DATABASE SYSTEMS GROUP

Ludwig-Maximilians-Universität München
Institut für Informatik
Lehr- und Forschungseinheit für Datenbanksysteme

LMU

Lecture notes

# Knowledge Discovery in Databases

Summer Semester 2012

## Lecture 10: Outlier detection

Lecture: Dr. Eirini Ntoutsi
Tutorials: Erich Schubert

http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_I_(KDD_I)

# Sources



- Previous KDD I lectures on LMU (P. Kröger, A. Zimek)
- “Outlier Detection Techniques” tutorial by H.-P. Kriegel, P. Kröger, A. Zimek at KDD’10.

- Tan P.-N., Steinbach M., Kumar V., *Introduction to Data Mining*, Addison-Wesley, 2006

- Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques, 3rd ed.,* Morgan Kaufmann, 2011.

# Outline



- Introduction
- Approaches for outlier detection
- Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches
- Things you should know
- Homework/tutorial

# Outlier detection/ anomaly detection



- Goal: find objects that are considerably different from most other objects or unusual or in some way inconsistent with other objects
- Outliers / anomalous objects / exceptions
- Anomaly detection/ Outlier detection / Exception mining
- It is used either as a
  - Standalone task (anomalies are the focus)
  - Preprocessing task (to improve data quality)

- Applications
  - Fraud detection (credit card, telco)
  - Intrusion detection
  - Ecosystem disturbances
  - Public health
  - Medicine
  - Fault detection

# Applications I



- 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 II

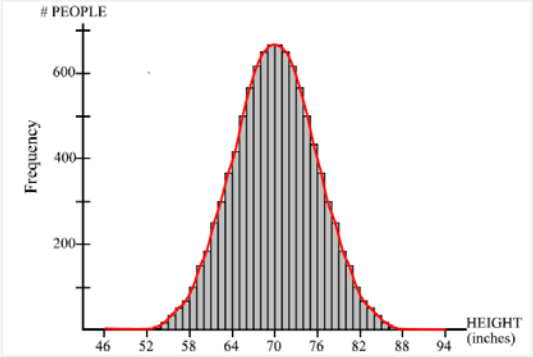


- Sports statistics
  - In many sports, various parameters are recorded for players in order to evaluate the players' performances
  - Outstanding (in a positive as well as a negative sense) players may be identified as having abnormal parameter values
  - Sometimes, players show abnormal values only 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."
- ....

# Causes of anomaly



- Data from different classes
  - An object might be different from other objects because its of another class.
  - E.g. an attack connection in a network has different characteristics from a normal connection. Or, a person who commits credit card fraud belongs to a different class than persons using credit cards legally.
  - Such anomalies are the focus in Data Mining
- Natural variation
  - Many datasets can be modeled by statistical distributions e.g. Gaussian (most of the objects are near the center and the likelihood that an object differs significantly from this avg object is small).
  - e.g., an exceptional tall person




- Data measurement and collection errors
  - erroneous measurements due to human/ measuring device errors, noise presence.
  - Such errors should be eliminated since they just reduce the quality of data
- Other causes …

# What is an outlier?



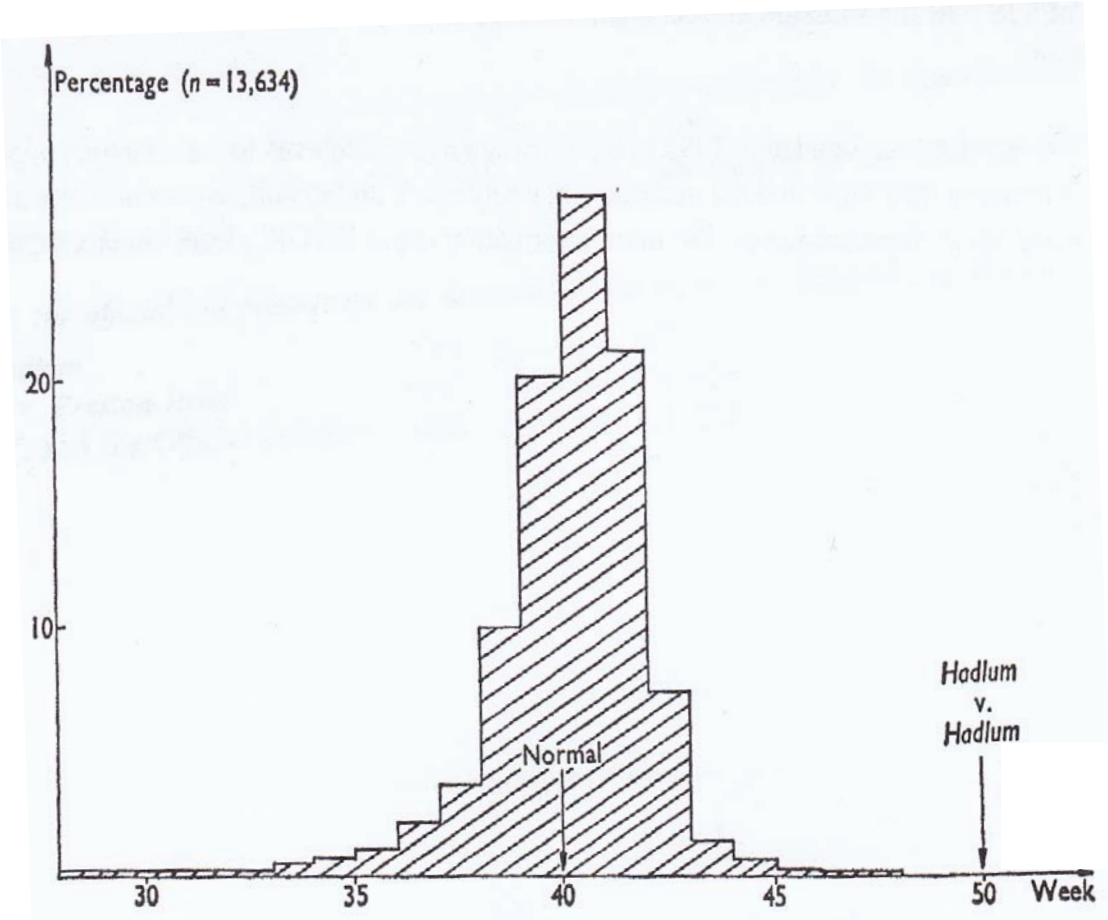
- Definition of Hawkins [Hawkins 1980]:

  "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 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 military service service.
- Average human gestation period is 280 days (40 weeks).
- Statistically, 349 days is an outlier.

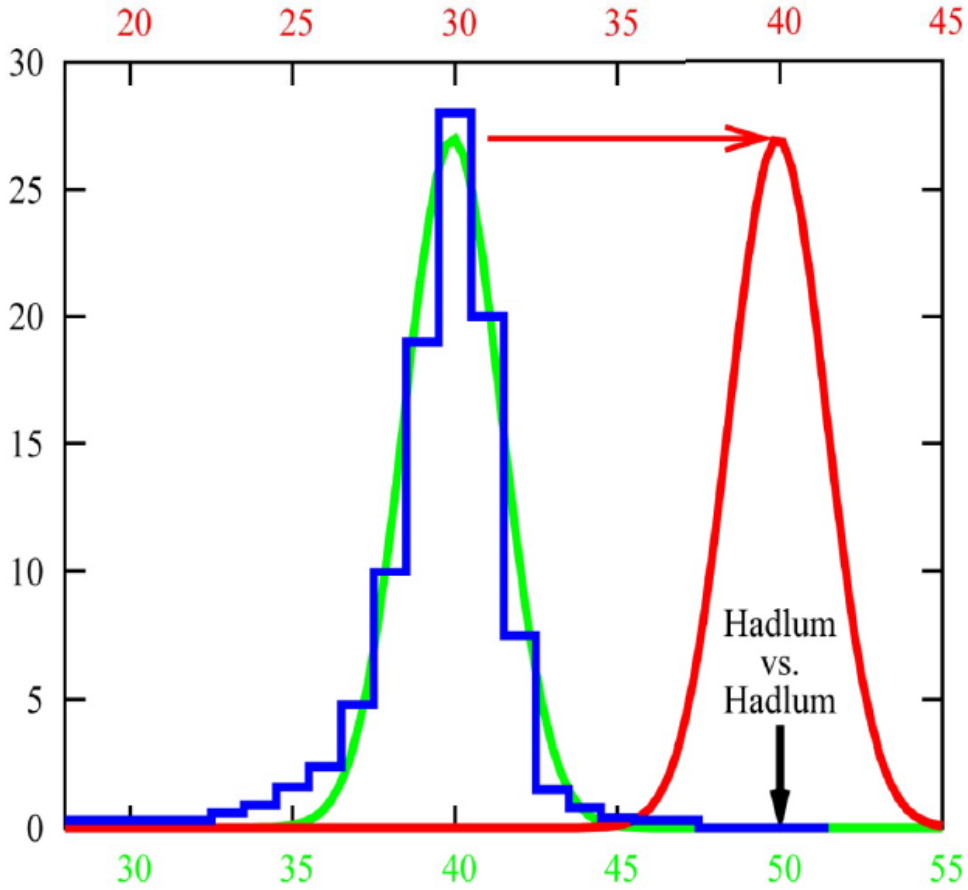


Distribution of human gestation periods.

# 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. 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)
  - Under this assumption the gestation period has an average duration and the specific birthday highest possible has highest-probability



# Discussion of the basic intuition based on Hawkins



- Data is usually multivariate, i.e., multi-dimensional

  => basic model is univariate, i.e., 1-dimensional
- There is usually more than one generating mechanism/statistical process underlying the “normal” data

  => basic model assumes only one “normal” generating mechanism
- Anomalies may represent a different class (generating mechanism) of objects, so there may be a large class of similar objects that are the outliers

  =>basic model assumes that outliers are rare observations

- A lot of models and approaches have evolved in the past years in order to exceed these assumptions

# Variants of outlier detection problems



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

- Given a database D, find all the data points $x \in D$ having the top-n largest anomaly scores f(x)

- Given a database D, containing mostly normal (but unlabeled) data points, and a test point x, compute the anomaly score of x with respect to D

# Outline



- Introduction
- Approaches for outlier detection
- Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches
- Things you should know
- Homework/tutorial

# Basic application scenarios for outlier detection



- Supervised anomaly detection
  - In some applications, training data with normal and abnormal data objects are provided
  - There may be multiple normal and/or abnormal classes
  - Often, the classification problem is highly imbalanced
- Semi-supervised anomaly detection
  - In some applications, only training data for the normal class(es) (or only the abnormal class(es)) are provided
- Unsupervised anomaly detection
  - In most applications there are no training data available
  - In such cases, the goal is to assign a score to each instance that reflects the degree to which the instance is anomalous.
  - In most applications, this is the case.

# Outlier detection vs clustering



- Are outliers just a side product of some clustering algorithms?
  - Many clustering algorithms do not assign all points to clusters but account for noise objects
  - Look for outliers by applying one of those algorithms and retrieve the noise set
- Problems
  - Clustering algorithms are optimized to find clusters rather than outliers
  - Accuracy of outlier detection depends on how good the clustering algorithm captures the structure of clusters
  - A set of many abnormal data objects that are similar to each other would be recognized as a cluster rather than as noise/outliers
- So, outlier is a problem on its own.

# Different classification approaches for outlier detection



- Global versus local outlier detection
  - Considers the set of reference objects relative to which each point's "outlierness" is judged
- Labeling versus scoring outliers
  - Considers the output of an algorithm
- Modeling properties
  - Considers the concepts based on which "outlierness" is modeled

- NOTE: we focus on models and methods for Euclidean data but many of those can be also used for other data types (because they only require a distance measure)

# Global versus local approaches



Considers the resolution of the reference set w.r.t. which the "outlierness" of a particular data object is determined

- Global approaches
  - The reference set contains all other data objects
  - Basic assumption: there is only one normal mechanism
  - Basic problem: other outliers are also in the reference set and may falsify the results
- Local approaches
  - The reference contains a (small) subset of data objects
  - No assumption on the number of normal mechanisms
  - Basic problem: how to choose a proper reference set
- NOTE: Some approaches are somewhat in between
  - The resolution of the reference set is varied e.g. from only a single object (local) to the entire database (global) automatically or by a user-defined input parameter

# Labeling versus scoring



Considers the output of an outlier detection algorithm

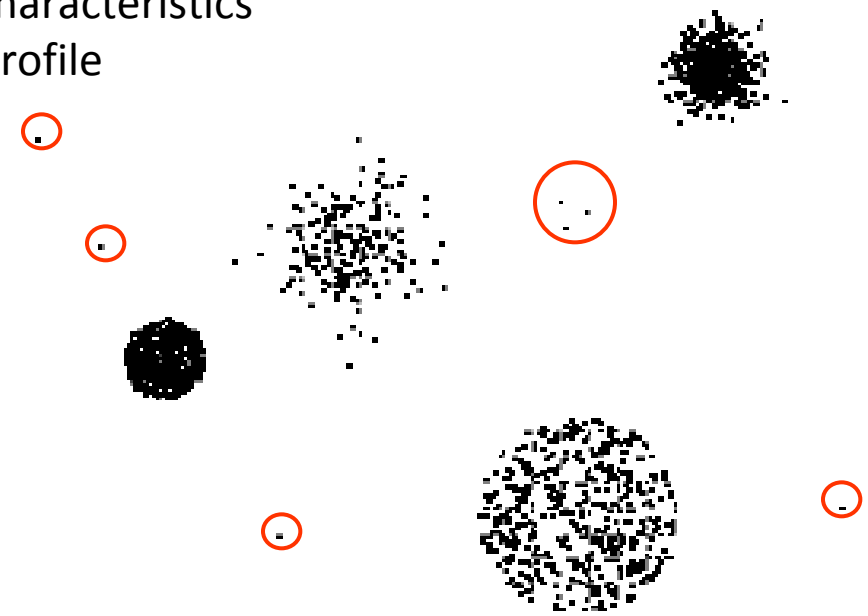
- Labeling approaches
  - Binary output
  - Data objects are labeled either as normal or outlier
- Scoring approaches
  - Continuous output
  - For each object an outlier score is computed (e.g. the probability for being an outlier)
  - Data objects can be sorted according to their scores
- Notes
  - Many scoring approaches focus on determining the top-n outliers (parameter n is usually given by the user)
  - Scoring approaches can usually also produce binary output if necessary (e.g. by defining a suitable threshold on the scoring values)

# Outlier detection schemes



w.r.t. modeling properties

- General steps
  - Build a profile of the “normal” behavior
    - Profile can be patterns or summary statistics for the overall population
  - Use the “normal” profile to detect anomalies
    - Anomalies are observations whose characteristics differ significantly from the normal profile



- Types of anomaly detection schemes
  - Model-based
  - Distance-based
  - Density-based
  - Clustering-based

# Outline



- Introduction
- Approaches for outlier detection
- Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches
- Things you should know
- Homework/tutorial

# Statistical approaches



- Are model-based approaches: a model is created for the data, and objects are evaluated w.r.t. how well they fit the model
- Most approaches are based on building a probability distribution model and considering how likely objects are under that model

An outlier is an object that has a low probability w.r.t. a probability distribution model of the data

# Statistical approaches



- 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
- A huge number of 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)

# Statistical tests



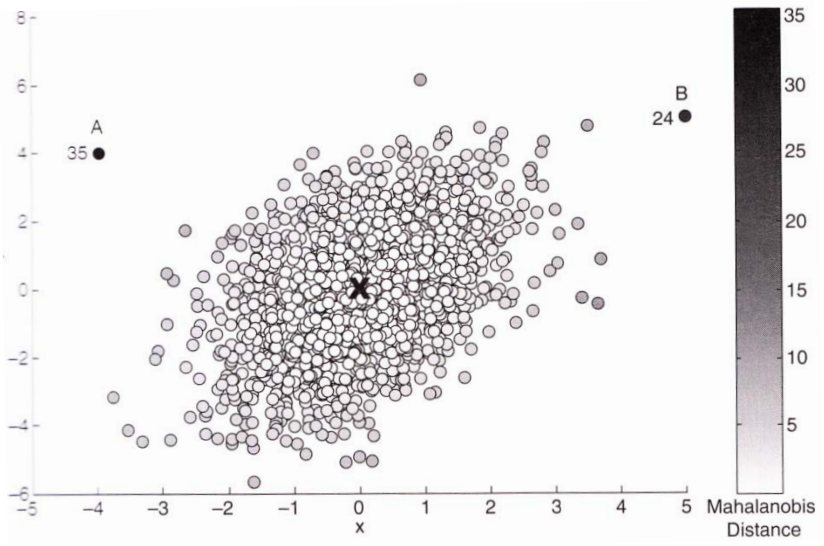
- Example below: Gaussian distribution, Multivariate, 1 model, parametric
- Probability density function of a multivariate normal distribution

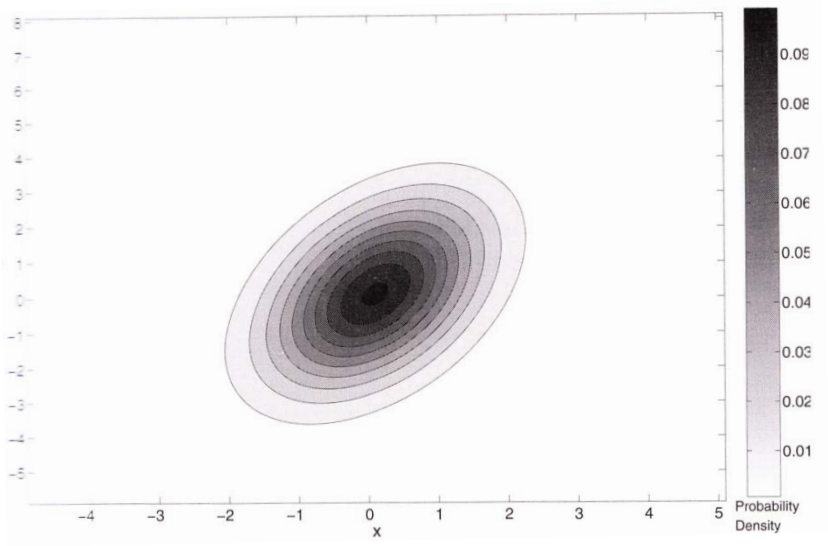
$$N(x) = \frac{1}{\sqrt{(2\pi)^d \, |\Sigma|}} e^{-\frac{(x-\mu)^{\mathbf{T}} \boldsymbol{\Sigma}^{-1} (x-\mu)}{2}}$$

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

# Example

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

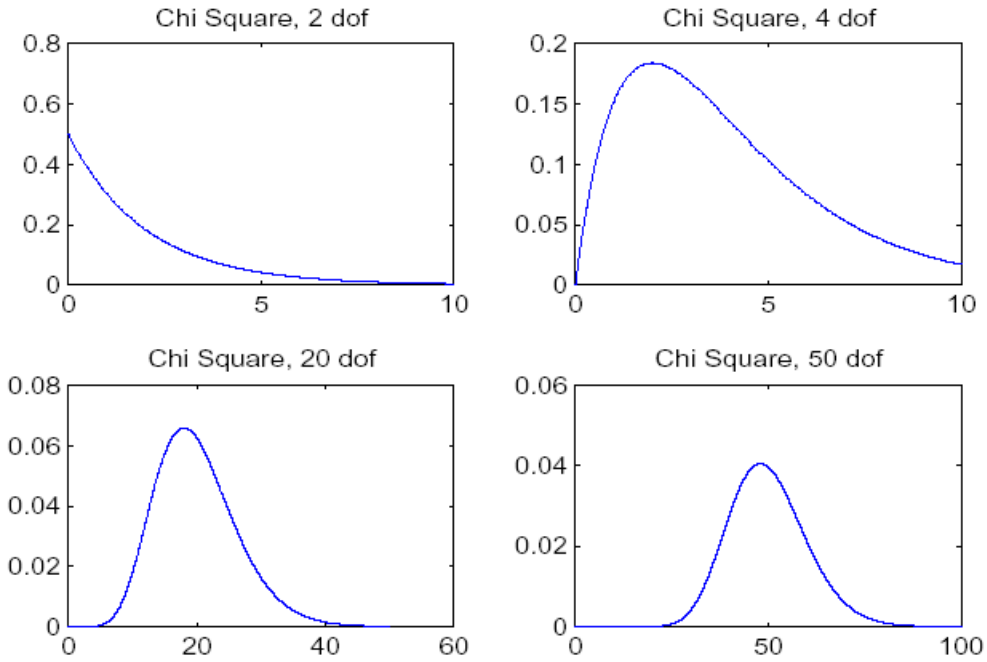


# Problems I



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




x-axis: observed MDist values

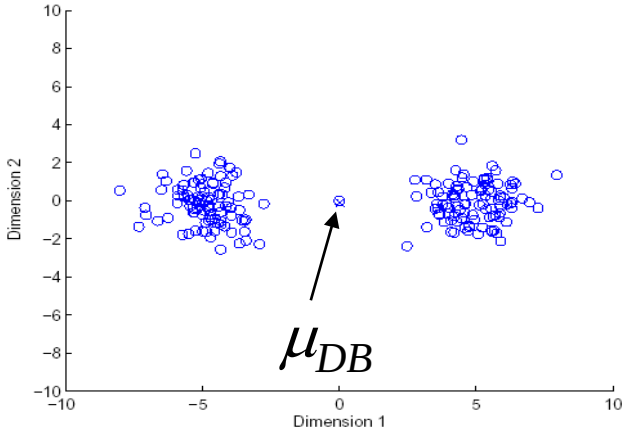
y-axis: frequency of observation

# Problems II



- 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

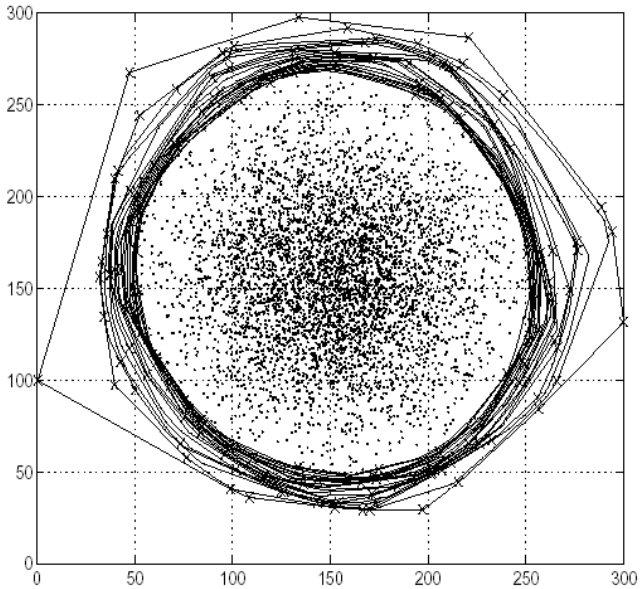
    => Minimum Covariance Determinant [Rousseeuw and Leroy 1987] minimizes the influence of outliers on the Mahalanobis distance
- Discussion
  - Data distribution is fixed
  - Low flexibility (no mixture model)
  - Global method
  - Outputs a label but can also output a score



# Depth-based approaches



- General idea
  - Search for outliers at the border of the data space but independent of statistical distributions
- Organize data objects in convex hull layers
- Outliers are objects on outer layers




Picture taken from [Johnson et al. 1998]

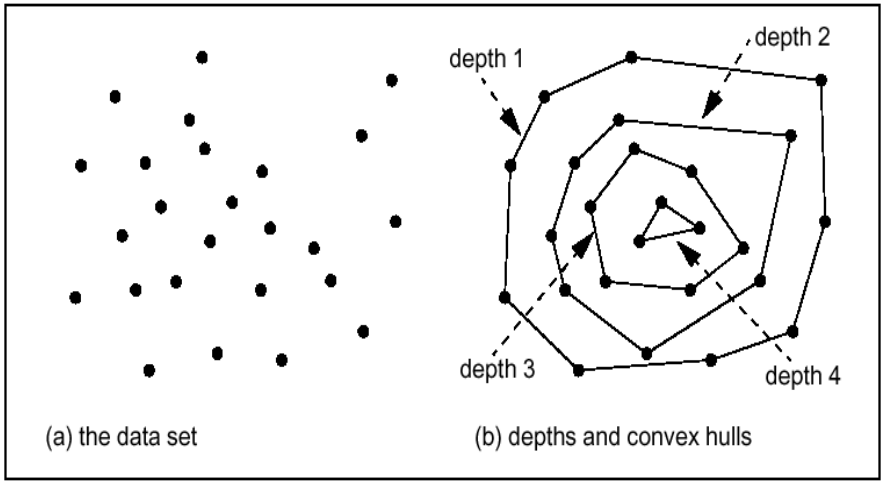
- Basic assumption
  - Outliers are located at the border of the data space
  - Normal objects are in the center of the data space

# Depth-based approaches



Model [Tukey 1977]

- Points on the convex hull of the full data space have depth = 1
- Points on the convex hull of the data set after removing all points with depth = 1 have depth = 2
- …
- Points having a depth $\leq k$ are reported as outliers

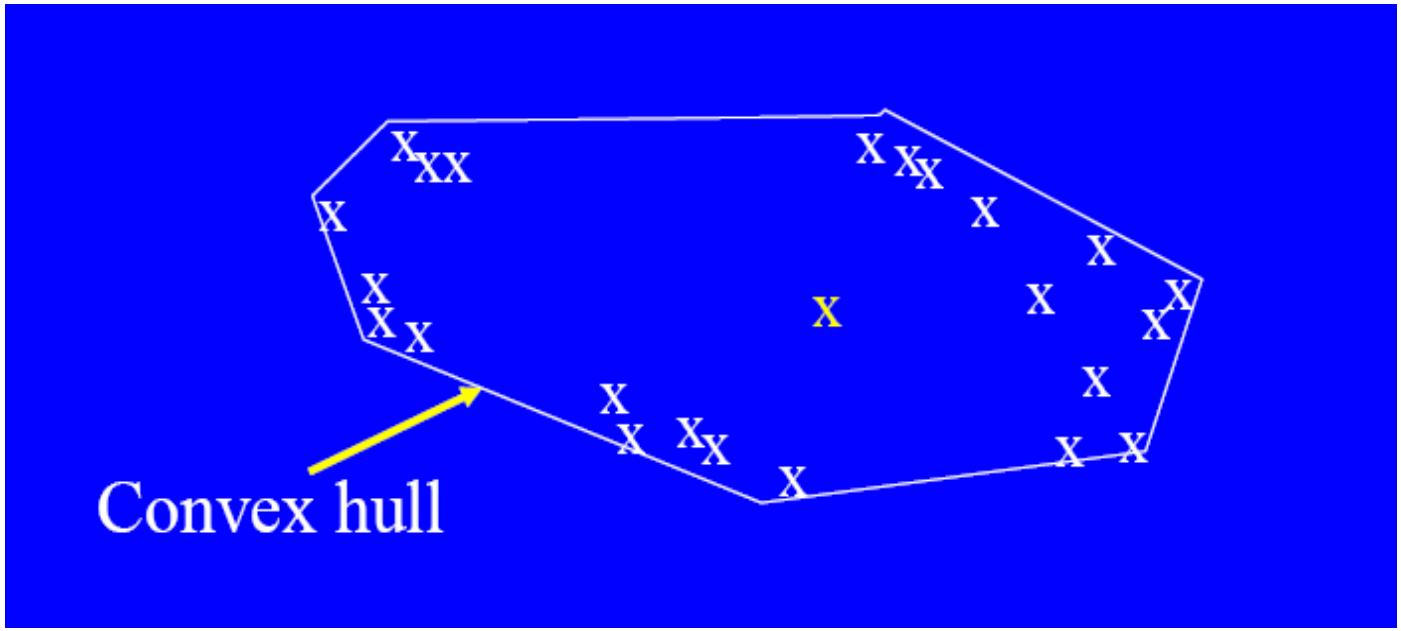



Picture taken from [Preparata and Shamos 1988]

# Depth-based approaches



What if the outlier occurs in the middle of the data?



# Discussion



- Sample algorithms
  - ISODEPTH [Ruts and Rousseeuw 1996]
  - FDC [Johnson et al. 1998]

- Discussion
  - Similar idea like classical statistical approaches (k = 1 distributions) but independent from the chosen kind of distribution
  - Convex hull computation is usually only efficient in 2D / 3D spaces
  - Originally outputs a label but can be extended for scoring easily (take depth as scoring value)
  - Uses a global reference set for outlier detection

# Deviation-based approaches



- General idea
  - Given a set of data points (local group or global set)
  - Outliers are points that do not fit to the general characteristics of that set, e.g., the variance of the set is minimized when removing the outliers
- Basic assumption
  - Outliers are the outermost points of the data set

# Deviation-based approaches



- Model
  - Given a smoothing factor SF($I$) that computes for each $I \subseteq DB$ how much the variance of $DB$ is decreased when $I$ is removed from $DB$
  - With equal decrease in variance, a smaller exception set is better
  - The outliers are the elements of the exception set $E \subseteq DB$ for which the following holds:

$$SF(E) \geq SF(I) \text{ for all } I \subseteq DB$$

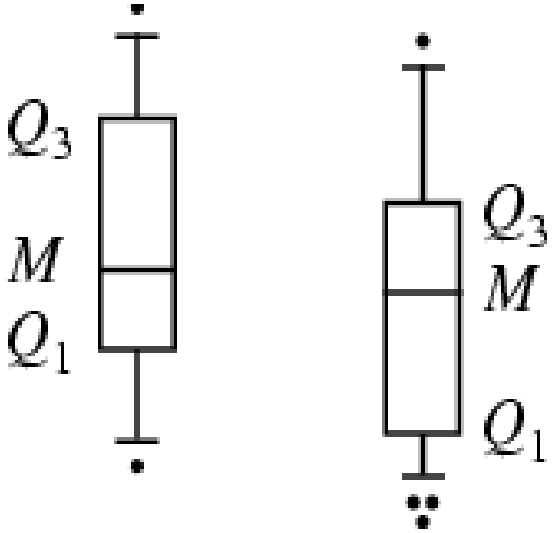
- Discussion:
  - Similar idea like classical statistical approaches (k = 1 distributions) but independent from the chosen kind of distribution
  - Naïve solution is in $O(2^n)$ for $n$ data objects
  - Heuristics like random sampling or best first search are applied
  - Applicable to any data type (depends on the definition of SF)
  - Originally designed as a global method
  - Outputs a labeling

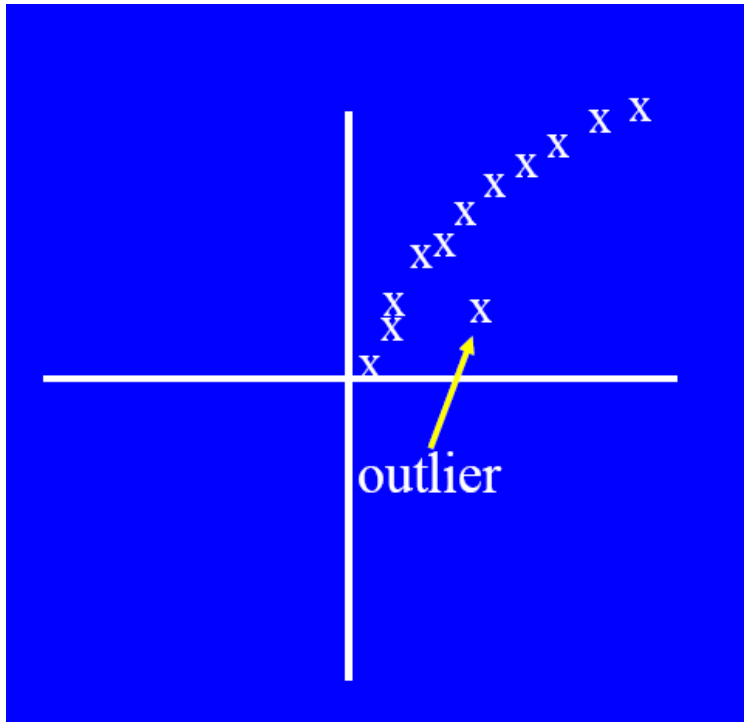
# Graphical approaches



- Boxplot (1-D), Scatter plot (2-D)

- Limitations
  - Time consuming
  - Subjective



# Outline



- Introduction
- Approaches for outlier detection
- Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches
- Things you should know
- Homework/tutorial

# Distance-based approaches



- Data is represented as a vector of features
- General idea:
  - Judge a point based on the distance(s) to its neighbors
- Basic assumption:
  - An object is an anomaly if it is distant from most points
  - Normal data objects have a dense neighborhood
  - Outliers are far apart from their neighbors, i.e., have a less dense neighborhood
- More general and more easily applied than statistical approaches since its easier to find a suitable proximity measure than to determine the statistical distribution
- Several variations
  - Data points for which there are fewer than p neighboring points within a distance D
  - The top n data points whose distance to the kth nearest neighbor is greatest
  - The top n data points whose average distance to the k nearest neighbors is greatest

# Basic model [Knorr and Ng 1997]

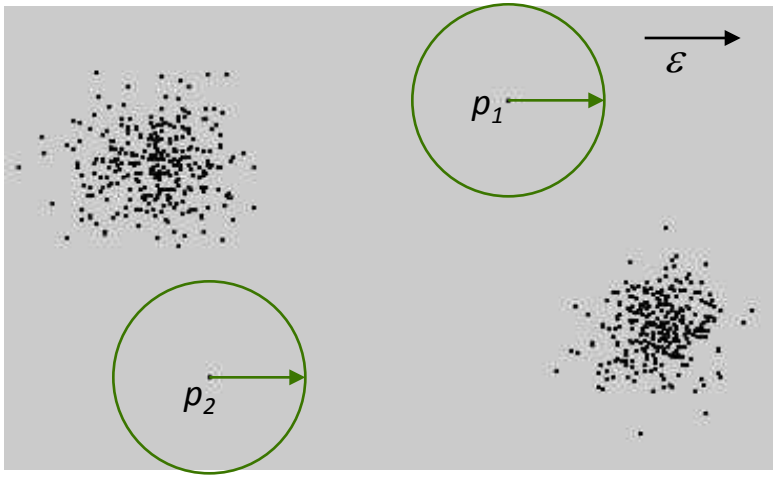


## DB(ε,π)-Outliers

- 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

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

range-query with radius ε



# $k^{th}$ nearest neighbor based I



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

- lowest outlier score 0.

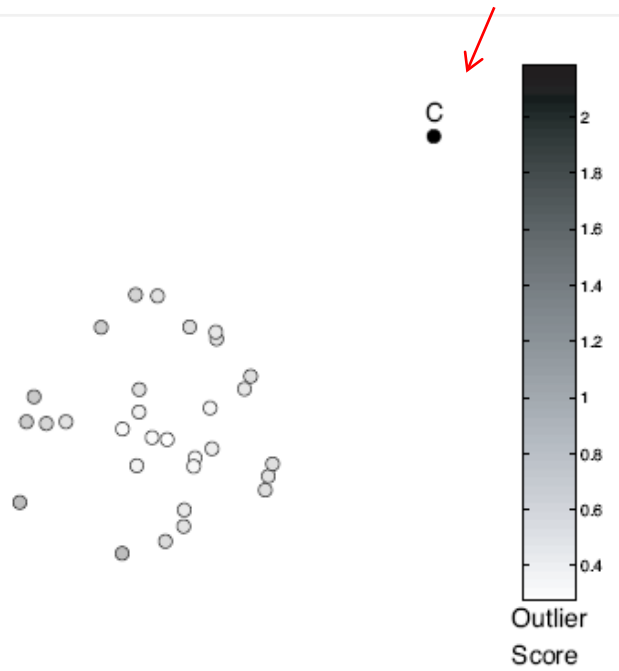



**Figure 10.4.** Outlier score based on the distance to fifth nearest neighbor.

k=5

# $k^{th}$ nearest neighbor based II



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

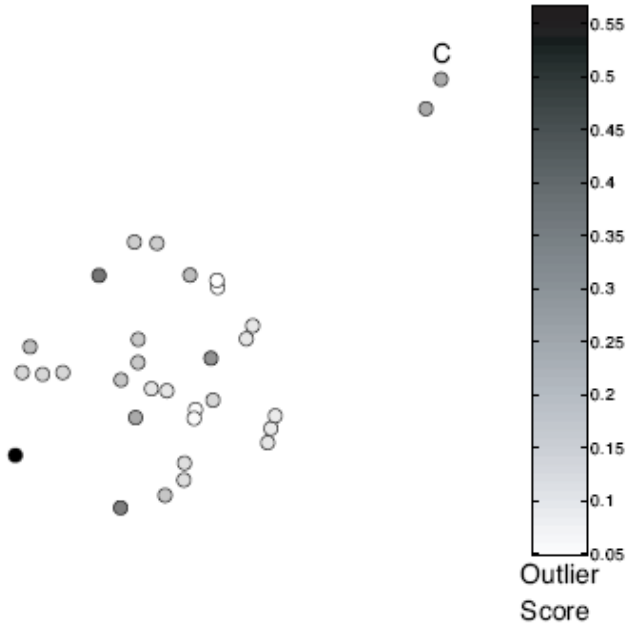



**Figure 10.5.** Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

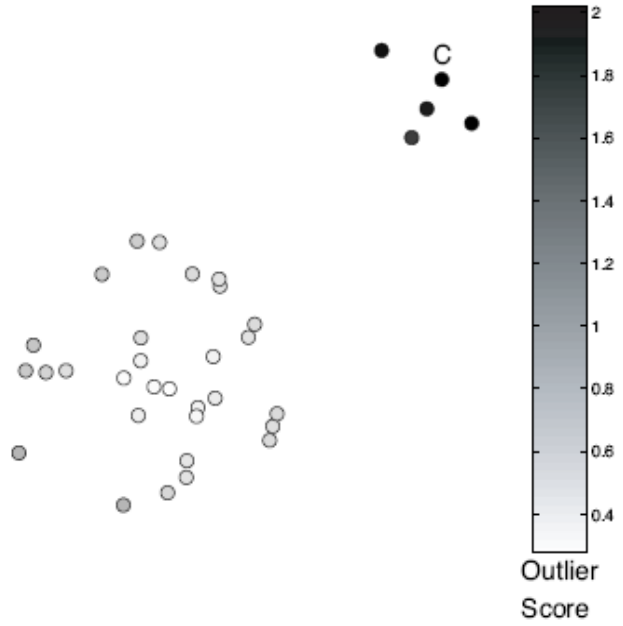



**Figure 10.6.** Outlier score based on distance to the fifth nearest neighbor. A small cluster becomes an outlier.

If k is to small, then a small number of nearby outliers can cause low outlier scores.

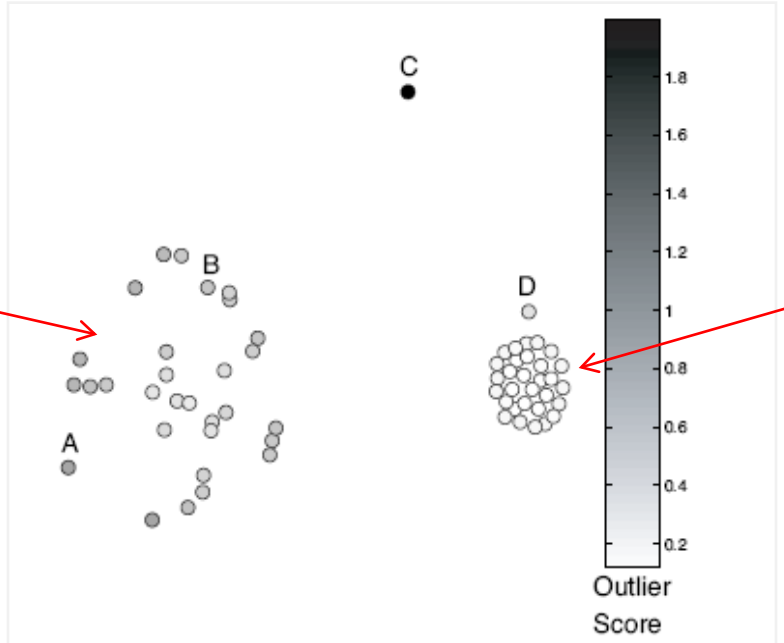
If k is to large, then all objects in a cluster with less than k objects might become outliers.

# $k^{th}$ nearest neighbor based III



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

Loose cluster




Dense cluster

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

# Distance-based approaches overview



- Simple schemes
- Expensive
  - Index structures or specialized algorithms have been proposed for performance improvement
- Sensitive to the choice of parameters
- In high-dimensional space, data is sparse and notion of proximity becomes meaningless
  - Every point is an almost equally good outlier from the perspective of proximity-based definitions
  - Lower-dimensional projection methods have been proposed

# Outline



- Introduction
- Approaches for outlier detection
- Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches
- Things you should know
- Homework/tutorial

# Density-based approaches I



- Outliers are objects in regions of low density
- 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 essentially 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
- Closely related to distance-based methods, since density is usually defined in terms of proximity.

# Density-based approaches II



The outlier score of an object is the inverse of the density around this object

- 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 (Recall our discussion on clustering)
  - Fail when data contain regions of different densities
  - Solution: use a notion of density that is relative to the neighborhood of the object

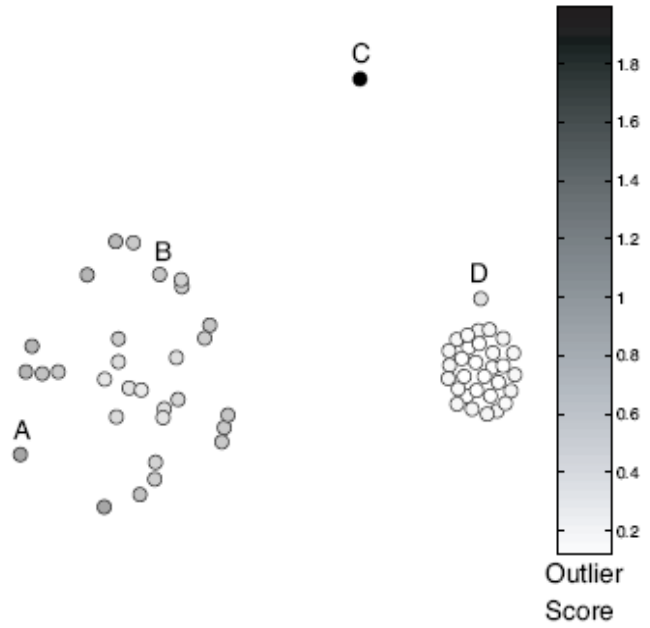


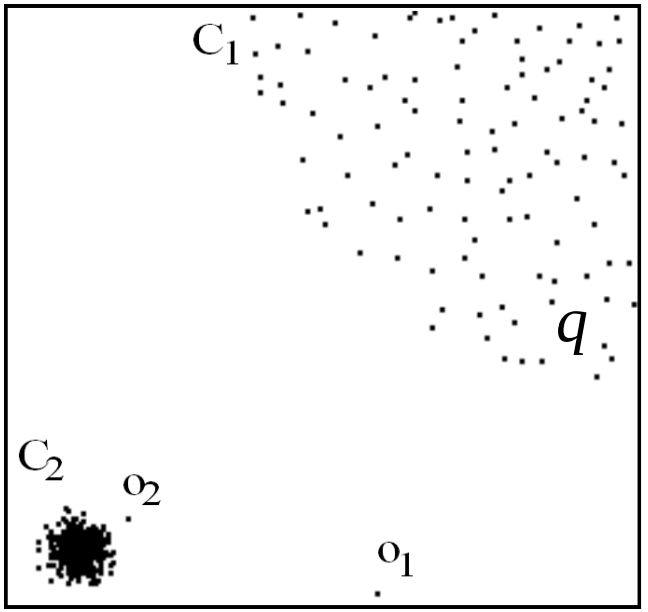

**Figure 10.7.** Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

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

# LOF(Local Outlier Factor)



- Local Outlier Factor (LOF) [Breunig et al. 1999], [Breunig et al. 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($\varepsilon$,$\pi$)-outlier model
      - Parameters $\varepsilon$ and $\pi$ 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

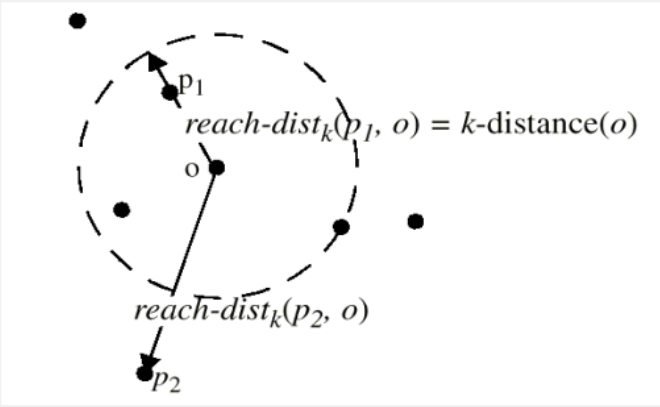


# LOF model



- 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*

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

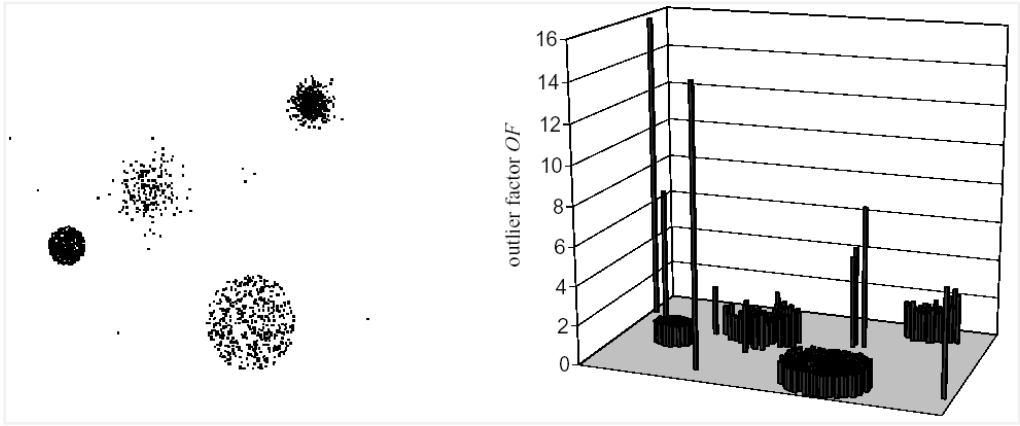
- Local outlier factor (LOF) of point *p*
  - Average ratio of lrds of neighbors of p and lrd of p

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

# LOF



- Properties
  - LOF ≈ 1: point is in a cluster (region with homogeneous density around the point and its neighbors)
  - LOF >> 1: point is an outlier
  - So, outliers are points with the largest LOF values




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)

# LOF example

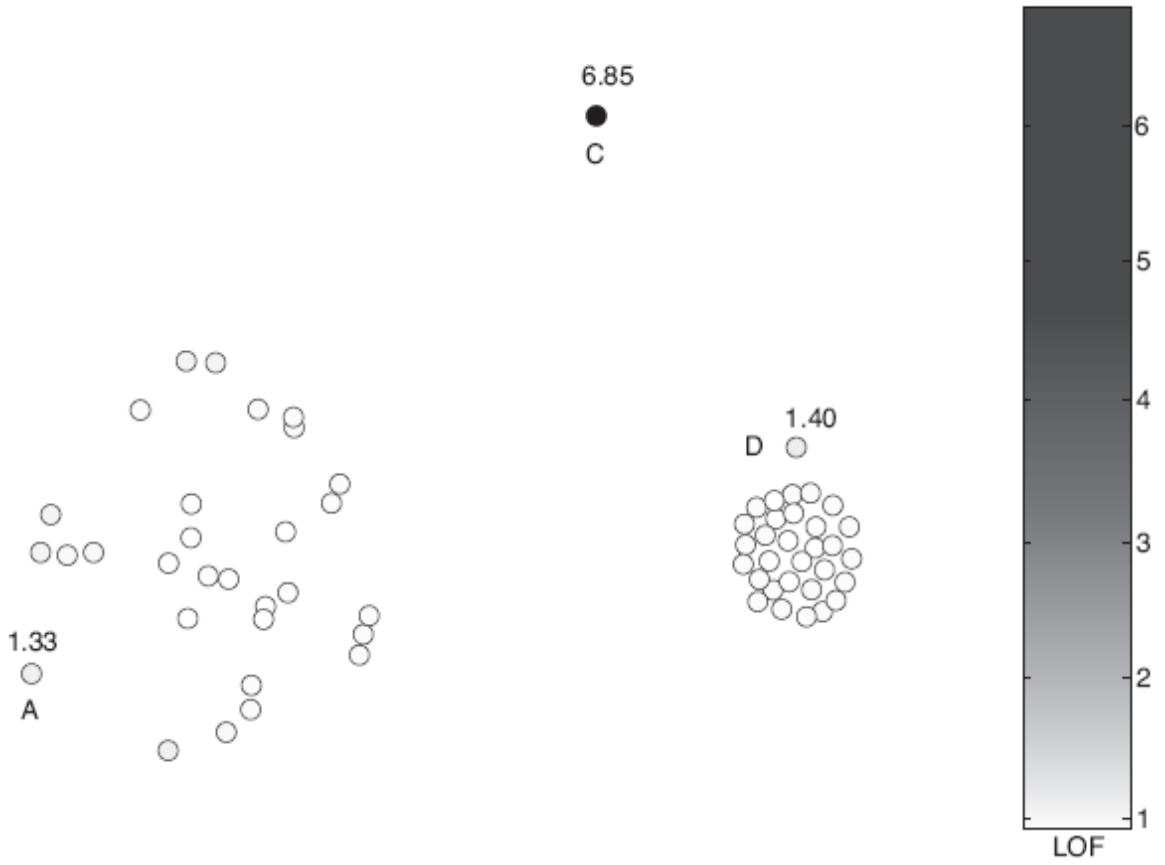



**Figure 10.8.** Relative density (LOF) outlier scores

# Variants of LOF



e.g., Mining top-n local outliers [Jin et al. 2001]

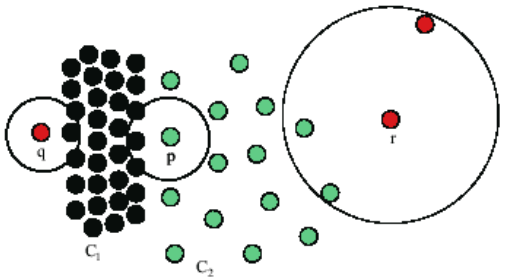
- Idea:
  - Usually, a user is only interested in the top-n outliers
  - Do not compute the LOF for all data objects => save runtime
- Method
  - Compress data points into micro clusters using the CFs of BIRCH [Zhang et al. 1996]
  - Derive upper and lower bounds of the reachability distances, lrd-values, and LOF-values for points within a micro clusters
  - Compute upper and lower bounds of LOF values for micro clusters and sort results w.r.t. ascending lower bound
  - Prune micro clusters that cannot accommodate points among the top-n outliers (n highest LOF values)
  - Iteratively refine remaining micro clusters and prune points accordingly

# Influenced Outlierness (INFLO)



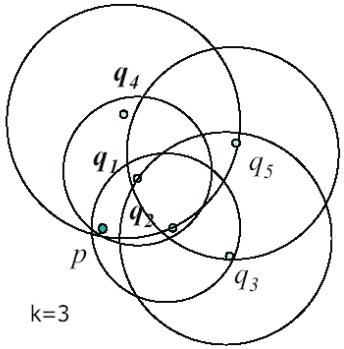
## Influenced Outlierness (INFLO) [Jin et al. 2006]

- Motivation
  - If clusters of different densities are not clearly separated, LOF will have problems




Point p will have a higher LOF than points q or r which is counter intuitive

- Idea
  - Take symmetric neighborhood relationship into account
  - Influence space (kIS(p)) of a point p includes its kNNs (kNN(p)) and its reverse kNNs (RkNN(p))




$$kIS(p) = kNN(p) \cup RkNN(p))$$

$$= \{q_1, q_2, q_4\}$$

# INFLO model



- Density is simply measured by the inverse of the kNN distance, i.e.,

  den(p) = 1/k-distance(p)

- Influenced outlierness of a point p

$$INFLO_k(p) = \frac{\sum_{o \in kIS(p)} den(o) \,/\, Card(kIS(p))}{den(p)}$$

- INFLO takes the ratio of the average density of objects in the neighborhood of a point p (i.e., in kNN(p) $\cup$ RkNN(p)) to p's density
- Proposed algorithms for mining top-n outliers
  - Index-based
  - Two-way approach
  - Micro cluster based approach
- Similar to LOF
  - INFLO ~ 1: point is in a cluster
  - INFLO >> 1: point is an outlier

# Local outlier correlation integral (LOCI)



Local outlier correlation integral (LOCI) [Papadimitriou et al. 2003]

- Idea is similar to LOF and variants
- Differences to LOF
  - Take the ε-neighborhood instead of kNNs as reference set
  - Test multiple resolutions (here called "granularities") of the reference set to get rid of any input parameter
- Model
  - ε-neighborhood of a point p: $N(p,\varepsilon) = \{q \mid dist(p,q) \leq \varepsilon\}$
  - Local density of an object p: number of objects in $N(p,\varepsilon)$
  - Average density of the neighborhood

$$den(p,\varepsilon,\alpha) = \frac{\sum_{q \in N(p,\varepsilon)} Card(N(q,\alpha \cdot \varepsilon))}{Card(N(p,\varepsilon))}$$

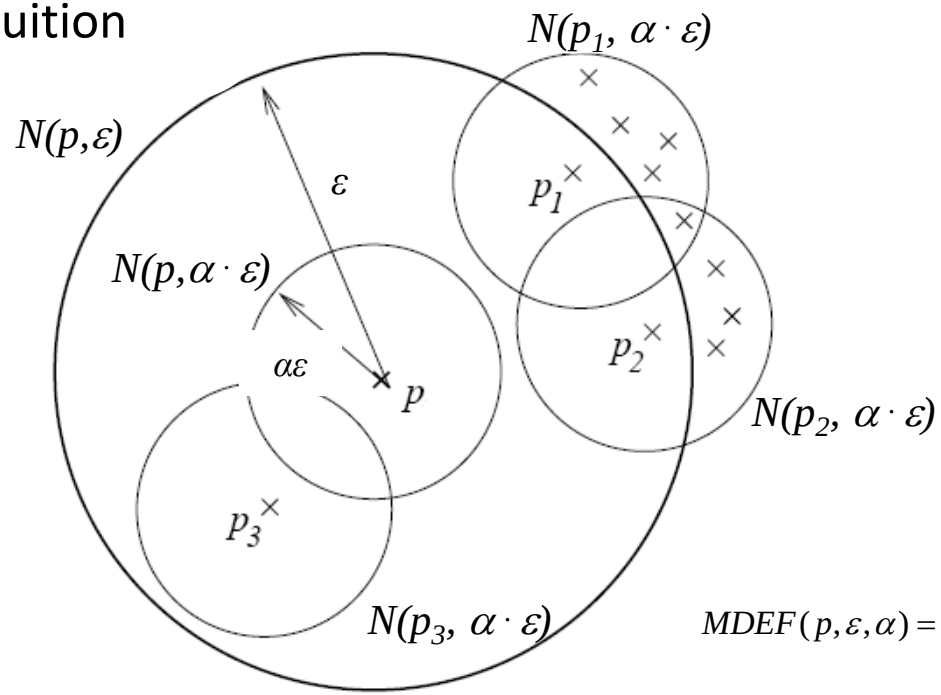
  - Multi-granularity Deviation Factor (MDEF)

$$MDEF(p,\varepsilon,\alpha) = \frac{den(p,\varepsilon,\alpha) - Card(N(p,\alpha \cdot \varepsilon))}{den(p,\varepsilon,\alpha)} = 1 - \frac{Card(N(p,\alpha \cdot \varepsilon))}{den(p,\varepsilon,\alpha)}$$

# LOCI



- Intuition




$$den(p,\varepsilon,\alpha) = \frac{\sum_{q \in N(p,\varepsilon)} Card(N(q,\alpha \cdot \varepsilon))}{Card(N(p,\varepsilon))}$$

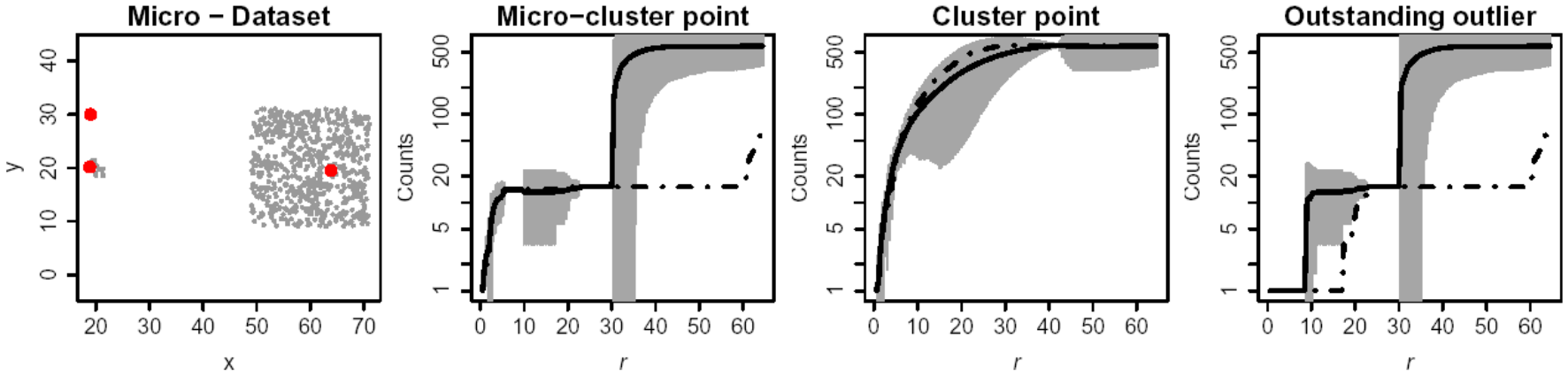
$$MDEF(p,\varepsilon,\alpha) = \frac{den(p,\varepsilon,\alpha) - Card(N(p,\alpha \cdot \varepsilon))}{den(p,\varepsilon,\alpha)} = 1 - \frac{Card(N(p,\alpha \cdot \varepsilon))}{den(p,\varepsilon,\alpha)}$$

- σMDEF(p,ε,a) is the normalized standard deviation of the densities of all points from N(p,ε)
- Properties
  - MDEF = 0 for points within a cluster
  - MDEF > 0 for outliers or MDEF > 3σ MDEF => outlier

# LOCI



- Features
  - Parameters α and ε are automatically determined
  - In fact, all possible values for ε are tested
  - LOCI plot displays for a given point p the following values w.r.t. ε
    - Card(N(p, αε))
    - den(p, ε, α) with a border of (+-) 3σden(p, ε, α)

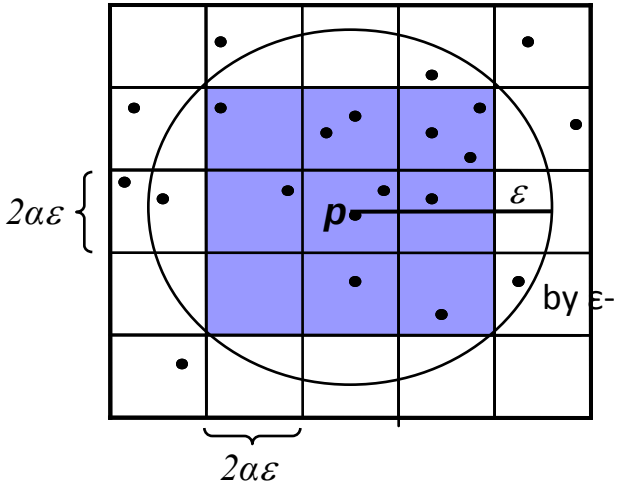


# LOCI algorithm



- Exact solution is rather expensive (compute MDEF values for all possible ? values)
- aLOCI: fast, approximate solution
  - Discretize data space using a grid with side length 2αε
  - Approximate range queries trough grid cells
  - ε - neighborhood of point p: ζ(p,ε)

    all cells that are completely covered
    sphere around p




- Then,

$$Card(N(q,\alpha\cdot\varepsilon)) = \frac{\sum_{c_j\in\zeta(p,\varepsilon)} c_j^{\,2}}{\sum_{c_j\in\zeta(p,\varepsilon)} c_j}$$

  where $c_j$ is the object count the corresponding cell

- Since different ε values are needed, different grids are constructed with varying resolution
- These different grids can be managed efficiently using a Quad-tree

# LOCI overview



- Exponential runtime w.r.t. data dimensionality
- Output:
  - Label: if MDEF of a point > 3σMDEF then this point is marked as outlier
- LOCI plot
  - At which resolution is a point an outlier (if any)
  - Additional information such as diameter of clusters, distances to clusters, etc.
- All interesting resolutions, i.e., possible values for ε, (from local to global) are tested

# Outline



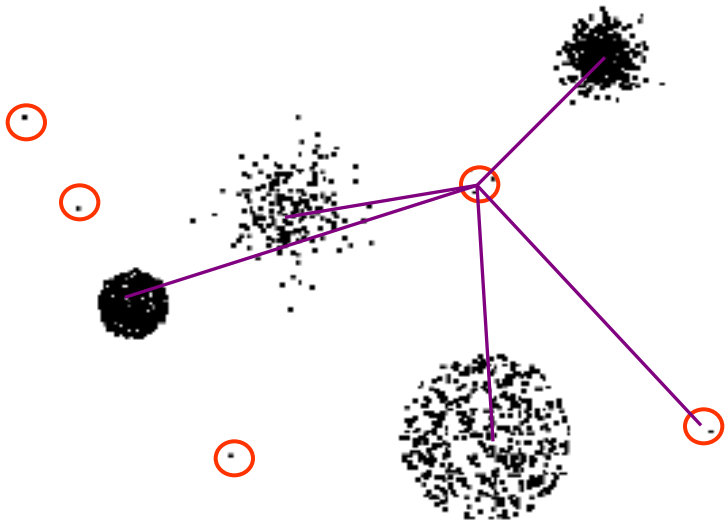
- Introduction
- Approaches for outlier detection
- Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches
- Things you should know
- Homework/tutorial

# Clustering-based approaches



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, 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.



# Prototype-based clusters



- Like k-Means, k-Medoids
- Several ways to assess the extent to which a point belongs to a cluster
  - Measure the distance of the object to the cluster prototype and take this as the outlier score
  - Or, if the clusters are of different densities, the outlier score could be the relative distance of an object from the cluster prototype w.r.t. the distances of the other objects in the cluster.

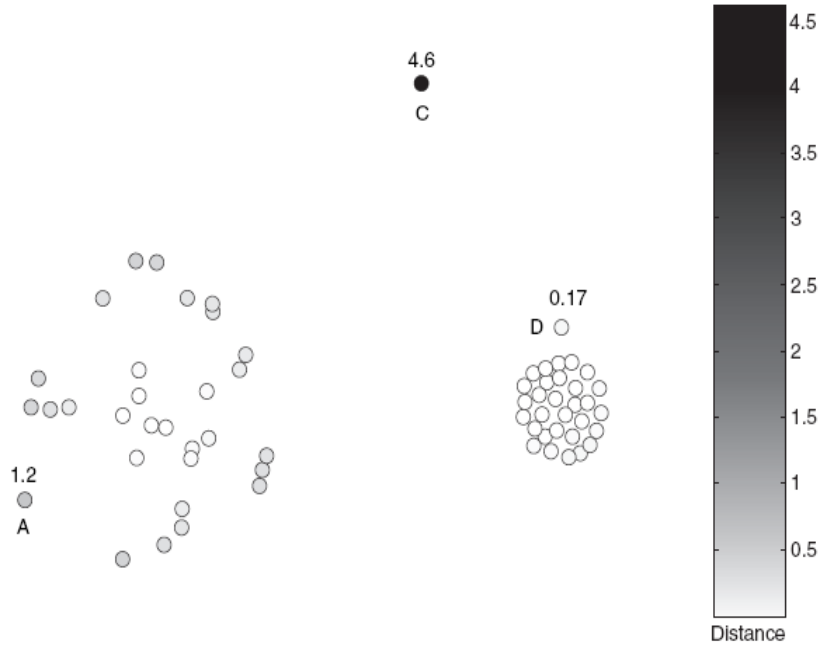



**Figure 10.9.** Distance of points from closest centroid.

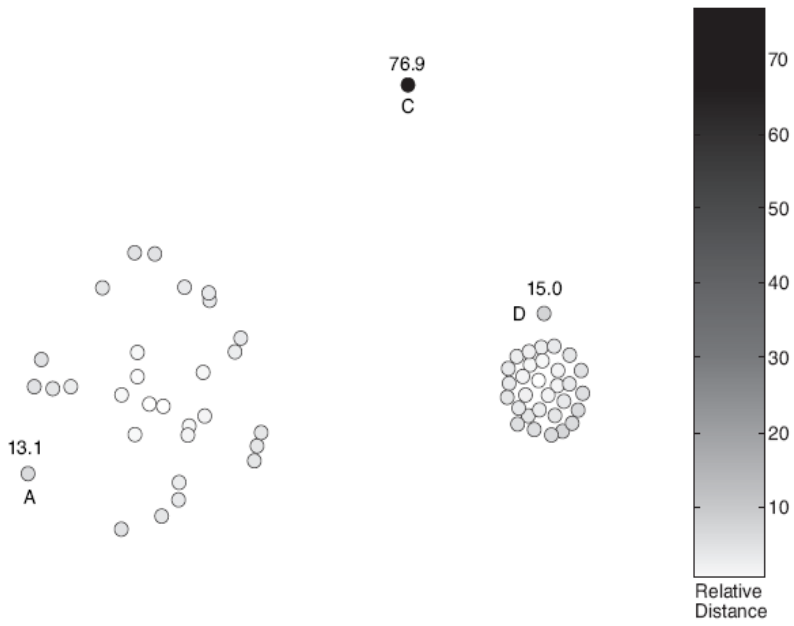



**Figure 10.10.** Relative distance of points from closest centroid.

# Outline



- Introduction
- Approaches for outlier detection
- Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches
- Things you should know
- Homework/tutorial

# Things you should know



- The notion of outliers
- Basic approaches to outlier detection
  - Supervised, unsupervised, semi-supervised
- Basic techniques
- Statistical-based
- Distance-based
  - $k^{th}$ nearest neighbor
- Density-based
  - LOF
- Clustering-based

# Homework/ Tutorial



**Tutorial:** Tutorial this Thursday on outlier detection

**Homework:**

- Try different outlier detection algorithms in Elki
  - Play with parameters
  - Interpret the charts

**Suggested reading:**

- Tan P.-N., Steinbach M., Kumar V., *Introduction to Data Mining*, Addison-Wesley, 2006 (Chapter 10).
- Han J., Kamber M., Pei J. *Data Mining: Concepts and Techniques 3rd ed., Morgan Kaufmann, 2011* (Chapter 12)