Data Mining, Lecture 1

Introduction & Data Preparation

S. Nõmm

[1]Department of Software Science, Tallinn University of Technology

03.09.2019

Course organization: (administrative part)

- For all correspondence concerning the course use email sven.nomm@ttu.ee Please avoid using phone.
- Lectures & Practices. Observe the fact that two different codes are assigned to the course!
- Grading:
  - 3x mandatory closed book tests. Each test gives 10% of the final grade. For each test one make-up attempt.
  - 3x mandatory home assignments (Computational assignment +short write up.) 10% of the final grade each. Assignments are accepted up to one week after the deadline with the penalty of 10% for each day except Saturday and Sunday.
  - final exam (gives 40 % of the final grade): Written report on assigned topic + discussion with lecturer. Prerequisites:
    - ⋆ all 3 closed book tests are accepted (graded as 51 or higher)
    - ⋆ all 3 home assignments are accepted (graded as 51 or higher)
- **No Plagiarism in any of assignments and final project!!!**. You should cite all the references, including software and extra libraries. The student should be able to explain the meaning of all the computations performed, interpret and present the results.

Course organization: Grading *vs.* expected knowledge

- **Excellent 91 -100** Able to apply all the methods and techniques, thought during the course, on practice. Interpret the results and explain theoretical foundations of the applied techniques Discuss achieved results with respect of possible further analysis. Able to learn new techniques independently and apply them on practice.
- **Very Good 81 -90** Able to apply all the methods and techniques, thought during the course, on practice. Interpret the results and explain theoretical foundations of the applied techniques. Discuss achieved results with respect of possible further studies.
- **Good 71 -80** Able to apply all the methods and techniques, thought during the course, on practice, interpret the results and explain theoretical foundations of the applied techniques.
- **Satisfactory 61-70** Able to apply all the methods and techniques, thought during the course, on practice. Interpret the results.
- **Acceptable 51-60** Able to apply all core methods and techniques, thought during the course, on practice. Interpret the results.

References

The structure of the present course, main notations and definitions are inherited from [1]. [7] provides basic knowledge of "R" and "Rattle" for the data mining assignments. Implementation of different data mining algorithms in "R" is discussed by [3]. Some data mining methods are borrowed from the neighbouring fields of research, such as Machine Learning [2], Pattern Recognition [6] and Feature Extraction [4]. Lectures related to the networked data mining are based on [5].

[1] C.C. Aggarwal. *Data Mining: The Textbook*. Springer International Publishing, 2015.

[2] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 2013.

[3] P. Cichosz. *Data Mining Algorithms: Explained Using R*. Wiley, 2015.

[4] I. Guyon, S. Gunn, M. Nikravesh, and L.A. Zadeh. *Feature Extraction: Foundations and Applications*. Studies in Fuzziness and Soft Computing. Springer Berlin Heidelberg, 2008.

[5] E.D. Kolaczyk and G. Csárdi. *Statistical Analysis of Network Data with R*. Use R! Springer New York, 2014.

[6] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Elsevier Science, 2008.

[7] G. Williams. *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*. Use R! Springer New York, 2011.

Course organization: administrative part (continued)

- You are expected to attend the lectures and practices. Lecture slides do not contain all the information. Also, this is the place where you can gain experience!
- Consultations: By appointment only! Please do not hesitate to ask if you need consultation.
- It is advisable to write your own notes!
- Mind academic 15 min!
- Many concepts introduced during the course require understanding of the probability theory and statistics.
- "R" and some related packages will be used to perform computational part of the assignments.
- Any questions?

It is advisable to refresh your knowledge of:

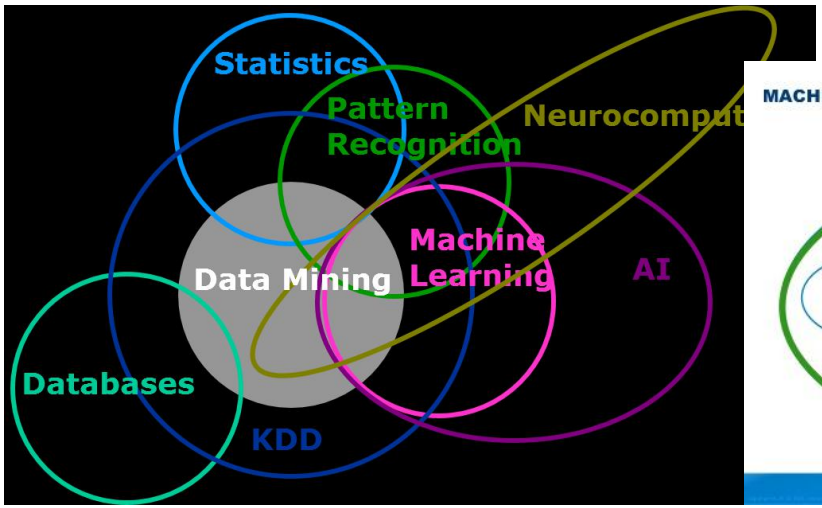
- Mathematics (calculus and linear algebra).
- Statistics.
- Programming.

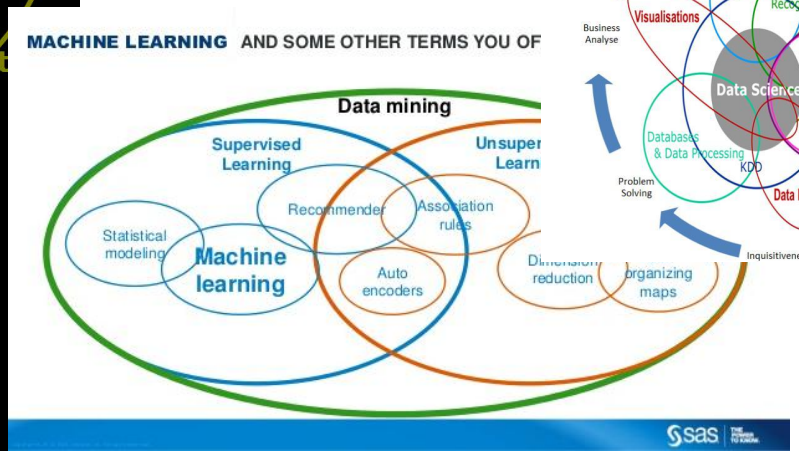
Course main topics(tentative)

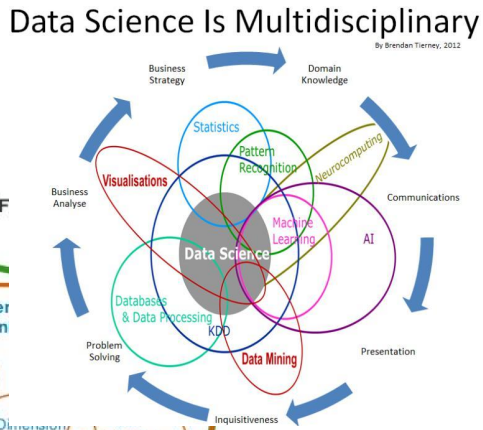
- What data mining is?
- Data mining vs. Machine learning vs. ?
- Four "super problems" of data mining:
  - Clustering
  - Classification
  - Association pattern mining
  - Outlier analysis and anomalies detection.
- Main Topics
  - Data types and Data Preparation
  - Similarity and Distances, Association Pattern Mining
  - Cluster Analysis, Classification, Outlier analysis
  - Data streams, Text Data, Time Series, Discrete Sequences
  - Spatial Data, Graph Data, Web Data, Social Network Analysis
  - Privacy-Preserving Data Mining

What data mining is?

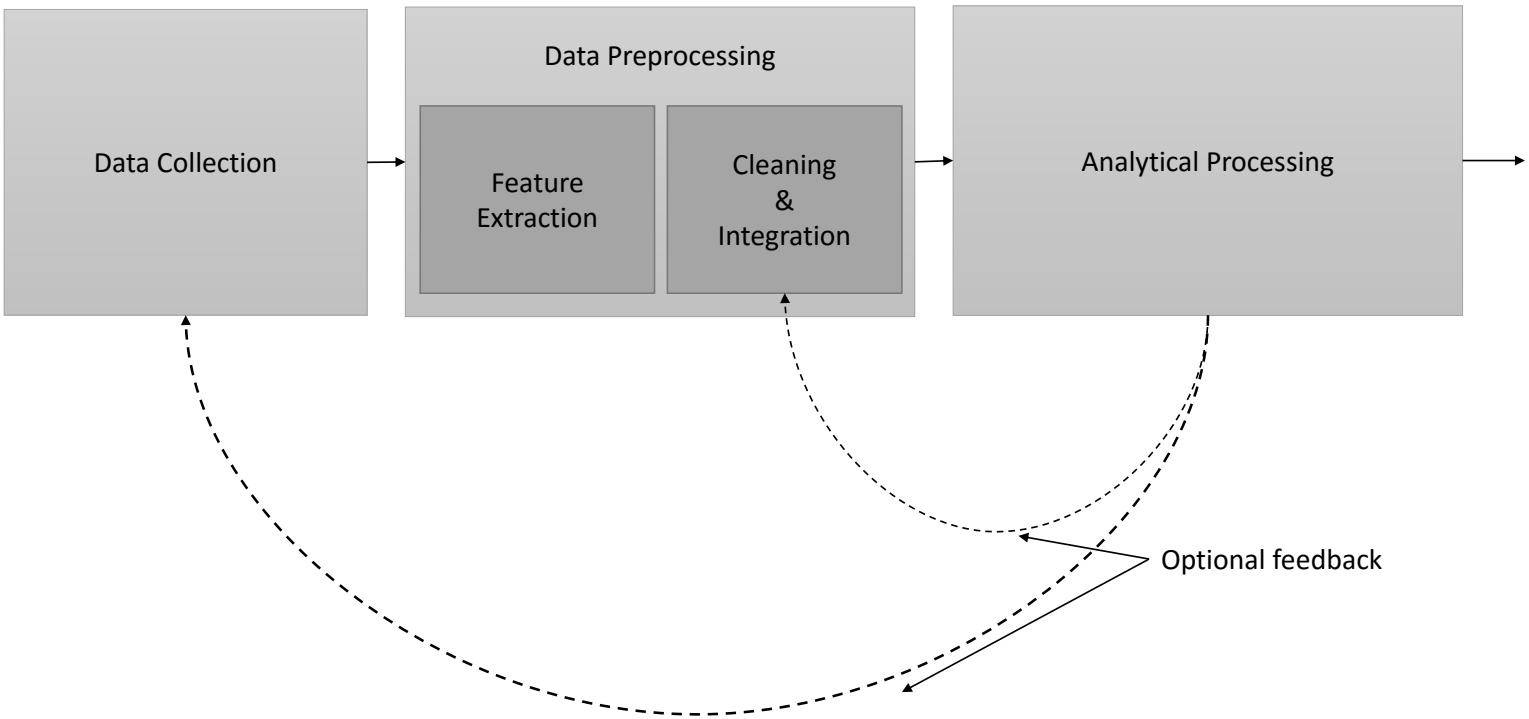
- Aggarwal: "Data mining is the study of collecting, cleaning, processing, analyzing and gaining useful insides from the data."
- Williams: "Data mining is the art and science of intelligent data analysis."







The Data Mining Process



Attribute, Feature, Dimensionality

- **Widely used explanation** Different measured properties of the process are referred as *features*, *attributes* or *dimensions*.
- In order to avoid confusion, here and after, single measured property of the process will be referred as *attribute*, sets or tuples of attributes will be referred as *features*. Note! That feature may contain just one attribute therefore attribute is always a feature but not vice verse! *Dimensionality* is the property of the process describing number of attributes.

Data Types

- **Nondependency-oriented data** The simplest form of data usually refers to multidimensional data.
  - Quantitative multidimensional data
  - Binary and set data
  - Text data
- **Dependency-oriented data**
  - Time-series data
  - Discrete sequences and strings
  - Spatial data
  - Network and graph data

Nondependency-oriented data / multidimensional data

Definition (1)

Multidimensional Data: *A multidimensional data set $\mathcal{D}$ is a set of n records, $\bar{X}_1 \ldots \bar{X}_n$, such that each record $\bar{X}_i$ contains a set of features denoted $(x_i^1 \ldots x_i^d)$.*

- **Quantitative multidimensional data.** If each element x_i^j in Definition 1 is quantitative, then corresponding data set $\mathcal{D}$ is referred as *quantitative multidimensional* data.
- **Categorical and mixed attribute data.** If each element x_i^j in Definition 1 categorical (unordered discrete), then corresponding data set $\mathcal{D}$ is referred as *unordered discrete-valued* or *categorical*
- **Binary and set data.** may be considered as a special case of either multidimensional categorical data (each attribute may take only one of two values) or multidimensional quantitative data (ordering exists between two values).
- **Text data** belong to the dependency oriented data types but its vector-space representation (words correspond to attributes and their frequencies to the values of these attributes).

Dependency-Oriented Data

It is assumed that at least between two records of the data set explicit or implicit relations may exist.

- **Time-series data**

Definition

A time series of length n and dimensionality d is a $n \times d$ matrix Y, where each string corresponds to the certain time instance and each row corresponds to a certain numeric feature.

- **Discrete sequences and strings.** Categorical analog of time-series data. Each element of the matrix Y may take discrete or categorical value.

Dependency-Oriented Data

- **Spatial data**

Definition

Spatial data: *A d - dimensional spatial data record contains d behavioral attributes and one or more contextual attributes containing the spatial location. d- dimensional spatial data set is a set of d -dimensional records $\bar{X}_1 \dots \bar{X}_n$, together with the set of locations $L_1 \dots L_n$, such that the record $\bar{X}_i$ is associated with the location L_i.*
Important subclass spatiotemporal data.

- **Network and graph data**

Definition

A network $G = (N, A)$ is defined by the set of nodes N and a set of edges A, where edges represent th relationships between the nodes.

In some cases an attribute sets $\bar{X}_i$ and $\bar{Y}_{i,j}$ may be associated with node i and edge i, j correspondingly.

Problems (major building blocks)

- **Association Pattern Mining** Frequent pattern mining

Definition

Given a binary $n \times d$ data matrix $\mathcal{D}$, determine all subsets of columns such that all values in this columns take on the value of 1 for at least a fraction s of the rows in the matrix.

- **Data Clustering**

Definition

Given a data matrix or database $\mathcal{D}$, partition its rows (records) into sets $C_1 \ldots C_k$ such the rows (records) in each cluster are similar to one another.

- **Outlier Detection**

Definition

Given a data matrix $\mathcal{D}$, determine the rows that are "very" different from the remaining rows of the matrix

Problems (major building blocks)

- **Data Classification**

Definition

Given an $n \times d$ training data matrix (database) $\mathcal{D}$, and a class label volume in $\{1, \ldots, k\}$ associated with each of the n rows (records in case of database) in $\mathcal{D}$, create a training model $\mathcal{M}$ which can be used to predict the class label of a d dimensional row (record) $\bar{Y} \notin \mathcal{D}$.

- **Complex Data Types**
- **Scalability Issues**

Data Preparation

- **Feature Extraction.** Feature Extraction, Conversion between different data types e.g. Numeric to Categorical data or Categorical to Numeric data
- **Data Cleaning.** Handling missing entries, handling incorrect and inconsistent entries, scaling and normalization.
- **Data Reduction and Transformation.** Sampling, feature subset selection dimensionality reduction.

Feature Extraction

Feature extraction is the process of selecting the attributes and features relevant to the goal of analysis.

- Sensor data: Fourier transform, time series cleaning
- Image data: Pixels, color histograms, *visual words*. Main challenge is the high dimensionality of the data.
- Web logs: convert web logs into multidimensional representation of the categorical or numeric format.
- Network traffic.
- Document data.

Conversion

- **Discretization:** Numeric to Categorical Data: Divide range of numeric attribute into finite number of intervals. To each data point assign categorical value of the interval containing its numerical attribute.
  - Equi-width ranges: the ranges have the same length.
  - Equi-log ranges: $\log(b) - \log(a)$ have the same length for all the intervals. Here a is a beginning and b is the ending of the intervals.
  - Equi-depth ranges: Each range contains the same number of intervals.
- **Biniarization:** Categorical to Numeric Data
- Text to Numeric Data
- Time Series to Discrete Sequence Data
- Time series to Numeric Data
- Discrete Sequence to Numeric Data
- Spatial to Numeric Data
- Graphs to Numeric Data

Data Cleaning

- Missing Entries
- Incorrect and Inconsistent Entries
- Scaling and Normalization: Different features represent different scales and not always comparable.
  - Normalization Let j^{th} attribute has mean μ_j and standard deviation σ_j then j^{th} attribute value x_i^j of the record $\bar{X}_i$ may be normalized as follows

$$z_i^j = \frac{x_i^j - \mu_j}{\sigma_j} \tag{1}$$

  - Min - max scaling:

$$y_i^j = \frac{x_i^j - \min(x^j)}{\max(x^j) - \min(x^j)} \tag{2}$$

Data Reduction and Transformation

- Sampling
  - Sampling for Static data
    - ★ Biased sampling
    - ★ Stratified sampling
  - Reservoir Sampling for Data Streams
- Feature Subset Selection
- Dimensionality Reduction
  - Principal Component Analysis
  - Singular Value Decomposition
  - Latent Semantic Analysis

Principal Component Analysis

Problem: Significant number of correlations may exist between different attributes. Usually used after the mean centering (subtracting the mean of the data set from each data point). The goal of PCA is to rotate the data into a coordinate system where the greatest amount of variance is captured in a smaller number of dimensions.
Let $\mathcal{D}$ be $n \times d$ data matrix and $\mathcal{C}$ $d \times d$ covariance matrix. Each element c_{ij} of the matrix $\mathcal{C}$ is the covariance between the columns i and j of matrix $\mathcal{D}$

$$c_{ij} = \frac{\sum_{k=1}^{n} x_k^i x_k^j}{n} - \mu_i \mu_j \qquad \forall i, j \in \{1 \dots d\} \tag{3}$$

Let $\bar{\mu} = (\mu_1 \dots \mu_d)$ then

$$\mathcal{C} = \frac{\mathcal{D}^T \mathcal{C}}{n} - \bar{\mu}^T \bar{\mu} \tag{4}$$

Principal Component Analysis

The covariance matrix $\mathcal{C}$ is positive semi-definite

$$\bar{v}^T \mathcal{C} \bar{v} = \frac{\left(\mathcal{D}\bar{v}\right)^T \mathcal{D}\bar{v}}{n} - (\bar{\mu}\bar{v})^2 \tag{5}$$

which is equal to the variance of 1 -dimensional points in $\mathcal{D}\bar{v} \geq 0$. PCA allows to determine orthonormal vectors $\bar{v}$ maximizing $\bar{v}^T \mathcal{C} \bar{v}$. Since $\mathcal{C}$ is positive semi-definite

$$\mathcal{C} = P \Lambda P^T \tag{6}$$

P contains orthonormal eigenvectors of $\mathcal{C}$ and diagonal matrix Λ - corresponding nonnegative eigenvalues.

Principal Component Analysis

- Both eigenvectors and eigenvalues have a geometric interpretation.
- It may be shown that $\binom{d}{2}$ covariances of transformed features are zero.
- Matrix Λ is the covariance matrix after axis rotations.
- Eigenvectors with large eigenvalues preserve greater variance and referred as principal components
- Transformed data matrix is computed as follows

$$\mathcal{D}' = \mathcal{D}P \tag{7}$$

Do you need to refresh your knowledge of math?

Positive semi-definite matrix? orthogonal vectors? Eigenvalues and eigenvectors? If any of these notions cause a problem please refresh your knowledge of algebra.