DM583

Data Mining

Unsupervised Outlier Detection

Authorized Adaption & Extension of the Lecture:

Outlier Detection

Dr. Arthur Zimek
University of Southern Denmark

Introduction

What is an Outlier?

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"

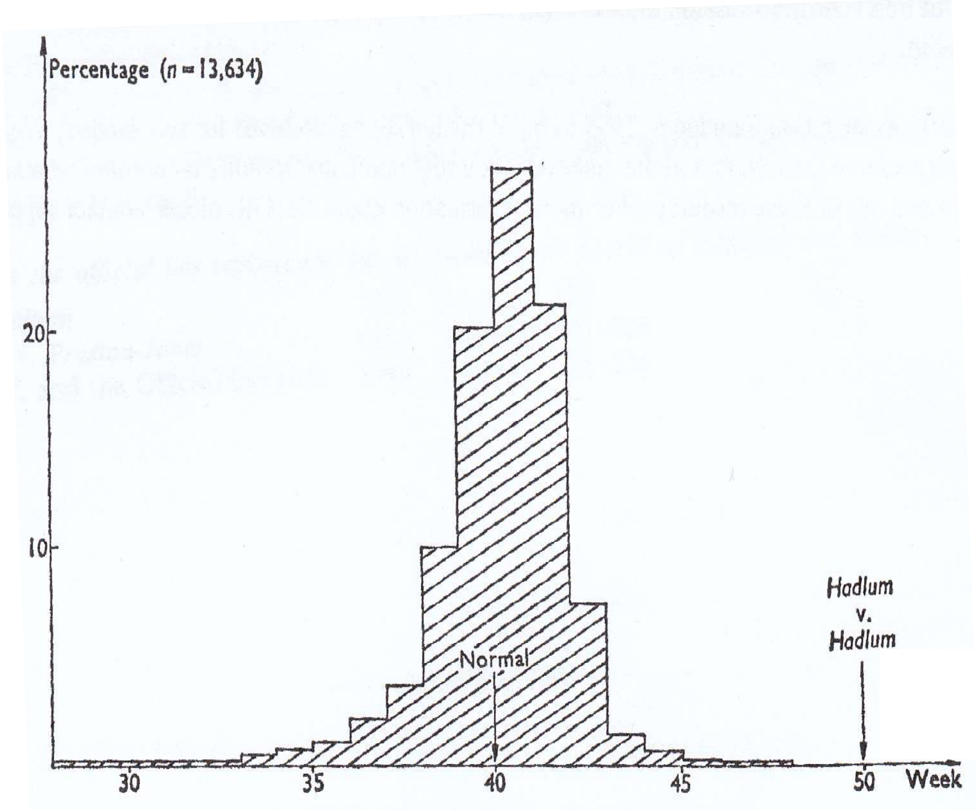
Statistics-Based Intuition

- Normal observations follow a "generating mechanism", e.g. some given statistical process
- Abnormal observations deviate from this generating mechanism

Introduction

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

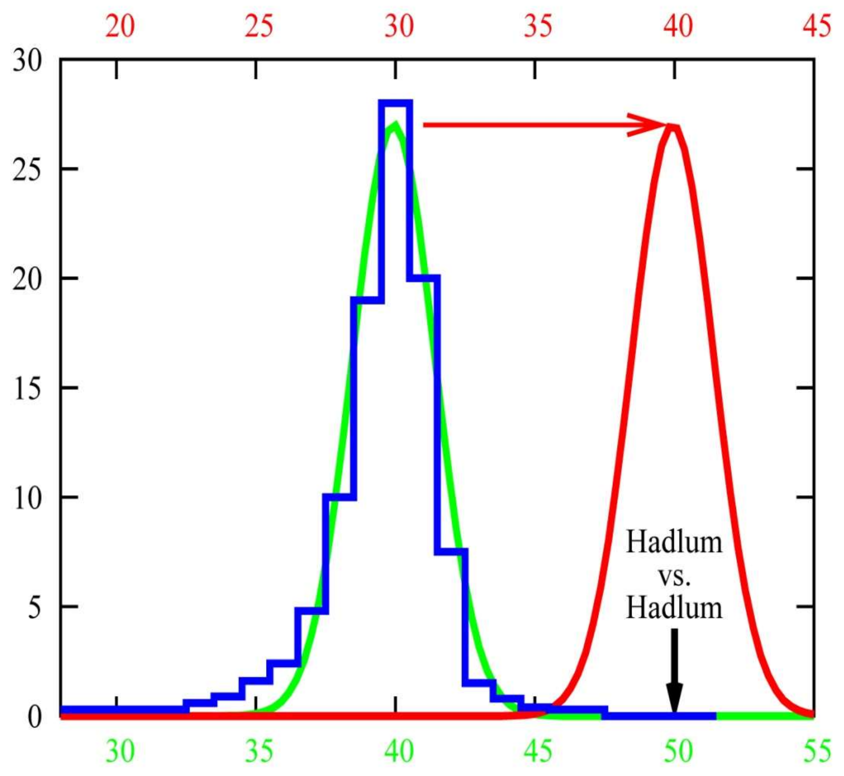
- The birth of a child to Mrs. Hadlum happened 349 days after Mr. Hadlum left for military service.
- Average human gestation period is 280 days (40 weeks).
- Statistically, 349 days is an outlier.



Distribution of human gestation periods.

Introduction

- Example: 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. Hadlum’s child for being generated by this process
  - red: Mr. Hadlum’s assumption (another Gaussian process is responsible for the observed birth; gestation period starts later)
    - Under this assumption, the gestation period has an average duration, and the specific birthday has highest-possible probability



Introduction

- Sample applications of outlier detection
  - 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, …)

Introduction

- Sample applications of outlier detection (cont.)
  - Sports statistics
    - 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 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 analysis tasks
  - …

Introduction

- General application scenarios
  - **Supervised Scenario**
    - In some applications, training data with normal and abnormal observations 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 one of the classes (usually the “normal” class) are available
      - one-class classification problem
  - **Unsupervised Scenario**
    - In many applications there are no training data available

Introduction

- General application scenarios
  - **Supervised Scenario**
    - In some applications, training data with normal and abnormal observations 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 one of the classes (usually the "normal" class) are available
      - one-class classification problem
  - **Unsupervised Scenario**
    - In many applications there are no training data available
- here, we focus on the *unsupervised scenario*

Introduction

- Main unsupervised outlier detection approaches:
  - **Statistical Approach**
    - Outliers are unusual observations according to an estimated (usually parametric) probability distribution
  - **Clustering-Based Approach**
    - Outliers are observations that do not match well any cluster, or they are those that constitute a relatively very small cluster
  - **Non-Parametric Density-Based Approach**
    - Outliers are observations with unusual (absolute or relative) density

Introduction

- Main unsupervised outlier detection approaches:
  - **Statistical Approach**
    - Outliers are unusual observations according to an estimated (usually parametric) probability distribution
  - **Clustering-Based Approach**
    - Outliers are observations that do not match well any cluster, or they are those that constitute a relatively very small cluster
  - **Non-Parametric Density-Based Approach**
    - Outliers are observations with unusual (absolute or relative) density

Statistical Tests

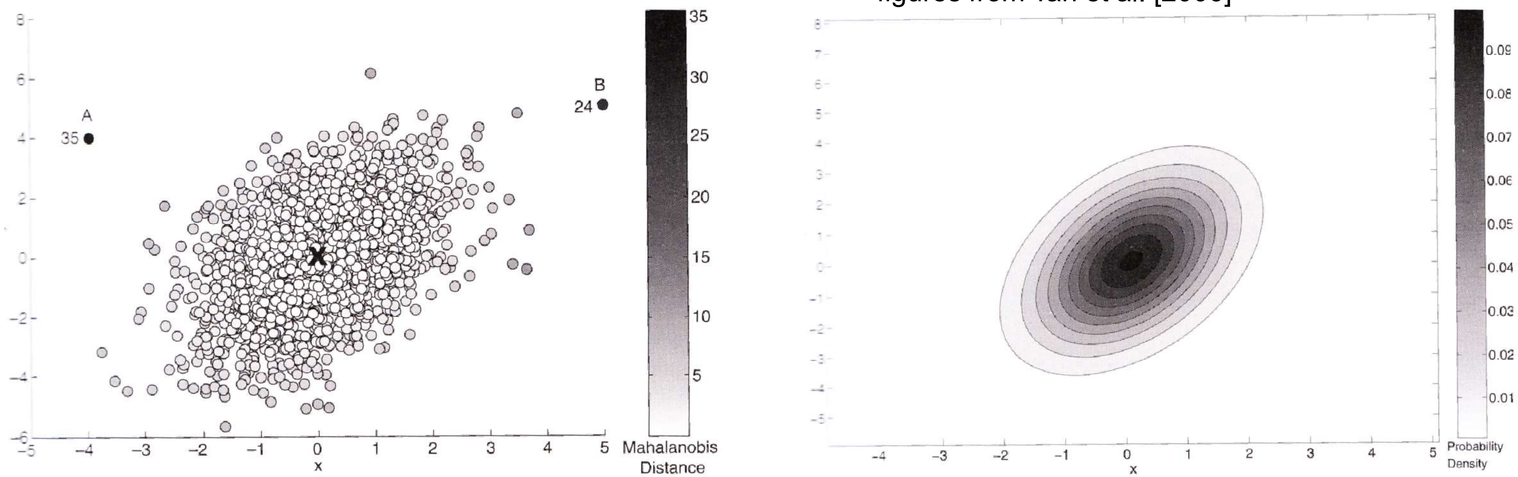
- General idea
  - Given a certain type of statistical distribution (e.g., Gaussian)
  - Compute the parameters assuming all data points have been generated by such a statistical distribution (e.g., mean and standard deviation)
  - Outliers are points that have a low probability to be generated by the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

- 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

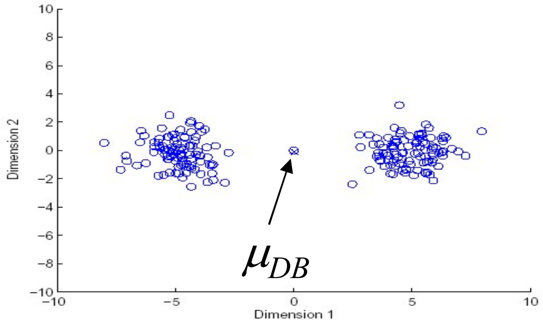
Statistical Tests

- Many 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)
  - …
- Example (2D):

figures from Tan et al. [2006]



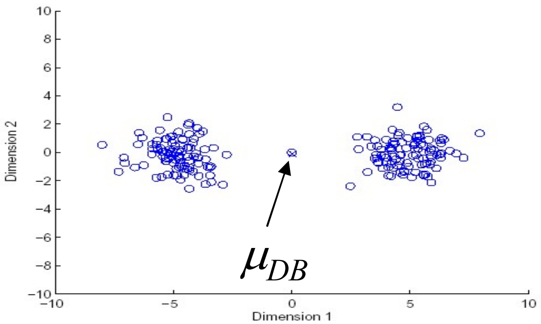
Statistical Tests

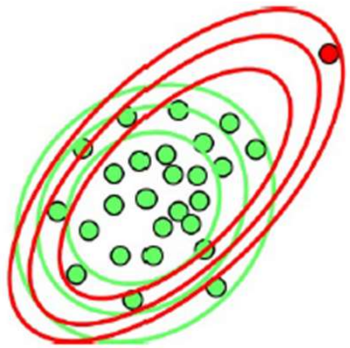


- Main Problems:
  - Parametric Assumption
    - A specific distribution must be assumed, but the assumption may not hold true for the data at hand. **True distribution is rarely known in practice**
    - For example, simple unimodal distributions that are well-studied and easy to apply / manipulate do not represent many real-world scenarios
    - On the other hand, non-parametric methods like histograms are usually not accurate and/or computationally feasible in high-dimensional data

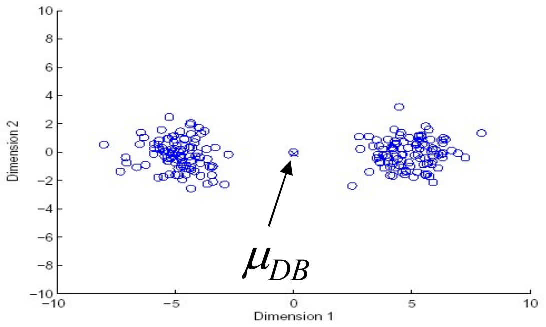
Statistical Tests

- Main Problems:
  - Parametric Assumption
    - A specific distribution must be assumed, but the assumption may not hold true for the data at hand. **True distribution is rarely known in practice**
    - For example, simple unimodal distributions that are well-studied and easy to apply / manipulate do not represent many real-world scenarios
    - On the other hand, non-parametric methods like histograms are usually not accurate and/or computationally feasible in high-dimensional data
  - Robustness
    - Distribution parameters (e.g. mean and standard deviation) may be very sensitive to outliers: **outliers may *mask* themselves as false negatives or *swamp* inliers as false positives by affecting distribution estimates**
      - Sophisticated (robust) distribution estimation techniques are required

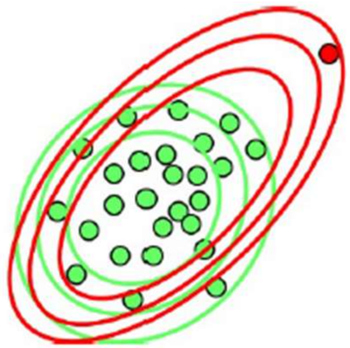




Statistical Tests



- Main Problems:
  - Parametric Assumption
    - A specific distribution must be assumed, but the assumption may not hold true for the data at hand. **True distribution is rarely known in practice**
    - For example, simple unimodal distributions that are well-studied and easy to apply / manipulate do not represent many real-world scenarios
    - On the other hand, non-parametric methods like histograms are usually not accurate and/or computationally feasible in high-dimensional data
  - Robustness
    - Distribution parameters (e.g. mean and standard deviation) may be very sensitive to outliers: **outliers may *mask* themselves as false negatives or *swamp* inliers as false positives by affecting distribution estimates**
      - Sophisticated (robust) distribution estimation techniques are required
    - **Estimation in higher dimensional data can be very difficult, if possible**



Introduction

- Main unsupervised outlier detection approaches:
  - **Statistical Approach**
    - Outliers are unusual observations according to an estimated (usually parametric) probability distribution
  - **Clustering-Based Approach**
    - Outliers are observations that do not match well any cluster, or they are those that constitute a relatively very small cluster
  - **Non-Parametric Density-Based Approach**
    - Outliers are observations with unusual (absolute or relative) density

Clustering-Based Approaches

- Main Problems:
  - Effectiveness depends highly on the clustering method used (**they may not be optimized for outlier detection**)
  - Computational cost may be prohibitive depending on the clustering method used (need to 1st find clusters, then outliers)
    - More sophisticate clustering algorithms, e.g., able to model the notion of noise and automatically determine the number of clusters in data, are often computationally more expensive

Introduction

- Main unsupervised outlier detection approaches:
  - **Statistical Approach**
    - Outliers are unusual observations according to an estimated (usually parametric) probability distribution
  - **Clustering-Based Approach**
    - Outliers are observations that do not match well any cluster, or they are those that constitute a relatively very small cluster
  - **Non-Parametric Density-Based Approach**
    - Outliers are observations with unusual (absolute or relative) density

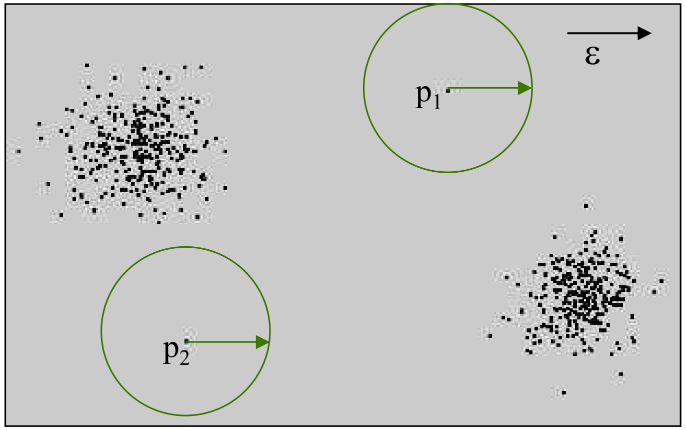
Introduction

- Main unsupervised outlier detection approaches:
  - **Statistical Approach**
    - Outliers are unusual observations according to an estimated (usually parametric) probability distribution
  - **Clustering-Based Approach**
    - Outliers are observations that do not match well any cluster, or they are those that constitute a relatively very small cluster
  - **Non-Parametric Density-Based Approach**
    - Outliers are observations with unusual (absolute or relative) density
- In the following, we provide a few classic examples of the non-parametric *density-based* approach

Non-Parametric Density-Based Global Approaches

- General idea
  - Perform a non-parametric density estimate
  - Take the density of an observation as an absolute (global) measure of "inlierness"
    - Measure of outlyingness can be derived, e.g., as its inverse or complement
- Basic assumption
  - The density around a normal data observation (an inlier) is high
  - The density around an outlier is low

Precursor of Non-Parametric Density-Based Global Approaches



- DB(ε,π)-Outliers (Distance-Based Approach)
  - 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

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

range-query with radius ε

- How can we modify the above definition, so we get rid of a parameter and at the same time provide *outlier scores* for each data object, rather than a binary classification (inliers vs outliers)??

Non-Parametric Density-Based Global Approaches

- Outlier scoring based on *k*NN distances (also deemed a **distance-based approach**)
  - One of the simplest approaches
  - ***Density** roughly estimated as inverse of distance(s) of a point to its k-nearest neighbor(s)*

  Two general models:

  - **KNN Method**: Take the *k*NN distance of a point as its outlier score
  - **Weighted KNN Method**: Aggregate (average) the distances of a point to all its 1NN, 2NN, …, *k*NN as an outlier score

  Result: "*outlier scores*"

Non-Parametric Density-Based Global Approaches

- Example (simple KNN Outlier):

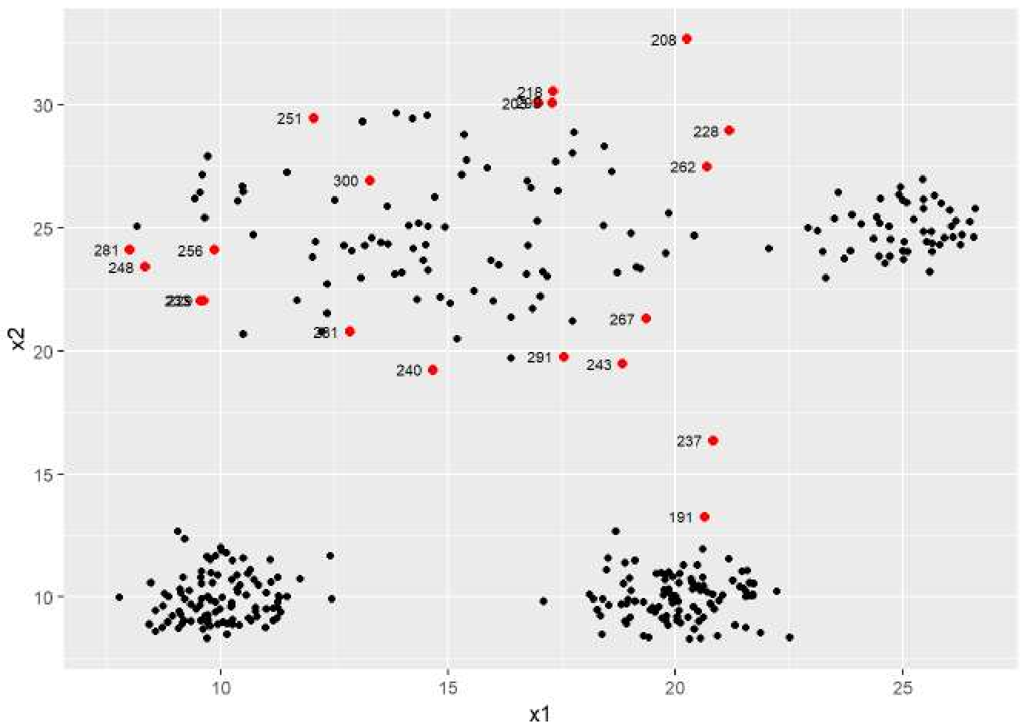


Fig. 2. KNN Outlier ($k = 4$): red points are the top 20 outliers, with their IDs

```
##      ID    score
## 1   237 4.423726
## 2   208 4.177243
## 3   228 3.411198
## 4   240 2.709398
## 5   218 2.518054
## 6   243 2.464709
## 7   262 2.380678
## 8   251 2.303902
## 9   299 2.127866
## 10  203 2.092628
## 11  233 2.079155
## 12  281 2.078306
## 13  229 2.071261
## 14  191 2.023404
## 15  267 2.012530
## 16  231 1.967966
## 17  291 1.956565
## 18  256 1.845032
## 19  248 1.832962
## 20  300 1.831324
```

Non-Parametric Density-Based Global Approaches

- Choosing the *k* Parameter:

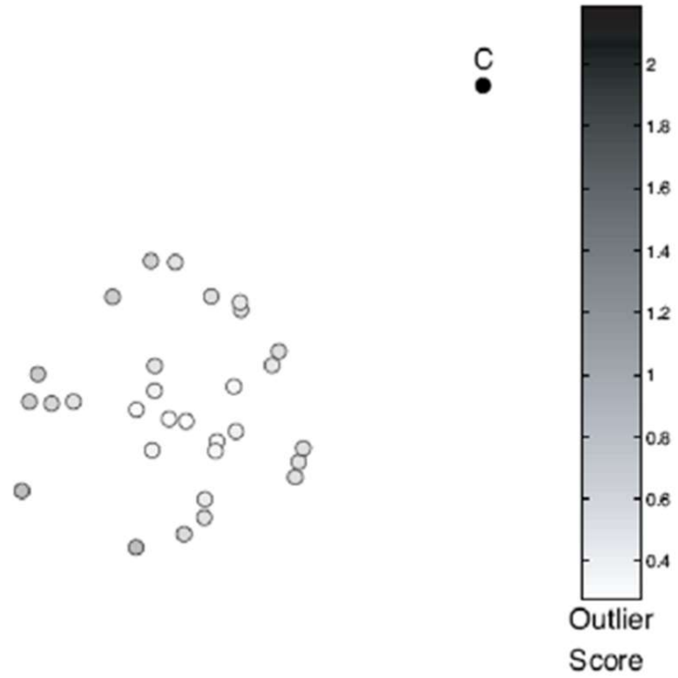


Figure 10.4. Outlier score based on the distance to fifth nearest neighbor.

(Figures from Tan et al. [2006].)

Non-Parametric Density-Based Global Approaches

- Choosing the *k* Parameter:

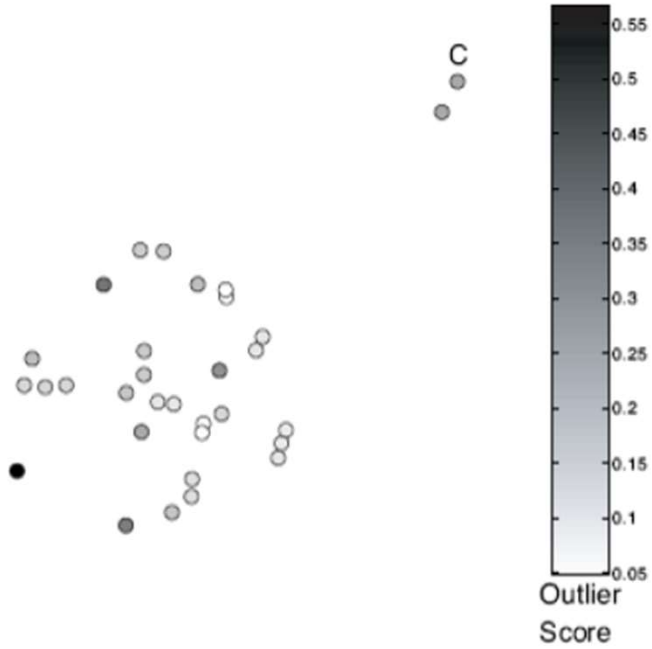


Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

(Figures from Tan et al. [2006].)

Non-Parametric Density-Based Global Approaches

- Choosing the *k* Parameter:

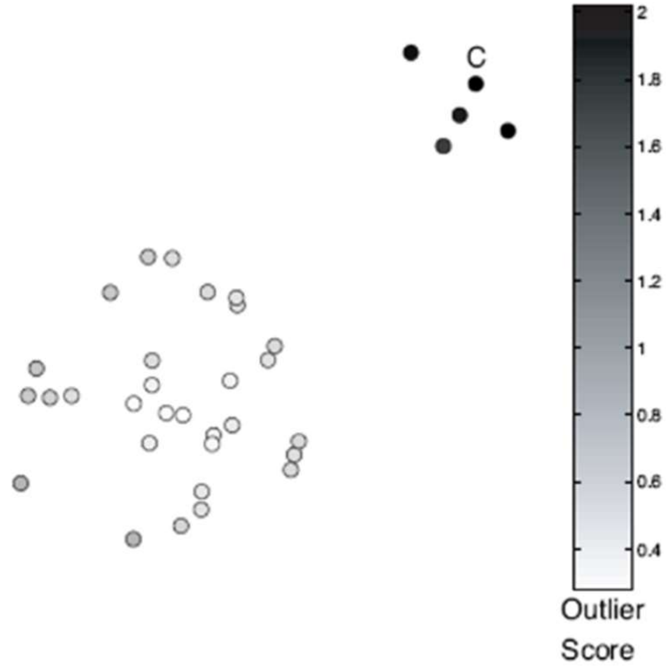


Figure 10.6. Outlier score based on distance to the fifth nearest neighbor. A small cluster becomes an outlier.

(Figures from Tan et al. [2006].)

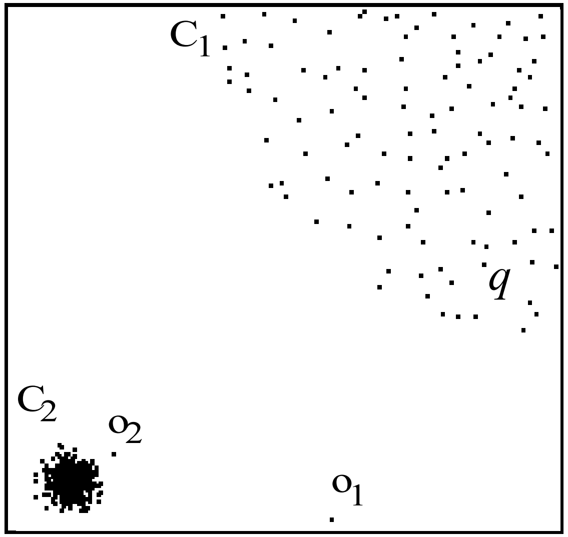
Non-Parametric Density-Based Local Approaches

- General idea
  - Compare the density around a point with the density around its local neighbors
  - The relative density of a point as contrasted 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 observation is similar to the density around its neighbors
  - The density around an outlier is considerably different to the density around its neighbors

Non-Parametric Density-Based Local Approaches

- Local Outlier Factor (LOF)
- Motivation:
  - How to detect observations that are outliers relative to a certain local subset of the data
  - Global models have problems, particularly when there are regions with different densities
  - Example
    - Outliers based on kNN-distance
      - kNN-distances of observations in C_1 (e.g. *q*) are larger than the kNN-distance of o_2



– Solution: consider relative density

Non-Parametric Density-Based Local Approaches

– Draft LOF Model (Oversimplistic, for Pedagogic Purposes)

- "Density" estimate for observation *p*
  - Inverse of the average distances from the *k*NNs of *p*

$$lrd_k(p) = 1 / \left(\frac{\sum_{o \in kNN(p)} dist(p,o)}{Card\left(kNN(p)\right)} \right)$$

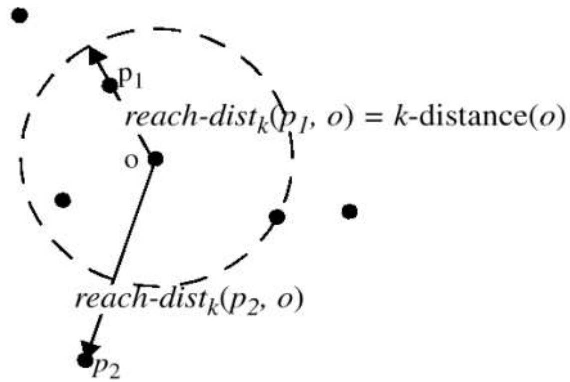
- (Draft) LOF score of observation *p*
  - *Average ratio of lrds of neighbors of p and lrd of p*

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

Non-Parametric Density-Based Local Approaches

- Original LOF Model:
  - Uses *reachability distance* rather than simple distance
    - This has the effect of reducing statistical fluctuations within clusters

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



  - Local reachability distance (lrd) of observation *p*
    - Inverse of the average reach_dists from the *k*NNs of *p*

$$lrd_k(p) = 1 / \left(\frac{\sum_{o \in kNN(p)} reach\_dist_k(p,o)}{Card\left(kNN(p)\right)} \right)$$

  - Local Outlier Factor (LOF) of observation *p*
    - Average ratio of lrds of neighbors of *p* and lrd of *p*

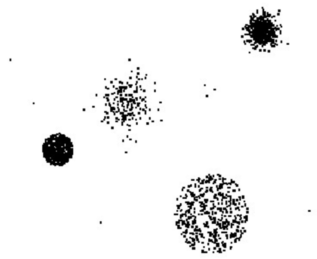
$$LOF_k(p) = \frac{\sum_{o \in kNN(p)} \frac{lrd_k(o)}{lrd_k(p)}}{Card\left(kNN(p)\right)}$$

Non-Parametric Density-Based Local Approaches

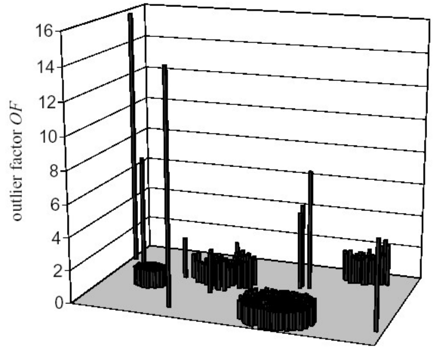
– Properties:

  • 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 (k = 40)

– Choice of k:
  • Specifies the reference local data subset
  • Works as a smoothing term of the density estimates

Non-Parametric Density-Based Local Approaches

- Example (LOF):

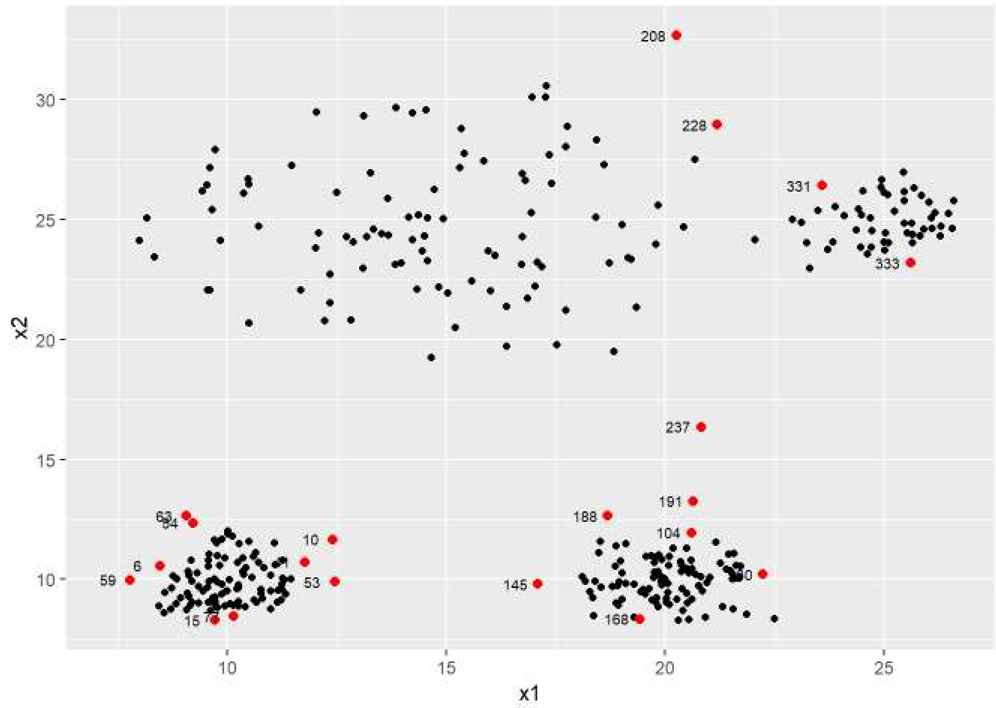
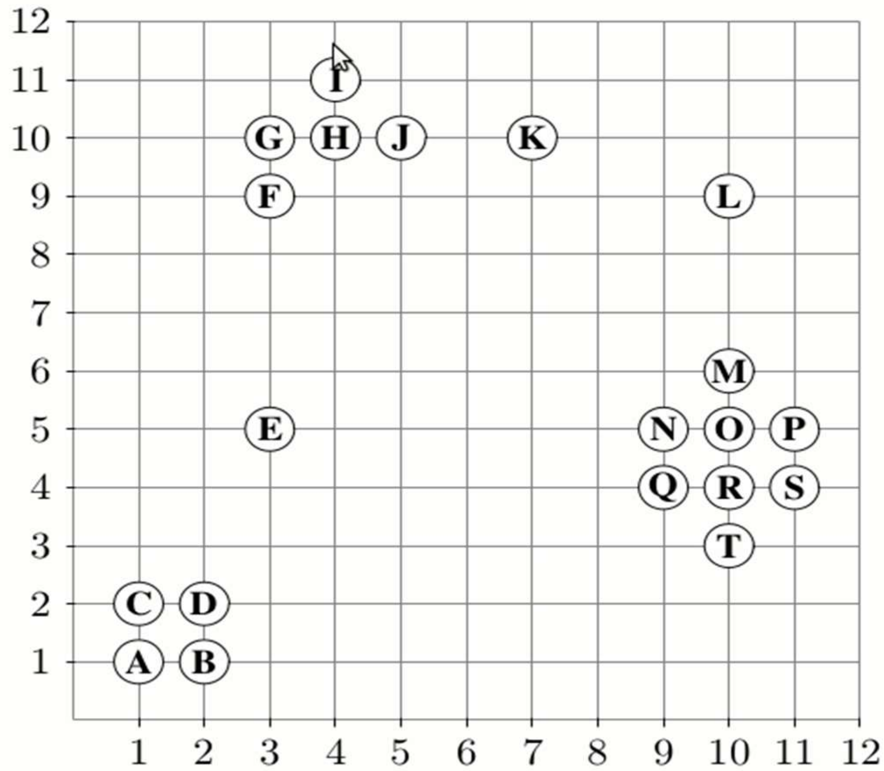


Fig. 4. LOF ($k = 4$): red points are the top 20 outliers, with their IDs

```
##      ID    score
## 1   237 2.916624
## 2   191 2.908333
## 3    53 2.615192
## 4   130 2.539595
## 5    63 2.439944
## 6    10 2.402330
## 7   145 2.289441
## 8    54 2.091639
## 9    59 1.998936
## 10  333 1.992135
## 11  228 1.944133
## 12  104 1.840832
## 13   77 1.801824
## 14  188 1.766500
## 15  208 1.753433
## 16  168 1.726824
## 17   41 1.663926
## 18  331 1.662891
## 19    6 1.619528
## 20   15 1.616930
```

Exercise

Exercise 1-1 *Outlier Scores*



As distance function, use Manhattan distance $L_1(a, b) := |a_1 - b_1| + |a_2 - b_2|$.

Compute the following (without including the query point when determining the kNN):

Exercise (continued)

- LOF using $k = 2$ for the points E, K and O.
- LOF using $k = 4$ for the points E, K and O.
- kNN distance using $k = 2$ for all points.
- kNN distance using $k = 4$ for all points.
- aggregated kNN distances for $k = 2$ and $k = 4$ for all points

Main References

- Tan, et al., “Introduction to Data Mining”, 2nd Edition, Pearson, 2018
- M. J. Zaki and W. Meira Jr. “Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge, 2nd Edition, 2020
- Han, J. and Kamber, M. “Data Mining: Concepts and Techniques”, 2nd edition, Morgan Kaufmann, 2006