Anomaly detection

Prof. Dr. Stephan Trahasch

Offenburg University of Applied Sciences

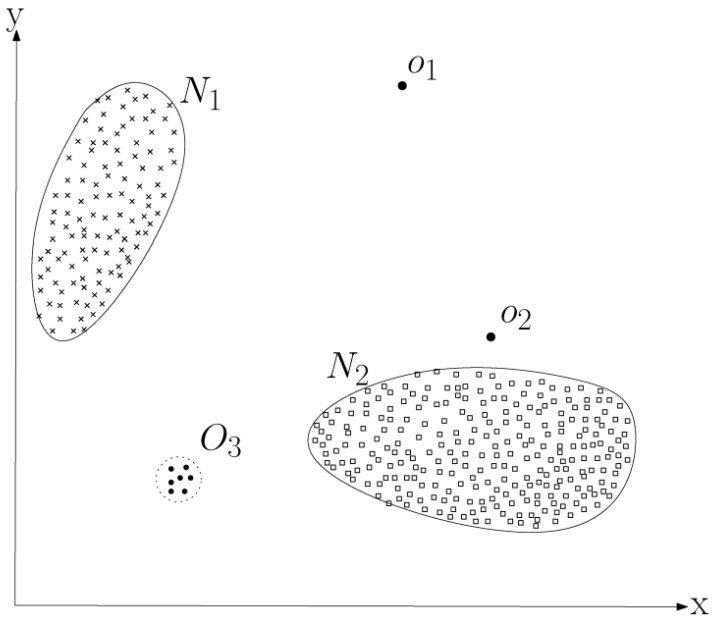
Some slides taken or adapted from:

"Anomaly Detection: A Tutorial" Arindam Banerjee, Varun Chandola, Vipin Kumar, Jaideep Srivastava, University of Minnesota
Aleksandar Lazarevic, United Technology Research Center

Outline

- Introduction
- Techniques for anomaly detection
  - Statistical
  - Proximity-based
  - Density-based
  - Cluster-based
- Summary

- Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior.
- These nonconforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities …



A simple example of anomalies in a two-dimensional data set.

Source: Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection. ACM Computing Surveys, 41(3), 1–58

What is an anomaly?

Definition of Hawkins [Hawkins 1980] :

> "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"

Statistics-based intuition

- Normal data objects follow a "generating mechanism", e.g. some given statistical process
- Abnormal objects deviate from this generating mechanism
- Anomaly: a data point or collection of data points that do not follow the same pattern or have the same structure as the rest of the data

- Historically, the field of statistics tried to find and remove outliers as a way to improve analyses.
- There are now many fields where the outliers / anomalies are the objects of greatest interest.
- The rare events may be the ones with the greatest impact, and often in a negative way.

Applications of anomaly detection

- Network intrusion
- Insurance / credit card fraud
- Healthcare informatics / medical diagnostics
- Industrial damage detection
- Image processing / video surveillance
- Novel topic detection in text mining
- …

Intrusion detection

- Intrusion detection
  - Monitor events occurring in a computer system or network and analyze them for intrusions
  - Intrusions defined as attempts to bypass the security mechanisms of a computer or network
- Challenges
  - Traditional intrusion detection systems are based on signatures of known attacks and cannot detect emerging cyber threats
  - Substantial latency in deployment of newly created signatures across the computer system
- Anomaly detection can alleviate these limitations

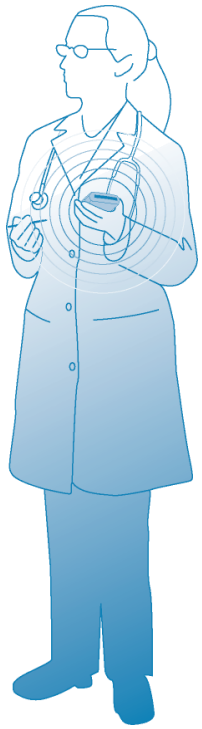
Fraud detection

- Detection of criminal activities occurring in commercial organizations.
- Malicious users might be:
  - Employees
  - Actual customers
  - Someone posing as a customer (identity theft)
- Types of fraud
  - Credit card fraud
  - Insurance claim fraud
  - Mobile / cell phone fraud
  - Insider trading

- Challenges
  - Fast and accurate real-time detection
  - Misclassification cost is very high
- Example: credit card transactions
  - €15.10 Amazon
  - €42.50 Deutsche Bahn tickets, Freiburg central station
  - €18.28 Frankfurt Airport
  - $500.00 Cash withdrawal. Muscat, Oman
  - $1000.00 Cash withdrawal. Cochin, India

Healthcare informatics

- Detect anomalous patient records
  - Indicate disease outbreaks, instrumentation errors, etc.
- Key challenges
  - Only normal labels available
  - Misclassification cost is very high
  - Data can be complex: spatio-temporal



Industrial damage detection

- Detect faults and failures in complex industrial systems, structural damages, intrusions in electronic security systems, suspicious events in video surveillance, abnormal energy consumption, etc.
  - Example: aircraft safety
    - anomalous aircraft (engine) / fleet usage
    - anomalies in engine combustion data
    - total aircraft health and usage management

- Key challenges
  - Data is extremely large, noisy, and unlabelled
  - Most of applications exhibit temporal behavior
  - Detected anomalous events typically require immediate intervention

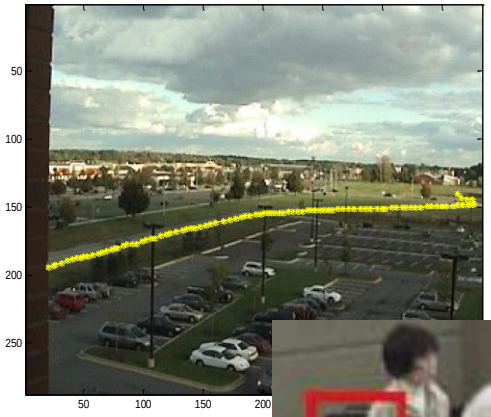
Errors in (Sensor) Data

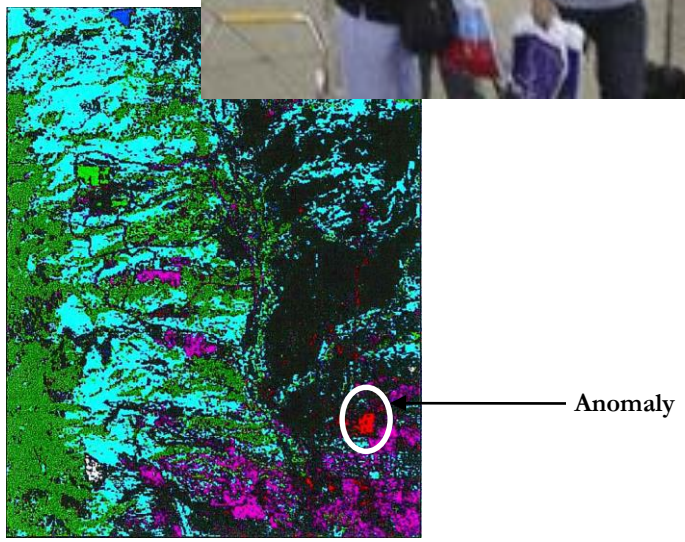
- malfunctioning sensors
- errors in manual data processing (e.g., twisted digits)
- storage/transmission errors
- encoding problems, misinterpreted file formats
- bugs in processing code
- …

https://www.flickr.com/photos/16854395@N05/3032208925/

Image processing

- Detecting outliers in a image mo
- Detecting anomalous regions wit
- Used in
  - mammography image analysis
  - video surveillance
  - satellite image analysis
- Key Challenges
  - Detecting collective anomalies
  - Data sets are very large





Causes of anomalies

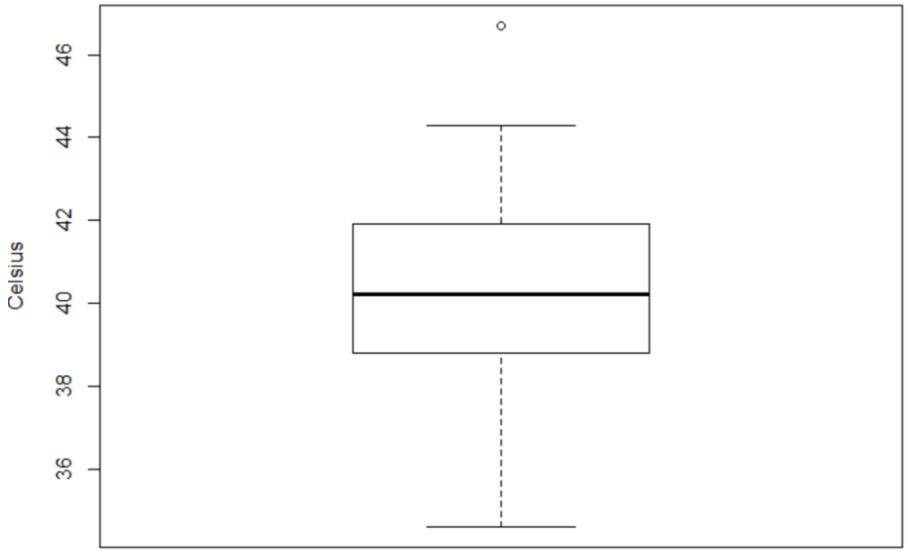
- Data from different class of object or underlying mechanism
  - disease vs. non-disease
  - fraud vs. not fraud
- Natural variation
  - tails on a Gaussian distribution
- Data measurement and collection errors

Types of anomalies: Point anomalies

An individual data instance is anomalous with respect to the data.

Examples:

A single 48C daily high temperature among a set of ordinary spring days.



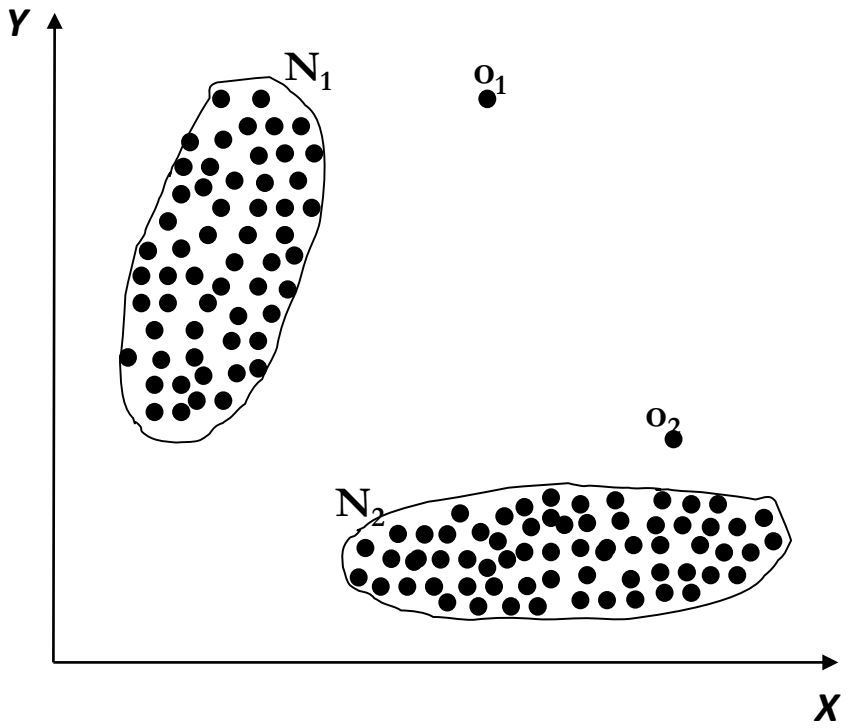

```
> data(maxtemp)
> boxplot(maxtemp, ylab = "Celsius")
```

Types of anomalies: Point anomalies

An individual data instance is anomalous with respect to the data.

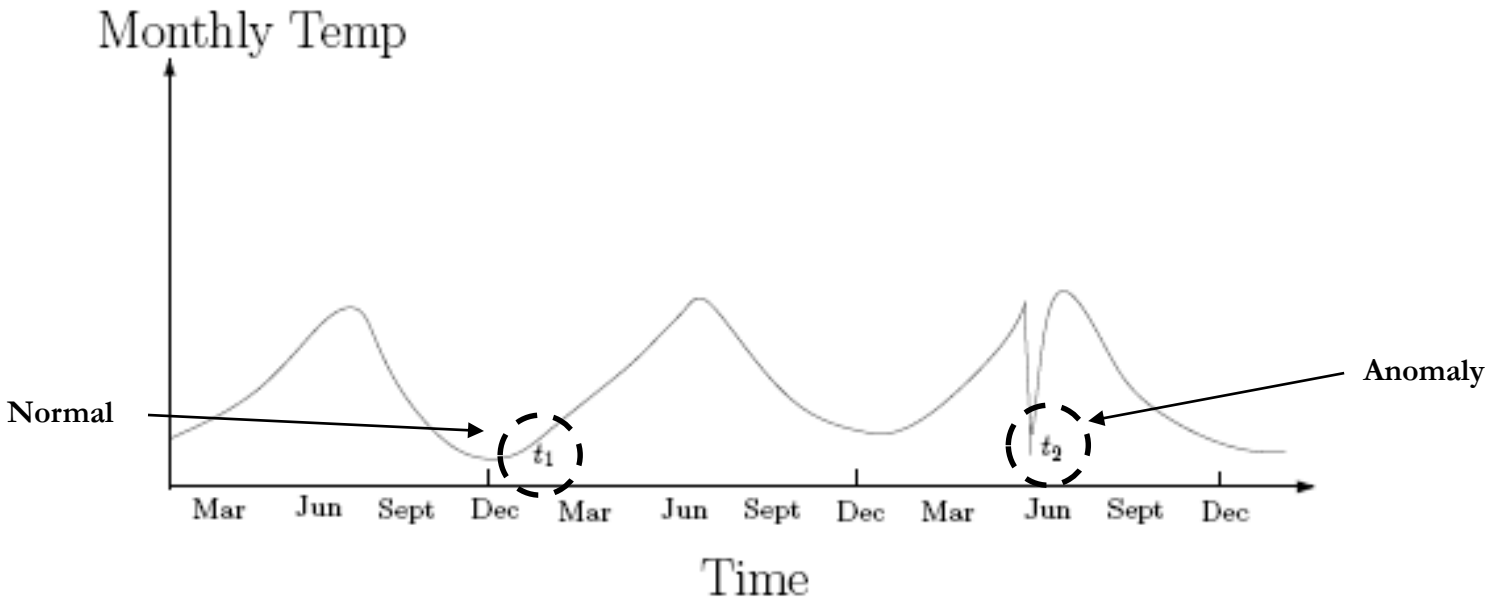
Examples:

A single 30C daily high temperature among a set of ordinary spring days.



Types of anomalies: Contextual anomalies

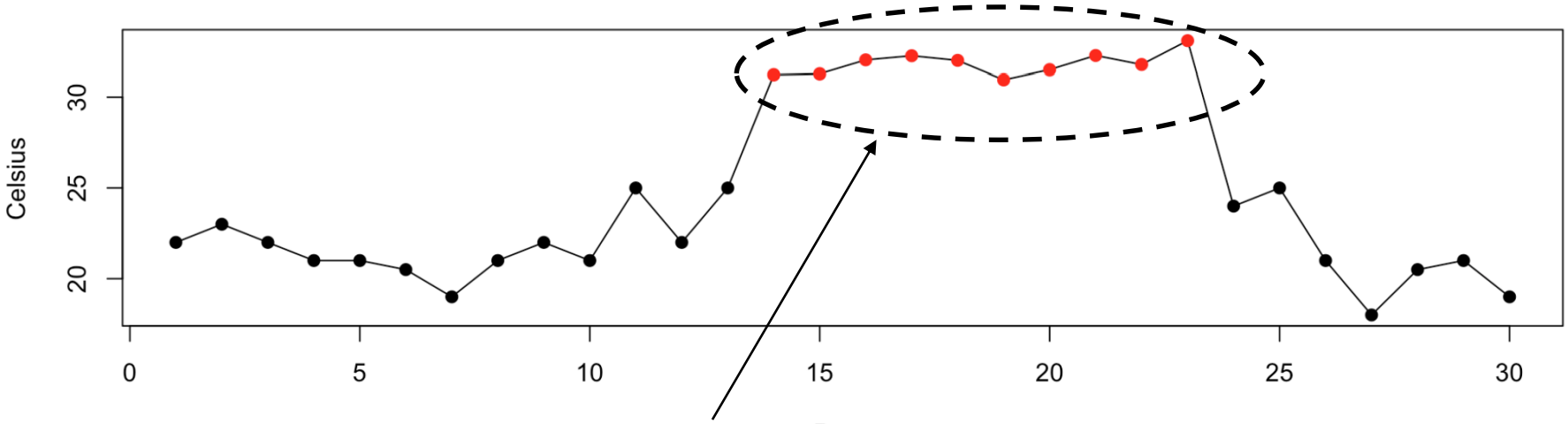
- An individual data instance is anomalous within a context
- Requires a notion of context
- Also referred to as conditional anomalies *



* Song, et al, "Conditional Anomaly Detection", IEEE Transactions on Data and Knowledge Engineering, 2006.

Types of anomalies: Collective anomalies

- A collection of related data instances is anomalous
- Requires a relationship among data instances
  - Sequential data
  - Spatial data
  - Graph data
- The individual instances within a collective anomaly are not anomalous by themselves



Use of data labels in anomaly detection

- Supervised anomaly detection
  - Labels available for both normal data and anomalies
  - Similar to classification with high class imbalance
- Semi-supervised anomaly detection
  - Labels available only for normal data
- Unsupervised anomaly detection
  - No labels assumed
  - Based on the assumption that anomalies are very rare compared to normal data

Output of anomaly detection

- Label
  - Each test instance is given a *normal* or *anomaly* label
  - Typical output of classification-based approaches

- Score
  - Each test instance is assigned an anomaly score
    - allows outputs to be ranked
    - requires an additional threshold parameter

Variants of anomaly detection problem

- Given a dataset D, find all the data points **x** $\in$ D with anomaly scores greater than some threshold t.

- Given a dataset D, find all the data points **x** $\in$ D having the top-n largest anomaly scores.

- Given a dataset D, containing mostly normal data points, and a test point **x**, compute the anomaly score of **x** with respect to D.

Unsupervised anomaly detection

- No labels available
- Based on assumption that anomalies are very rare compared to "normal" data

- General steps
  - Build a profile of "normal" behavior
    - summary statistics for overall population
    - model of multivariate data distribution
  - Use the "normal" profile to detect anomalies
    - anomalies are observations whose characteristics differ significantly from the normal profile

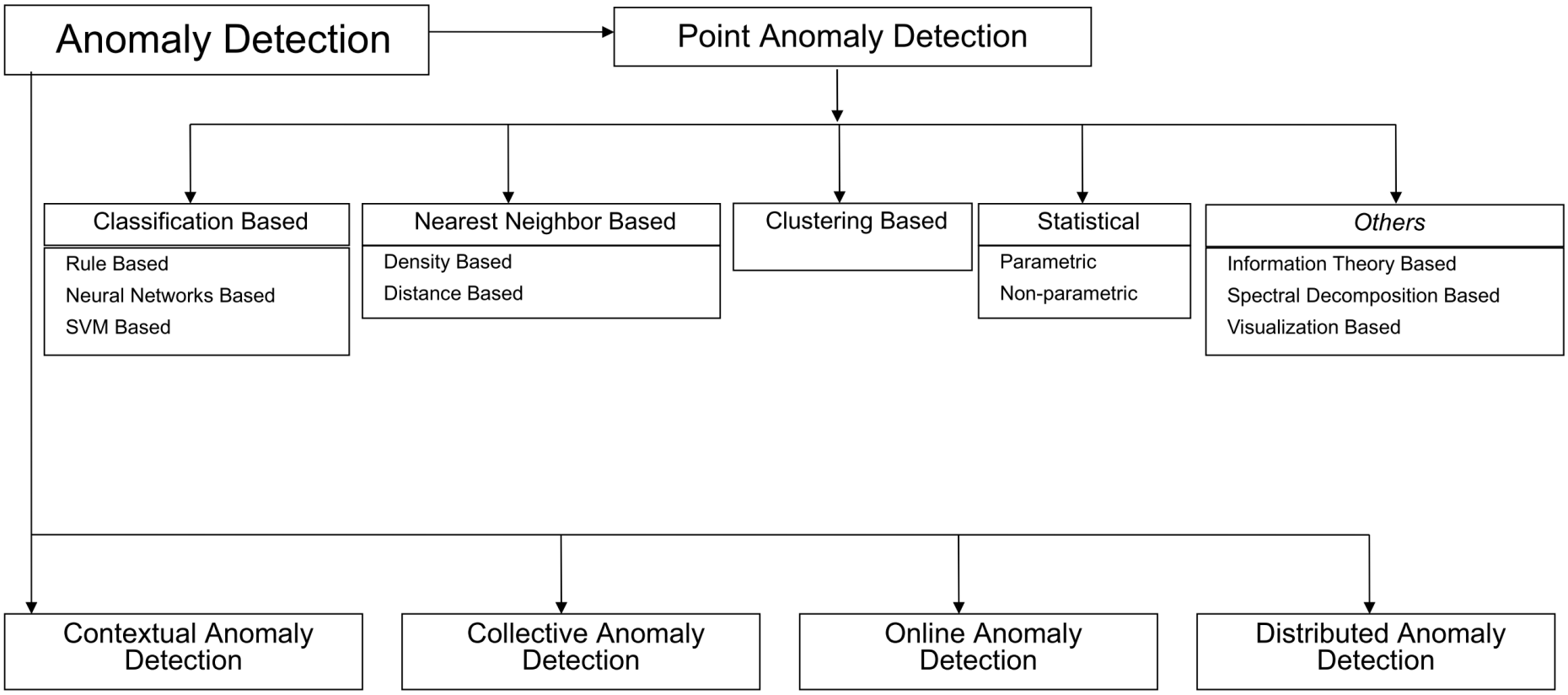
Outline

- Introduction
- Techniques for anomaly detection
  - Statistical
  - Proximity-based
  - Density-based
  - Cluster-based
- Summary

General application scenarios

- Supervised scenario
  - In some applications, training data with normal and abnormal data objects are provided objects are provided
  - There may be multiple normal and/or abnormal classes
  - Often, the classification problem is highly imbalanced
- Semi-supervised Scenario
  - In some applications, only training data for the normal class(es) (or only the abnormal class(es)) are provided
- Unsupervised Scenario
  - In most applications there are no training data available

Taxonomy


Anomaly Detection
Point Anomaly Detection
Classification Based
Rule Based
Neural Networks Based
SVM Based
Nearest Neighbor Based
Density Based
Distance Based
Clustering Based
Statistical
Parametric
Non-parametric
Others
Information Theory Based
Spectral Decomposition Based
Visualization Based
Contextual Anomaly Detection
Collective Anomaly Detection
Online Anomaly Detection
Distributed Anomaly Detection

Outline

- Introduction
- Techniques for anomaly detection
  - Statistical
  - Proximity-based
  - Density-based
  - Cluster-based
- Summary

Statistical outlier detection

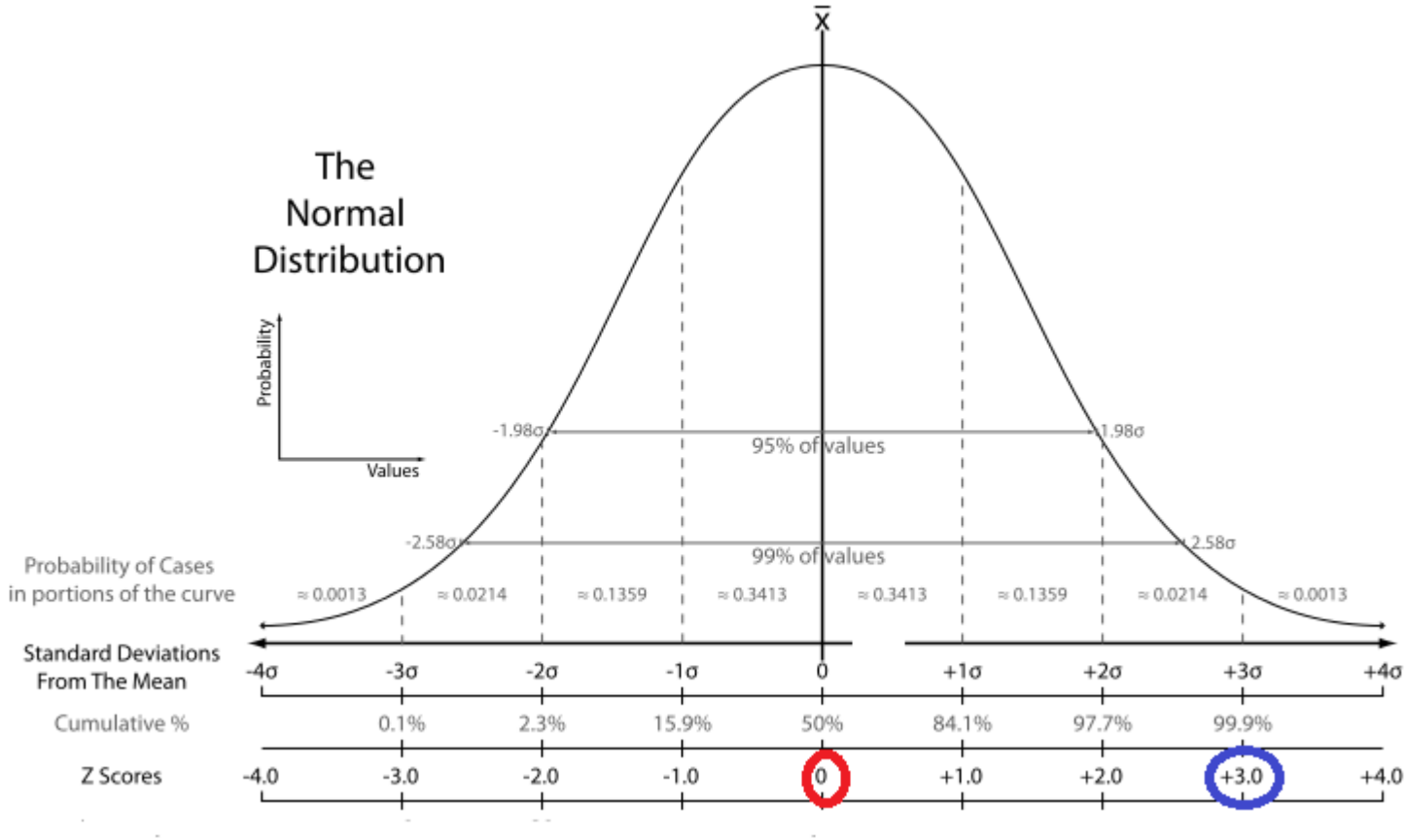
Outliers are objects that are fit poorly by a statistical model.

- Estimate a parametric model describing the distribution of the data
- Apply a statistical test that depends on
  - Properties of test instance
  - Parameters of model (e.g., mean, variance)
  - Confidence limit (related to number of expected outliers)

Statistical outlier detection

Univariate Gaussian distribution

Outlier defined by z-score > threshold

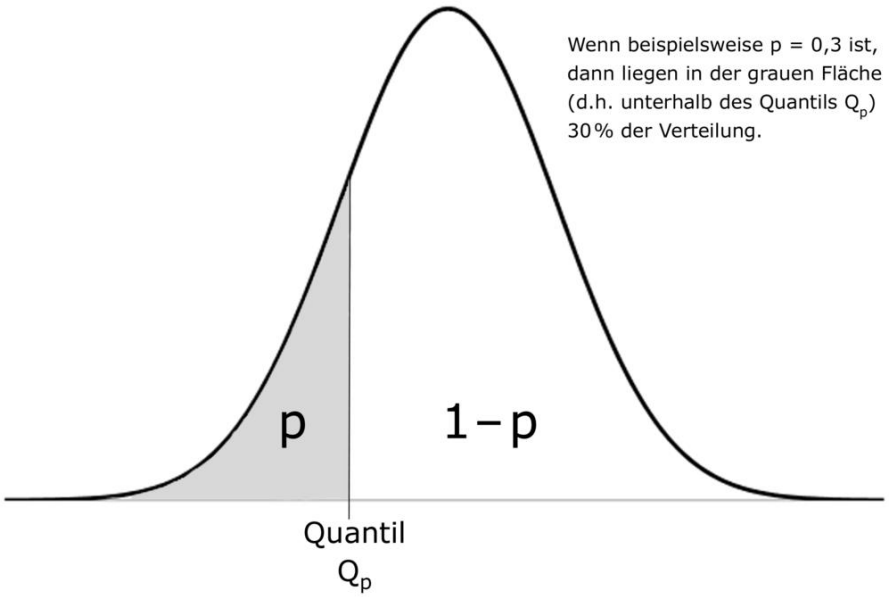


https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2013/09/The_Normal_Distribution.svg_1.png

Quantile

In analogy to the median, any quantiles can be defined as:

p% of the data values (ordered by size) are smaller or equal, 100-p% are greater or equal to the quantile in question.



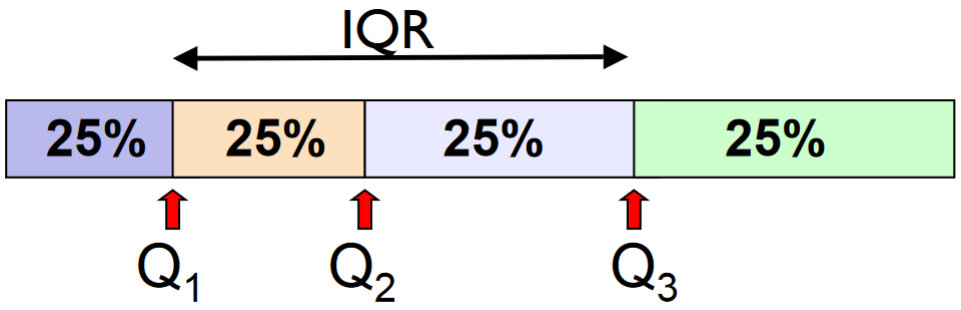
With the help of several quantiles you can easily characterize a distribution.

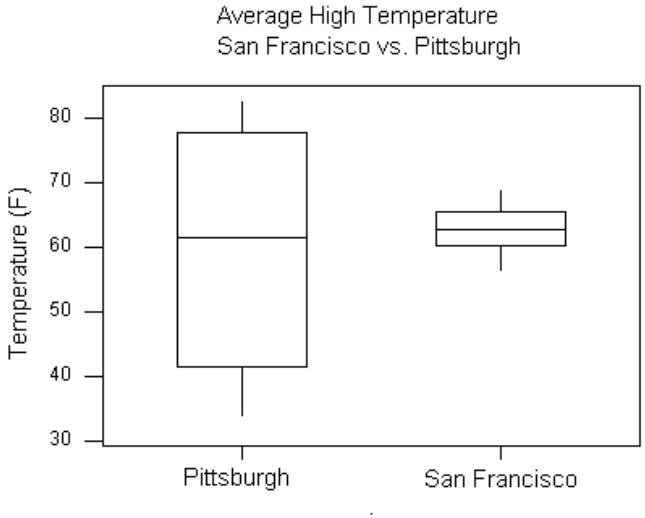
Interquartile Range: IQR

Between Q1 and Q3 is the middle 50% of the data. The range these data span is called the interquartile range, or the IQR.

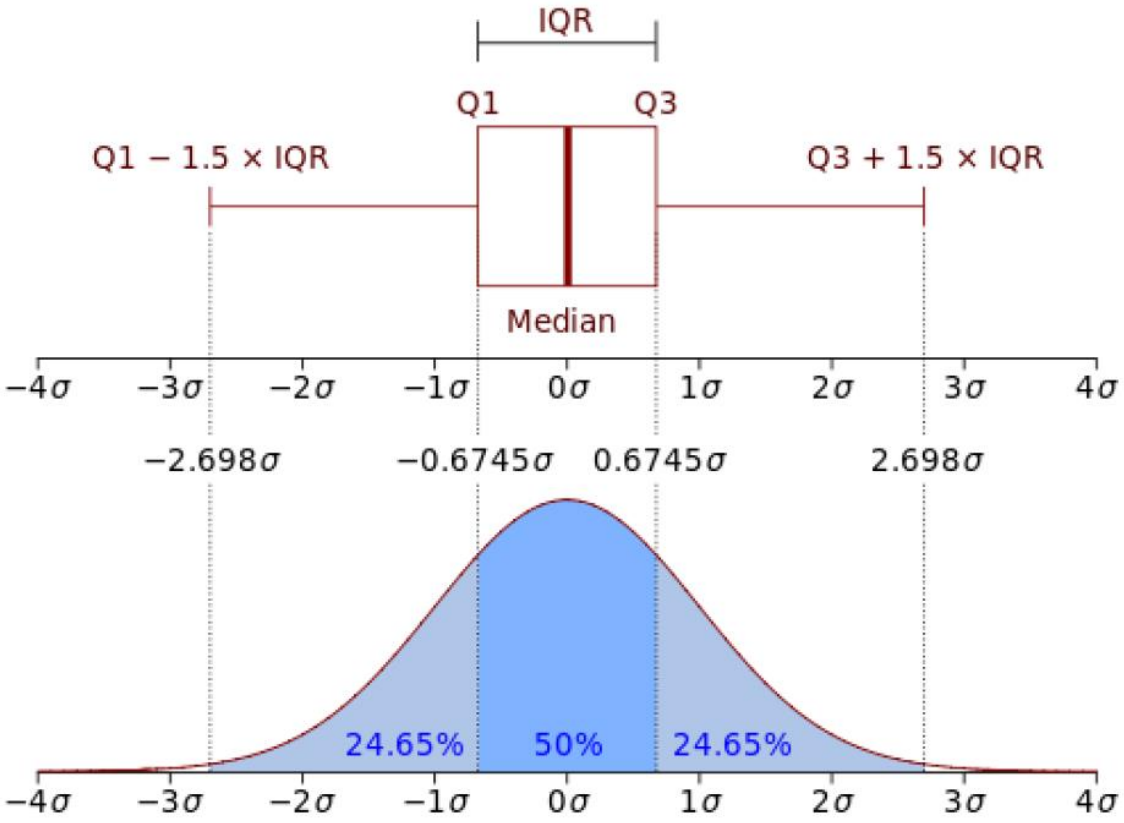
IQR = Q3 - Q1

Like SD, the IQR is a measure of spread. The IQR tells us the spread of the middle 50% of the data.





Interquartile Range assumes a normal distribution



Outliers

An outlier is an observation that appears extreme relative to the rest of the data.

One numerical rule of thumb for identifying outliers is values that are further than 1.5 x IQR below Q1 or above Q3. These are denoted with a dot on a box plot.

Outlier on upper end > Q3 + 1.5 x IQR

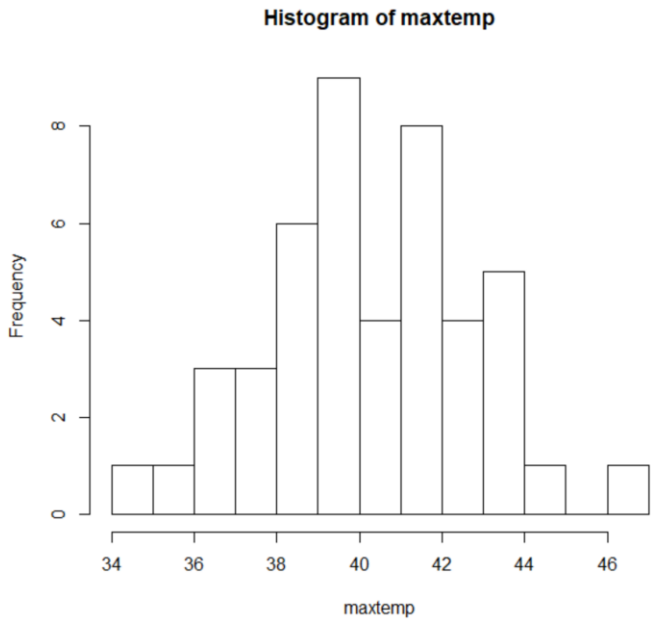
Outlier on lower end < Q1 - 1.5 x IQR

Given a data set with: Q1 = 10, Q3 = 20;

Testing the extremes with Grubbs' test

- Statistical test to decide if a point is outlying
- Requires checking the normality assumption first
- Assume data comes from normal distribution
- Detects one outlier at a time, remove the outlier, and repeat

Invented by Frank E. Grubbs (1913-2000)




```
> hist(maxtemp, breaks = 10)
```

Grubbs' test

- Detect outliers in univariate data $x_1, \ldots, x_n$
- H_0: There is no outlier in data
- H_A: There is at least one outlier

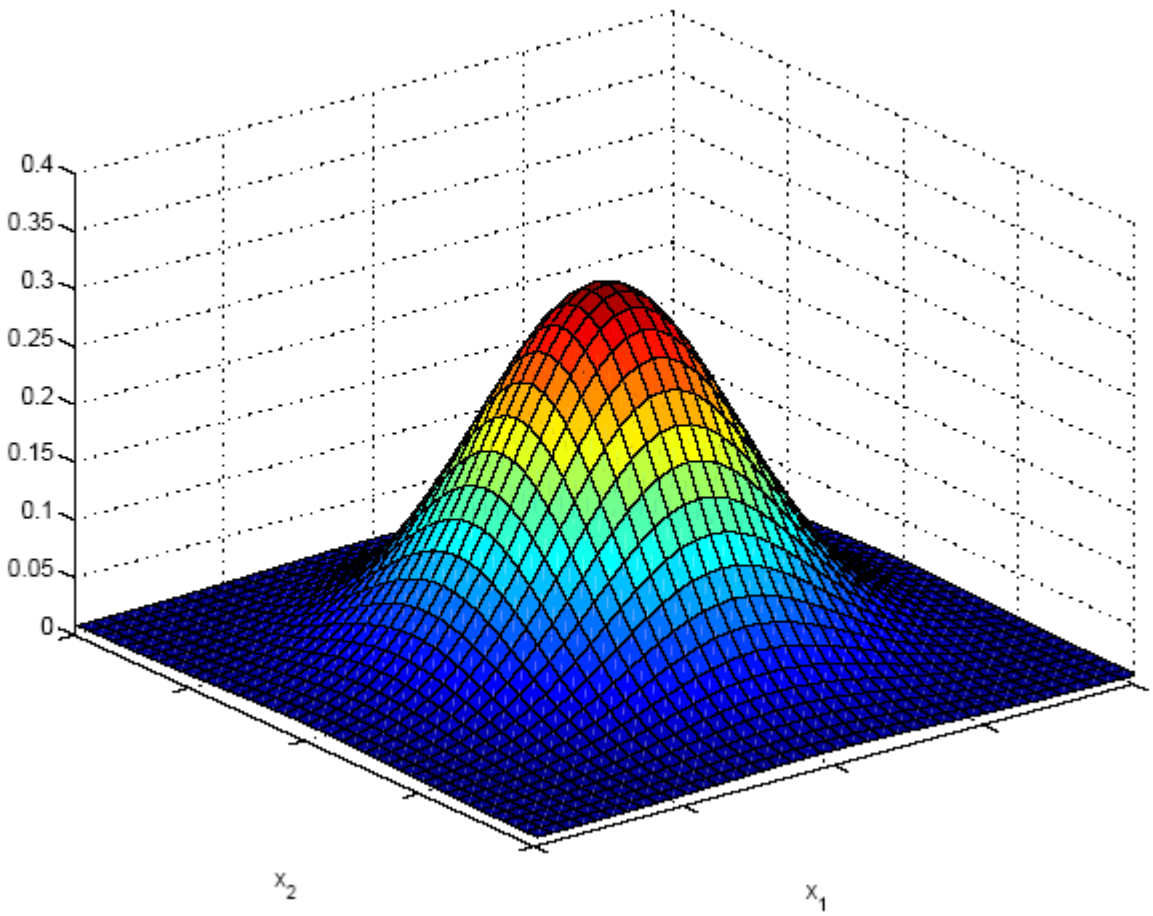
Grubbs' test statistic: $G = \frac{\max |X - \bar{X}|}{s}$

Reject H_0 if: $G > \frac{(n-1)}{\sqrt{n}} \sqrt{\frac{t^2_{(\frac{\alpha}{2n}, n-2)}}{n-2+t^2_{(\frac{\alpha}{2n}, n-2)}}}$

```
> grubbs.test(maxtemp)
        Grubbs test for one outlier
data: maxtemp
G.2009 = 2.52920, U = 0.85469, p-value = 0.206
alternative hypothesis: highest value 46.7 is an outlier
```

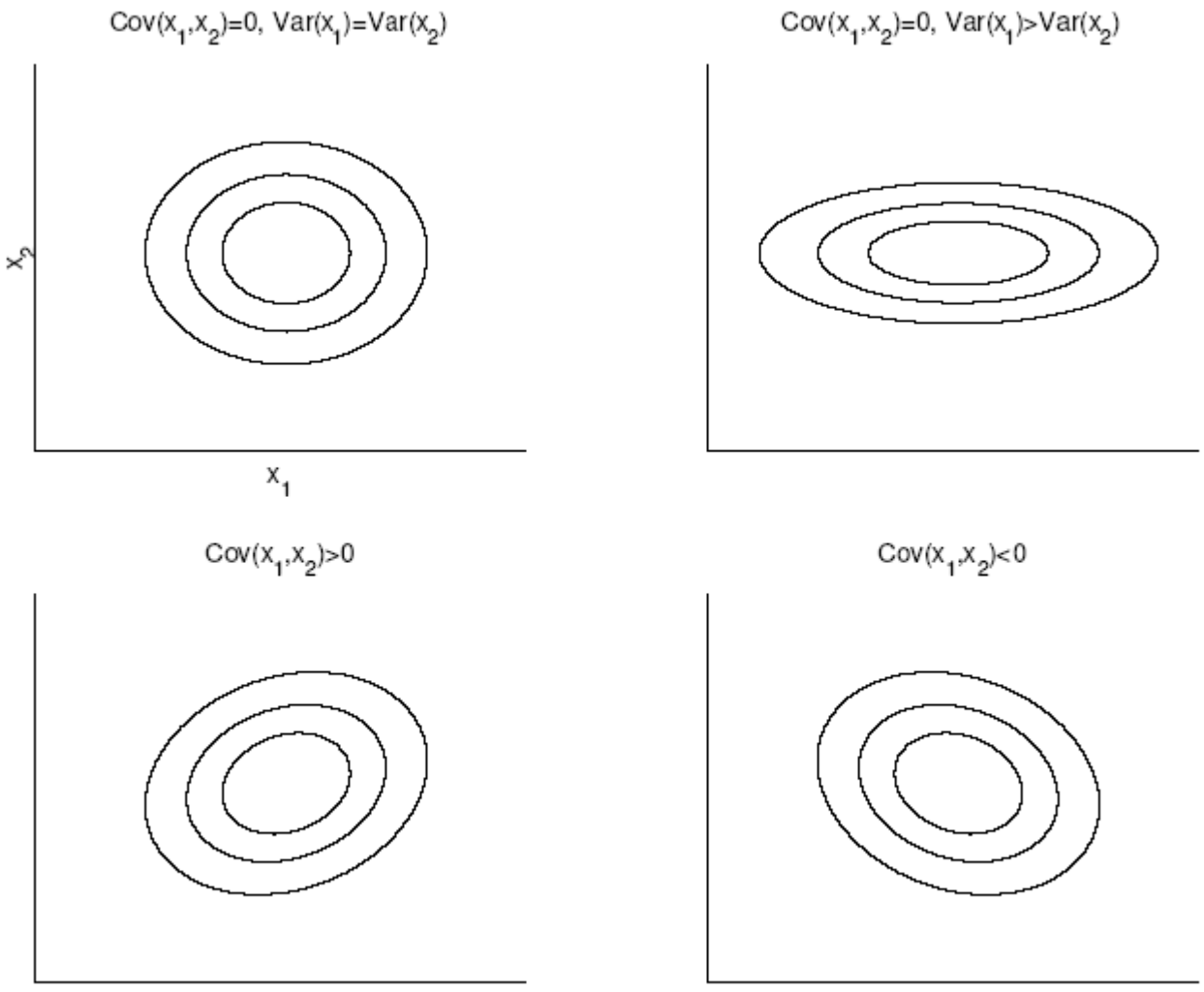
Multivariate Normal Distribution

$$p(x) = \frac{1}{(2\pi)^{\frac{d}{2}} * \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \qquad x \sim N_d(\mu, \Sigma)$$



Source: Lecture Notes for E Alpaydın
2004 Introduction to Machine Learning

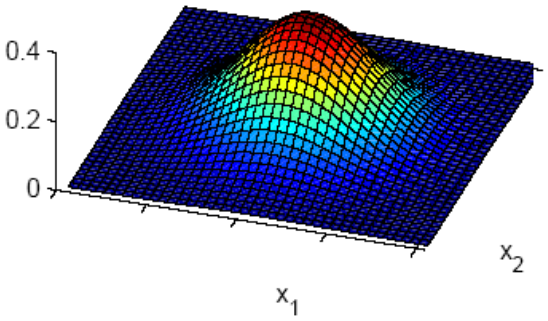
Bivariate Normal



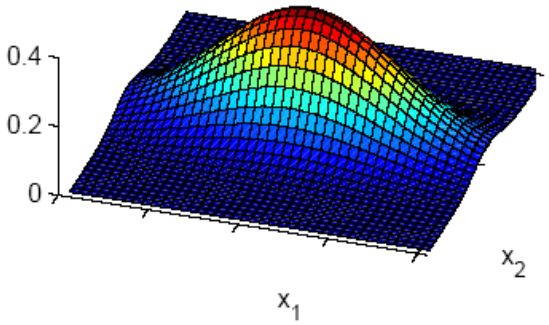
Source: Lecture Notes for E Alpaydın
2004 Introduction to Machine Learning

3 Dimensions

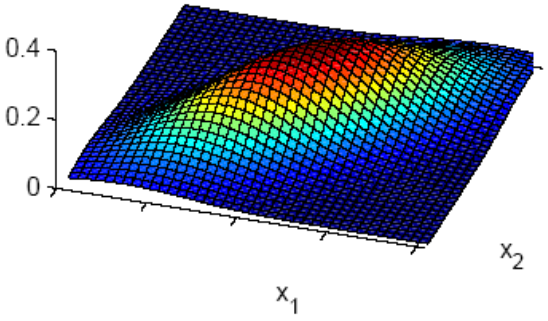
$\text{Cov}(x_1,x_2)=0$, $\text{Var}(x_1)=\text{Var}(x_2)$



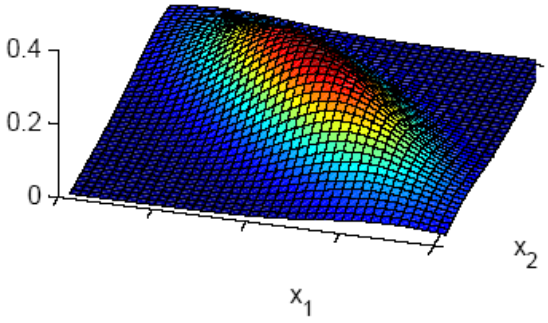
$\text{Cov}(x_1,x_2)=0$, $\text{Var}(x_1)>\text{Var}(x_2)$



$\text{Cov}(x_1,x_2)>0$



$\text{Cov}(x_1,x_2)<0$



Source: Lecture Notes for E Alpaydın 2004 Introduction to Machine Learning

Mahalanobis Distance

To account for differences in variance between the variables, and to account for correlations between variables, we use the Mahalanobis distance.

Mahalanobis distance: $D^2 = (x - \mu)^T \Sigma^{-1} (x - \mu)$

measures the distance from x to μ in terms of Σ
(normalizes for difference in variances and correlations)

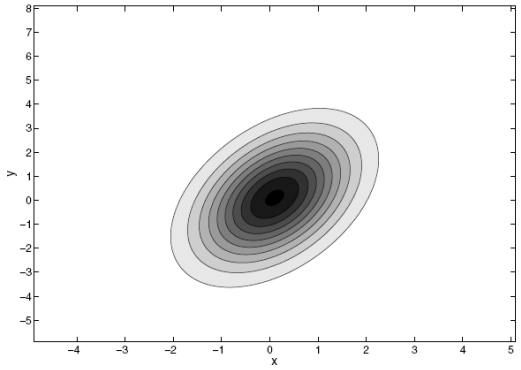
Bivariate: $d = 2$ $\quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \left(z_1^2 - 2\rho z_1 z_2 + z_2^2 \right) \right]$$

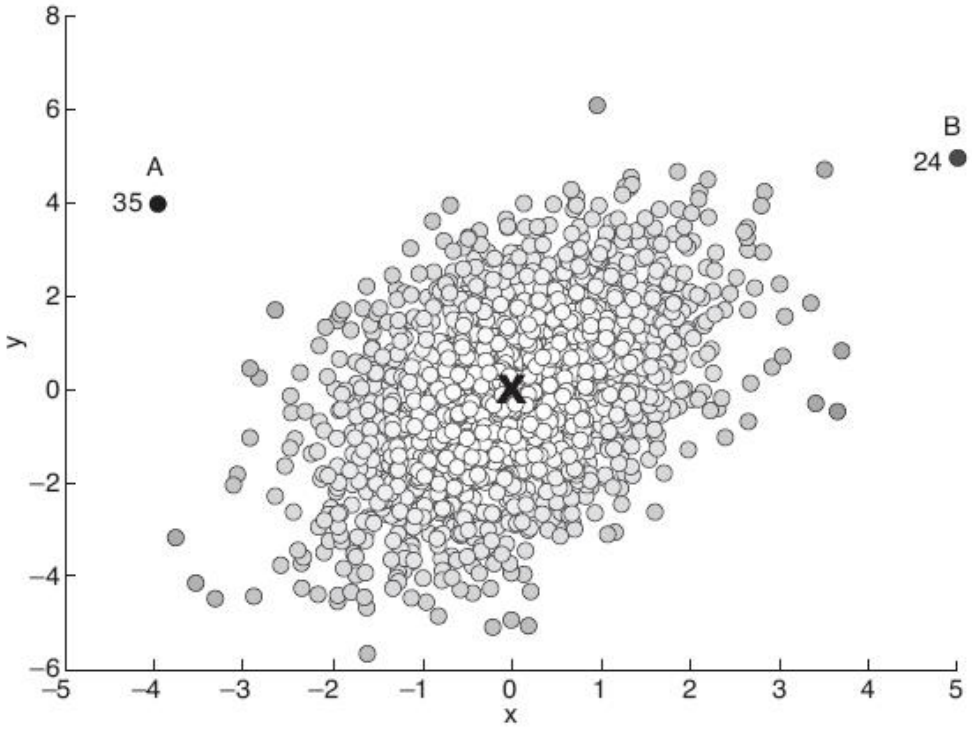
$$z_i = (x_i - \mu_i) / \sigma_i$$

Statistical anomaly detection
Multivariate Gaussian distribution

Outlier defined by Mahalanobis distance > threshold



	Distance	
	Euclidean	Mahalanobis
A	5.7	35
B	7.1	24



Likelihood approach

- Assume the dataset D contains samples from a mixture of two probability distributions:
  - M (majority distribution)
  - A (anomalous distribution)
- General approach:
  - Initially, assume all the data points belong to M
  - Let $L_t(D)$ be the log likelihood of D at time t
  - For each point x_t that belongs to M, move it to A
    - Let $L_{t+1}(D)$ be the new log likelihood.
    - Compute the difference, $\Delta = L_t(D) - L_{t+1}(D)$
    - If $\Delta > c$ (some threshold), then x_t is declared as an anomaly and moved permanently from M to A

Likelihood approach

- Data distribution, D = (1 – λ) M + λ A
- M is a probability distribution estimated from data
  - Can be based on any modeling method (naïve Bayes, maximum entropy, etc)
- A is initially assumed to be uniform distribution
- Likelihood at time t:

$$L_t(D) = \prod_{i=1}^{N} P_D(x_i) = \left((1-\lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i) \right) \left(\lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i) \right)$$

$$LL_t(D) = |M_t| \log(1-\lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

Statistical outlier detection

- Pros
  - Statistical tests are well-understood and well-validated.
  - Quantitative measure of degree to which object is an outlier.
- Cons
  - Data may be hard to model parametrically.
    - multiple modes
    - variable density
  - In high dimensions, data may be insufficient to estimate true distribution.

Outline

- Introduction
- Techniques for anomaly detection
  - Statistical
  - Proximity-based
  - Density-based
  - Cluster-based
- Summary

Proximity-based outlier detection

Outliers are objects far away from other objects.

- Common approach:
  - Outlier score is distance to k^{th} nearest neighbor.
  - Score sensitive to choice of k.

Proximity-based outlier detection

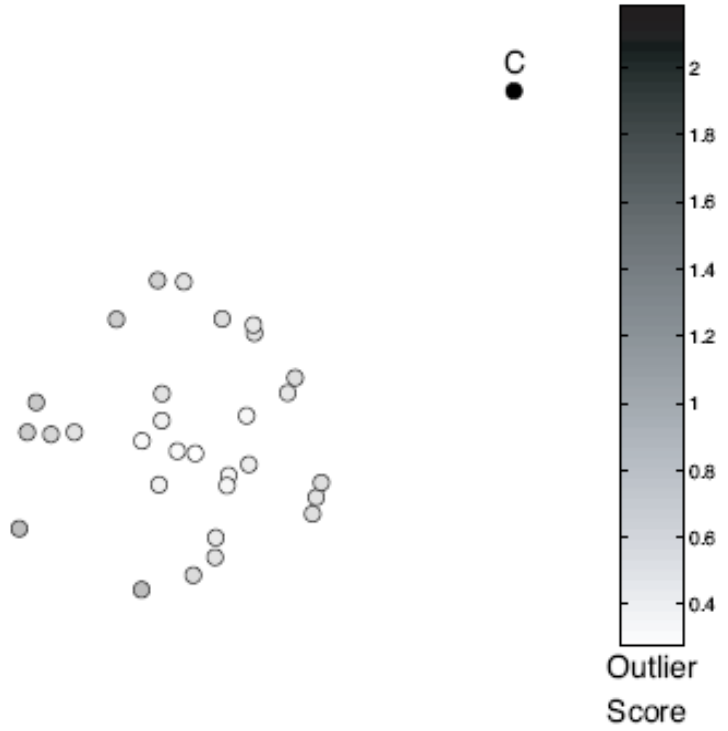


Figure 10.4. Outlier score based on the distance to fifth nearest neighbor.

Proximity-based outlier detection

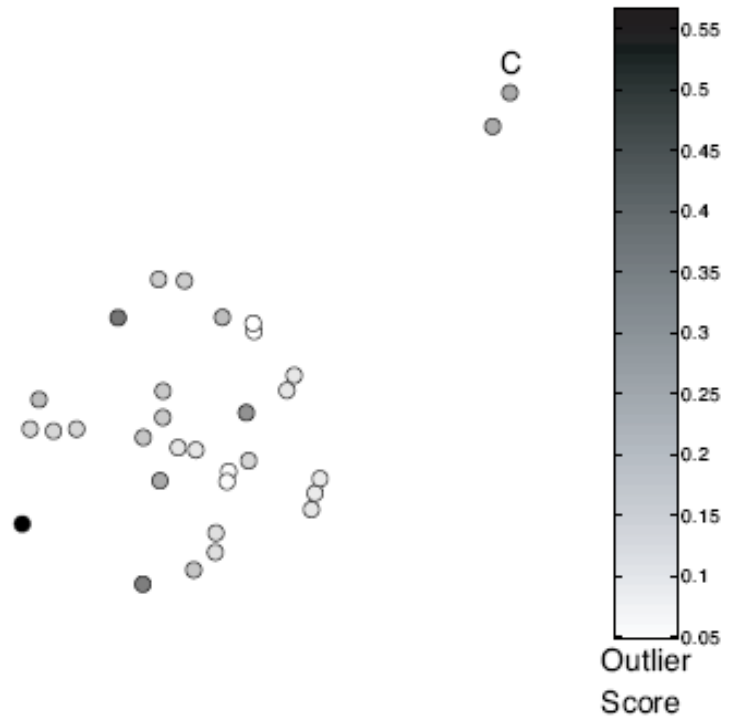


Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

Proximity-based outlier detection

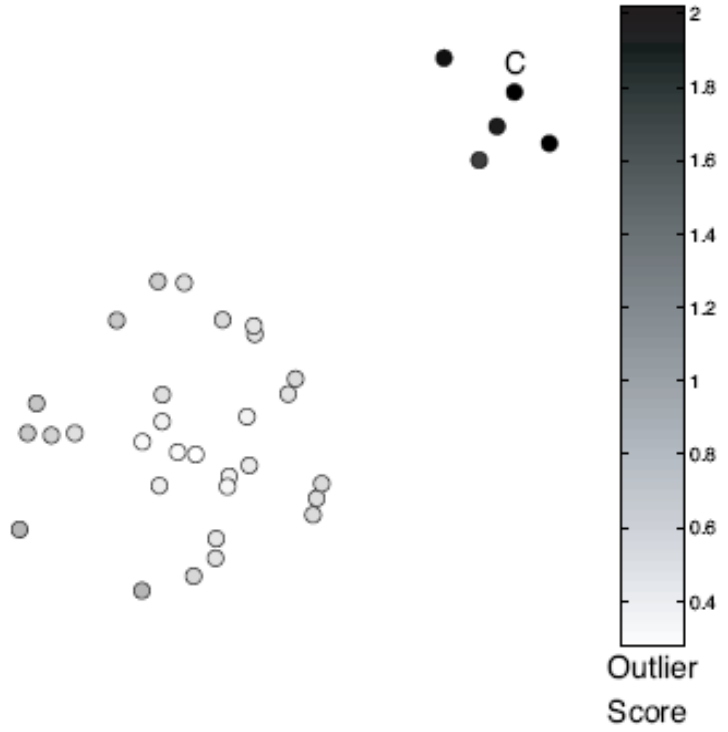


Figure 10.6. Outlier score based on distance to the fifth nearest neighbor. A small cluster becomes an outlier.

Proximity-based outlier detection

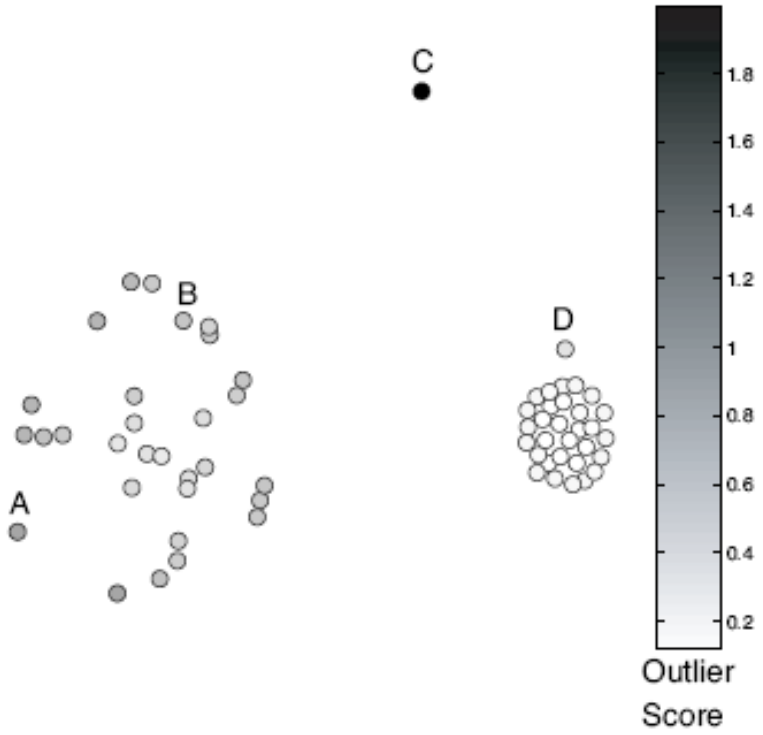
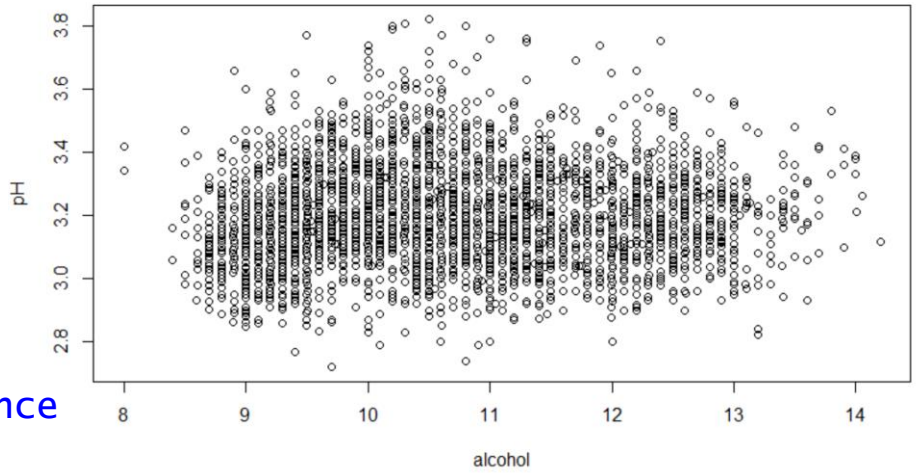


Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

Example

```
> # Scatterplot of wine pH against alcohol
> plot(pH ~ alcohol, data = wine)
```

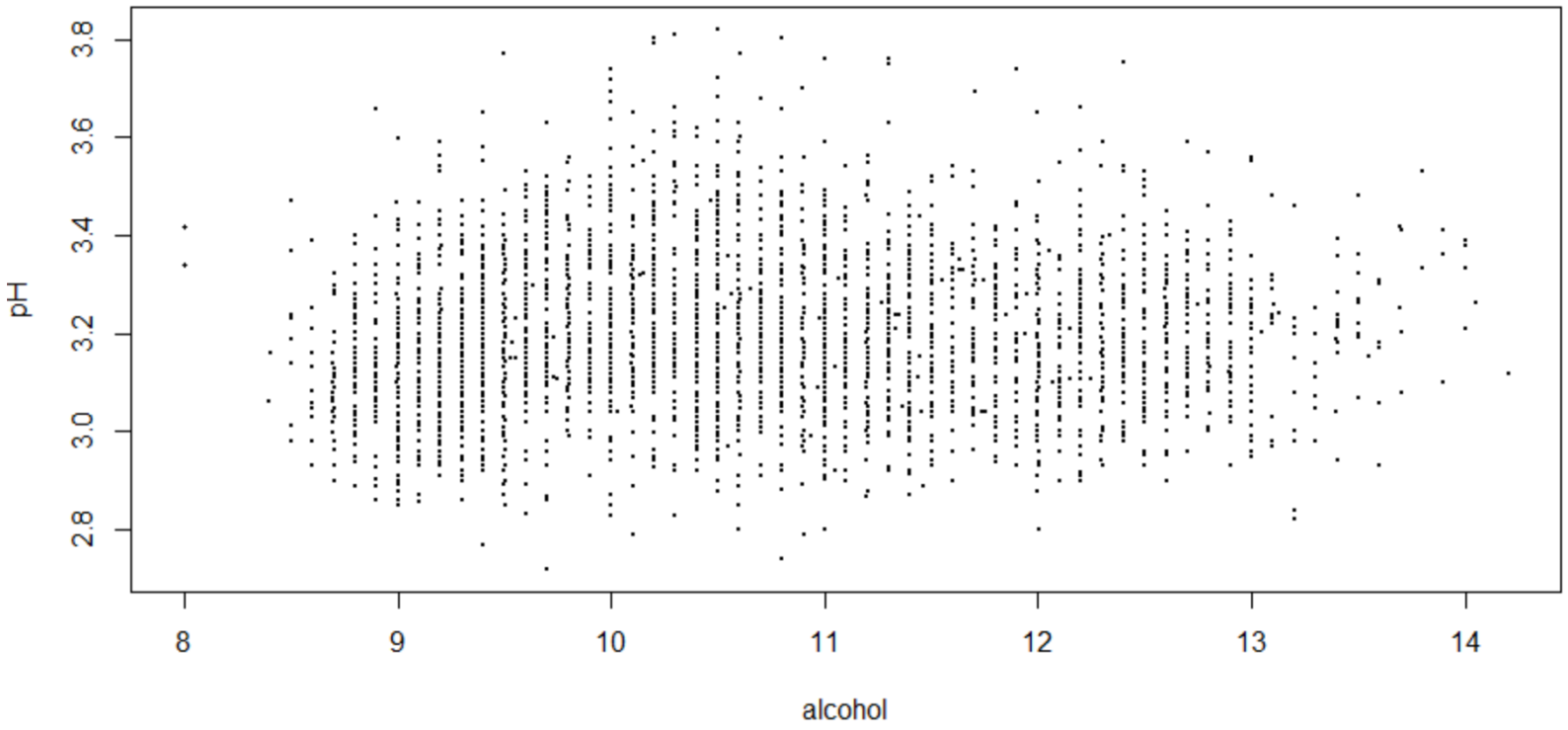



```
# Calculate the 5 nearest neighbors distance
> wine_nn <- get.knn(wine, k = 5)
>
> # View the distance matrix
> head(wine_nn$nn.dist)
  [,1] [,2] [,3] [,4] [,5] [1,]
  0.010067709 0.012404661 0.02233020 0.02743504 0.03287523
  [2,] 0.008050633 0.020066736 0.02058273 0.02776547 0.03029859
  [3,] 0.008990051 0.009987045 0.01736302 0.02212557 0.02986789
  [4,] 0.007606011 0.009860886 0.02110356 0.02199006 0.02822645
  [5,] 0.007579141 0.008917606 0.02081595 0.02083676 0.02927575
  [6,] 0.009465852 0.013015261 0.02019596 0.02098357 0.02471888
> # Create score by averaging distances
> wine_nnd <- rowMeans(wine_nn$nn.dist)
>
> # Print row index of the most anomalous point
> which.max(wine_nnd)
  [1] 1520
```

```
>  # Standardize the wine columns
>  wine_scaled <- scale(wine)
>  # Observe standardized column scales
>  summary(wine_scaled)
       pH                alcohol
 Min.   :-2.74170   Min.   :-2.16657
 1st Qu.:-0.71846   1st Qu.:-0.84910
 Median :-0.05055   Median :-0.07542
 Mean   : 0.00000   Mean   : 0.00000
 3rd Qu.: 0.68136   3rd Qu.: 0.77698
 Max.   : 3.44366   Max.   : 2.63531
>

> # Print the 5-nearest neighbor distance score
> wine_nnd[1:5]
[1] 0.02102257 0.02135283 0.01766671 0.01775739 0.01748504
> # Add the score as a new column
> wine$score <- wine_nnd
```

```
> # Scatterplot showing pH, alcohol and kNN score
> plot(pH ~ alcohol, data = wine, cex = sqrt(score), pch = 20)
```



Proximity-based outlier detection

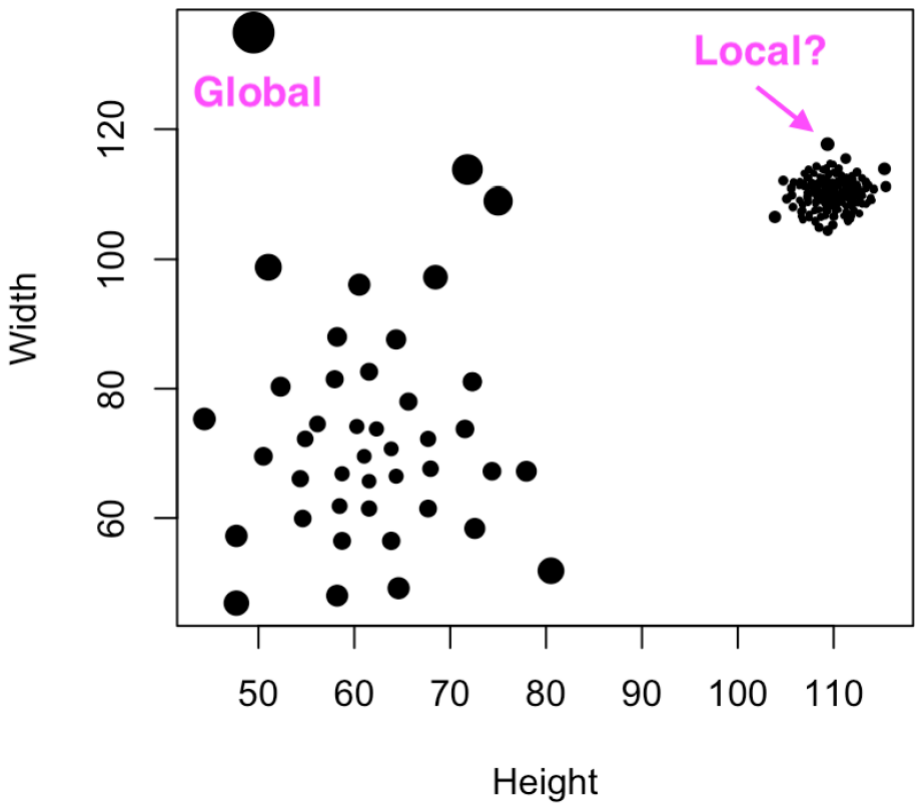
- Pros
  - Easier to define a proximity measure for a dataset than determine its statistical distribution.
  - Quantitative measure of degree to which object is an outlier.
  - Deals naturally with multiple modes.

- Cons
  - $O(n^2)$ complexity.
  - Score sensitive to choice of k.
  - Does not work well if data has widely variable density.

Outline

- Introduction
- Techniques for anomaly detection
  - Statistical
  - Proximity-based
  - Density-based
  - Cluster-based
- Summary

Problem



Density-based outlier detection

Outliers are objects in regions of low density.

- Outlier score is inverse of density around object.
- Scores usually based on proximities.
- Example scores:
  - Reciprocal of average distance to *k* nearest neighbors:

$$\text{density}(x,k) = \left(\frac{1}{k}\sum_{y \in N(x,k)} \text{distance}(x,y)\right)^{-1}$$

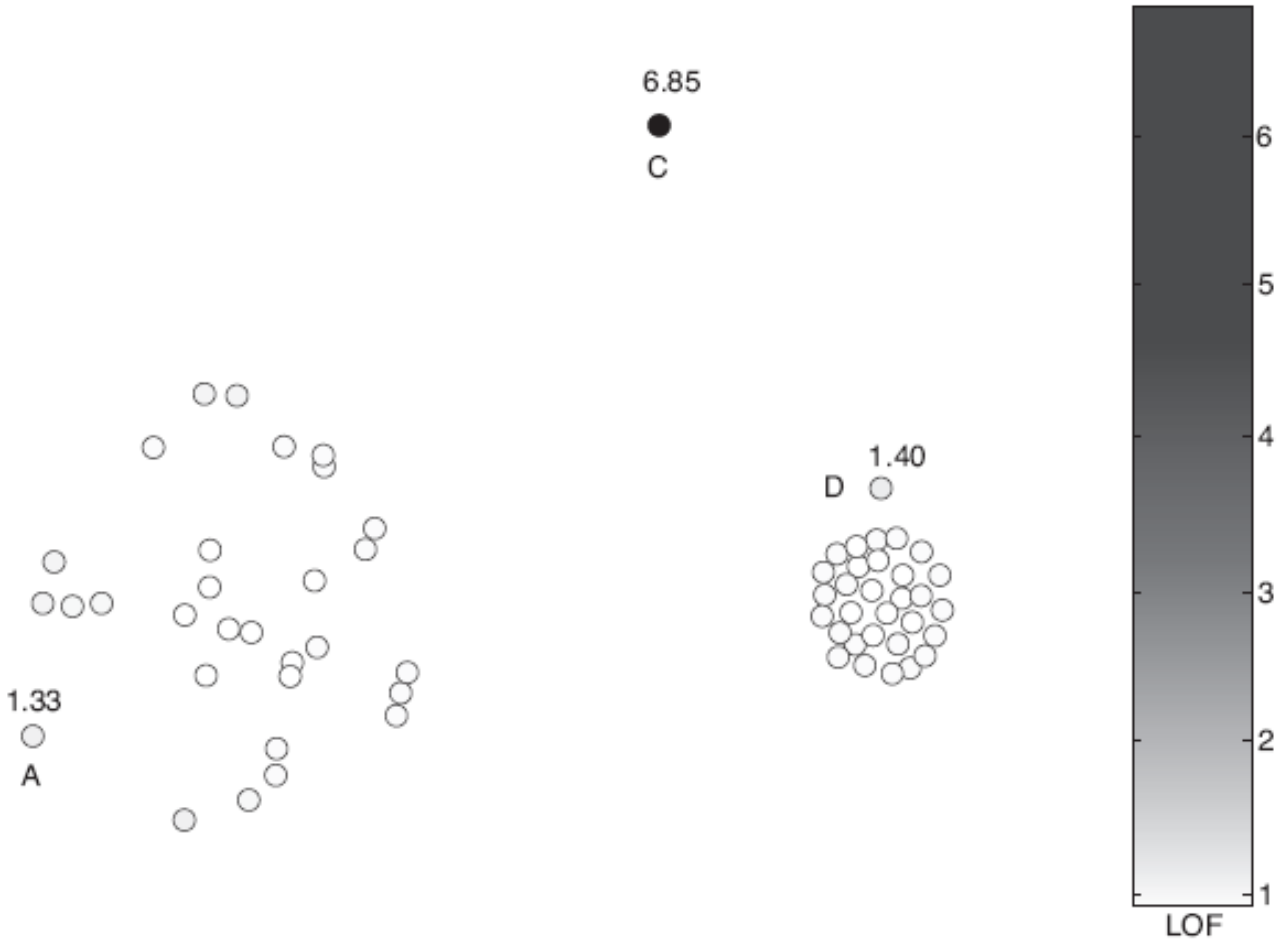
  - Number of objects within fixed radius *d* (DBSCAN).
  - These two example scores work poorly if data has variable density.

Density-based outlier detection

- Relative density outlier score (Local Outlier Factor, LOF)
  - Reciprocal of average distance to *k* nearest neighbors, relative to that of those *k* neighbors.

$$\text{relativedensity}(x,k) = \frac{\text{density}(x,k)}{\frac{1}{k}\sum_{y \in N(x,k)} \text{density}(y,k)\ )}$$

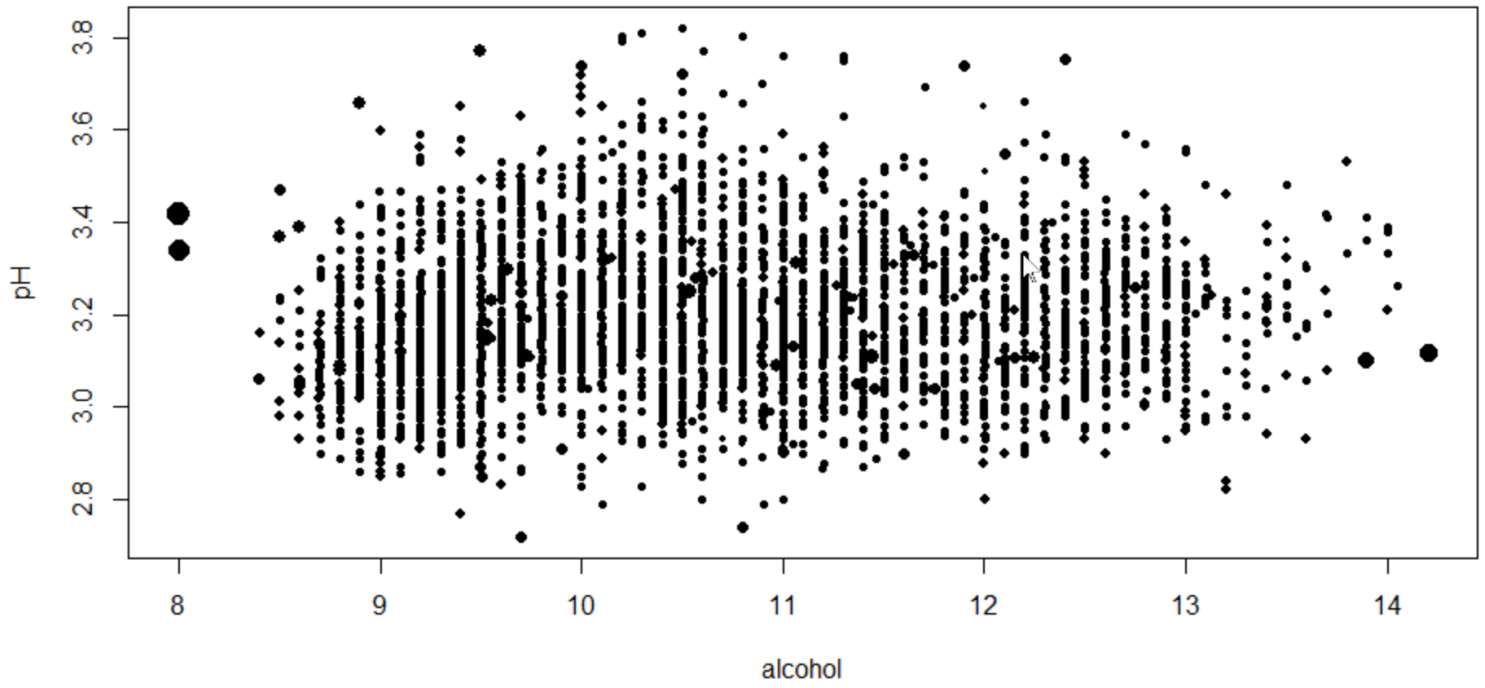
Density-based outlier detection



relative density (LOF) outlier scores

Example with package Rlof

```
> # Calculate the LOF for wine data
> wine_lof <- lof(scale(wine), k = 5)
> # Append the LOF score as a new column
> wine$score <- wine_lof
> # Scatterplot showing pH, alcohol and LOF score
> plot(pH ~ alcohol, data = wine, cex = score, pch = 20)
```



Density-based outlier detection

- Pros
  - Quantitative measure of degree to which object is an outlier.
  - Can work well even if data has variable density.

- Cons
  - $O(n^2)$ complexity
  - Must choose parameters
    - k for nearest neighbor
    - d for distance threshold

Outline

- Introduction
- Techniques for anomaly detection
  - Statistical
  - Proximity-based
  - Density-based
  - Cluster-based
- Summary

Cluster-based outlier detection

Outliers are objects that do not belong strongly to any cluster.

- Approaches:
  - Assess degree to which object belongs to any cluster.
  - Eliminate object(s) to improve objective function.
  - Discard small clusters far from other clusters.

- Issue:
  - Outliers may affect initial formation of clusters.

Cluster-based outlier detection

Assess degree to which object
belongs to any cluster.

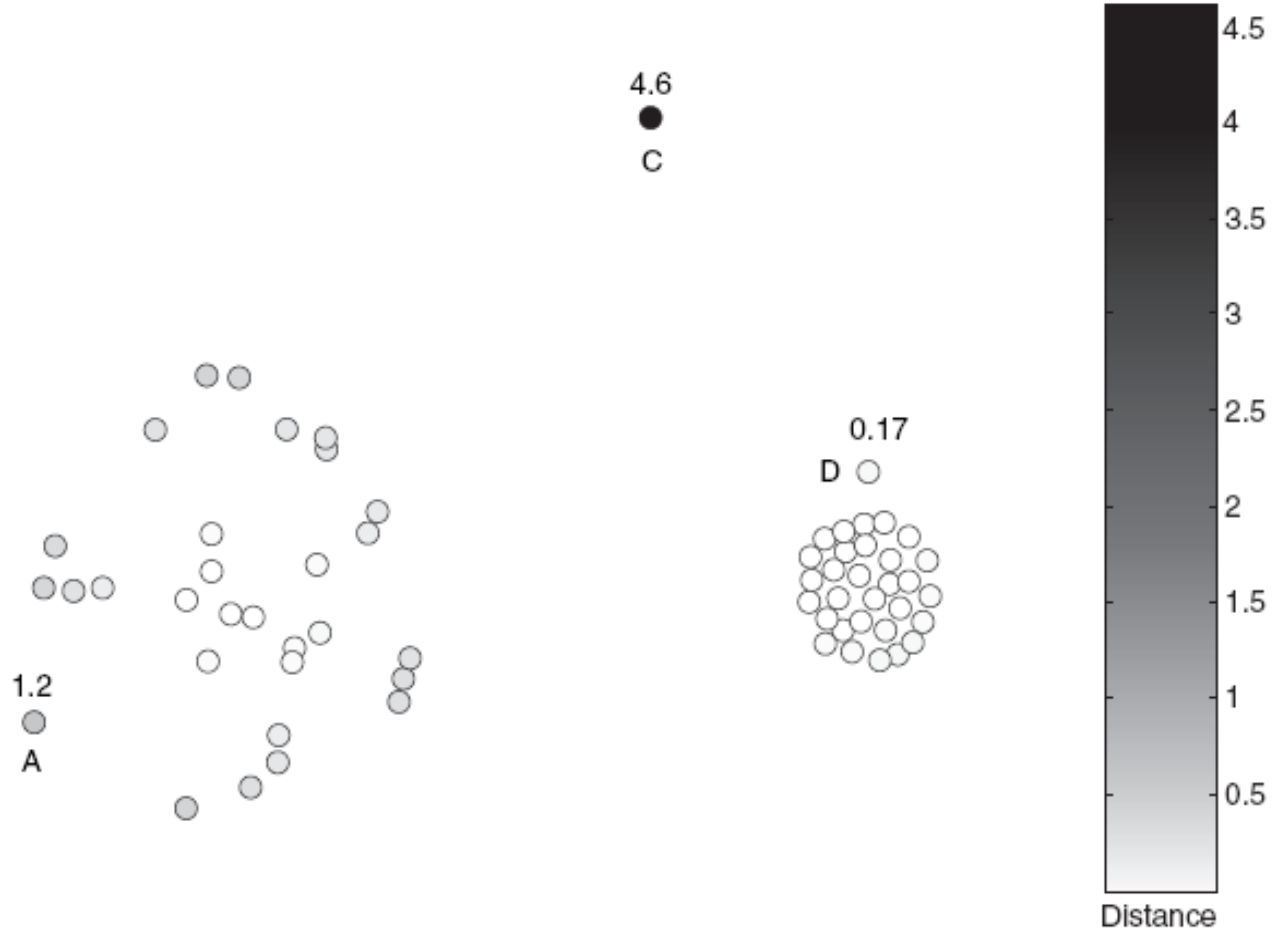
For prototype-based clustering (e.g. k-means), use distance to cluster centers.

- To deal with variable density clusters, use relative distance:

$$\frac{\text{distance}(x, \text{centroid}_c)}{\text{median}(\{\forall_{x' \epsilon C} \text{distance}(x', \text{centroid}_c)\})}$$

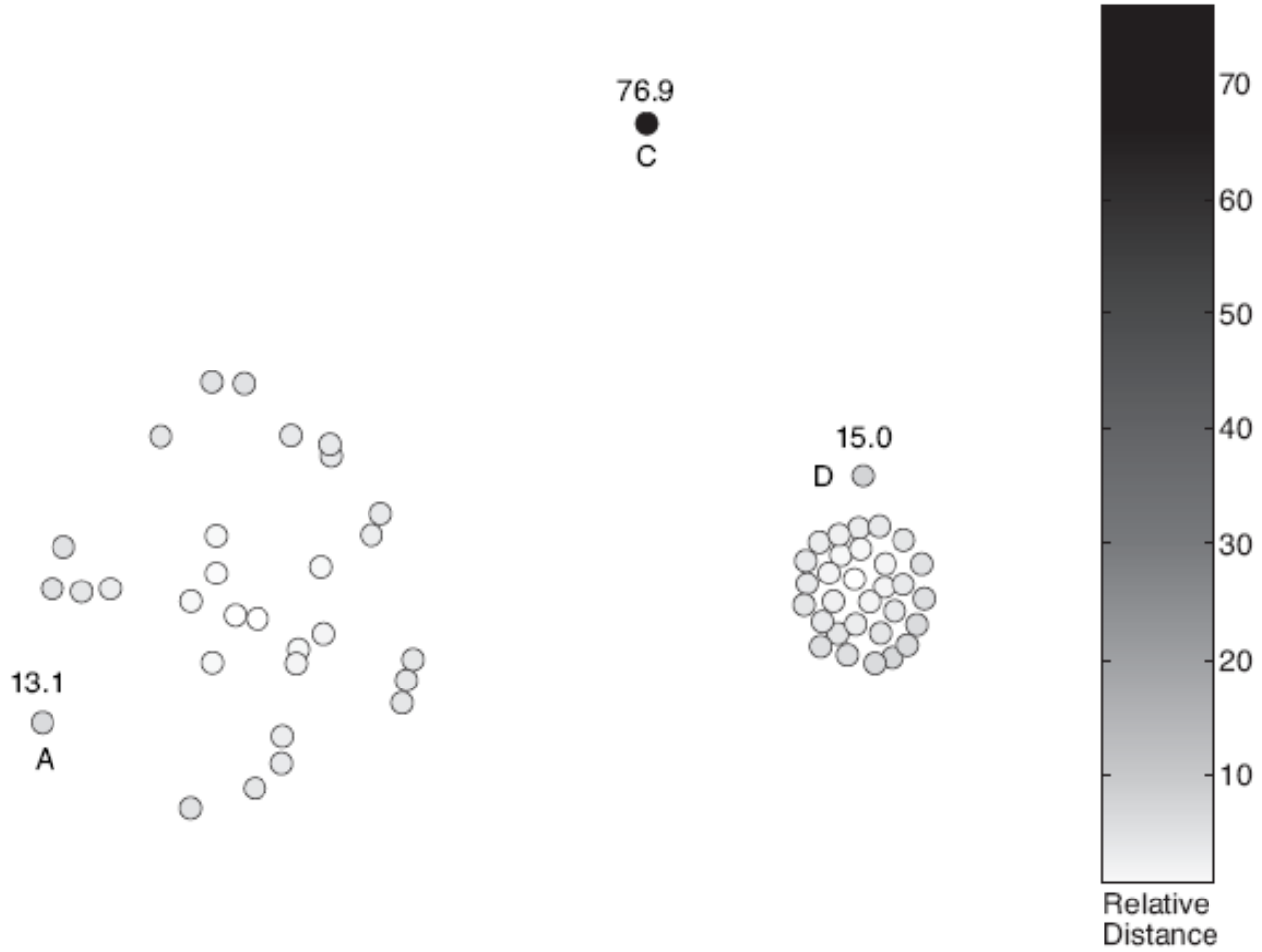
Similar concepts for density-based or connectivity-based clusters.

Cluster-based outlier detection



distance of points from nearest centroid

Cluster-based outlier detection



relative distance of points from nearest centroid

Cluster-based outlier detection

Eliminate object(s) to improve objective function.

1) Form initial set of clusters.
2) Remove the object which most improves objective function.
3) Repeat step 2) until …

Cluster-based outlier detection

Discard small clusters far from other clusters.

- Need to define thresholds for “small” and “far”.

Cluster-based outlier detection

- Pro:
  - Some clustering techniques have $O(n)$ complexity.
  - Extends concept of outlier from single objects to groups of objects.

- Cons:
  - Requires thresholds for minimum size and distance.
  - Sensitive to number of clusters chosen.
  - Hard to associate outlier score with objects.
  - Outliers may affect initial formation of clusters.

Summary

- Real-world issues in anomaly detection
  - Data often streaming, not static
    - Credit card transactions
  - Anomalies can be *bursty*
    - Network intrusions

Literature

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection. ACM Computing Surveys, 41(3), 1–58.

Hawkins, D. 1980. Identification of Outliers. Chapman and Hall.

KDD Topics: Outlier and Anomaly Detection

Curated by: Varun Chandola and Vipin Kumar

http://www.kdd.org/kdd2016/topics/view/outlier-and-anomaly-detection