# Data Mining Lab Course SoSe 2023

## Data Mining Basics

L. Richter

School of CIT
Technische Universität München

Wednesday, May 17th

ROSTLAB.

# Overview



ROSTLAB.

# Schedule

**Table:** Intended Schedule

| Date | Topic | Date | Topic |
|---|---|---|---|
| Apr 26th | Kick-off | Jun 14th | Descriptive Mining 5 ? |
| May 3rd | Data Set Presentation | Jun 21st | Predictive Mining 1 |
| May 10th | Data Set Selection | Jun 28th | Predictive Mining 2 |
| | Group Formation/EDA1 | | |
| ▷May 17h | Descriptive Mining 2 | Jul 5th | Predictive Mining 3 |
| May 24th | Descriptive Mining 3 | Jul 12th | Final Presentation 1 |
| May 31st | Descriptive Mining 4 | Jul 19th | Final Presentation 2 |
| Jun 7th | Descriptive Mining 5 | | |

ROSTLAB.

# From Observations to Data

- an object/observation is described by variable values that corresponds to certain properties of the object
- this is an encoding and comes already with a(n unavoidable) loss of information
- in our case these variables are called attributes or features
- for example: people can be described by height, body weight, gender, age, a.s.f.
- the set of feature values describing one object/observation is called an instance
- the set of features is called feature space
- the encoding can be heavily influenced and can be improved by domain knowledge

ROSTLAB.

# Attribute Types

- we can distinguish different feature types:
  - boolean/binary
  - nominal
  - ordinal
  - integer
  - interval-scaled
  - ratio-scaled

ROSTLAB.

# Categorial Types

- boolean: see below
- nominal: a variable to put an object into categories: like color, gender, profession, a.s.f. It might come in numerical form, but allows NO arithmetic operations! Binary attributes can be seen as a special case with only the categories true/false, male/female, passed/failed, a.s.f. for example.
- ordinal: nominal variables with an order relationship, like small, medium, large or new borne, infant, pupil, student, adult
- this is often indicated by a data set's codebook

# Continuous Types

- integer: unlike true nominal variables arithmetic is meaningful, even if they have only discrete values, e.g. number of children
- interval-scaled: this is a variable that takes numerical values which are measured at equal intervals from an arbitrary origin. An example is the temperature in °C. A value of 0 does not necessarily mean the absence of temperature! You can define an order on theses values.
- ratio-scaled: these are similar to interval-scaled variables, but 0 means an absence of the property. Weight or size is an example for this. A value of 0 means not existing.

ROSTLAB.

A few fundamental considerations for all kinds of preprocessing:

- lost information can never be recovered
- what is the aim of the preprocessing?
- to which extent are my strategies biased by my skills?
- (if I have a hammer, everything is a nail. ;-))

# Instance Related Issues

- data used for data mining is usually not the result of an dedicated experiment but a by-product of other activities
- data can be assembled from different experiments/sources/periods, e.g. the layout may differ
- data inspection/visualization gives a quick, first impression
- because data can be noisy, faulty or missing

# Erroneous Data

- noisy: non-consistent instances, i.e. instances disagree in certain features, typically the labels
- faulty: recorded values do not match the feature type or are wrong, due to input errors or merging
- outliers: true exceptions or input typos? this may depend on the context of your work
- missing values: some features are not applicable or were not recorded at this time
- strategies (without any priority or preference): correct, exclude, ignore or impute

# Feature Related Issues and Strategies

Since the amount of possible transformations is just too high for brute force exploration you should try to get some hints from direct inspection:

- what is the type and distribution of an attribute (for all types)
- for nominal attribute: how many values, value frequencies?
- for numerical attributes: uniform or normal distributed or different?
- are there unexpected concentrations?
- are there unusual values, e.g. missing values are coded by a special value ($\rightarrow$ codebook)

ROSTLAB.

# Dependencies and Redundancies

- are the attributes independent
- determine correlation resp. the similarity between attributes (Pearson, Manhattan, Cosine, Tanimoto, a.s.f)
- check the codebook for synonymous or hierarchical attributes (e.g. address and latitude/longitude, a.s.f.)
- if you consider features as target variables: dual or multi-class problems, what is the class distribution, i.e. the partition sizes?
- do not become biased by pre-specified class labels

ROSTLAB.

# Task of the week

- Structure the wiki for your group
- Get the data
- Start with the EDA