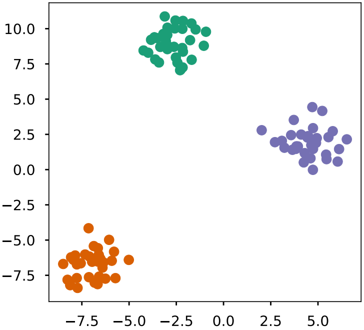
Ludwig-Maximilians-Universität München
Lehrstuhl für Datenbanksysteme und Data Mining
Prof. Dr. Thomas Seidl

Knowledge Discovery and Data Mining 1

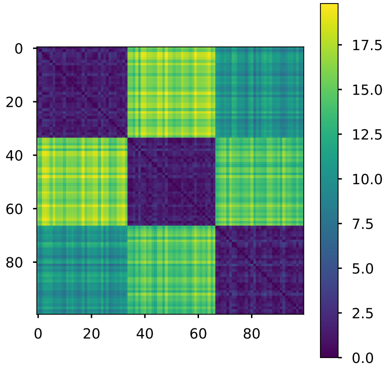
(Data Mining Algorithms 1)

Winter Semester 2019/20

Evaluating the Distance Matrix



dataset
(well separated)



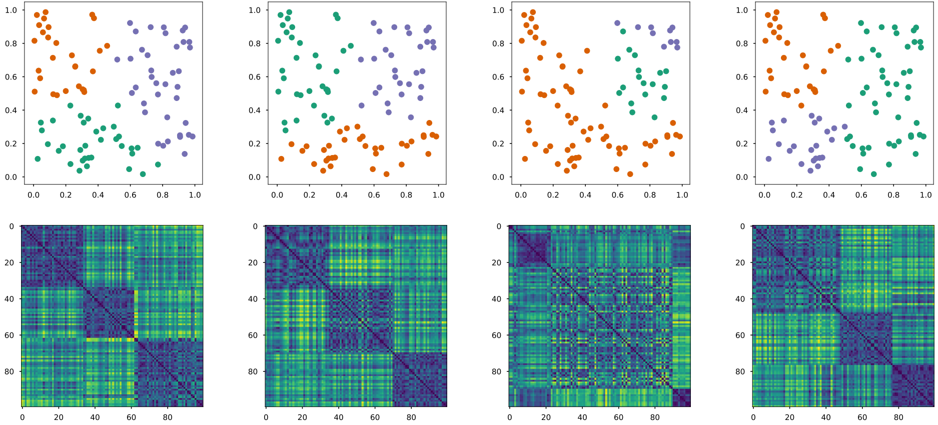
Distance matrix
(sorted by *k*-means cluster label)

after: Tan, Steinbach, Kumar: Introduction to Data Mining (Pearson, 2006)

Evaluating the Distance Matrix

Distance matrices differ for different clustering approaches (here on random data)

k-means EM DBSCAN Complete Link

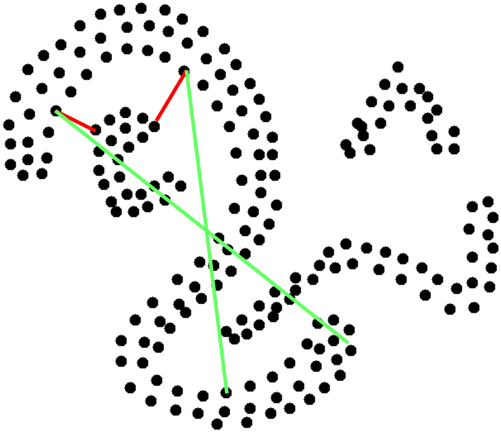


after: Tan, Steinbach, Kumar: Introduction to Data Mining (Pearson, 2006)

Cohesion and Separation

Problem

Suitable for convex cluster, but not for stretched clusters (cf. silhouette coefficient).



Ambiguity of Clusterings

- Clustering according to: Color of shirt, direction of view, glasses, . . .

Ambiguity of Clusterings

- Clustering according to: Color of shirt, direction of view, glasses, ...

Ambiguity of Clusterings

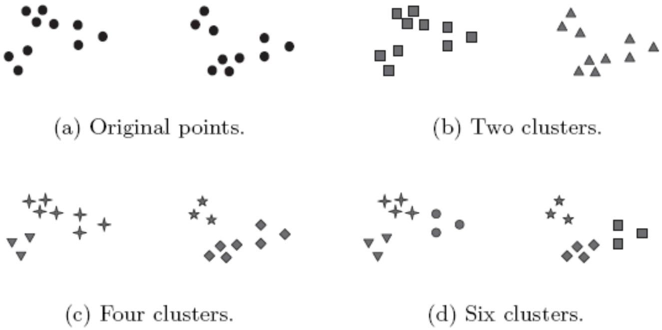


Figure 8.1. Different ways of clustering the same set of points.

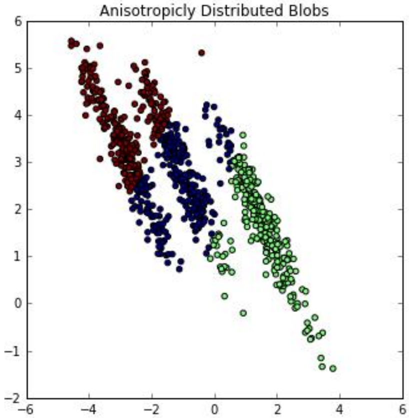
from: Tan, Steinbach, Kumar: Introduction to Data Mining (Pearson, 2006)

Ambiguity of Clusterings

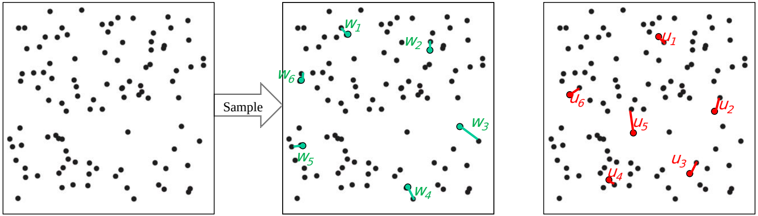
"Philosophical" Problem

"What is a correct clustering?"

- Most approaches find clusters in every dataset, even in uniformly distributed objects
- Are there clusters?
  - Apply clustering algorithm
  - Check for reasonability of clusters
- Problem: No clusters found $\neq$ no clusters existing
  - Maybe clusters exists only in certain models, but can not be found by used clustering approach



Hopkins Statistics



dataset
(n objects)

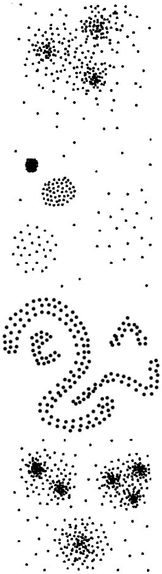
Random selection
(m objects) $m << n$

m uniformly
distributed objects

$$H = \frac{\sum_{i=1}^{m} u_i}{\sum_{i=1}^{m} u_i + \sum_{i=1}^{m} w_i}$$

- w_i: distance of selected objects to the next neighbor in dataset
- u_i: distances of uniformly distributed objects to next neighbor in dataset
- $0 \leq H \leq 1$;
  - $H \approx 0$: very regular data (e.g. grid);
  - $H \approx 0.5$: uniformly distributed data;
  - $H \approx 1$: strongly clustered,

Recap: Observed Clustering Methods



- Partitioning Methods: Find k partitions, minimizing some objective function
- Probabilistic Model-Based Clustering (EM)
- Density-based Methods: Find clusters based on connectivity and density functions
- Mean-Shift: Find modes in the point density
- Spectral Clustering: Find global minimum cut
- Hierarchical Methods: Create a hierarchical decomposition of the set of objects
- Evaluation: External and internal measures

Agenda

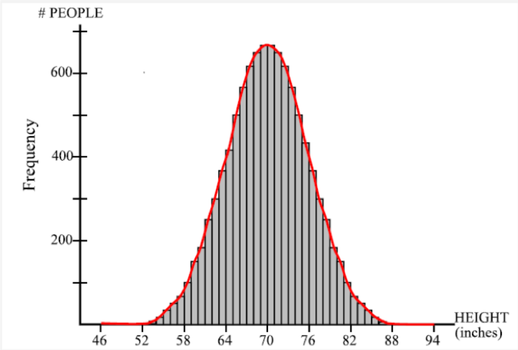
Agenda

Introduction

What is an outlier?

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



Introduction

Applications

- Fraud detection
  - Purchasing behavior of a credit card owner usually changes when the card is stolen
  - Abnormal buying patterns can characterize credit card abuse
- Medicine
  - Whether a particular test result is abnormal may depend on other characteristics of the patients (e.g. gender, age, ...)
  - Unusual symptoms or test results may indicate potential health problems of a patient
- 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.

Introduction

Applications (cont'd)

- 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."*

Introduction

Important Properties of Outlier Models

- Global vs. local approach
  - "Outlierness" regarding whole dataset (global) or regarding a subset of data (local)?
- Labeling vs. Scoring
  - Binary decision or outlier degree score?
- Assumptions about "Outlierness"
  - What are the characteristics of an outlier object?

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

Agenda

Density-Based Approaches

General Idea

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

Basic Assumption

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

Density-Based Approaches

Problems

- Different definitions of density: e.g., #points within a specified distance ϵ from the given object
- The choice of ϵ is critical (too small $\Longrightarrow$ normal points considered as outliers; too big $\Longrightarrow$ outliers considered normal)
- A global notion of density is problematic (as it is in clustering); fails when data contain regions of different densities

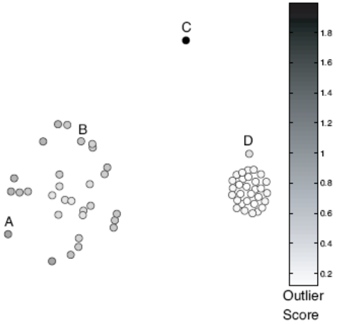


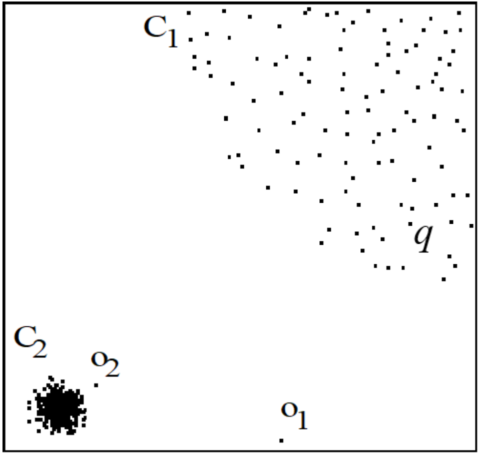
Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

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

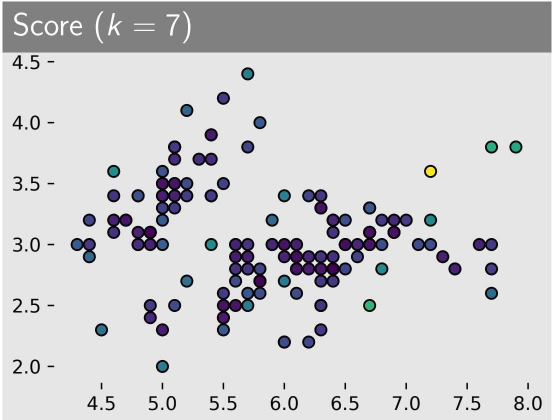
Density-Based Approaches

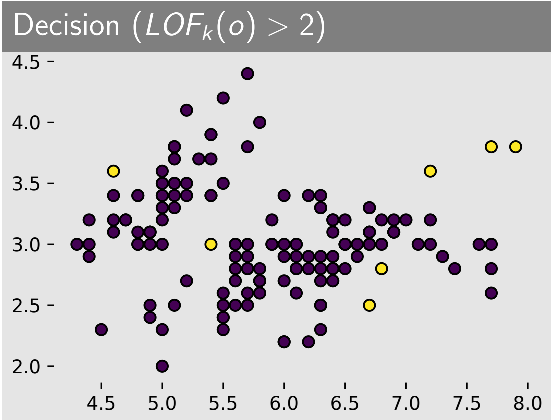
Failure Case of Distance-Based

- $D(\epsilon, \pi)$: parameters ϵ, π cannot be chosen s.t. o_2 is outlier, but none of the points in C_1 (e.g. q)
- kNN-distance: kNN-distance of objects in C_1 (e.g. q) larger than the kNN-distance of o_2.



Density-Based Approaches





Density-Based Approaches

Solution

Consider the relative density w.r.t. to the neighbourhood.

Model

- Local Density (ld) of point p (inverse of avg. distance of kNNs of p)

$$ld_k(p) = \left(\frac{1}{k} \sum_{o \in kNN(p)} dist(p, o) \right)^{-1}$$

- Local Outlier Factor (LOF) of p (avg. ratio of lds of kNNs of p and ld of p)

$$LOF_k(p) = \frac{1}{k} \sum_{o \in kNN(p)} \frac{ld_k(o)}{ld_k(p)}$$

Density-Based Approaches

Extension **(Smoothing factor)**

- Reachability "distance"

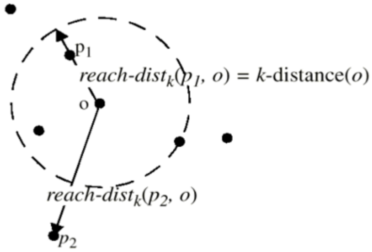
$$rd_k(p,o) = \max\{kdist(o), dist(p,o)\}$$

- Local reachability distance lrd_k

$$lrd_k(p) = \left(\frac{1}{k} \sum_{o \in kNN(p)} rd(p,o) \right)^{-1}$$

- Replace ld by lrd

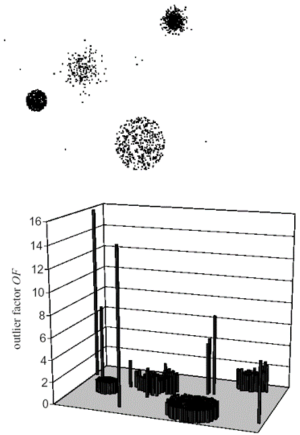
$$LOF_k(p) = \frac{1}{k} \sum_{o \in kNN(p)} \frac{lrd_k(o)}{lrd_k(p)}$$



Density-Based Approaches

Discussion

- $LOF \approx 1 \implies$ point in cluster
- $LOF \gg 1 \implies$ outlier.
- Choice of k defines the reference set

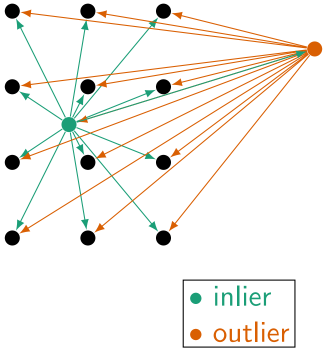


Agenda

Angle-Based Approach

General Idea

- Angles are more stable than distances in high dimensional spaces
- *o outlier* if most other objects are located in similar directions
- *o no outlier* if many other objects are located in varying directions



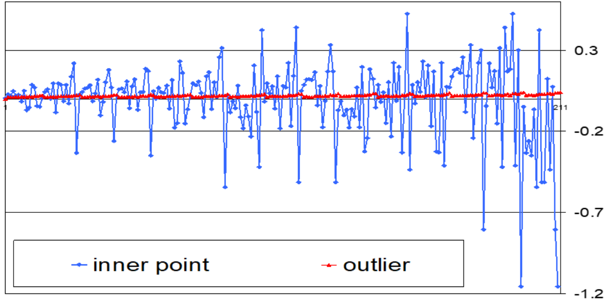
Basic Assumption

- Outliers are at the border of the data distribution
- Normal points are in the center of the data distribution

Angle-Based Approach

Model

- Consider for a given point p the angle between $\overrightarrow{px}$ and $\overrightarrow{py}$ for any two x, y from the database
- Measure the variance of the angle spectrum



Angle-Based Approach

Model (cont'd)

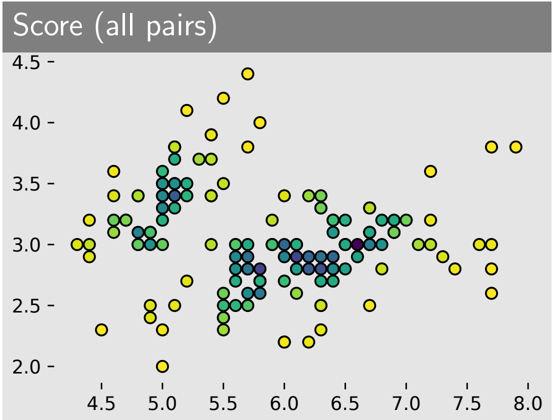
- Weighted by the corresponding distances (for lower dimensional data sets where angles are less reliable)
  Angle-based Outlier Detection[5]:

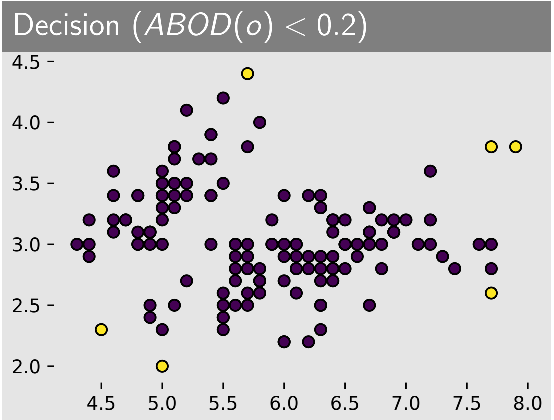
$$ABOD(p) = \mathrm{VAR}_{x,y \in D}\left(\frac{1}{\|\overrightarrow{xp}\|_2 \|\overrightarrow{yp}\|_2} \cos\left(\overrightarrow{xp}, \overrightarrow{yp}\right)\right) = \mathrm{VAR}_{x,y \in D}\left(\frac{\langle \overrightarrow{xp}, \overrightarrow{yp} \rangle}{\|\overrightarrow{xp}\|_2^2 \|\overrightarrow{yp}\|_2^2}\right)$$

- Small ABOD $\iff$ outlier

[5] Kriegel, Hans-Peter, Matthias Schubert, and Arthur Zimek. "Angle-based outlier detection in high-dimensional data." Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008.

Angle-Based Approaches





Agenda

Tree-Based Approaches: Isolation Forest

General Idea

Outlierness = how easy it is to separate a point from the rest by random space splitting?

Basic Assumption

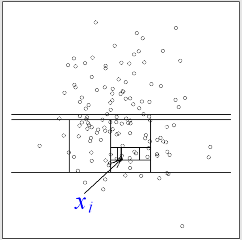
- Anomalies are the minority consisting of fewer instances
- Anomalies have attribute-values that are very different from those of normal instances

Tree-Based Approaches

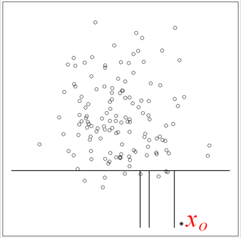
Isolation Tree - Training

1. Randomly select one dimension
2. Randomly select a split position in that dimension
3. Repeat until: a) only one point left or b) height reaches predefined threshold h

Normal point path length=10 splits



Outlier point path length=4 splits

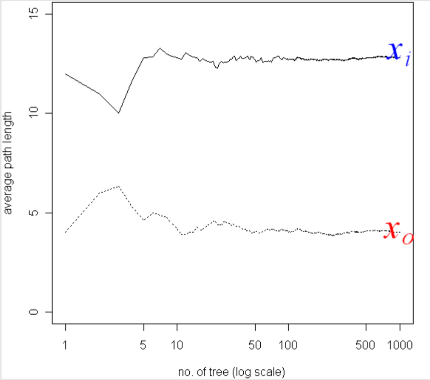


Tree-Based Approaches: Training

Isolation Forest - Training

1. Random sample ψ points, build an isolation tree
2. Repeat for t times $\Rightarrow$ a forest of t isolation trees

Average path lengths converge



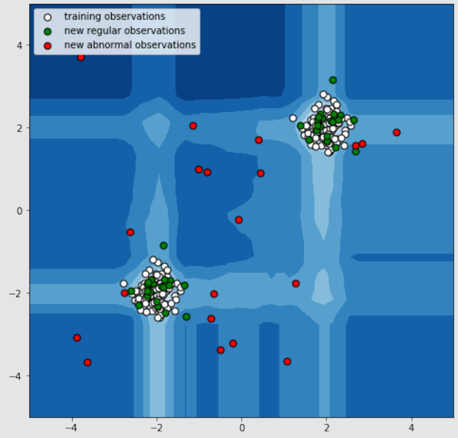
Tree-Based Approaches: Anomaly Score

- Let $h(x)$ be the path length of x on an isolation tree, and estimate $E(h(x))$ by the *average path length* among t isolation trees.
- Let $c(\psi) = 2H(\psi - 1) - 2(\psi - 1)/\psi$, which is the expected path length of unsuccessful search in BST of ψ points; $H(\cdot)$ is the harmonic number.
- Define the anomaly score of a point x as $s(x) = 2^{-\frac{E(h(x))}{c(\psi)}}$
- Observe $s(x) \in (0, 1)$
  - $E(h(x)) \to c(\psi)$ yields $s \to 0.5$,
  - $E(h(x)) \to 0$ yields $s \to 1$,
  - $E(h(x)) \to n - 1$ yields $s \to 0$.
- Usually, set $s = 0.5$ as threshold, i.e. the average of the expected path length

Tree-Based Approaches: Discussion

- Advantages:
  - Anomaly score between 0 and 1
  - Very efficient, especially on large dataset
  - A model (the forest) is learned from the training dataset
  - Easy for parallelization
  - Can be adapted to categorical data
- Disadvantages:
  - Only detects global outliers (of course, follow-up approaches are available)
  - Not efficient on high-dimensional data

iForest anomaly score contour



Recap - Outlier Detection

- Properties: global vs. local, labeling vs. scoring
- *Clustering-Based* Outliers: Identification as non-(cluster-members)
- *Statistical* Outliers: Assume probability distribution; outliers = unlikely to be generated by distribution
- *Distance-Based* Outliers: Distance to neighbors as outlier metric
- *Density-Based* Outliers: Relative density around the point as outlier metric
- *Angle-Based* Outliers: Angles between outliers and random point pairs vary only slightly

Agenda

Agenda

What is Frequent Pattern Mining?

Setting: Transaction Databases

A database of transactions, where each transaction comprises a set of items, e.g. one transaction is the basket of one customer in a grocery store.

Frequent Pattern Mining

Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories.

Applications

Basket data analysis, cross-marketing, catalogue design, loss-leader analysis, clustering, classification, recommendation systems, etc.

What is Frequent Pattern Mining?

Task 1: Frequent Itemset Mining

Find all subsets of items that occur together in many transactions.

Example

Which items are bought together frequently?

$$
\begin{aligned}
D = \{ \quad & \{ \quad butter, bread, milk, sugar\}, \\
& \{ \quad butter, flour, milk, sugar\}, \\
& \{ \quad butter, eggs, milk, salt\}, \\
& \{ \quad eggs\}, \\
& \{ \quad butter, flour, milk, salt, sugar\}\}
\end{aligned}
$$

$\rightsquigarrow$ 80% of transactions contain the itemset {milk, butter}

What is Frequent Pattern Mining?

Task 2: Association Rule Mining

Find all rules that correlate the presence of one set of items with that of another set of items in the transaction database.

Example

98% of people buying tires and auto accessories also get automotive service done

Agenda

Mining Frequent Itemsets: Basic Notions

- **Items** $I = \{i_1, \ldots, i_m\}$: a set of literals (denoting items)
- **Itemset** X: Set of items $X \subseteq I$
- **Database** D: Set of *transactions* T, each transaction is a set of items $T \subseteq I$
- Transaction T contains an itemset X: $X \subseteq T$
- **Length** of an itemset X equals its cardinality $|X|$
- k-**itemset**: itemset of length k
- (Relative) **Support** of an itemset: $supp(X) = |\{T \in D \mid X \subseteq T\}|/|D|$
- X is **frequent** if $supp(X) \geq minSup$ for threshold $minSup$.

Goal

Given a database D and a threshold $minSup$, find all frequent itemsets $X \in Pot(I)$.

Mining Frequent Itemsets: Basic Idea

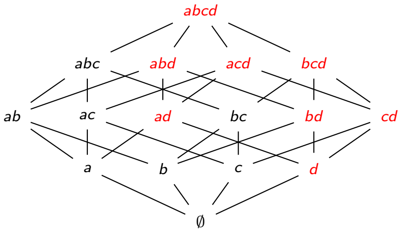
Naïve Algorithm

Count the frequency of all possible subsets of I in the database D.

Problem

Too expensive since there are 2^m such itemsets for m items (for $|I| = m$, $2^m =$ cardinality of the powerset of I).

Mining Frequent Patterns: Apriori Principle



- frequent
- non-frequent

Apriori Principle (anti-monotonicity)

- Any non-empty subset of a frequent itemset is frequent, too!

$$A \subseteq I : supp(A) \geq minSup \implies \forall \emptyset \neq A' \subset A : supp(A') \geq minSup$$

- Any superset of a non-frequent itemset is non-frequent, too!

$$A \subseteq I : supp(A) < minSup \implies \forall A' \supset A : supp(A') < minSup$$

Apriori Algorithm

Idea

- First count the 1-itemsets, then the 2-itemsets, then the 3-itemsets, and so on
- When counting $(k+1)$-itemsets, only consider those $(k+1)$-itemsets where all subsets of length k have been determined as frequent in the previous step

Apriori Algorithm

variable C_k: candidate itemsets of size k
variable L_k: frequent itemsets of size k
$L_1 = \{\text{frequent items}\}$
for $(k = 1;\ L_k \neq \emptyset;\ k{+}{+})$ **do**

Produce candidates.

join L_k with itself to produce C_{k+1} ▷ JOIN STEP
discard $(k+1)$-itemsets from C_{k+1} that ... ▷ PRUNE STEP
...contain non-frequent k-itemsets as subsets

C_{k+1} = candidates generated from L_k

Prove candidates.

for each transaction $T \in D$ **do**
Increment the count of all candidates in C_{k+1} ...
...that are contained in T

L_{k+1} = candidates in C_{k+1} with *minSupp*

return $\bigcup_k L_k$

Apriori Algorithm: Generating Candidates – Join Step

Requirements for Candidate $(k+1)$-itemsets

- *Completeness*: Must contain all frequent $(k+1)$-itemsets (superset property $C_{k+1} \supseteq L_{k+1}$)
- *Selectiveness*: Significantly smaller than the set of all $(k+1)$-subsets

Suppose the itemsets are sorted by any order (e.g. lexicographic)

Step 1: Joining ($C_{k+1} = L_k \bowtie L_k$)

- Consider frequent k-itemsets p and q
- p and q are joined if they share the same first $(k-1)$ items.

Apriori Algorithm: Generating Candidates – Join Step

Example

- $k = 3$ ($\implies k + 1 = 4$)
- $p = (a, c, f) \in L_k$
- $q = (a, c, g) \in L_k$
- $r = (a, c, f, g) \in C_{k+1}$

SQL example

insert into C_{k+1}

select $p.i_1, p.i_2, \ldots, p.i_k, q.i_k$

from $L_k : p, L_k : q$

where $p.i_1 = q.i_1, \ldots, p.i_{k-1} = q.i_{k-1}, p.i_k < q.i_k$

Apriori Algorithm: Generating Candidates – Prune Step

Step 2: Pruning ($L_{k+1} = \{X \in C_{k+1} \mid supp(X) \geq minSup\}$)

- *Naïve*: Check support of every itemset in C_{k+1} $\rightsquigarrow$ inefficient for huge C_{k+1}
- *Better*: Apply Apriori principle first: Remove candidate $(k+1)$-itemsets which contain a non-frequent k-subset s, i.e., $s \notin L_k$

Pseudocode

for all $c \in C_{k+1}$ **do**
    for all k-subsets s of c **do**
        if $s \notin L_k$ **then**
            Delete c from C_{k+1}

Apriori Algorithm: Generating Candidates – Prune Step

Example

- $L_3 = \{acf, acg, afg, afh, cfg\}$
- Candidates after join step: $\{acfg, afgh\}$
- In the pruning step: delete $afgh$ because $fgh \notin L_3$, i.e. fgh is not a frequent 3-itemset (also $agh \notin L_3$)
- $C_4 = \{acfg\} \rightsquigarrow$ check the support to generate L_4

Apriori Algorithm: Full example

Database

TID	items
0	acdf
1	bce
2	abce
3	aef

minSup = 0.5

Alphabetic Ordering

k	candidate	prune	count	threshold
1	a		3	a
	b		2	b
	c		3	c
	d		1	
	e		3	e
	f		2	f
2	ab		1	
	ac		2	ac
	ae		2	ae
	af		2	af
	bc		2	bc
	be		2	be
	bf		0	
	ce		2	ce
	cf		1	
	ef		1	
3	ace		1	
	acf	with cf		
	aef	with ef		
	bce		2	bce

Frequency-Ascending Ordering

k	candidate	prune	count	threshold
1	d		1	
	b		2	b
	f		2	f
	a		3	a
	c		3	c
	e		3	e
2	bf		0	
	ba		1	
	bc		2	bc
	be		2	be
	fa		2	fa
	fc		1	
	fe		1	
	ac		2	ac
	ae		2	ae
	ce		2	ce
3	bce		2	bce
	ace		1	