1. Introduction to Data Mining

Grading

- 60% during the semester:
  - 10% Course activity (course attendance)
  - 20% Midterm exam (questions with multiple choice answers)
  - 30% Project (project attendance, algorithm presentation, project delivery)
- 40% Final exam (questions with multiple choice answers)
- Course site:
  - http://cs.curs.pub.ro

Road Map

- ❑ What is data mining
- ❑ Steps in data mining process
- ❑ Data mining methods and subdomains
- ❑ Summary

Definition ([Liu 11])

- Data mining is also called Knowledge Discovery in Databases (KDD).
- It is commonly defined as the process of discovering useful patterns or knowledge from data sources, e.g., databases, texts, images, the web, etc.
- The patterns must be valid, potentially useful and understandable.

Definition ([Ullman 09, 10])

- ❑Discovery of useful, possibly unexpected, patterns in data.
- ❑Discovery of “models” for data:
  - – Statistical modeling
  - – Machine learning
  - – Computational Approaches to Modeling
  - – Summarization
  - – Feature Extraction

Definition ([Wikipedia])

- Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets ("big data") involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

Definition ([Wikipedia])

- ❑ The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.
- ❑ Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

Definition ([Kimball, Ross 02])

- A class of undirected queries, often against the most atomic data, that seek to find unexpected patterns in the data.
- The most valuable results from data mining are clustering, classifying, estimating, predicting, and finding things that occur together.
- There are many kinds of tools that play a role in data mining, including decision trees, neural networks, memory- and case-based reasoning tools, visualization tools, genetic algorithms, fuzzy logic, and classical statistics.
- Generally, data mining is a client of the data warehouse.

Conclusions

- The data mining process converts data into valuable knowledge that can be used for decision support
- Data mining is a collection of data analysis methodologies, techniques and algorithms for discovering new patterns
- Data mining is used for large data sets
- Data mining process is automated (no need for human intervention)
- Data mining and Knowledge Discovery in Databases (KDD) are considered by some authors to be the same thing. Other authors list data mining as the analysis step in the KDD process - after data cleaning and transformation and before results visualization / evaluation.

Success stories (1)

Some early success stories in using data mining (from [Ullman 03]):

- Decision trees constructed from bank-loan histories to produce algorithms to decide whether to grant a loan.
- Patterns of traveler behavior mined to manage the sale of discounted seats on planes, rooms in hotels, etc.
- “Diapers and beer" Observation that customers buying diapers are more likely to buy beer than average, allowing supermarkets to place beer and diapers nearby, knowing that many customers would walk between them. Placing potato chips between increased sales of all three items.

Success stories (2)

- Skycat and Sloan Sky Survey: clustering sky objects by their radiation levels in different bands allowed astronomers to distinguish between galaxies, nearby stars, and many other kinds of celestial objects.
- Comparison of the genotype of people with/without a condition allowed the discovery of a set of genes that together account for many cases of diabetes. This sort of mining will become much more important as the human genome is constructed.

What is not Data Mining

- ❑ Find a certain person in an employee database
- ❑ Compute the minimum, maximum, sum, count or average values based on table/tables columns
- ❑ Use a search engine to find your name occurrences on the web

DM software (1)

In ([Mikut, Reischl 11]) DM software programs are classified in 9 categories:

- **Data mining suites** (DMS) focus on data mining and include numerous methods and support feature tables and time series. Examples:
  - Commercial: IBM SPSS Modeler, SAS Enterprise Miner, DataEngine, GhostMiner, Knowledge Studio, NAG Data Mining Components, STATISTICA
  - Open source: RapidMiner
- **Business intelligence packages** (BIs) include basic data mining functionality - statistical methods in business applications, are often restricted to feature tables and time series and large feature tables are supported. Examples:
  - Commercial: IBM Cognos 8 BI, Oracle DataMining, SAPNetweaver Business Warehouse, Teradata Database, DB2 Data Warehouse from IBM
  - Open source: Pentaho

DM software (2)

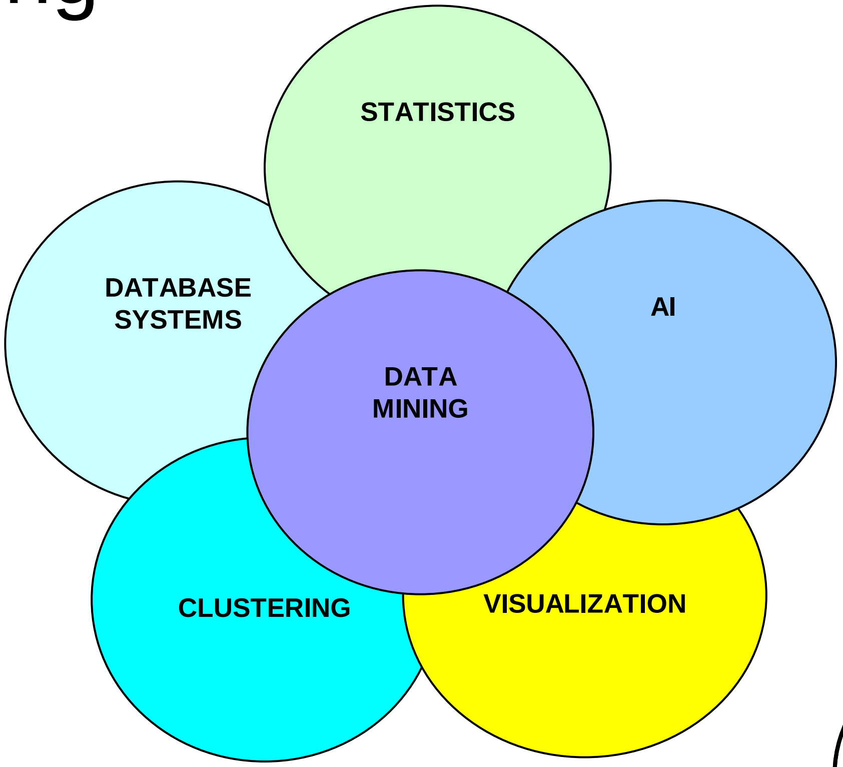
- **Mathematical packages** (MATs) provide a large and extendable set of algorithms and visualization routines. Examples:
  - Commercial: MATLAB, R-PLUS
  - Open source: R, Kepler
- **Integration packages** (INTs) are extendable bundles of many different open-source algorithms
  - Stand-alone software (KNIME, the GUI-version of WEKA, KEEL, and TANAGRA)
  - Larger extension package for tools from the MAT type
- **Extensions** (EXT) are smaller add-ons for other tools such as Excel, Matlab, R, with limited but quite useful functionality. Examples:
  - Artificial neural networks for Excel (Forecaster XL and XLMiner)
  - MATLAB (Matlab Neural Networks Toolbox).

DM software (3)

- **Data mining libraries** (LIBs) implement data mining methods as a bundle of functions and can be embedded in other software tools using an Application Programming Interface. Examples: Neurofusion for C++, WEKA, MLC++, JAVA Data Mining Package, LibSVM
- **Specialties** (SPECs) are similar to DMS tools, but implement only one special family of methods such as artificial neural networks. Examples: CART, Bayesia Lab, C5.0, WizRule, Rule Discovery System, MagnumOpus, JavaNNS, Neuroshell, NeuralWorks Predict, RapAnalyst.
- **Research** (RES) are usually the first implementations of new algorithms, with restricted graphical support and without automation support. RES tools are mostly open source. WEKA and RapidMiner started in this category.
- **Solutions** (SOLs) describe a group of tools that are customized to narrow application fields. Examples: for text mining: GATE, image processing: ITK, ImageJ, drug discovery: Molegro Data Modeler

Communities involved

The most important communities involved in data mining



Road Map

- What is data mining
- Steps in data mining process
- Data mining methods and subdomains
- Summary

Data mining steps (1)

1. **Data collection**: Data gathering from existing databases or (for Internet documents) from Web crawling.
2. **Data preprocessing**, including:
   - Data cleaning: replace (or remove) missing values, smooth noisy data, remove or just identify outliers, remove inconsistencies.
   - Data integration: integration of data from multiple sources, with possible different data types and structures and also handling of duplicate or inconsistent data.
   - Data transformation: data normalization (or standardization), summarizations, generalization, new attributes construction, etc.

Data mining steps (2)

2. **Data preprocessing (cont):**
   - Data reduction (called also feature extraction): not all the attributes are necessary for the particular Data Mining process we want to perform. Only relevant attributes are selected for further processing reducing the total size of the dataset (and the time needed for running the algorithm).
   - Discretization: some algorithms work only on discrete data. For that reason the values for continuous attributes must be replaced with discrete ones from a limited set. One example is replacing age (number) with an attribute having only three values: Young, Middle-age and Old.

Data mining steps (3)

3. **Pattern extraction and discovery**. This is the stage where the data mining algorithm is used to obtain the result. Some authors consider that Data Mining is reduced only at this step, the whole process being called KDD.
4. **Visualization**: because data mining extracts hidden properties/information from data it is necessary to visualize the results for a better understanding and evaluation. Also needed for the input data.
5. **Evaluation of results**: not everything that outputs from a data mining algorithm is a valuable fact or information. Some of them are statistic truths and others are not interesting/useful for our activity. Expert judgment is necessary in evaluating the results

Bonferroni principle (1)

A true information discovered by a ‘data mining’ process can be a statistical truth. Example (from [Ullman 03]):

- In 1950’s David Rhine, a parapsychologist, tested students in order to find if they have or not extrasensorial perception (ESP).
- He asked them to guess the color of 10 successive cards – red or black. The result was that 1/1000 of them guessed all 10 cards (he declared they have ESP).
- Re-testing only these students he found that they have lost ESP after knowing they have this feature
- David Rhine did not realize that the probability of guessing 10 successive cards is 1/1024 = $1/2^{10}$, because the probability for each of these 10 cards is ½ (red or black).

Bonferroni principle (2)

- This kind of results may be included in the output of a data mining algorithm but must be recognized as a statistical truth and not a real data mining output.
- This fact is also the object of the Bonferroni principle. This can be synthesized as below:

- if your method of finding significant items returns significantly more items that you would expect in the actual population, you can assume most of the items you find with it are bogus [rationalwiki.org]

Road Map

- ❑ What is data mining
- ❑ Steps in data mining process
- ❑ Data mining methods and subdomains
- ❑ Summary

Method types

- **Prediction methods**. These methods use some variables to predict the values of other variables. A good example for that category is classification. Based on known, labeled data, classification algorithms build models that can be used for classifying new, unseen data.
- **Description methods**. Algorithms in this category find patterns that can describe the inner structure of the dataset. For example clustering algorithms find groups of similar objects in a dataset (called clusters) and possible isolated objects, far away from any cluster, called outliers.

Algorithms

- **Prediction algorithm types:**
  - Classification
  - Regression
  - Deviation detection
- **Description algorithm types:**
  - Clustering
  - Association rule discovery
  - Sequential pattern discovery

Classification

Input:

- A set of **k** classes $C = \{c_1, c_2, \ldots, c_k\}$
- A set of **n** labeled items $D = \{(d_1, c_{i1}), (d_2, c_{i2}), \ldots, (d_n\ c_{in})\}$. The items are $d_1, \ldots, d_n$, each item d_j being labeled with class $c_j \in C$. D is called the **training set**.
- For calibration of some algorithms a **validation set** is required. This validation set contains also labeled items not included in the training set.

Output:

- A **model** or **method** for classifying new items (a classifier). The set of new items that will be classified using the model/method is called the **test set**

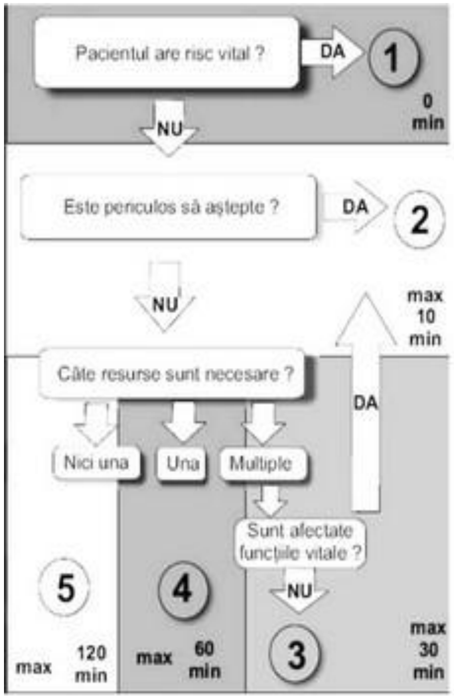
Example

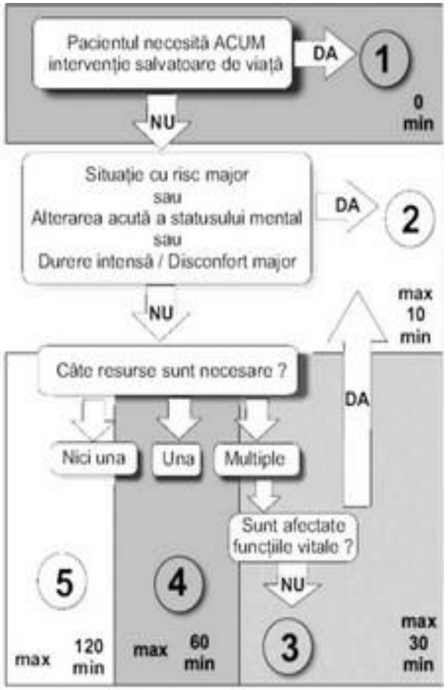
- Let us consider a medical set of items where each item is a patient of a hospital emergency unit (RO: UPU).
- There are 5 classes, representing maximum waiting time categories: C0, C10, C30, C60 and C120, Ck meaning the patient waits maximum k minutes.
- We may represent these data in tabular format
- The output of a classification algorithm using this training set may be for example a decision tree or a set of ordered rules.
- The model may be used to classify future patients and assign a waiting time label to them

Emergency unit training set

Name (or ID)	Vital risk?	Danger if waits?	0 resource needed	1 resource needed	>1 resource needed	>1 resource needed and vital function s affected	Waiting time (class label)
John	Yes	Yes	No	Yes	No	No	C0
Maria	No	Yes	No	No	Yes	No	C10
Nadia	Yes	Yes	Yes	No	No	No	C0
Omar	No	No	No	No	Yes	Yes	C30
Kiril	No	No	No	Yes	No	Yes	C60
Denis	No	No	No	No	Yes	No	C10
Jean	No	No	Yes	Yes	No	No	C120
Patricia	Yes	Yes	No	No	Yes	Yes	C60

Result: decision tree


Pacientul are risc vital ?
DA
1
0 min
NU
Este periculos să aştepte ?
DA
2
NU
max 10 min
Câte resurse sunt necesare ?
DA
Nici una
Una
Multiple
Sunt afectate funcţiile vitale ?
NU
5
4
3
max 120 min
max 60 min
max 30 min


Pacientul necesită ACUM intervenţie salvatoare de viaţă
DA
1
0 min
NU
Situaţie cu risc major
sau
Alterarea acută a statusului mental
sau
Durere intensă / Disconfort major
DA
2
NU
max 10 min
Câte resurse sunt necesare ?
DA
Nici una
Una
Multiple
Sunt afectate funcţiile vitale ?
NU
5
4
3
max 120 min
max 60 min
max 30 min

- The result for the example:

Felix	**Yes**	**Yes**	**No**	**No**	**No**	**Yes**	**?????**

will be C0

Regression (1)

- ❑ Regression is related with statistics.
- ❑ Meaning: predicting a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency ([Tan, Steinbach, Kumar 06]).
- ❑ Used in prediction and forecasting - its use overlaps machine learning.
- ❑ Regression analysis is also used to understand relationship between independent variables and dependent variable and can be used to infer causal relationships between them.

Regression (2)

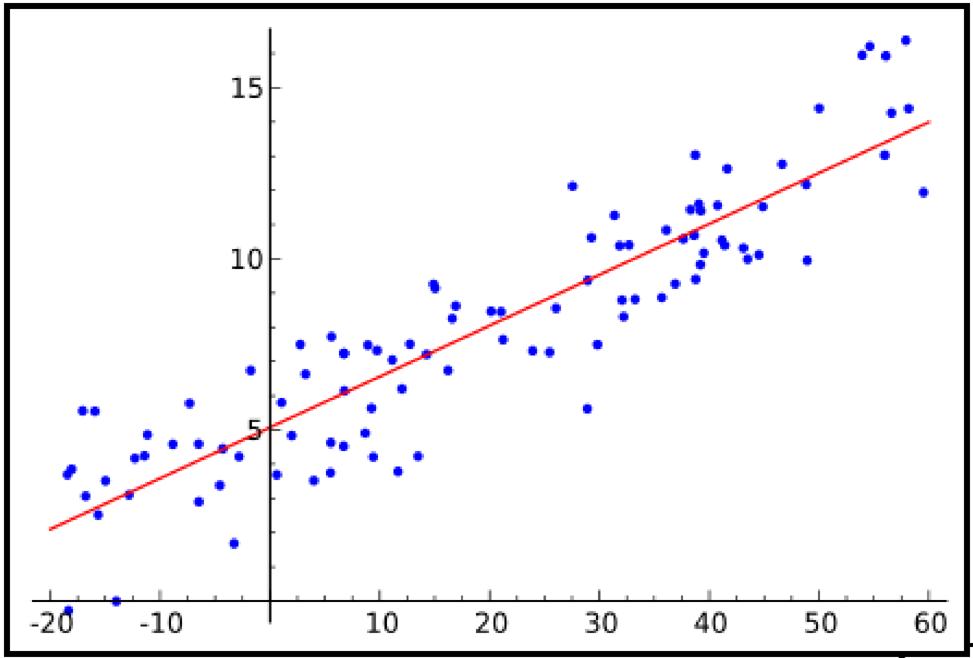
There are many types of regression. For example, Wikipedia lists:

- ❑Linear regression model
- ❑Simple linear regression
- ❑Logistic regression
- ❑Nonlinear regression
- ❑Nonparametric regression
- ❑Robust regression
- ❑Stepwise regression

Example

Linear regression example

(from http://en.wikipedia.org/wiki/File:Linear_regression.svg)



Deviation detection

- Deviation detection or anomaly detection means discovering significant deviation from the normal behavior. Outliers are a significant category of abnormal data.
- Deviation detection can be used in many circumstances:
  - ✓ Data mining algorithm running stage: often such information may be important for business decisions and scientific discovery.
  - ✓ Auditing: such information can reveal problems or mal-practices.
  - ✓ Fraud detection in a credit card system: fraudulent claims often carry inconsistent information that can reveal fraud cases.
  - ✓ Intrusion detection in a computer network may rely on abnormal data.
  - ✓ Data cleaning (part of data preprocessing): such information can be detected and possible mistakes may be corrected in this stage.

Deviation detection techniques

- Distance based techniques (example: k-nearest neighbor).
- One Class Support Vector Machines.
- Predictive methods (decision trees, neural networks).
- Cluster analysis based outlier detection.
- Pointing at records that deviate from association rules
- Hotspot analysis

Algorithms

- **Prediction algorithm types:**
  - Classification
  - Regression
  - Deviation Detection
- **Description algorithm types:**
  - Clustering
  - Association Rule Discovery
  - Sequential Pattern Discovery

Clustering

Input:

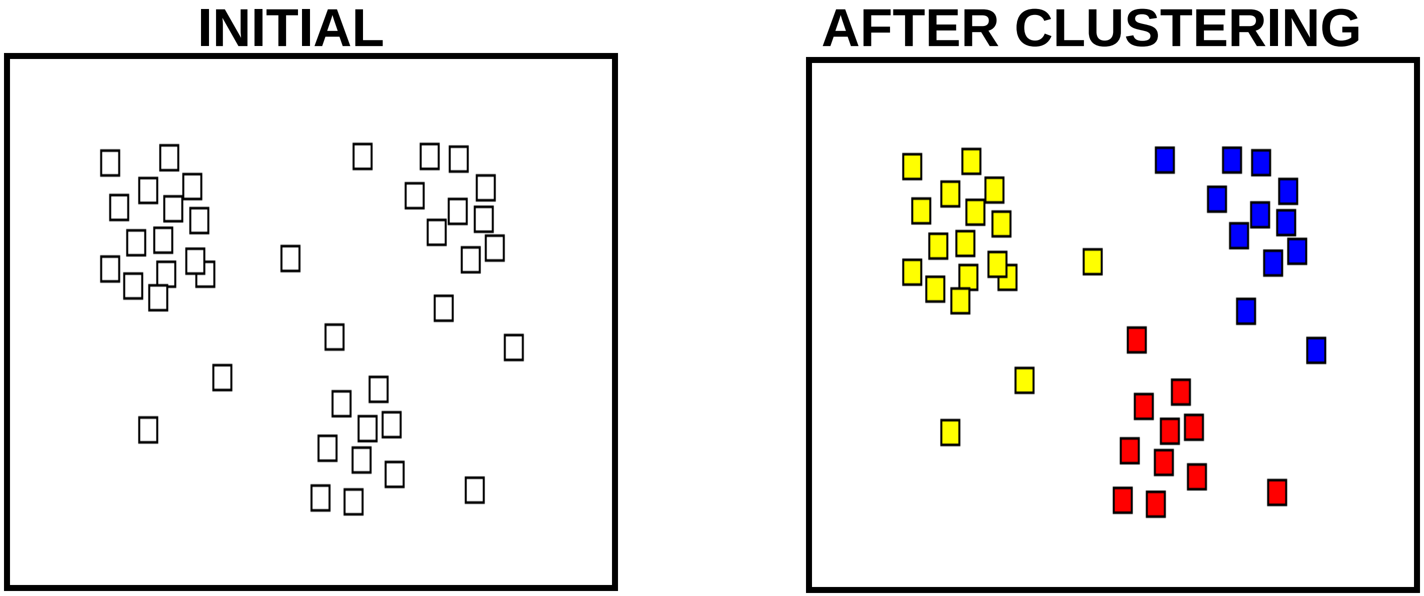
- ❑ A set of n objects D = {d_1, d_2, …, d_n} (called usually points). The objects are not labeled and there is no set of class labels defined.
- ❑ A distance function (dissimilarity measure) that can be used to compute the distance between any two points. Low valued distance means ‘near’, high valued distance means ‘far’.
- ❑ Some algorithms also need a predefined value for the number of clusters in the produced result.

Output:

- ❑ A set of object (point) groups called clusters where points in the same cluster are near one to another and points from different clusters are far one from another, considering the distance function.

Example

- Having a set of points in a 2 dimensional space, find the natural clusters formed by these points.



Source: http://en.wikipedia.org/wiki/File:Cluster-1.svg, http://en.wikipedia.org/wiki/File:Cluster-2.svg

Association Rule Discovery

Let us consider:

- A set of **m** items $I = \{i_1, i_2, \ldots, i_m\}$.
- A set of **n** transactions $T = \{ t_1, t_2, \ldots, t_n\}$, each transaction containing a subset of I, so if $t_k \in T$ then $t_k = \{i_{k1}, i_{k2}, \ldots, i_{kj}\}$ where j depends on k.

Then:

- ❑ A **rule** is a construction $X \rightarrow Y$ where X and Y are itemsets.

Association Rule Discovery

- The **support** of a rule is the number/proportion of transactions containing the union between the left and the right part of the rule (and is equal with the support of this union as an itemset):

$$\mathbf{support(X \rightarrow Y) = support(X \cup Y)}$$

- The **confidence** of a rule is the proportion of transactions containing Y in the set of transactions containing X:

$$\mathbf{confidence(X \rightarrow Y) = support(X \cup Y) \; / \; support(X).}$$

- We accept a rule as a valid one if the support and the confidence of the rule are at least equal with some given **thresholds**.

Association Rule Discovery

Input:

- ❑ A set of **m** items $I = \{i_1, i_2, \ldots, i_m\}$.
- ❑ A set of **n** transactions $T = \{ t_1, t_2, \ldots, t_n\}$, each transaction containing a subset of I, so if $t_k \in T$ then $t_k = \{i_{k1}, i_{k2}, \ldots, i_{kj}\}$ where j depends on k.
- ❑ A threshold s for the support, given either as a percent or in absolute value. If an itemset $X \subseteq I$ is part of w transactions then w is the support of X. If w >= s then X is called frequent itemset
- ❑ A second threshold c for rule confidence.

Output:

- ❑ The set of frequent itemsets in T, having support >= s
- ❑ The set of rules derived from T, having support >= s and confidence >= c

Example

- Consider the following set of transactions:

Transaction ID	Items
1	Bread, Milk, Butter, Orange Juice, Onion, Beer
2	Bread, Milk, Butter, Onion, Garlic, Beer, Orange Juice, Shirt, Pen, Ink, Baby diapers
3	Milk, Butter, Onion, Garlic, Beer
4	Orange Juice, Shirt, Shoes, Bread, Milk
5	Butter, Onion, Garlic, Beer, Orange Juice

- If s = 60% then {Bread, Milk, Orange Juice} or {Onion, Garlic, Beer} are frequent itemsets. Also if s = 60% and c=70% then the rule {Onion, Beer} → {Garlic} is a valid one because its support is 60% and the confidence is 75%.

Sequences

The model:

- **Itemset**: a set of n distinct items

$$I = \{i_1, i_2, \ldots, i_n \}$$

- **Event**: a non-empty collection of items; we can assume that items are in a given order (e.g. lexicographic): $(i_1, i_2 \ldots i_k)$
- **Sequence** : an ordered list of events:

$$< e_1 \;\; e_2 \;\; \ldots \;\; e_m >$$

Sequential Pattern Discovery

Input:

- ❑ A set of sequences **S** (or a sequence database).
- ❑ A Boolean function that can test if a sequence S_1 is included (or is a subsequence) of a sequence S_2. In that case S_2 is called a super sequence of S_1.
- ❑ A threshold **s** (percent or absolute value) needed for finding frequent sequences.

Output:

- ❑ The set of frequent sequences, i.e. the set of sequences that are included in at least **s** sequences from S.
- ❑ Sometimes a set of rules can be derived from the set of frequent sequences, each rule being of the form $S_1 \to S_2$ where S_1 and S_2 are sequences.

Examples

- ❑ In a bookstore we can find frequent sequences like:

  {(Book_on_C, Book_on_C++), (Book_on_Perl)}

- ❑ From this sequence we can derive a rule like that: after buying books about C and C++, a customers buys books on Perl:

  Book_on_C, Book_on_C++ → Book_on_Perl

Summary

This first course presented:

- ❑ A list of alternative definitions of Data Mining and some examples of what is Data Mining and what is not Data Mining
- ❑ A discussion about the researchers communities involved in Data Mining and about the fact that Data Mining is a cluster of subdomains
- ❑ The steps of the Data Mining process from collecting data located in existing repositories (data warehouses, archives or operational systems) to the final evaluation step.
- ❑ A brief description of the main subdomains of Data Mining with some examples for each of them.

Next week: Data preprocessing

References

- [Liu 11] Bing Liu, 2011. Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data, Second Edition, Springer, 1-13.
- [Tan, Steinbach, Kumar 06] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, 2006. Introduction to Data Mining, Adisson-Wesley, 1-16.
- [Kimbal, Ross 02] Ralph Kimball, Margy Ross, 2002. The Data Warehouse Toolkit, Second Edition, John Wiley and Sons, 1-16, 396
- [Mikut, Reischl 11] Ralf Mikut and Markus Reischl, Data mining tools, 2011, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 1, Issue 5, http://onlinelibrary.wiley.com/doi/10.1002/widm.24/pdf
- [Ullman] Jeffrey Ullman, Data Mining Lecture Notes, 2003-2009, web page: http://infolab.stanford.edu/~ullman/mining/mining.html
- [Wikipedia] Wikipedia, the free encyclopedia, en.wikipedia.org