Leibniz Universität Hannover

Fakultät für Elektrotechnik und Informatik
Institut für Verteilte Systeme
AG Intelligente Systeme - Data Mining group

Data Mining I

Summer semester 2019

Lecture 12.b: Evaluation & Outlier Detection

Lectures: Prof. Dr. Eirini Ntoutsi

TAs: Tai Le Quy, Vasileios Iosifidis, Maximilian Idahl, Shaheer Asghar

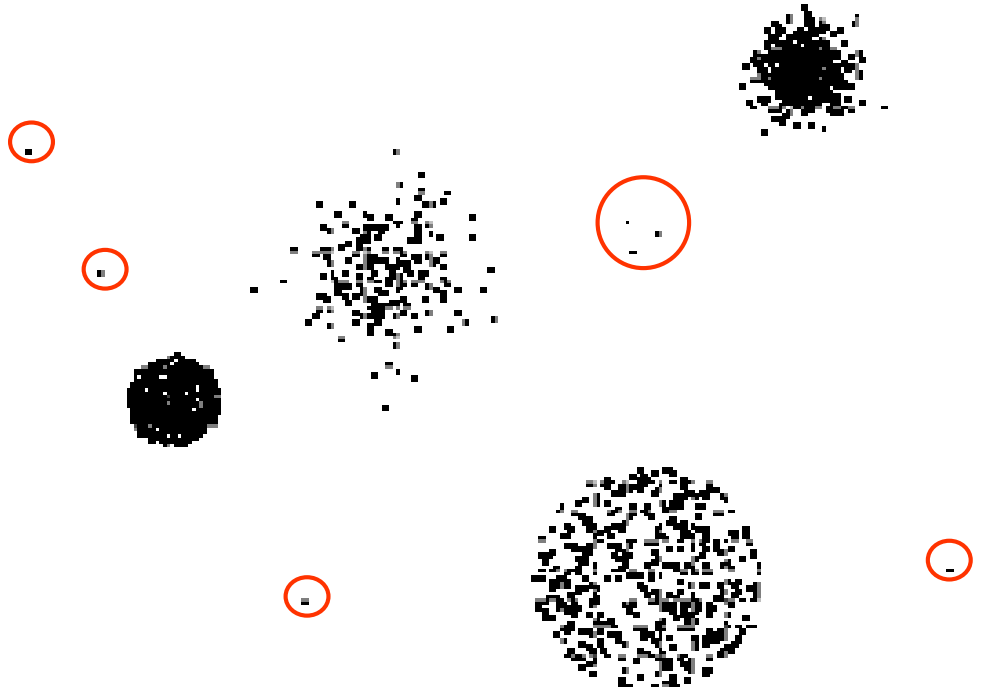
Outline

- 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

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)
  - Post-processing task (to improve pattern quality)



Applications 1/2

- Fraud detection, e.g., credit cards
  - 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 2/2

- 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
- Intrusion detection, e.g., in networks
  - Attack connections different than normal connections
- 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

An example from sports

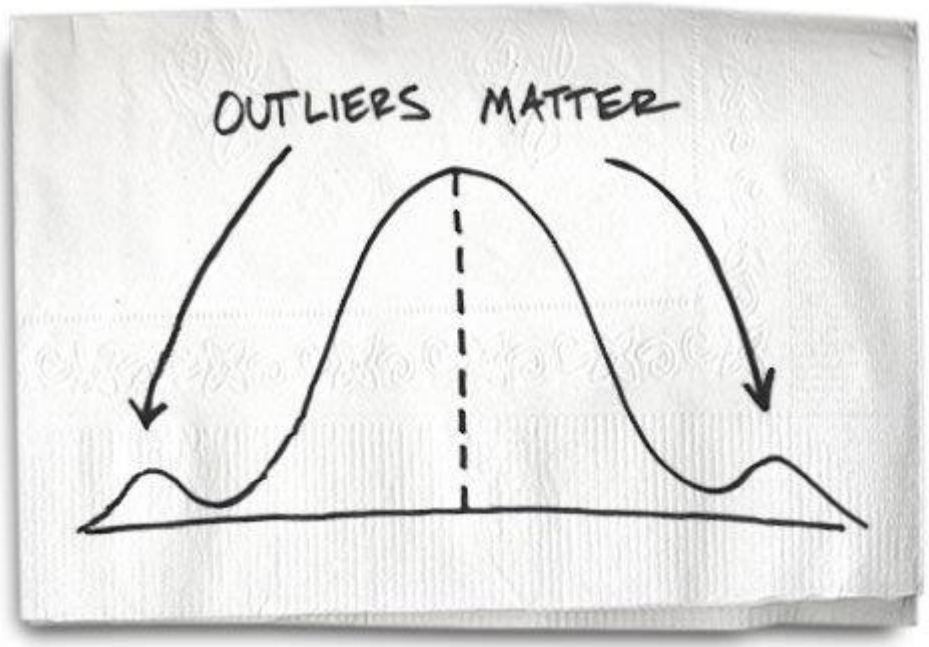
- Analysis of the SAT.1-Ran-Soccer-Database (Season 1998/99)
  - 375 players
  - Primary attributes: Name, #games, #goals, playing position (goalkeeper, defense, midfield, offense),
  - Derived attribute: Goals per game
  - Outlier analysis (playing position, #games, #goals)
- Result: Top 5 outliers

Rank	Name	# games	#goals	position	Explanation
1	Michael Preetz	34	23	Offense	Top scorer overall
2	Michael Schjönberg	15	6	Defense	Top scoring defense player
3	Hans-Jörg Butt	34	7	Goalkeeper	Goalkeeper with the most goals
4	Ulf Kirsten	31	19	Offense	2nd scorer overall
5	Giovanne Elber	21	13	Offense	High #goals/per game

Being an “outlier” is not necessarily a negative term.

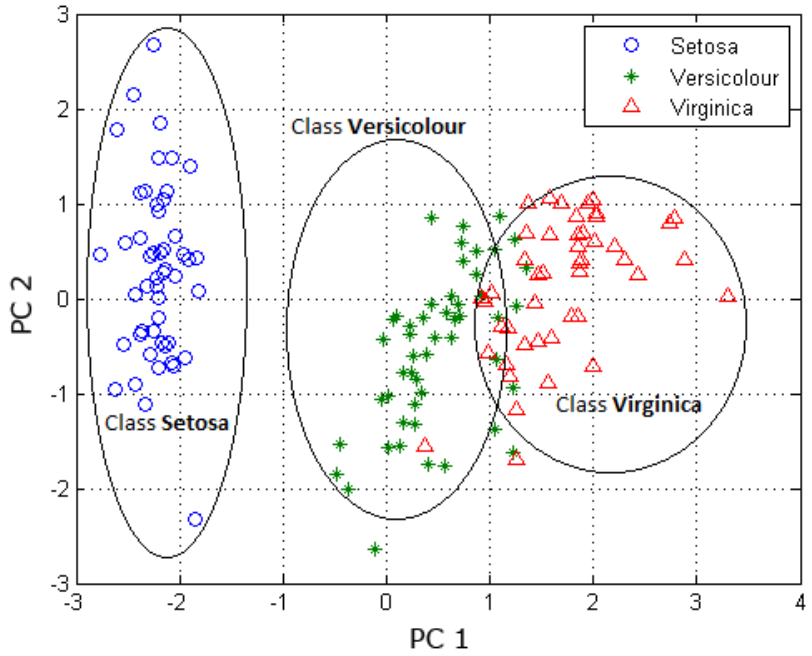
Outliers matter

One person‘s noise could be another person‘s signal.



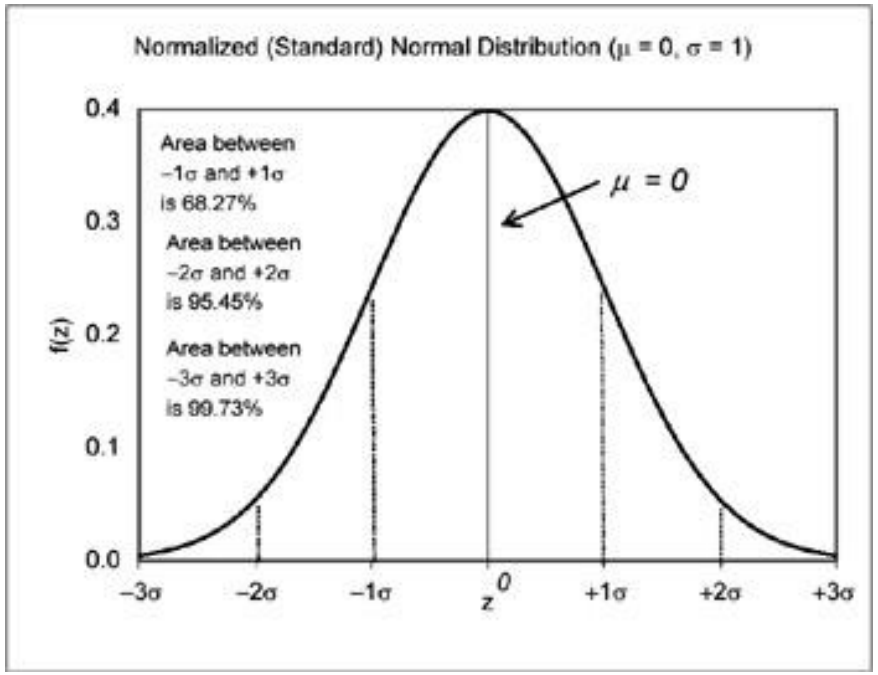
Causes of anomaly 1/3

- Data from different classes
  - An object might be different from other objects because it is from another class (it comes from another distribution).
    - 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.



Causes of anomaly 2/3

- Natural variation
  - Many datasets can be modeled by statistical distributions
    - e.g., Gaussian distribution (most objects are near the center, the likelihood that an object differs significantly from the avg object is small).
      - In that case an exceptional tall person is not anomalous (not from another distribution), but it has an extreme value



Causes of anomaly 3/3

- 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 can be used for all those 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

Variants of outlier detection problems

- Compute anomaly score for a query object
  - Given a database D, containing mostly normal (but unlabeled) data points, and a query point $x \in D$, compute the anomaly score of x with in D
- 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 $f(x)$

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 ML approaches 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*
- 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.
- Semi-supervised anomaly detection
  - In some applications, only training data for the normal class(es) (or only the abnormal class(es)) are provided

Outlier detection and clustering

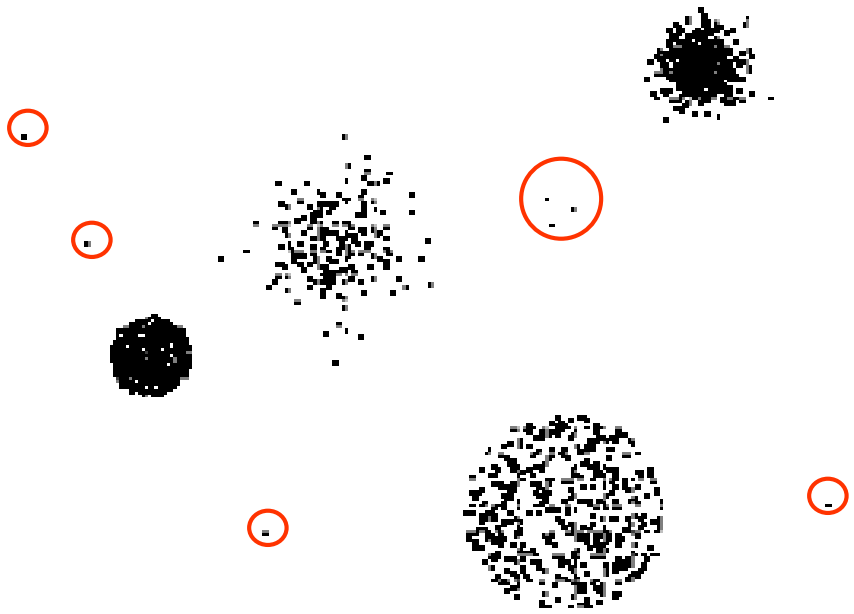
- Many clustering algorithms do not assign all points to clusters but account for noise objects
- Naïve approach: Look for outliers by applying one of those algorithms and retrieve the noise set
  - In this case, outliers are just a side product of the clustering algorithms

- 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 detection is a problem on its own.

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
  1. Model-based (or, statistical approaches)
  2. Distance-based
  3. Density-based
  4. 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
- Homework/tutorial

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

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

- Most approaches here follow the following 2 step approach:
  1. building a probability distribution model (by learning the distribution params from the data) and
  2. considering how likely objects are under that model

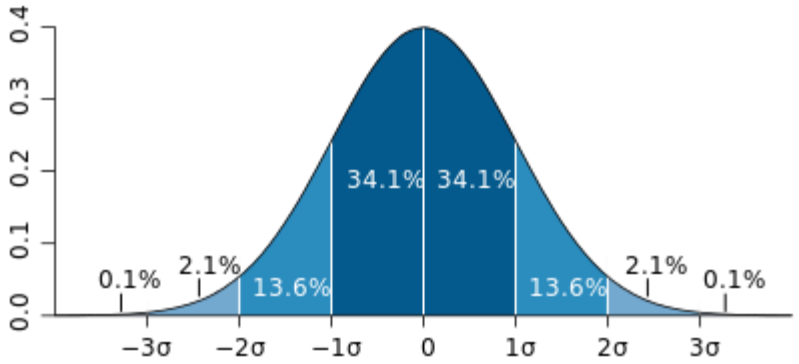
Statistical test example

- 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

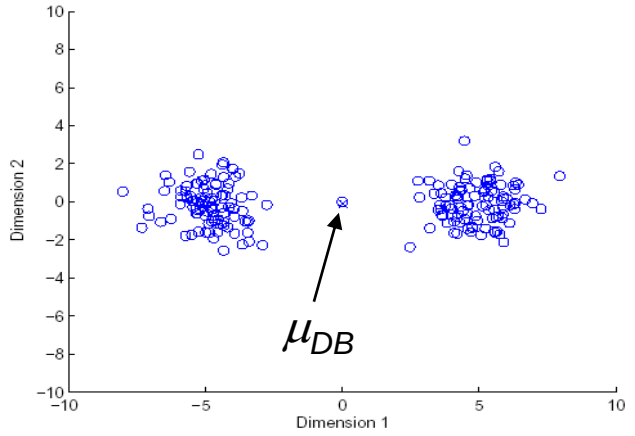
- Most of the objects are within +-3σ (99.7%)



- There is a little chance that an object will occur at the tails of the distribution
- Score outliers based on the distance of the points to the mean μ

Model-based approaches overview

- Robustness
  - Mean and standard deviation are very sensitive to outliers
  - These values are computed for the complete data set (including potential outliers)
- Discussion
  - Data distribution is fixed
  - Low flexibility (no mixture model)
  - Global method
  - Outputs a label but it can also output a score



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

Distance-based approaches

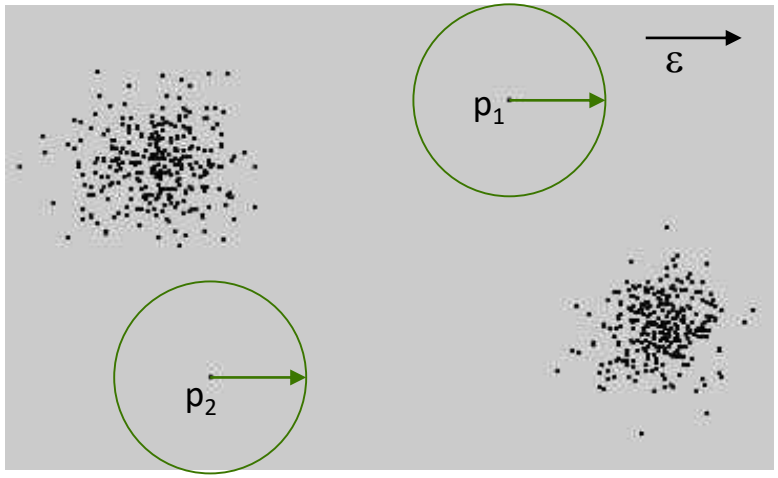
- General idea: Judge a point based on the distance(s) to its neighbor(s)
- Basic assumption:
  - An object is an anomaly if it is distant from most points
  - Normal data objects have a dense neighborhood
  - Outliers are far apart from their neighbors, i.e., they have a less dense neighborhood
- More 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 of what is an outlier
  - 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

Distance-based approaches: 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



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

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

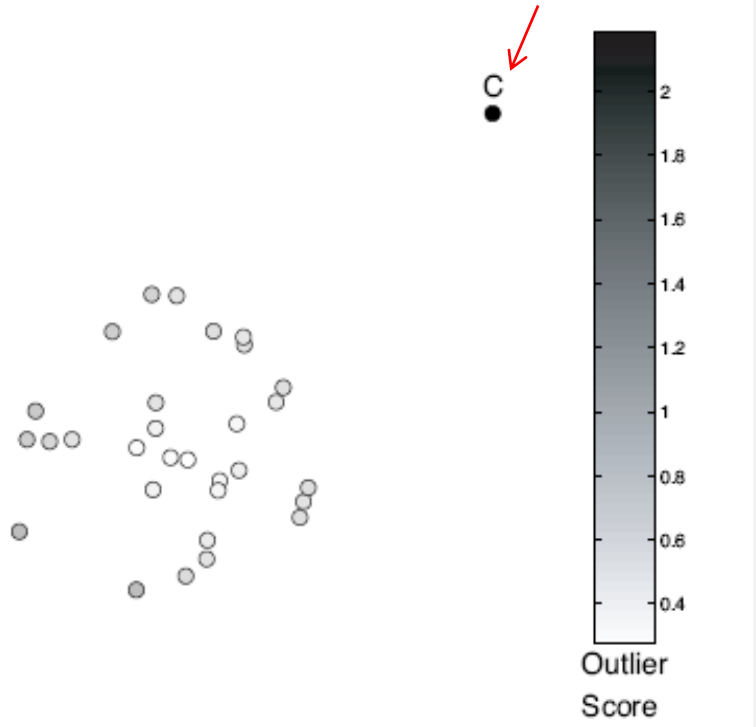


Figure 10.4. Outlier score based on the distance to fifth nearest neighbor.

k=5

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

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

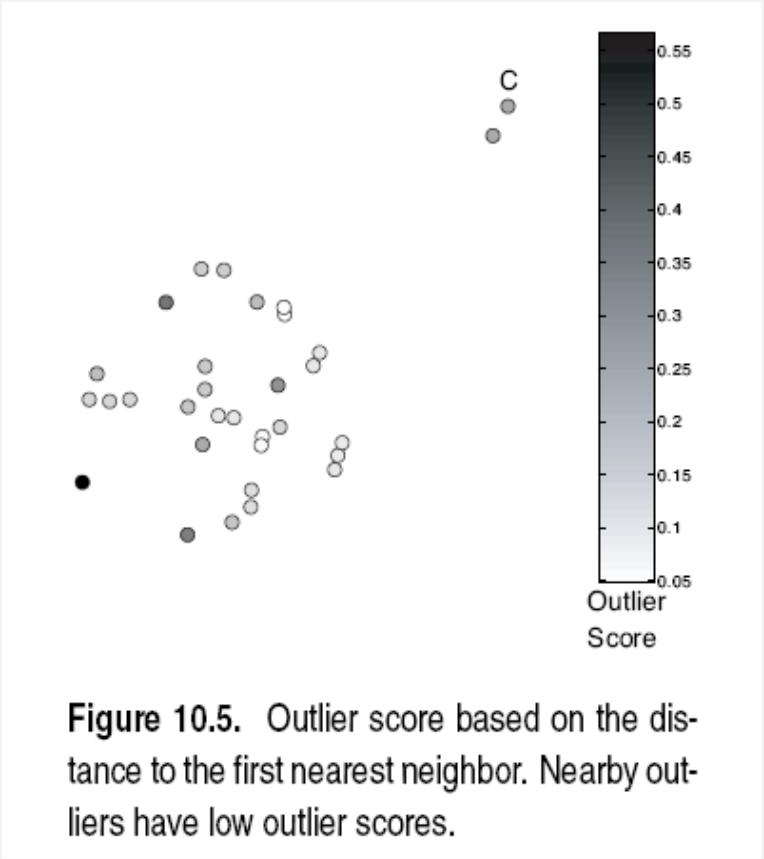


Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

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

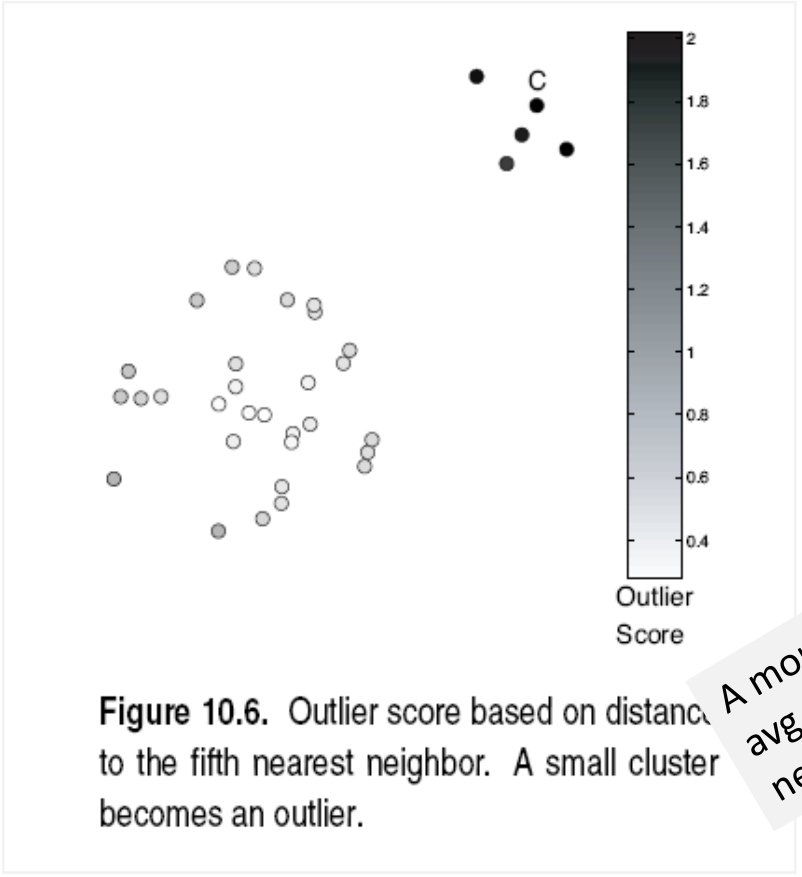


Figure 10.6. Outlier score based on distanc to the fifth nearest neighbor. A small cluster becomes an outlier.

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

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

Distance-based approaches: k^{th} nearest neighbor-based 3/3

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

Loose cluster

Dense cluster

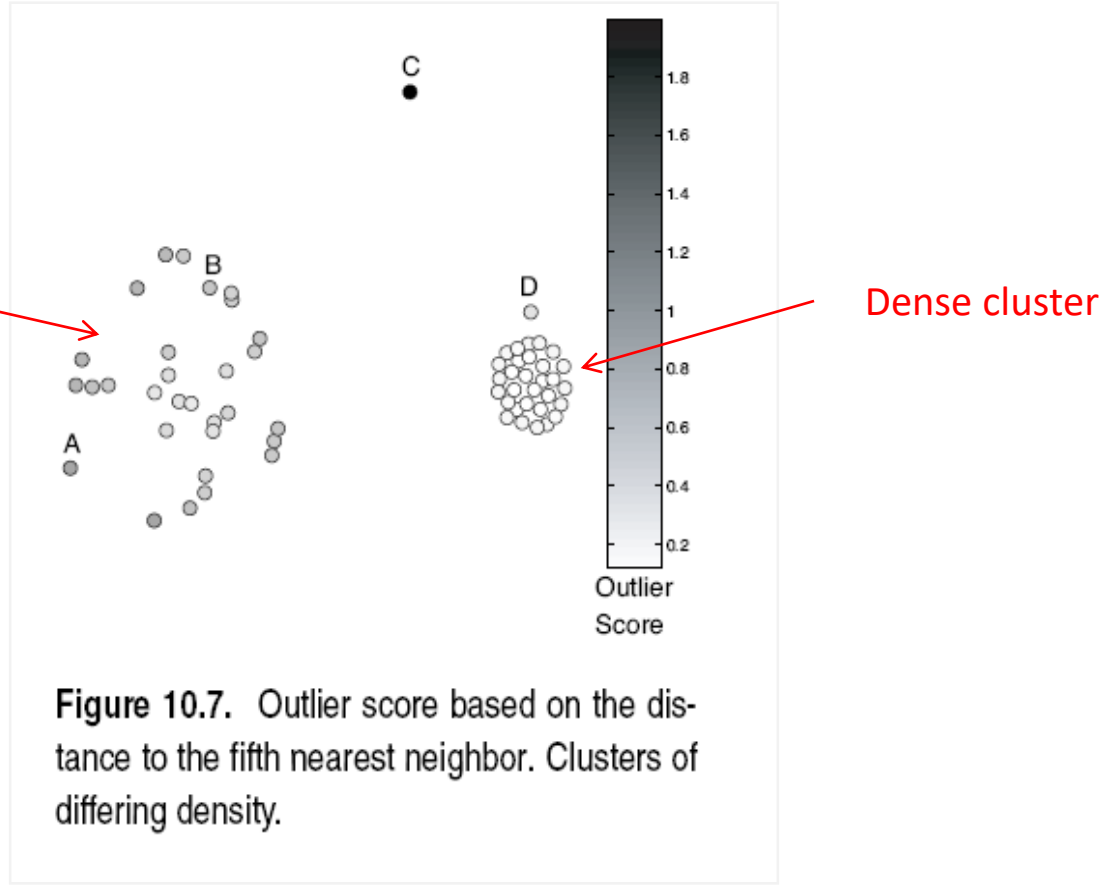


Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

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 to tackle this issue.

Outline

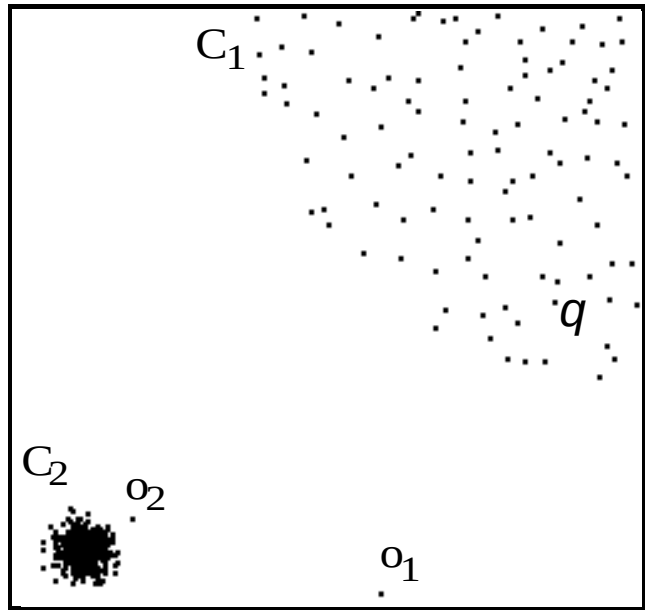
- Introduction
- Approaches for outlier detection
- Model-based or Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches
- Things you should know
- Homework/tutorial

Density-based approaches

- 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.

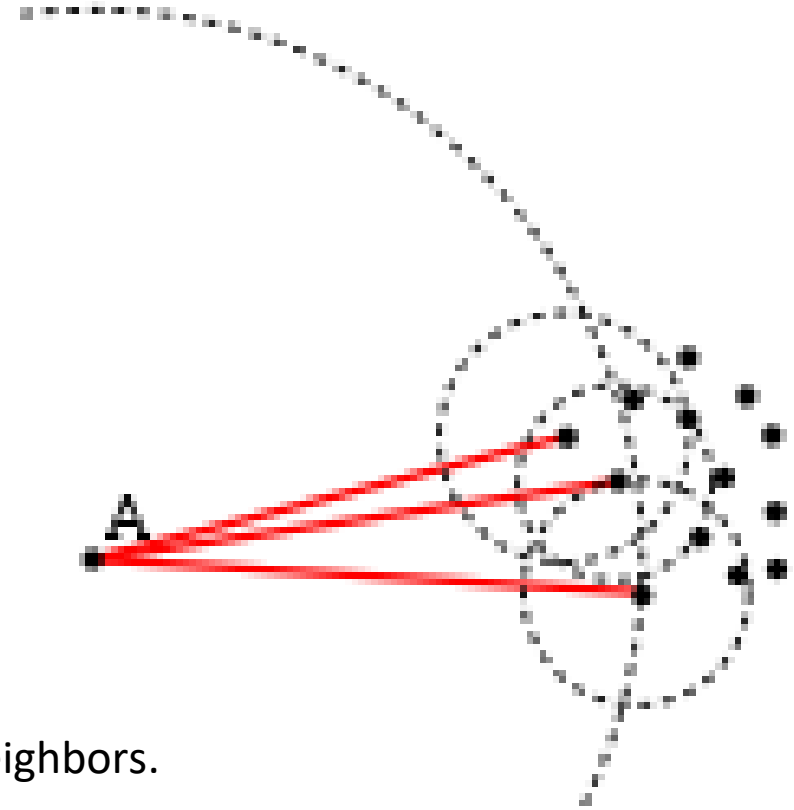
LOF(Local Outlier Factor) 1/6

- 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?
- Existing solutions
  - *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(Local Outlier Factor) 2/6

- Basic idea of LOF: comparing the local density of a point with the densities of its neighbors.



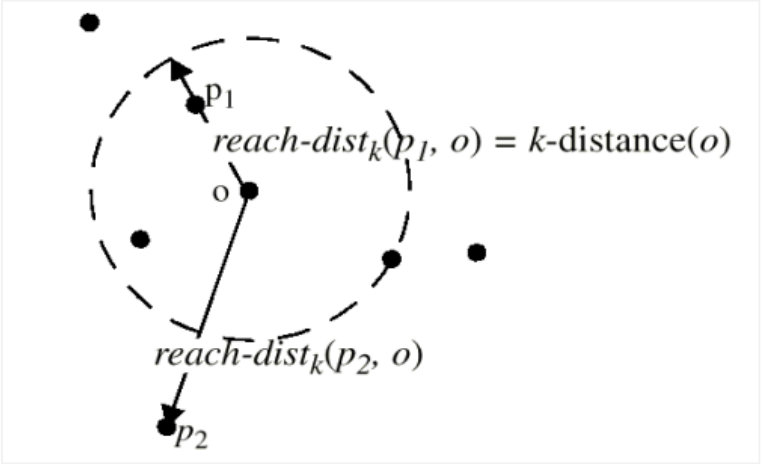
  - In our example, A has a much lower density than its neighbors.

LOF(Local Outlier Factor) 3/6

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

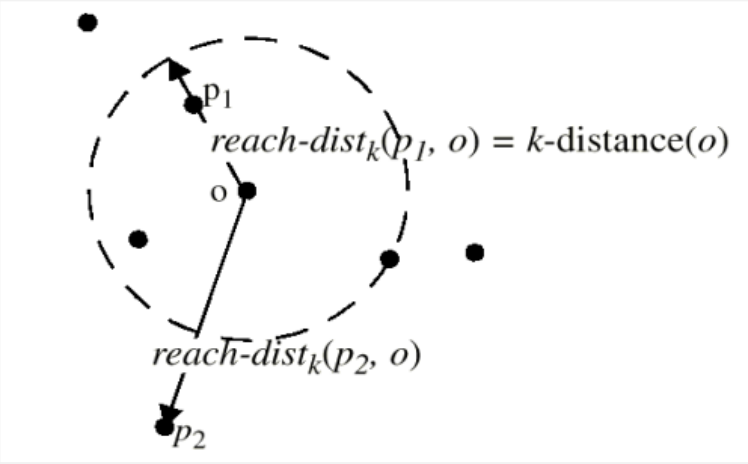
$$reach\text{–}dist_k(p,o) = \max\{k\text{–distance}(o), dist(p,o)\}$$

- *k*-distance(o) is the distance of o to its *k*-th nearest neighbor
- dist(p,o) is the actual distance between *p* and *o*



- Note that the set of the *k* nearest neighbors includes all objects at this distance, which can in the case of a "tie" be more than *k* objects
- This is not symmetric!

LOF(Local Outlier Factor) 4/6



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

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

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

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

$$\Rightarrow \frac{1}{lrd_k(p)} = \frac{\sum_{o \in kNN(p)} reach\text{–}dist_k(p,o)}{|kNN(p)|}$$

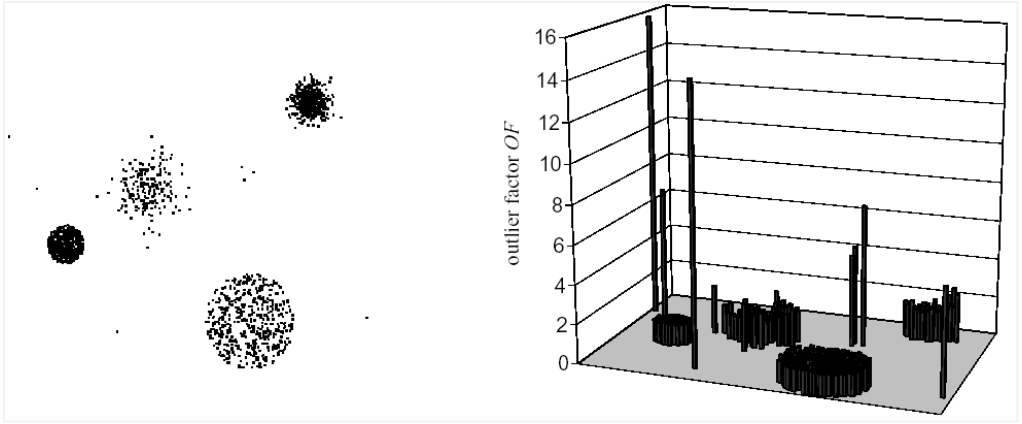
- Local outlier factor (LOF) of point *p*
  - Average ratio of *lrd*s of neighbors of *p* and *lrd* of *p*

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

relative density

LOF(Local Outlier Factor) 5/6

- 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(Local Outlier Factor) 6/6

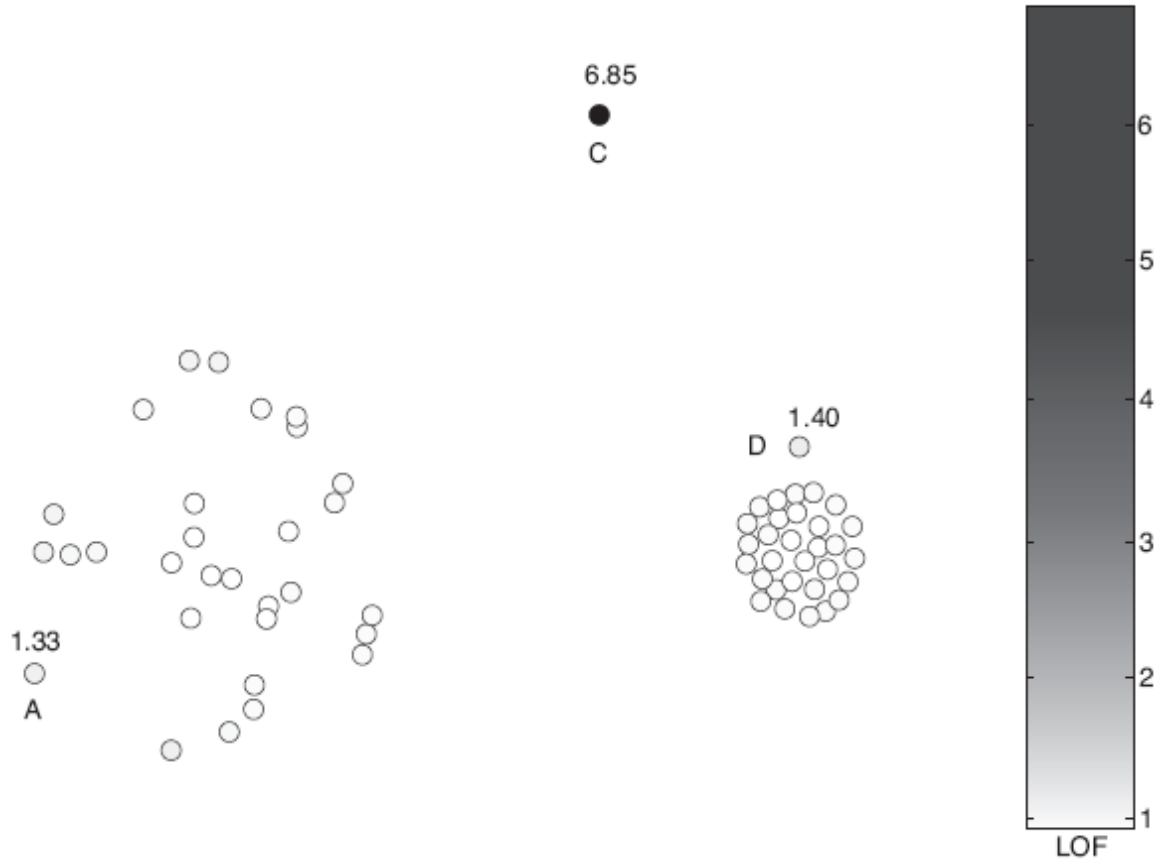


Figure 10.8. Relative density (LOF) outlier scores for two-dimensional points of Figure 10.7.

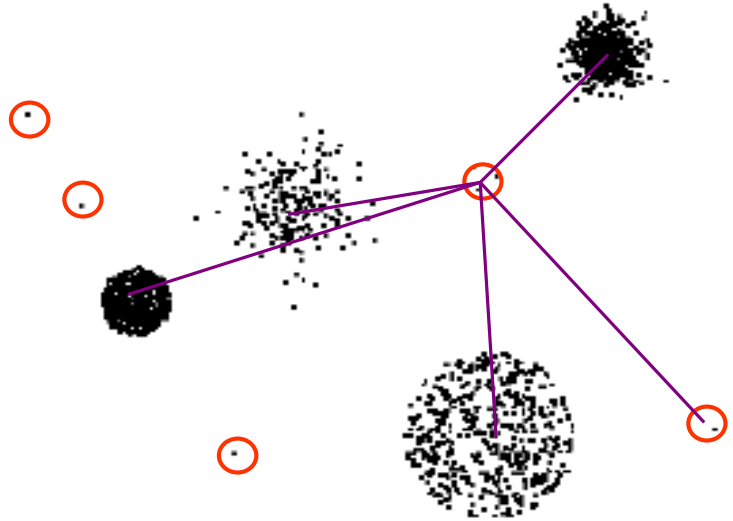
Outline

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

- 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.

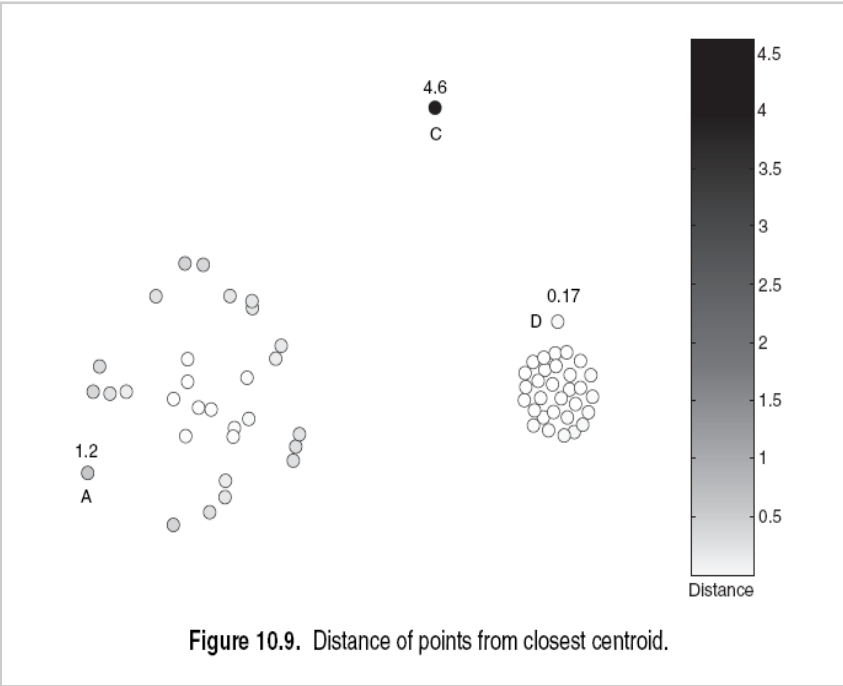


Figure 10.9. Distance of points from closest centroid.

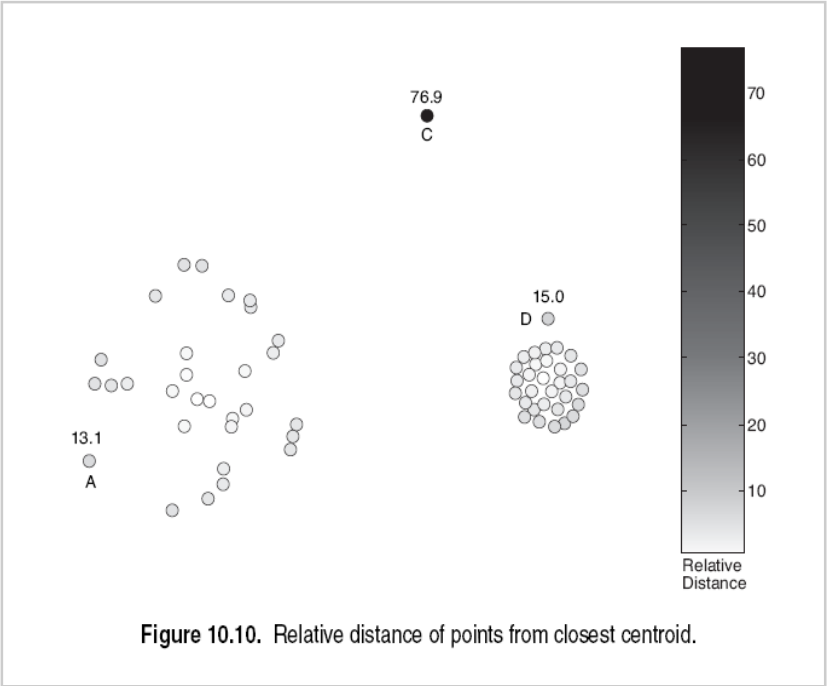


Figure 10.10. Relative distance of points from closest centroid.

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

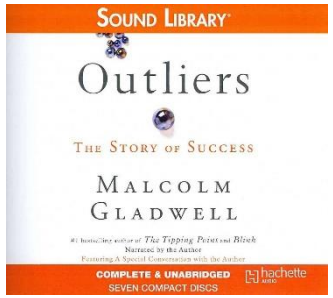
- Introduction
- Approaches for outlier detection
- Model-based or Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches
- Things you should know
- Homework/tutorial

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

Homework

- Homework
  - Use the Elki data mining tool to experiment with different outlier detection approaches http://elki.dbs.ifi.lmu.de/
    - Experiment with the parameters
    - Interpret the charts
- Readings:
  - Tan P.-N., Steinbach M., Kumar V book, Chapter 10.
  - “Outlier Detection Techniques” tutorial by H.-P. Kriegel, P. Kröger, A. Zimek at KDD’10.
- An interesting book on outliers



Exam

- Wednesday August 28, 2019 from 08:30 – 11:00
- Based on the lectures and tutorials
  - Mainly algorithmic exercises, e.g., Apriori, kMeans
    - Combined with some theoretical part
  - True/false exercise with different questions covering the material

- You are allowed to bring your own DIN A4 sheet of paper with hand written notes on both sides .
  - your own=> no copies are allowed; hand-written

- Q&A lecture & presentations of projects
  - E.g., Friday 16.8.2019?

Thank you

&

See you in the exam

Acknowledgement

- The slides are based on
  - KDD I lecture at LMU Munich (Johannes Aßfalg, Christian Böhm, Karsten Borgwardt, Martin Ester, Eshref Januzaj, Karin Kailing, Peer Kröger, Eirini Ntoutsi, Jörg Sander, Matthias Schubert, Arthur Zimek, Andreas Züfle)
  - Thank you to all TAs contributing to their improvement, namely Vasileios Iosifidis, Damianos Melidis, Tai Le Quy, Han Tran.