al. "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study." Data Mining and Knowledge Discovery 30.4 (2016): 891-927.

Görnitz, Nico, et al. "Toward supervised anomaly detection." Journal of Artificial Intelligence Research 46 (2013): 235-262.

The End

Thank you for your attention!

References I

Barnett, V. and Lewis, T. (1974). *Outliers in statistical data*. Wiley.

Farman, J. C., Gardiner, B. G., and Shanklin, J. D. (1985). Large losses of total ozone in antarctica reveal seasonal clox/nox interaction. *Nature*, 315(6016):207.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21.

Hawkins, D. M. (1980). *Identification of outliers*, volume 11. Springer.

References II

Quinn, T. P., Nguyen, T., Lee, S. C., and Venkatesh, S. (2019). Cancer as a tissue anomaly: classifying tumor transcriptomes based only on healthy data. *Frontiers in genetics*, 10.

Zimek, A. and Filzmoser, P. (2018). There and back again: Outlier detection between statistical reasoning and data mining algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(6):e1280.