Knowledge Discovery in Databases

SS 2016

Chapter 5: Outlier Detection

Lecture: Prof. Dr. Thomas Seidl

Tutorials: Julian Busch, Evgeniy Faerman,
Florian Richter, Klaus Schmid

Überblick

- Introduction
- Clustering based approach
- Statistical approaches
- Distance-based Outliers
- Density-based Outliers und Local Outliers
- Angle-based Outliers
- Summary

What is an outlier?

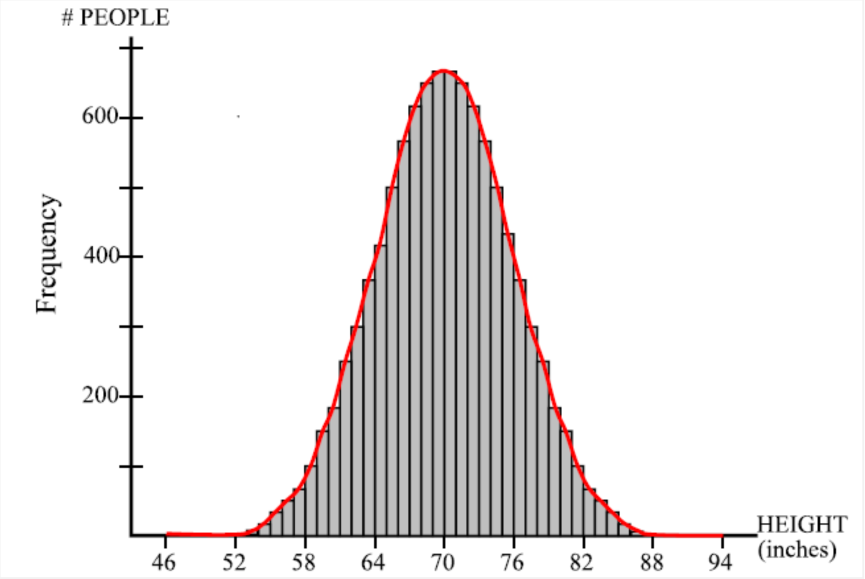
Definition nach Hawkins [Hawkins 1980]:

"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:

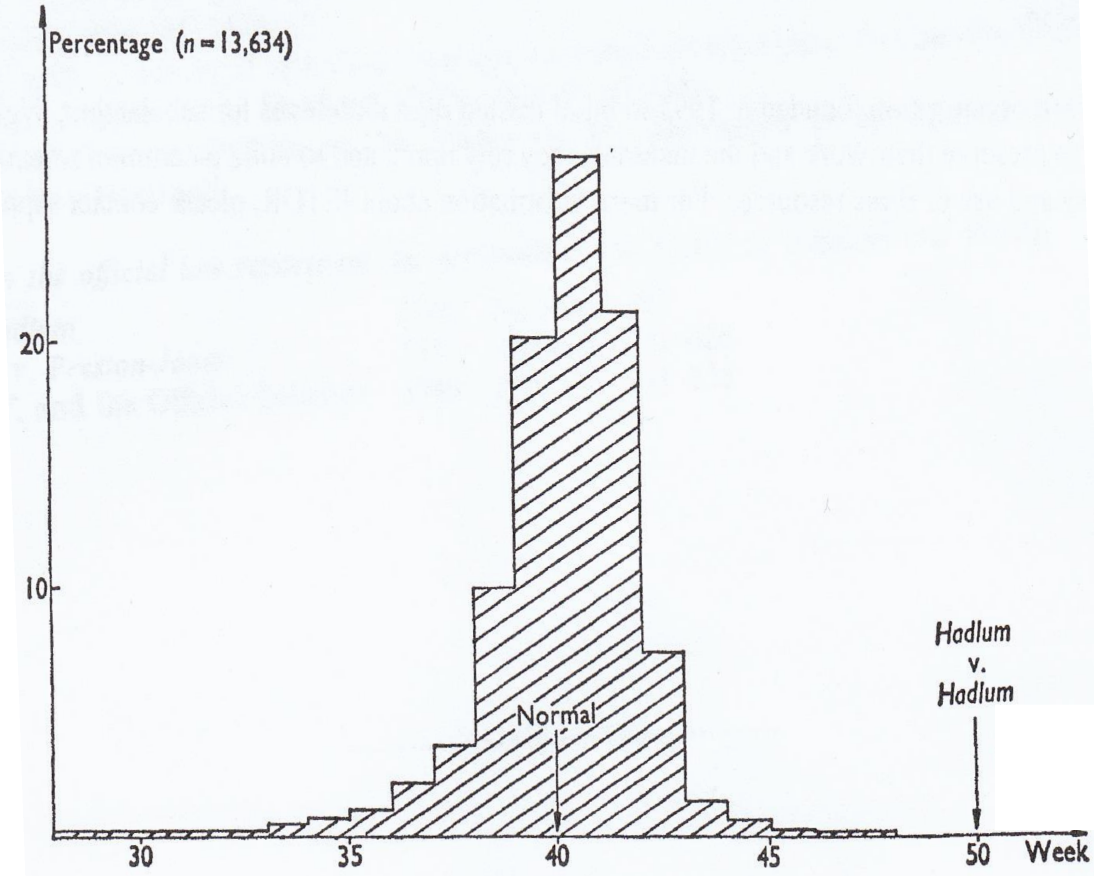
Normal data objects follow a "generating mechanism", e.g. some given statistical process

Abnormal objects deviate from this generating mechanism



Example: Hadlum vs. Hadlum (1949) [Barnett 1978]

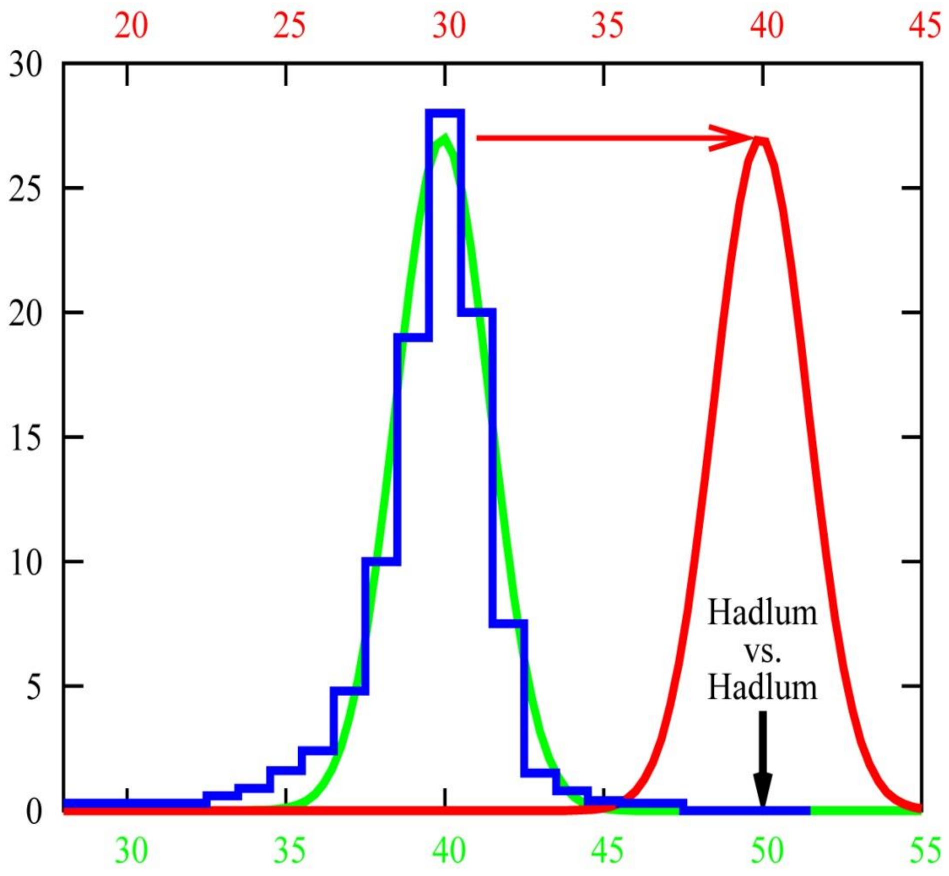
- The birth of a child to Mrs. Hadlum happened 349 days after Mr. Hadlum left for military service.
- Average human gestation period is 280 days (40 weeks).
- Statistically, 349 days is an outlier.



Distribution of human gestation periods.

Beispiel: Hadlum vs. Hadlum (1949) [Barnett 1978]

- blue: statistical basis (13634 observations of gestation periods)
- green: assumed underlying Gaussian process
  - Very low probability for the birth of Mrs. Hadlums child being generated by this process
- red: assumption of Mr. Hadlum (another Gaussian process responsible for the observed birth, where the gestation period starts later)



Applications:

- Fraud detection
  - Purchasing behavior of a credit card owner usually changes when the card is stolen
  - Abnormal buying patterns can characterize credit card abuse
- Medicine
  - Unusual symptoms or test results may indicate potential health problems of a patient
  - Whether a particular test result is abnormal may depend on other characteristics of the patients (e.g. gender, age, …)
- Public health
  - The occurrence of a particular disease, e.g. tetanus, scattered across various hospitals of a city indicate problems with the corresponding vaccination program in that city
  - Whether an occurrence is abnormal depends on different aspects like frequency, spatial correlation, etc.

Applications:

- Sports statistics
  - In many sports, various parameters are recorded for players in order to evaluate the players' performances
  - Outstanding (in a positive as well as a negative sense) players may be identified as having abnormal parameter values
  - Sometimes, players show abnormal values only on a subset or a special combination of the recorded parameters
- Detecting measurement errors
  - Data derived from sensors (e.g. in a given scientific experiment) may contain measurement errors
  - Abnormal values could provide an indication of a measurement error
  - Removing such errors can be important in other data mining and data analysis tasks
  - "One person's noise could be another person's signal."

Important properties of Outlier Models:

- Global vs. local approach:
  "Outlierness" regarding whole dataset (global) or regarding a subset of data (local)?

- Labeling vs. Scoring
  Binary decision or outlier degree score?

- Assumptions about "Outlierness":
  What are the characteristics of an outlier object?

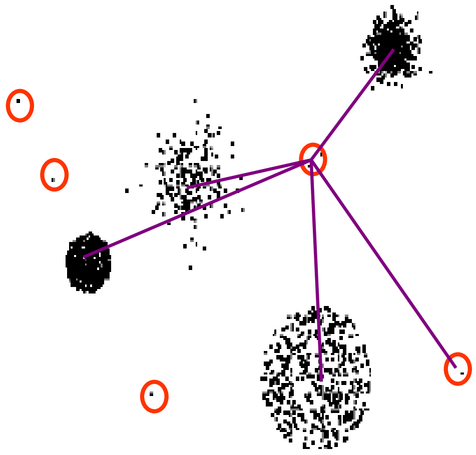
Überblick

- Introduction
- Clustering based approach
- Statistical approaches
- Distance-based Outliers
- Density-based Outliers und Local Outliers
- Angle-based Outliers
- Summary

An object is a cluster-based outlier if it does not strongly belong to any cluster:

Basic idea:

- Cluster the data into groups
- Choose points in small clusters as candidate outliers. Compute the distance between candidate points and non-candidate clusters.
  - If candidate points are far from all other non-candidate points and clusters, they are outliers



A more systematic approach

- Find clusters and then assess the degree to which a point belongs to any cluster
- e.g. for k-Means distance to the centroid
- In case of k-Means (or in general, clustering algorithms with some objective function), if the elimination of a point results in substantial improvement of the objective function, we could classify it as an outlier
  - i.e., clustering creates a model of the data and the outliers distort that model.

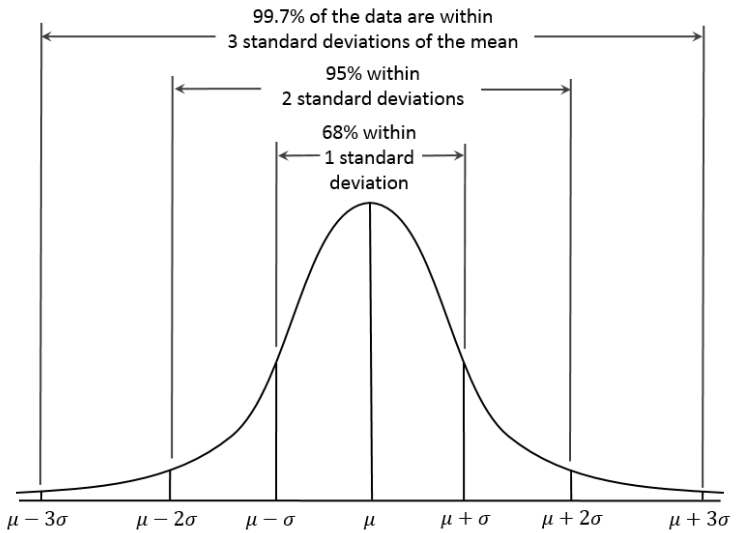
Überblick

- Introduction
- Clustering based approach
- Statistical approaches
- Distance-based Outliers
- Density-based Outliers und Local Outliers
- Angle-based Outliers
- Summary

Statistical Tests

General idea

- Given a certain kind of statistical distribution (e.g., Gaussian)
- Compute the parameters assuming all data points have been generated by such a statistical distribution (e.g., mean and standard deviation)
- Outliers are points that have a low probability to be generated by the overall distribution (e.g., deviate more than 3 times the standard deviation from the mean)



Basic assumption

- Normal data objects follow a (known) distribution and occur in a high probability region of this model
- Outliers deviate strongly from this distribution

Statistical Tests

A huge number of different tests are available differing in

- Type of data distribution (e.g. Gaussian)
- Number of variables, i.e., dimensions of the data objects (univariate/multivariate)
- Number of distributions (mixture models)
- Parametric versus non-parametric (e.g. histogram-based)

Example on the following slides

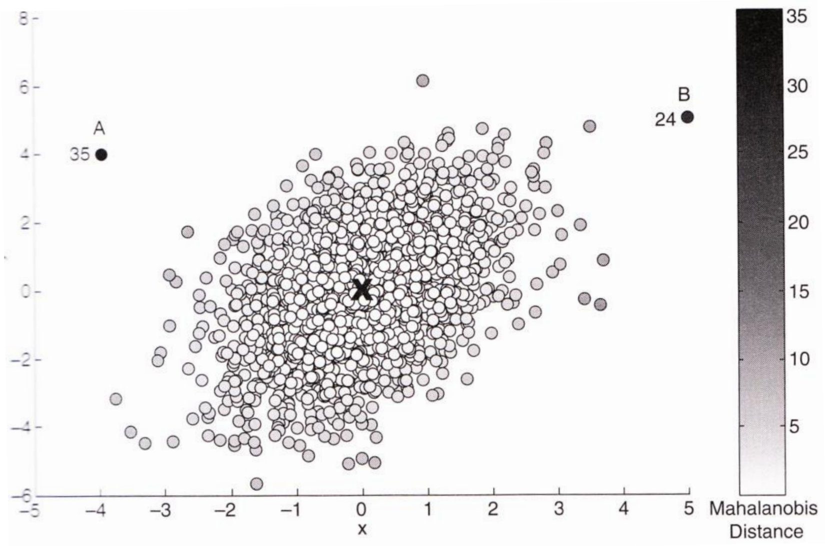
- Gaussian distribution
- Multivariate
- 1 model
- Parametric

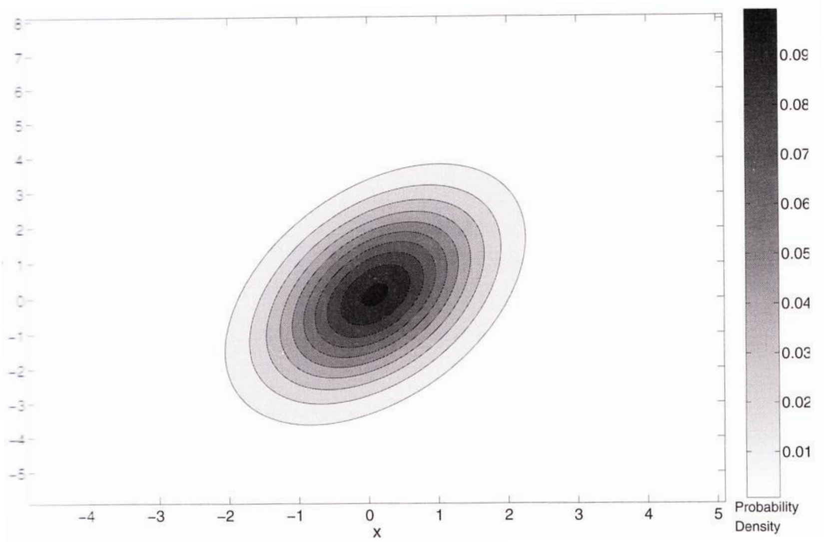
Probability density function of a multivariate normal distribution

$$N(x) = \frac{1}{\sqrt{(2\pi)^d \,|\Sigma|}} e^{-\frac{(x-\mu)^{\mathbf{T}} \mathbf{\Sigma}^{-1} (x-\mu)}{2}}$$

- μ is the mean value of all points (usually data are normalized such that $\mu=0$)
- Σ is the covariance matrix from the mean
- $MDist(x,\mu) = \sqrt{(x-\mu)^{\mathbf{T}} \mathbf{\Sigma}^{-1} (x-\mu)}$ is the Mahalanobis distance of point *x* to μ
- MDist follows a χ^2-distribution with *d* degrees of freedom (*d* = data dimensionality)
- All points *x*, with $MDist(x,\mu) > \chi^2(0{,}975)$ [$\approx 3\cdot\sigma$]

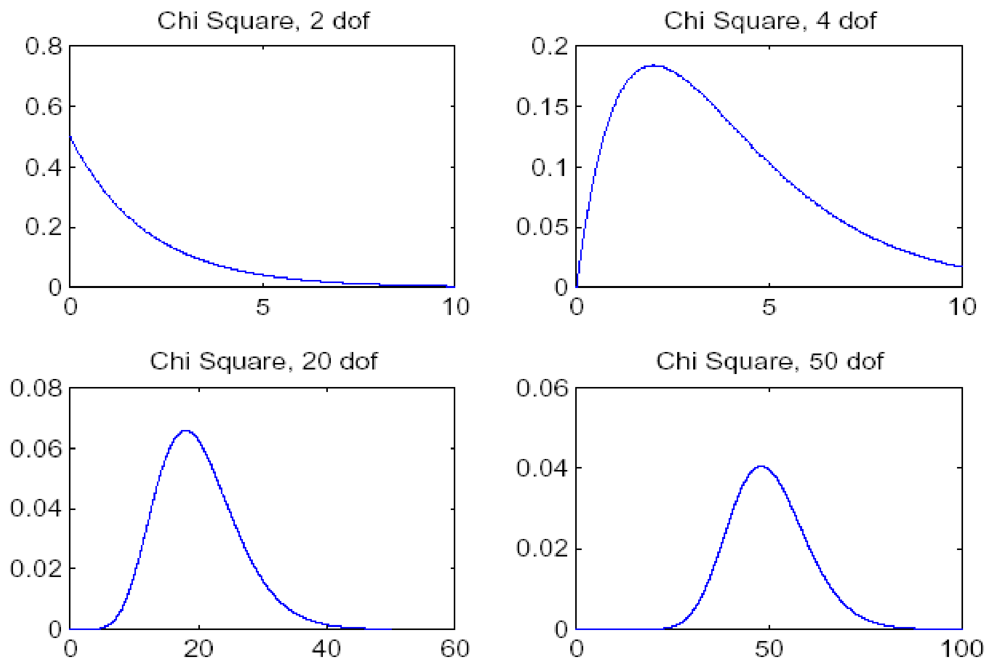
Visualization (2D) [Tan et al. 2006]





Problems

- Curse of dimensionality
  - The larger the degree of freedom, the more similar the MDist values for all points



x-axis: observed *MDist* values

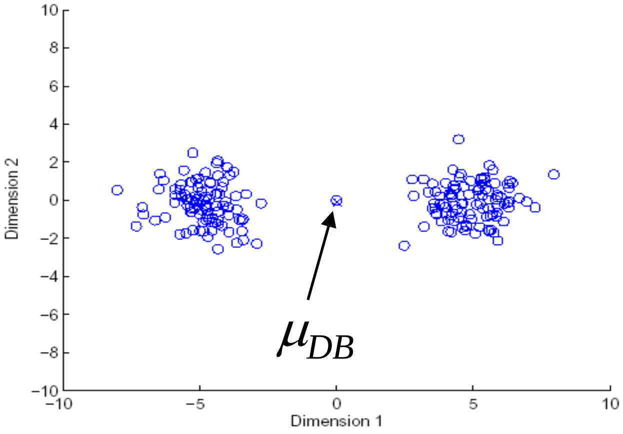
y-axis: frequency of observation

Problems (cont.)

- Robustness
  - Mean and standard deviation are very sensitive to outliers
  - These values are computed for the complete data set (including potential outliers)
  - The *MDist* is used to determine outliers although the *MDist* values are influenced by these outliers

Discussion

- Data distribution is fixed
- Low flexibility (if no mixture models)
- Global method



Überblick

- Introduction
- Clustering based approach
- Statistical approaches
- Distance-based Outliers
- Density-based Outliers und Local Outliers
- Angle-based Outliers
- Summary

Distance-based Approaches

General Idea

- Judge a point based on the distance(s) to its neighbors
- Several variants proposed

Basic Assumption

- Normal data objects have a dense neighborhood
- Outliers are far apart from their neighbors, i.e., have a less dense neighborhood

Distance-based Approaches

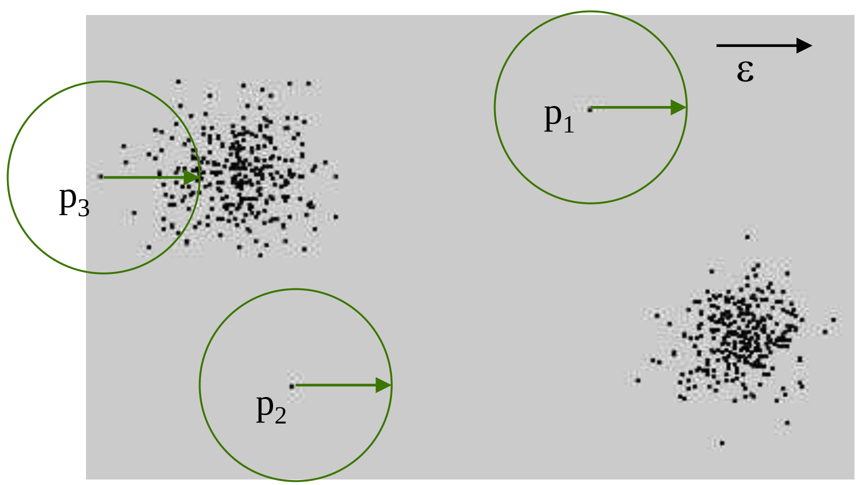
DB(ε,π)-Outliers

– Basic model [Knorr and Ng 1997]

- Given a radius ε and a percentage π
- A point *p* is considered an outlier if at most π percent of all other points have a distance to *p* less than ε

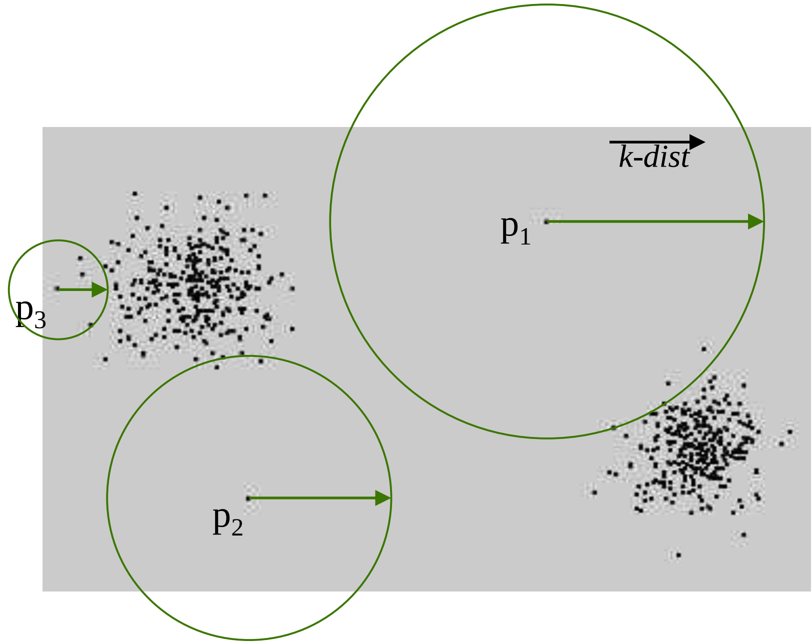
$$OutlierSet(\varepsilon,\pi) = \{ p \mid \frac{Card(\boxed{\{q \in DB \mid dist(p,q) < \varepsilon\}})}{Card(DB)} \leq \pi \}$$

range-query with radius ε



Outlier scoring based on *k*NN distances

- General models
  - Take the *k*NN distance of a point as its outlier score



k^{th} nearest neighbor based

The outlier score of an object is given by the distance to its *k*-nearest neighbor.

- theoretically lowest outlier score: 0

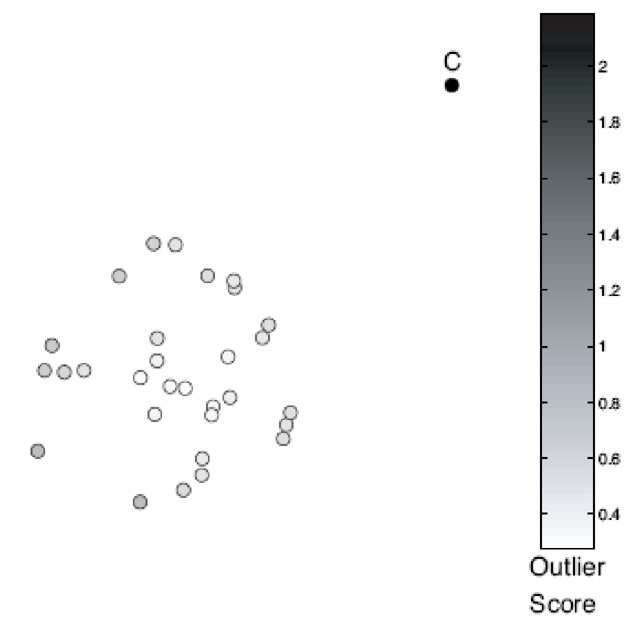


Figure 10.4. Outlier score based on the distance to fifth nearest neighbor.

[Tan, Steinbach, Kumar 2006]

k=5

k^{th} nearest neighbor based

- The outlier score is highly sensitive to the value of k

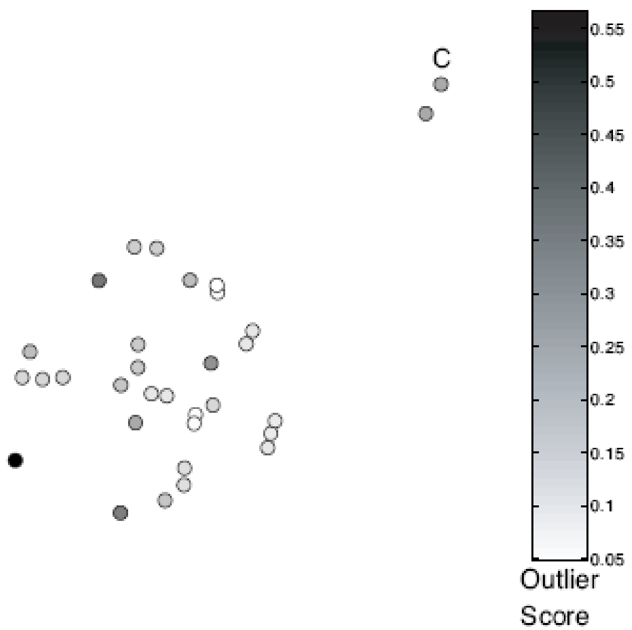


Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

If k is too small, then a small number of close neighbors can cause low outlier scores.

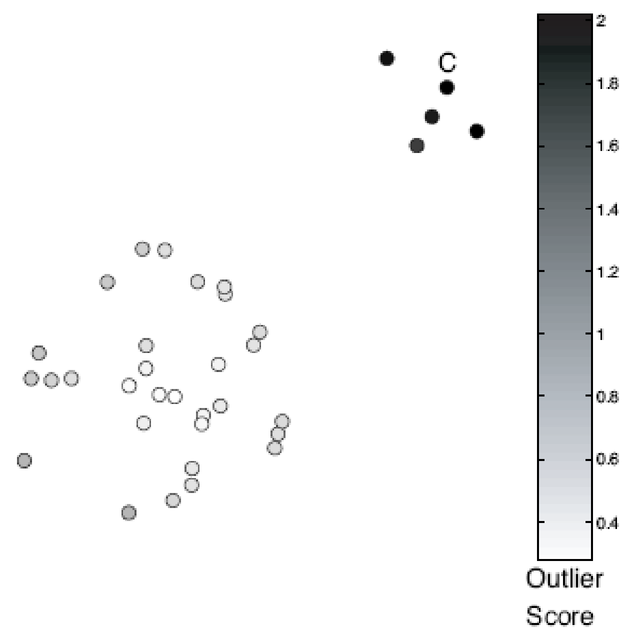


Figure 10.6. Outlier score based on distance to the fifth nearest neighbor. A small cluster becomes an outlier.

If k is too large, then all objects in a cluster with less than k objects might become outliers.

[Tan, Steinbach, Kumar 2006]

k^{th} nearest neighbor based

- cannot handle datasets with regions of widely different densities due to the global threshold

sparse cluster

dense cluster

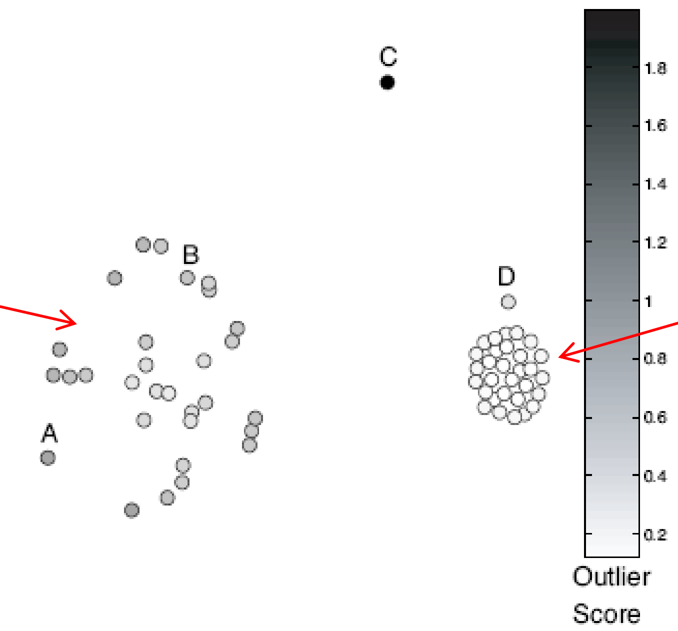


Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density. [Tan, Steinbach, Kumar 2006]

Überblick

Density-based Approaches

General idea

- Compare the density around a point with the density around its local neighbors.
- The relative density of a point compared to its neighbors is computed as an outlier score.
- Approaches also differ in how to estimate density.

Basic assumption

- The density around a normal data object is similar to the density around its neighbors.
- The density around an outlier is considerably different to the density around its neighbors.

Density-based Approaches

- Different definitions of density:
  - e.g., # points within a specified distance d from the given object
- The choice of d is critical
  - If d is to small many normal points might be considered outliers
  - If d is to large, many outlier points will be considered as normal
- A global notion of density is problematic (as it is in clustering)
  - fails when data contain regions of different densities
- Solution: use a notion of density that is relative to the neighborhood of the object

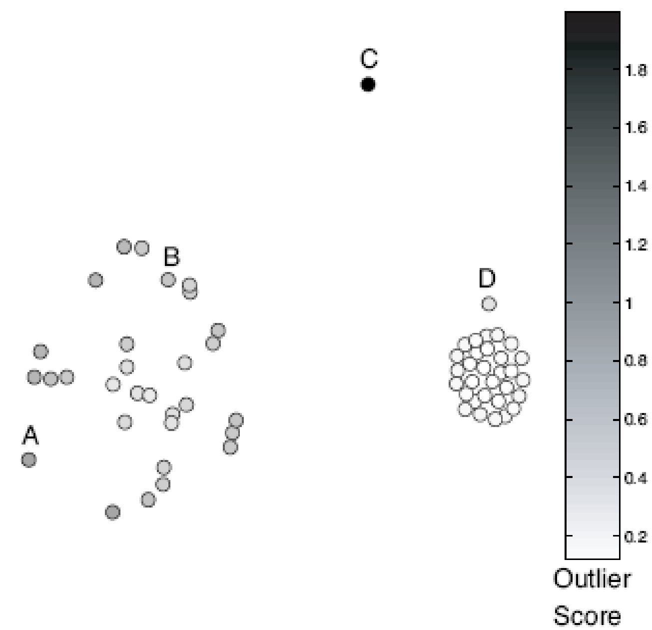


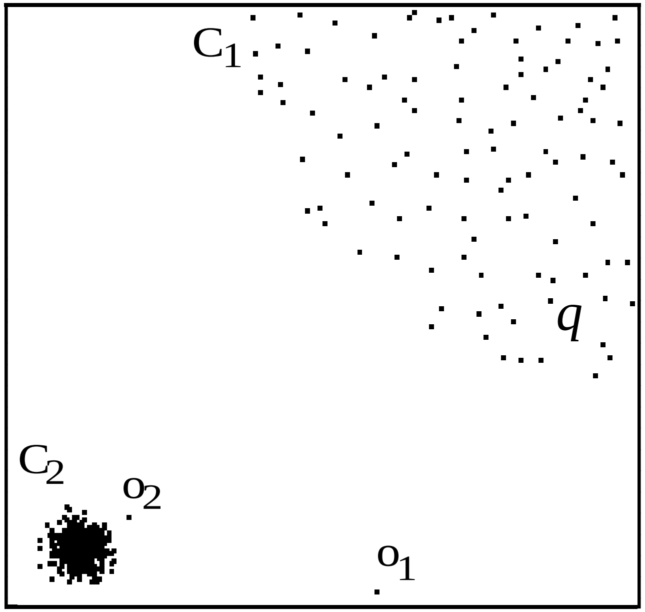
Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

[Tan, Steinbach, Kumar 2006]

D has a higher absolute density than A but compared to its neighborhood, D's density is lower.

Local Outlier Factor (LOF) [Breunig et al. 1999, 2000]

- Motivation:
  - Distance-based outlier detection models have problems with different densities
  - How to compare the neighborhood of points from areas of different densities?
  - Example
    - DB(ε,π)-outlier model
      - » Parameters ε and π cannot be chosen so that o_2 is an outlier but none of the points in cluster C_1 (e.g. *q*) is an outlier
    - Outliers based on kNN-distance
      - » kNN-distances of objects in C_1 (e.g. *q*) are larger than the kNN-distance of o_2

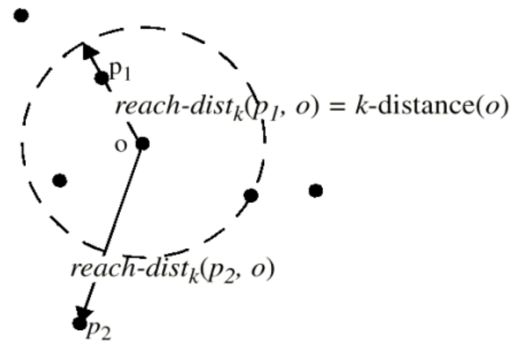


- Solution: consider relative density

Density-based Approaches

– Model
  - Reachability "distance"
    – Introduces a smoothing factor

$$reach\text{-}dist_k(p,o) = \max\{k\text{-distance}(o), dist(p,o)\}$$



  - Local reachability density (*lrd*) of point *p*
    – Inverse of the average reach-dists of the *k*NNs of *p*

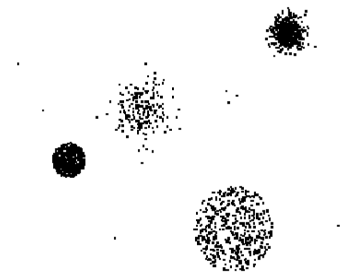
$$lrd_k(p) = \left(\frac{\sum_{o \in kNN(p)} reach\text{-}dist_k(p,o)}{Card(kNN(p))} \right)^{-1}$$

  - Local outlier factor (LOF) of point *p*
    – Average ratio of *lrd*s of neighbors of *p* and *lrd* of *p*

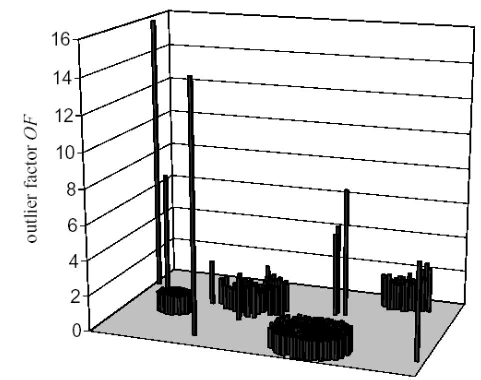
$$LOF_k(p) = \frac{\sum_{o \in kNN(p)} \frac{lrd_k(o)}{lrd_k(p)}}{Card(kNN(p))}$$

Density-based Approaches

– Properties
  - LOF ≈ 1: point is in a cluster (region with homogeneous density around the point and its neighbors)
  - LOF >> 1: point is an outlier



Data set



LOFs (*MinPts* = 40)

– Discussion
  - Choice of *k* (*MinPts* in the original paper) specifies the reference set
  - Originally implements a local approach (resolution depends on the user's choice for *k*)
  - Outputs a scoring (assigns an LOF value to each point)

Density-based Approaches

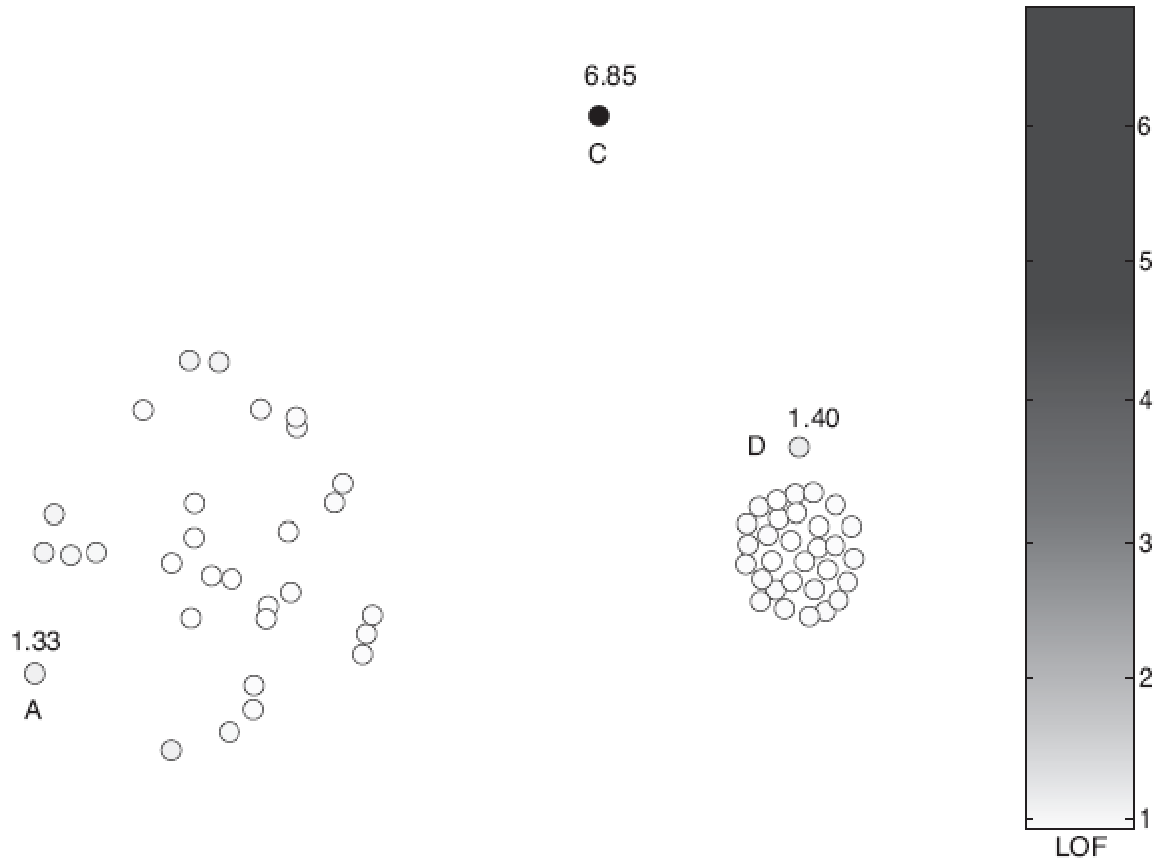


Figure 10.8. Relative density (LOF) outlier scores for two-dimensional points of Figure 10.7.

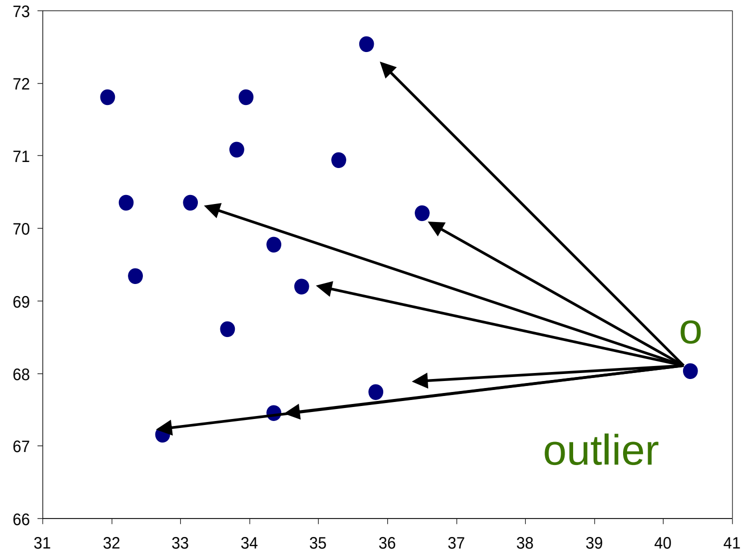
[Tan, Steinbach, Kumar 2006]

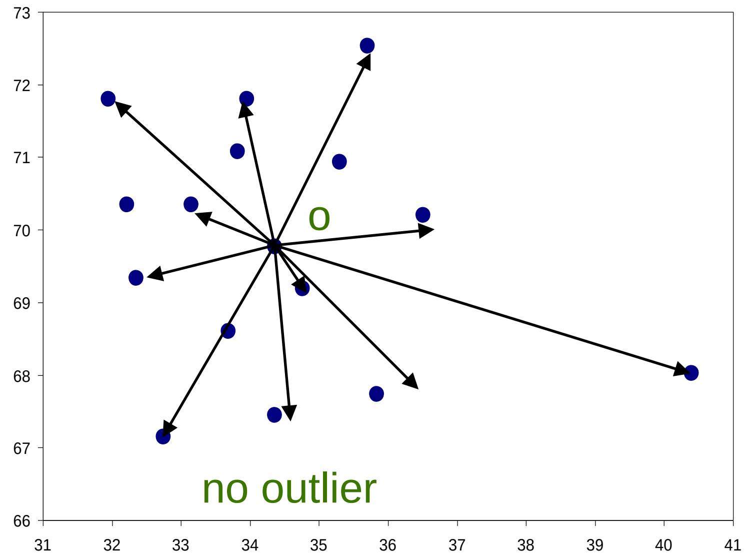
Überblick

- Introduction
- Clustering based approach
- Statistical approaches
- Distance-based Outliers
- Density-based Outliers und Local Outliers
- Angle-based Outliers
- Summary

ABOD – angle-based outlier degree [Kriegel et al. 2008]

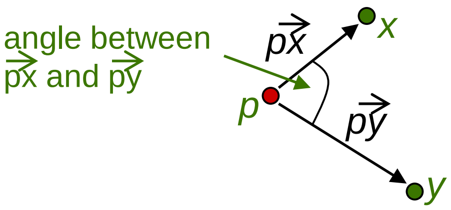
– Rational
  - Angles are more stable than distances in high dimensional spaces (cf. e.g. the popularity of cosine-based similarity measures for text data)
  - Object o is an outlier if most other objects are located in similar directions
  - Object o is no outlier if many other objects are located in varying directions

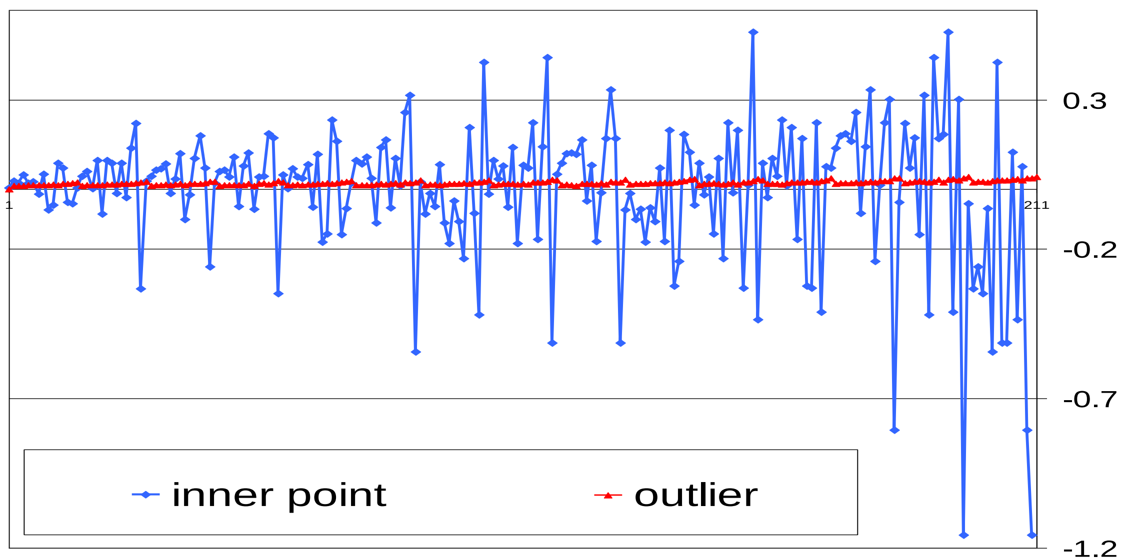




Angle-based Approach

– Basic assumption
  - Outliers are at the border of the data distribution
  - Normal points are in the center of the data distribution
– Model
  - Consider for a given point p the angle between $\overrightarrow{px}$ and $\overrightarrow{py}$ for any two x,y from the database
  - Consider the spectrum of all these angles
  - The broadness of this spectrum is a score for the outlierness of a point





Angle-based Approach

- Model (cont.)
  - Measure the variance of the angle spectrum
  - Weighted by the corresponding distances (for lower dimensional data sets where angles are less reliable)

$$ABOD(p) = \mathrm{VAR}_{\mathrm{x,y} \in DB}\left(\frac{\langle \overrightarrow{xp}, \overrightarrow{yp} \rangle}{\|\overrightarrow{xp}\|^2 \cdot \|\overrightarrow{yp}\|^2}\right)$$

  - Properties
    - Small ABOD => outlier
    - High ABOD => no outlier

Angle-based Approach

- Algorithms
  - Naïve algorithm is in $O(n^3)$
  - Approximate algorithm based on random sampling for mining top-*n* outliers
    - Do not consider all pairs of other points x,y in the database to compute the angles
    - Compute ABOD based on samples => lower bound of the real ABOD
    - Filter out points that have a high lower bound
    - Refine (compute the exact ABOD value) only for a small number of points
- Discussion
  - Global approach to outlier detection
  - Outputs an outlier score

    (inversely scaled:

      high ABOD score => inlier,

      low ABOD score => outlier)

Überblick

- Introduction
- Clustering-based approach
- Statistical approaches
- Distance-based Outliers
- Density-based Outliers und Local Outliers
- Angle-based Outliers
- Summary

Summary

- Algorithm properties:
  - global / local
  - labeling / scoring
  - model assumptions

- Clustering-based outliers:
  - Identification of not cluster members

- Statistical outliers:
  - Assumed probability distribution
  - The probability for the objects to be generated by this distribution is small

Summary

- Distance-based outliers:
  - Distance to the neighbors as outlier metric

- Density-based outliers:
  - Density around the point as outlier metric

- Angle-based outliers:
  - Angles between outliers and random point pairs vary slightly