C | A | U Christian-Albrechts-Universität zu Kiel Technische Fakultät

Inf-KDDM: Knowledge Discovery and Data Mining

Winter Term 2019/20

Lecture 7: Outlier Detection

Lectures: Prof. Dr. Matthias Renz

Exercises: Christian Beth

Recap from previous lecture

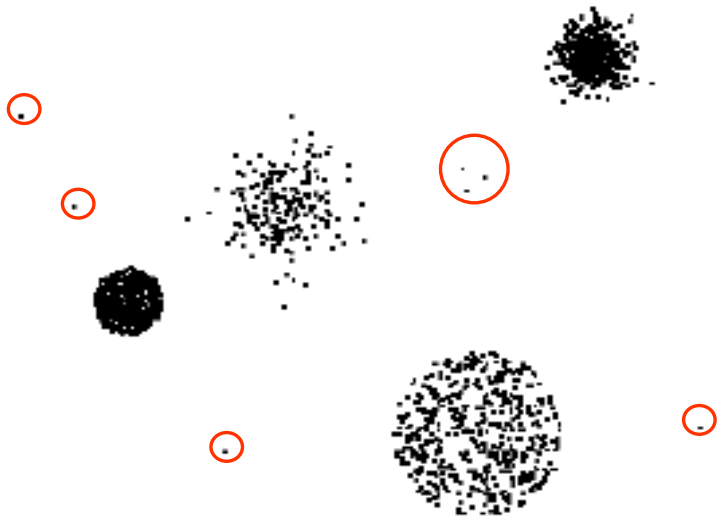
- Density-based clustering cont'
- EM clustering
- Cluster validity
  - Internal & external measures of cluster validity

Outline

- Introduction
- Approaches for outlier detection
- Model-based or Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches

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

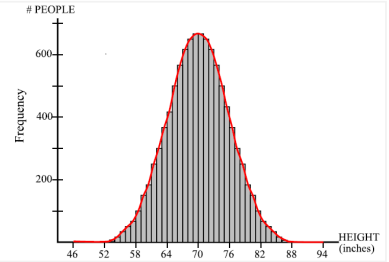
- Fraud detection
  - Purchasing behavior of a credit card owner usually changes when the card is stolen
  - Abnormal buying patterns can characterize credit card abuse
- Medicine
  - Unusual symptoms or test results may indicate potential health problems of a patient
  - Whether a particular test result 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 objects are near the center, s.t. the likelihood that an object differs significantly from this avg object is small).
  - e.g., an exceptional tiny elephant is not anomalous (not from another distribution), but in the sense o an extreme size value.



- 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 …
- In practice, the techniques are not affected by the source of anomaly

What is an outlier?

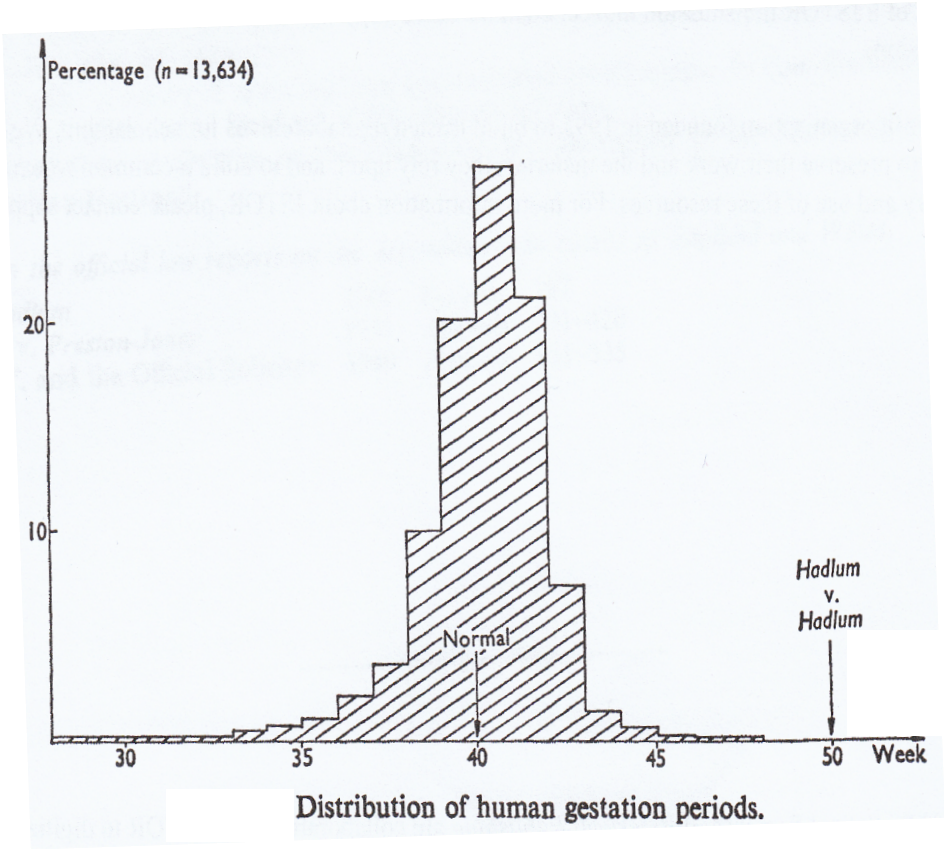
- Definition of Hawkins [Hawkins 1980]:

> *"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]

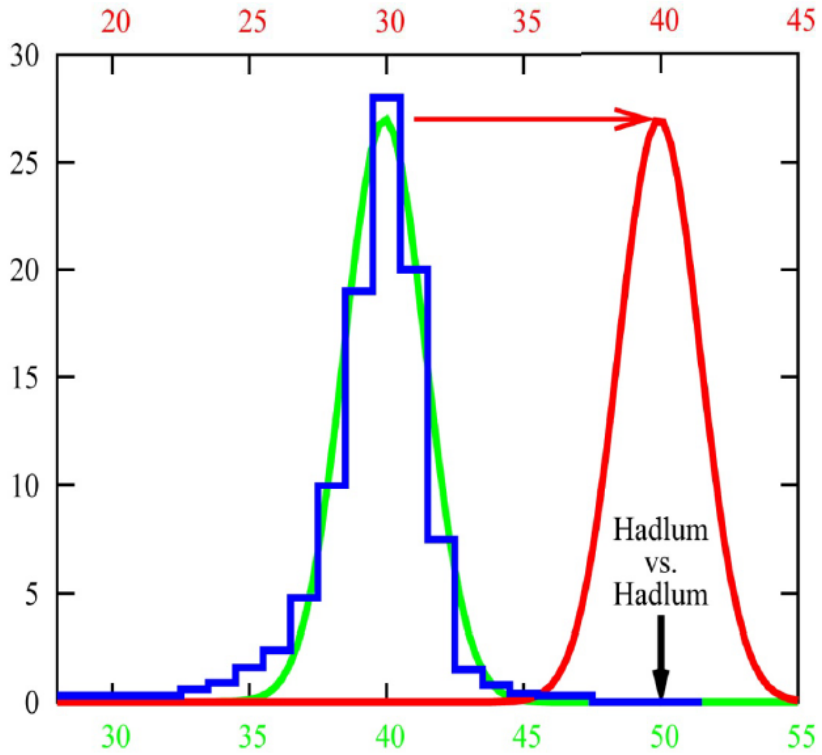
- The birth of a child to Mrs. Hadlum happened 349 days after Mr. Hadlum left for military service.
- Average human gestation period is 280 days (40 weeks).
- Statistically, 349 days is an outlier.



Distribution of human gestation periods.

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

- **blue**: statistical basis (13.634 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
  - but, basic model is univariate, i.e., 1-dimensional
- There is usually more than one generating mechanism/statistical process underlying the "normal" data (comp. EM-Clustering)
  - but, 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
  - but, 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

- Detect all anomalies in the database w.r.t. an anomaly threshold *t*: Given a database *D*, find all the data points $x \in D$ with anomaly scores $f(x)$ greater than some threshold *t*
- Detect top-*n* anomalies in the database: Given a database *D*, find all the data points $x \in D$ having the top-*n* largest anomaly scores *f(x)*
- Compute anomaly score for a query object: 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
- Model-based or Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches
- Things you should know
- Homework/tutorial

Basic application scenarios for outlier detection

Distinction based on the availability of class labels (for anomalies or normal instances)

- Supervised anomaly detection
  - In some applications, training data *with both 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.
  - This is the most common 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 (e.g. DBSCAN, OPTICS, etc.)
  - 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 (E.g. DBSCAN)
  - A set of many abnormal data objects that are similar to each other would be recognized as a cluster rather than as noise/outliers (E.g. OPTICS)
- So, outlier detection is a problem on its own.

Different classification approaches for outlier detection

- Global vs local outlier detection
  - Considers the set of reference objects relative to which each point's "outlierness" is judged
- Labeling vs 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 vs local outlier detection approaches

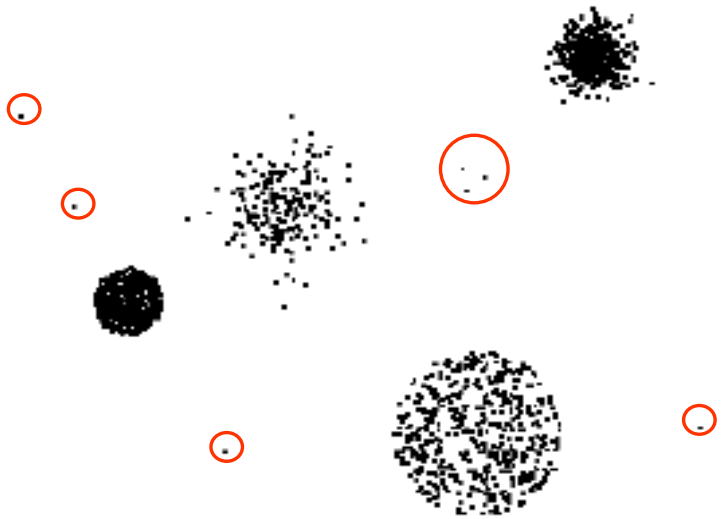
- Considers the resolution of the reference set w.r.t. which the “outlierness” of a particular data object is determined
- Global approaches
  - The reference set contains **all** other data objects
  - Basic assumption: there is only one normal mechanism
  - Basic problem: other outliers are also in 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 (hybrid)
  - The resolution of the reference set varies e.g. from only a single object (local) to the entire database (global) automatically or by a user-defined input parameter

Labeling vs scoring outlier detection approaches

- 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 of 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 be turned into labeling approaches if necessary (e.g. by defining a suitable threshold on the scoring values)

Outlier detection approaches w.r.t. modeling properties

- General steps
  - Build a profile of the "normal" behavior
    - i.e., 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 (or, statistical approaches)
  - Distance-based
  - Density-based
  - Clustering-based



Outline

- Introduction
- Approaches for outlier detection
- Model-based or Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches
- Things you should know

Model-based or Statistical approaches

- A model of the data is built, and objects are evaluated w.r.t. how well they fit the model
- Most approaches are based on
  - building a probability distribution model (by learning the distribution params from the data) (comp. EM-Clustering) 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.

Model-based or Statistical approaches at a glance

- 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 data distributions (mixture models)

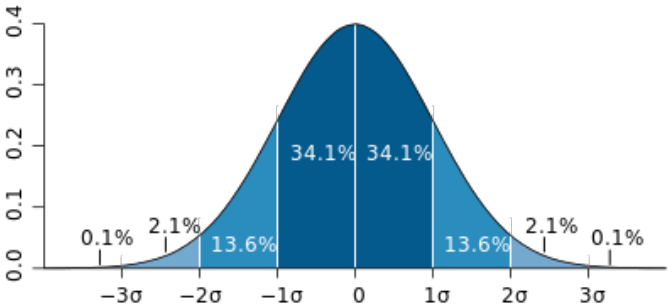
Statistical test example 1

- Example: Gaussian distribution, univariate, 1 model, parametric
- Probability density function of a univariate normal distribution

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- μ is the mean value of all points
- σ^2 is the variance

- Most of the objects are within +-3σ (99.7%)



- There is a little chance that an object will occur at the tails of the distribution
- The distance from the center of N(0,1) can be used as the outlierness test

Statistical test example 2

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

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

- μ is the mean value of all points
- Σ is the covariance matrix from the mean

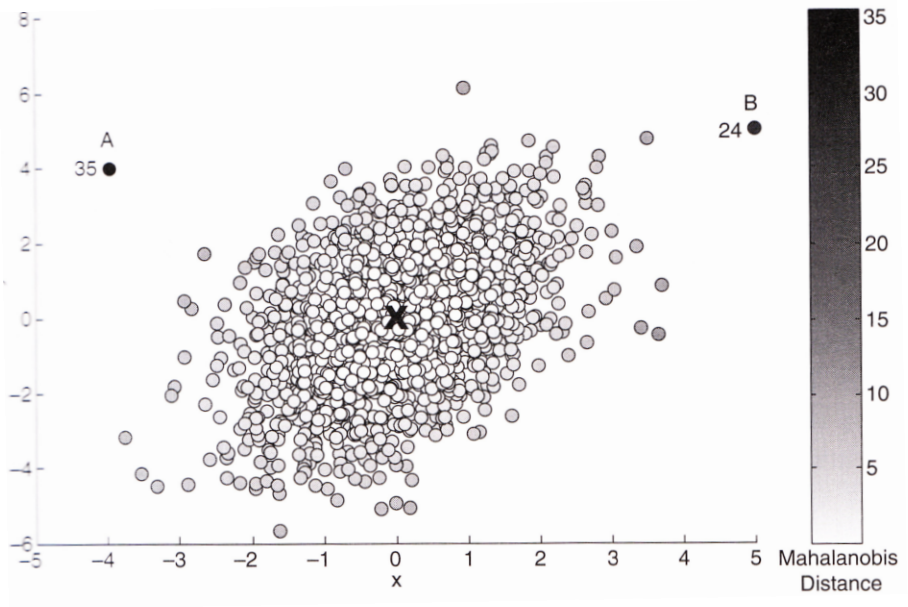
- Score outliers based on Mahalanobis distance of the point *x* to μ

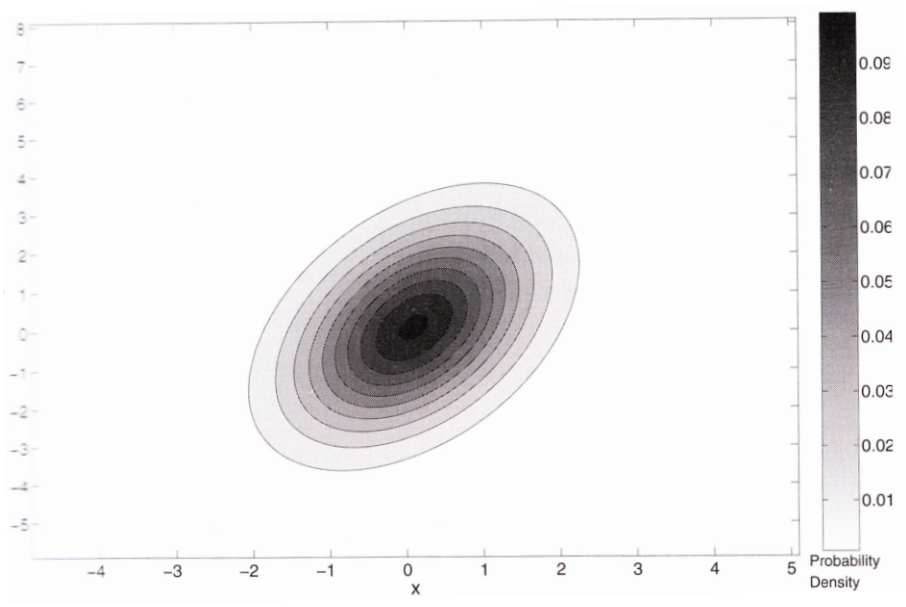
$$MDist(x, \mu) = (x - \mu)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (x - \mu)$$

  - the MDist is directly related to the probability of the point being generated from the distribution
  - For multivariate normally distributed data, the values are approximately chi-square (χ^2) distributed with *d* degrees of freedom
  - Compute the 97.5% quantile Q of the chi-square distribution with *d* degrees of freedom
  - All points *x*, with *MDist*(*x*,μ) > Q are declared as outliers

Example

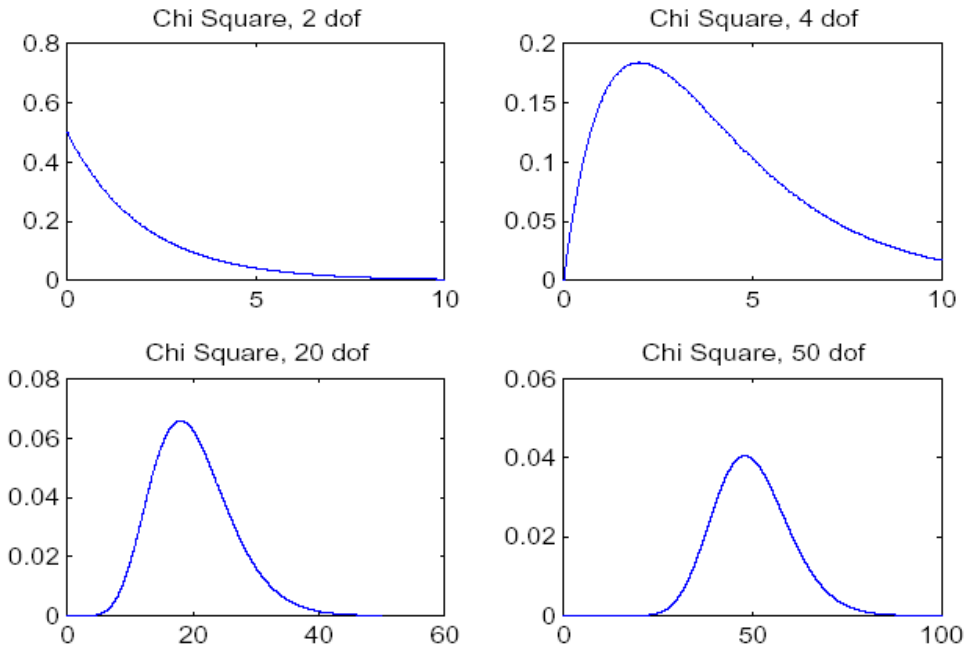
- Visualization (2D) [Tan et al. 2006]





Problems I

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

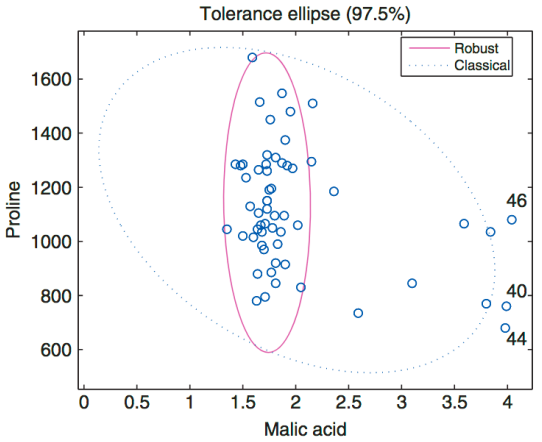


x-axis: observed MDist values

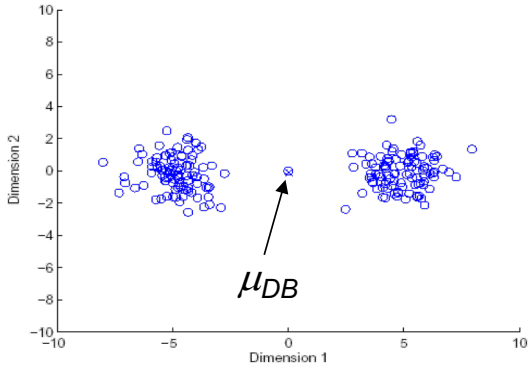
y-axis: frequency of observation

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 (MCD) [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 it can also output a score



Bivariate wine data with classical (MDist) and robust (MCD) tolerance ellipse.



Outline

- Introduction
- Approaches for outlier detection
- Model-based or Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches
- Things you should know

Distance-based approaches

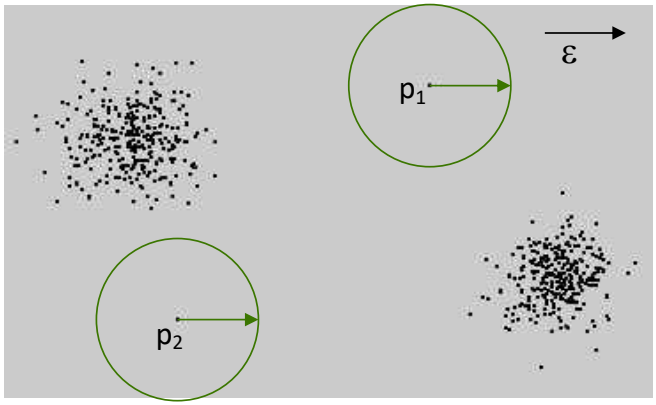
- Data is represented as a vector of features
- General idea: Judge a point based on the distance(s) to its neighbor(s)
- Basic assumption:
  - An object is an anomaly if it is distant from most points
  - Normal data objects have a dense neighborhood
  - Outliers are far apart from their neighbors, i.e., they 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 *k*-th nearest neighbor is greatest
  - The top-*n* data points whose average distance to the *k*-th nearest neighbors is greatest

Basic model [Knorr and Ng 1997] 1/2

- *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



Basic model [Knorr and Ng 1997] 2/2

Algorithms for efficient computation

- Index-based [Knorr and Ng 1998]
  - Compute distance range join using spatial index structure
  - Exclude point from further consideration if its ε-neighborhood contains more than $Card(DB)*\pi$ points
- Nested-loop based [Knorr and Ng 1998]
  - Divide buffer in two parts
  - Use second part to scan/compare all points with the points from the first part
- Grid-based [Knorr and Ng 1998]
  - Build a grid such that any two points from the same grid cell have a distance of at most ε to each other
  - Points need only compared with points from neighboring cells

k^{th} nearest neighbor-based 1/3

- The outlier score of an object is given by its distance to its *k*-nearest neighbor.
  - Lowest outlier score 0.

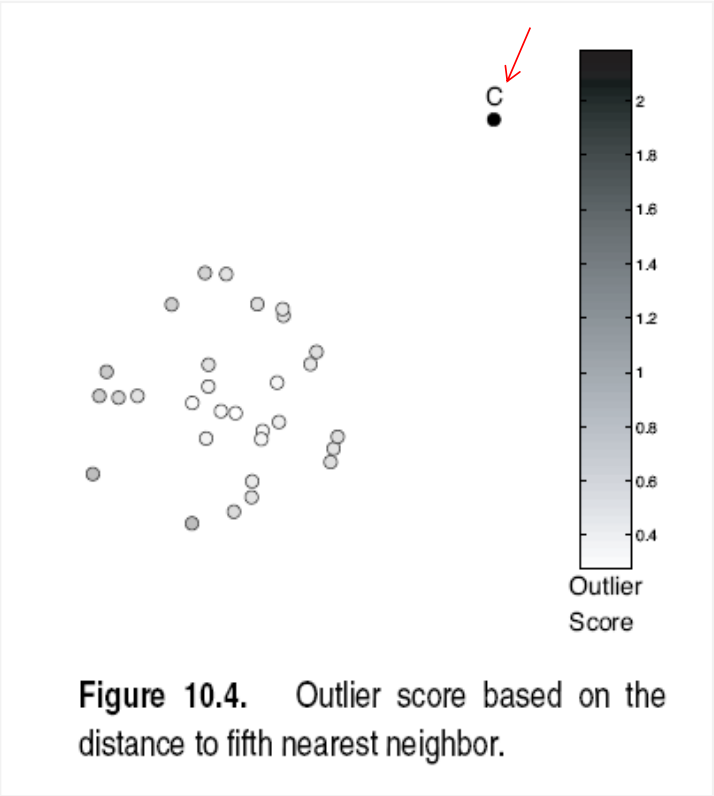


Figure 10.4. Outlier score based on the distance to fifth nearest neighbor.

k=5

k^{th} nearest neighbor-based 2/3

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

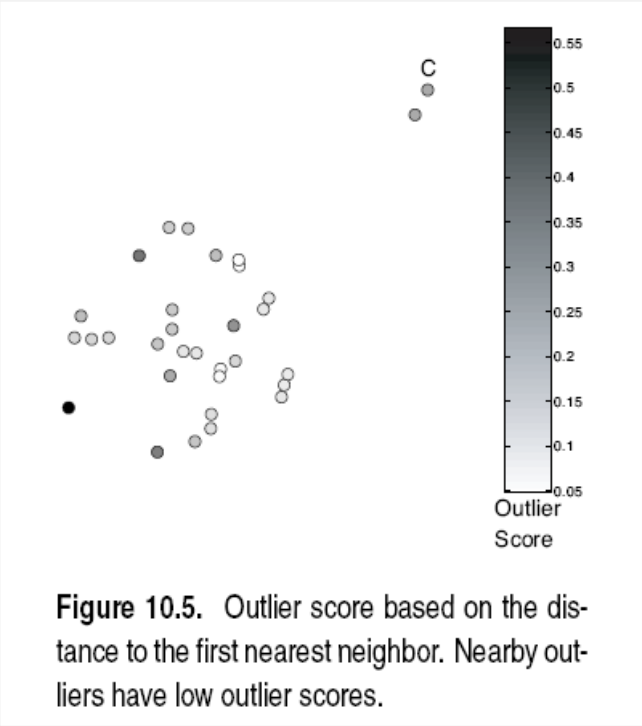


Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

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

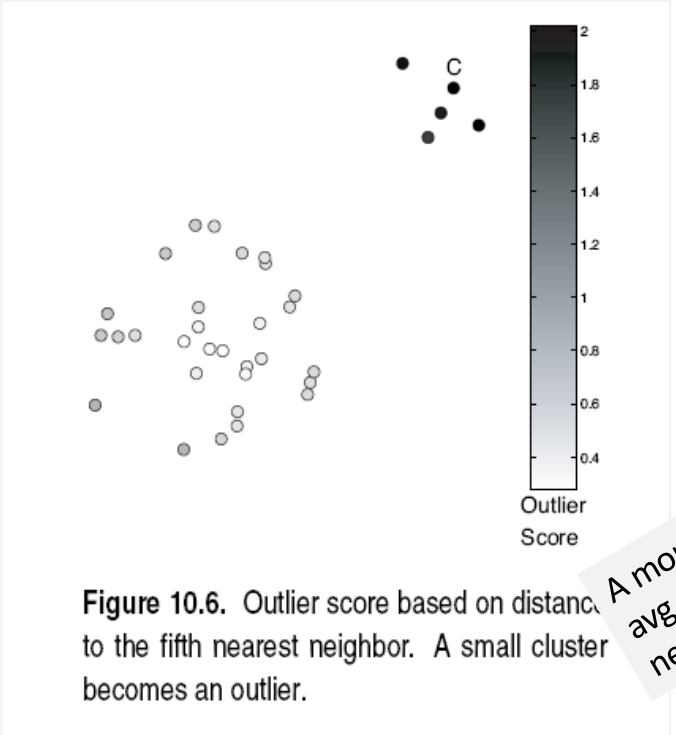


Figure 10.6. Outlier score based on distanc to the fifth nearest neighbor. A small cluster becomes an outlier.

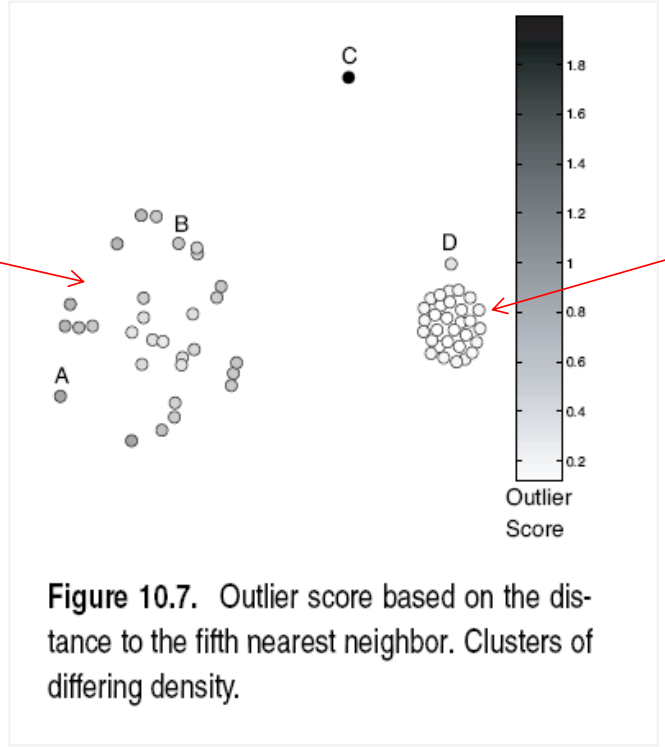
A more robust approach: avg distance to the first k nearest neighbors

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

k^{th} nearest neighbor-based 3/3

- 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 spaces, data is sparse and the 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
- Model-based or Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches
- Things you should know

Density-based approaches 1/2

- 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' density is computed as an outlier score
  - Approaches essentially differ on how they 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 from the density around its neighbors
- Closely related to distance-based methods, since density is usually defined in terms of proximity.

Density-based approaches 2/2

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
    - Too small *d* → many normal points will be considered outliers
    - Too larger *d* → many outlier points will be considered normal
- A global definition of density is problematic (recall our discussion on the clustering lectures)
  - Fail when data contains regions of different densities
  - Solution: use a notion of density that is relative to the neighborhood of the object

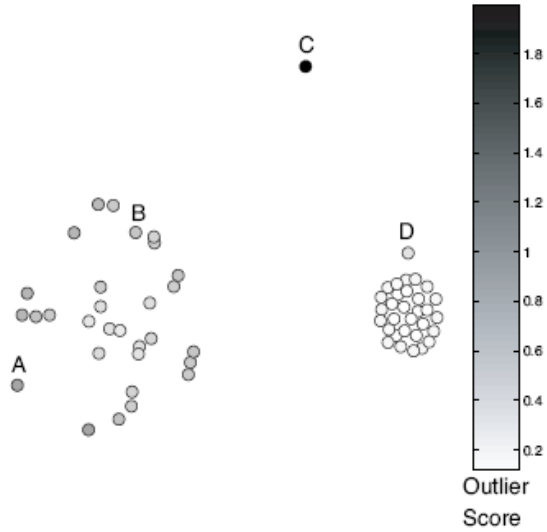
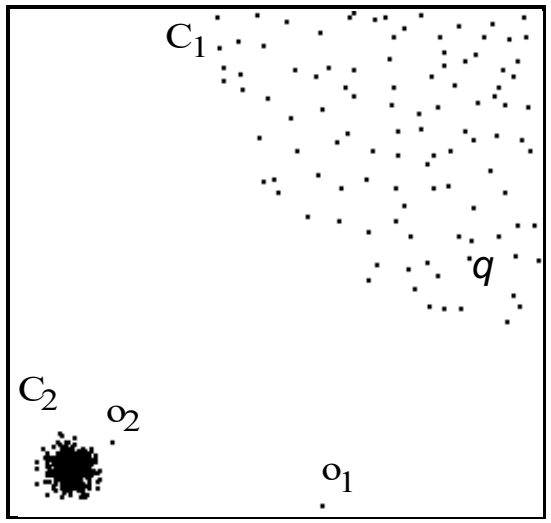


Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

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

LOF(Local Outlier Factor) 1/4

- 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*(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

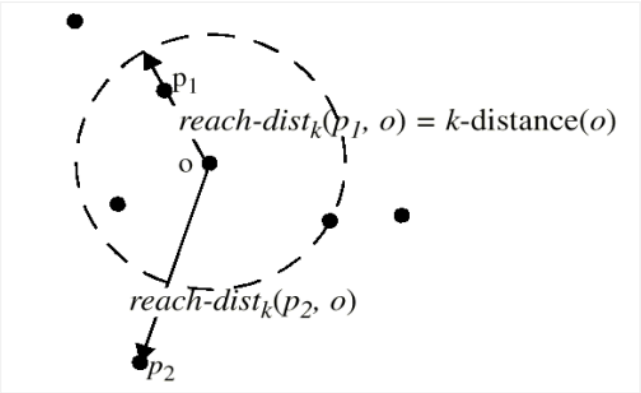


LOF model 2/4

- Reachability distance of an object *p* w.r.t. an object *o*

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

  - This is not symmetric!



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

$$lrd_k(p) = 1/\left(\frac{\sum_{o\in kNN(p)} reach\text{-}dist_k(p,o)}{|kNN(p)|}\right) \quad \Rightarrow \frac{1}{lrd_k(p)} = \frac{\sum_{o\in kNN(p)} reach\text{-}dist_k(p,o)}{|kNN(p)|}$$

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

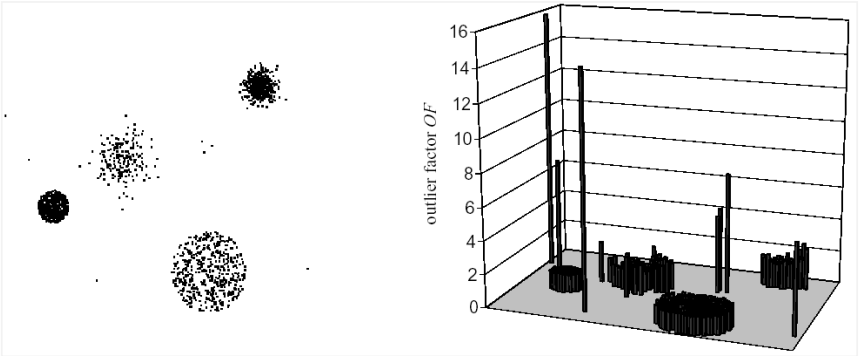
$$LOF_k(p) = \frac{1}{|kNN(p)|} * \sum_{o\in kNN(p)} \frac{lrd_k(o)}{lrd_k(p)}$$

relative density

average

LOF 3/4

- 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
  - Implements a local approach (resolution depends on the user's choice for *k*)
  - Outputs a scoring (assigns a LOF value to each point)

LOF example 4/4

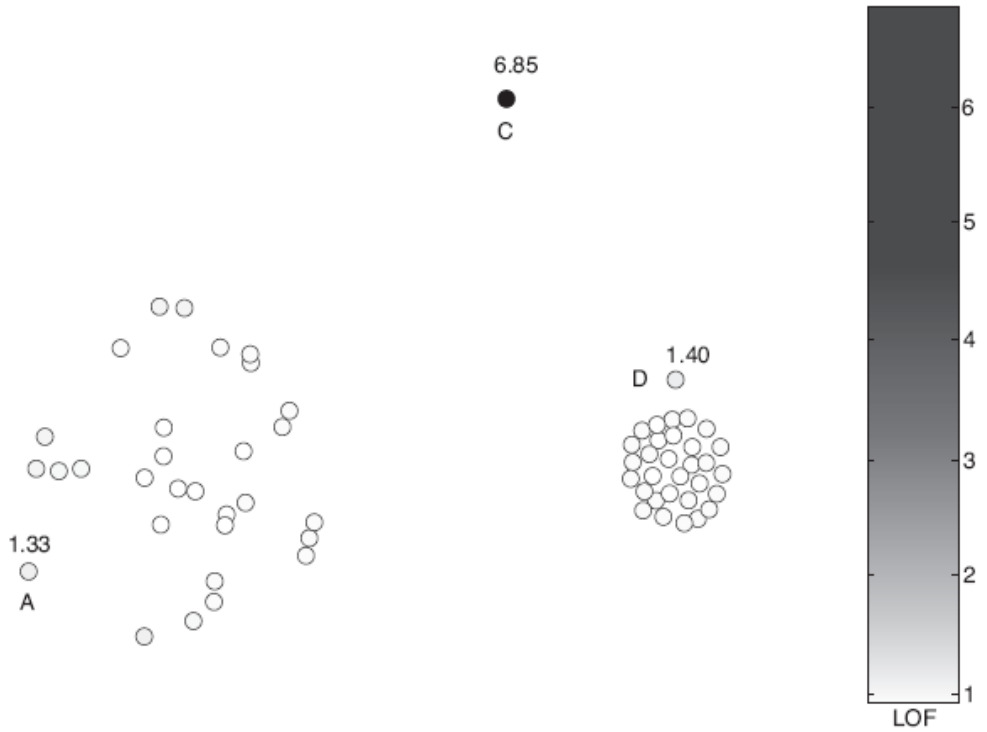


Figure 10.8. Relative density (LOF) outlier scores for two-dimensional points of Figure 10.7.

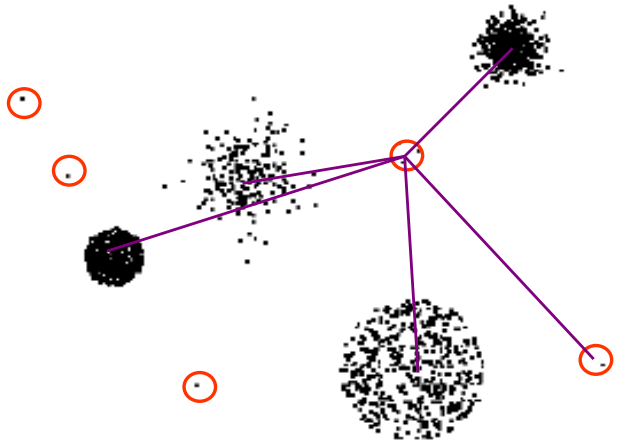
Outline

- Introduction
- Approaches for outlier detection
- Model-based or Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches
- Things you should know

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

- Methods 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.

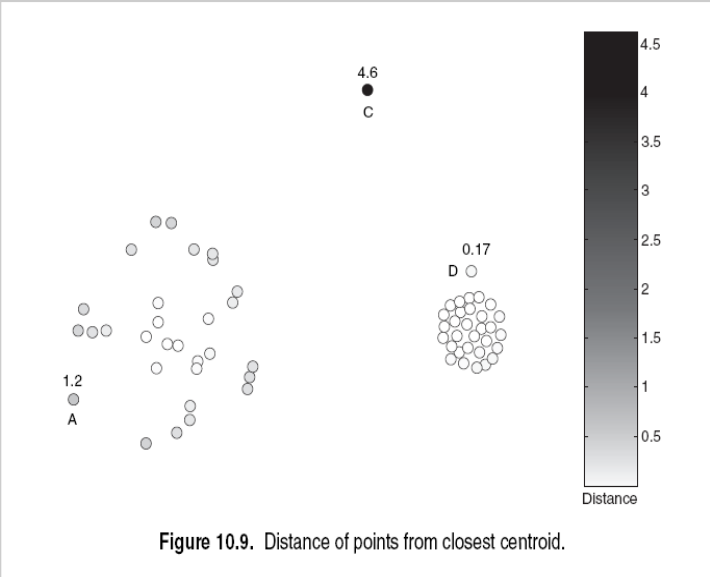


Figure 10.9. Distance of points from closest centroid.

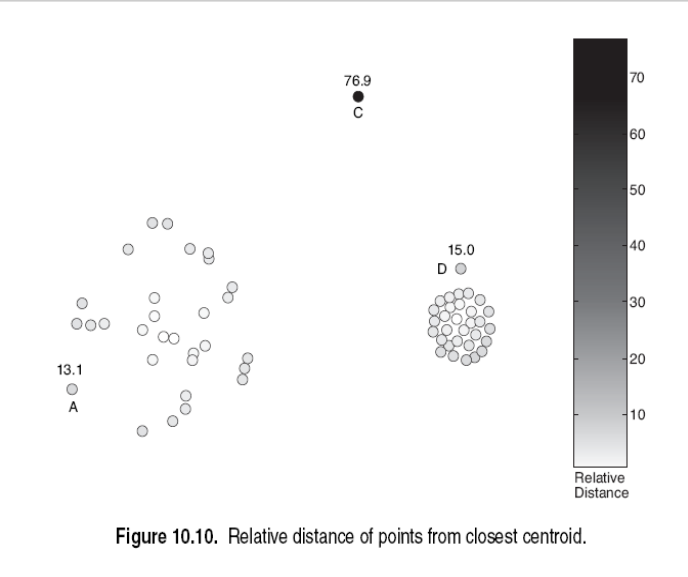


Figure 10.10. Relative distance of points from closest centroid.

Outlier evaluation

- If there are class labels
  - Similar to classifier evaluation, but outlier class is typically smaller than the normal class
  - Measures such as precision and recall are more appropriate than e.g., accuracy or error rate
- In the absence of class labels
  - More difficult
  - For model-based approaches, one could check model improvement after outlier removal

Outline

- Introduction
- Approaches for outlier detection
- Model-based or Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches
- Things you should know

Things you should know

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