UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Knowledge Discovery in Databases

Erich Schubert, Michael Gertz

Winter Semester 2017/18

Part II

Foundations: Data, Probability, and Statistics

Structured Data

For data analysis, we often require structured data:

- Collection of data objects
- Objects are described by their attributes
- Attributes have a specific data type
- Distinguish: *attribute* (feature, dimension, variable) versus *attribute value* (characteristics)
  - one attribute $\longrightarrow$ multiple attribute values
    e.g., currency in € or $; height in meter, centimeter, or feet?
  - different attributes $\longrightarrow$ same attribute value
    e.g., customerID, age, count: `int`

Attribute Types

Nominal attributes (categorical attributes)
E.g., a category
- No order
- No arithmetics
- $=$ and $\neq$

Special cases:
- **binary**: only two categories, e.g., employed / unemployed
- **identifier**: unique ID, e.g., customer number

Ordinal attributes
Ordered non-numeric data
E.g., "high" / "medium" / "low"
- Ordered
- No arithmetics
- $<, >, =, \neq$

Special cases:
- **Likert-type scale** data: strongly disagree, disagree, neutral, agree, strongly agree

Numeric attributes
Quantitative / measureable
E.g., price, shoe size
- Ordered
- Arithmetics: $+, -$, etc.
- $<, >, =, \neq$

Special cases:
- **interval-scaled**: measured on a scale with equal-size units (e.g., date, time)
- **ratio-scaled**: has an inherent zero point; a value is a multiple of another value (e.g., size, price)

Further Attribute Types

First proposed by Stevens [Ste46], but criticised (e.g., [VW93]), and extended [MT77]:

- Names (Categorical attributes)
- Grades (Ordinal attributes)
- Counted fractions bound by 0 and 1 (including percentage points)
- Counts (Non-negative integers)
- Amounts (Non-negative real numbers)
- Balances (any real number)

These types may require different handling.

Beware: often we can interpret an attribute in multiple ways!
For a detailed discussion, see [Han96].

Nominal Attributes

Often categories (A/B/C), but *can* be numeric (zip codes!) or binary.

Numbers do not reflect a quantity (e.g., user number)

Appropriate statistics:

- Frequency counts
- Mode (most frequent value)
- Frequency tests, such as χ^2 test

Be careful: sometimes encoded with integers (1=red, 2=blue, …)

Binary variables *can often* be considered ordinal (e.g., customer $>$ no customer)

Ordinal Attributes

Values with defined order (High $>$ Medium $>$ Low; "on a scale of $1 \ldots 5$" questions).

But no meaningful arithmetic: High $-$ Medium $\neq$ Medium $-$ Low

Appropriate statistics (additionally):

- Percentiles, quantiles, median
- Rank correlation (e.g., Spearman correlation)

Be careful: often encoded with numerical values!

If we sort the values, the rank can sometimes be considered ordinal.

Numeric Attributes: Interval Scale

Values where differences are comparable, *but* where zero is not special.

Example: Temperature in Celsius or Fahrenheit: $20C - 10C = 30C - 20C$ but $20C \neq 2 \cdot 10C$.

Appropriate statistics (additionally):

- Mean, Variance
- Pearson Correlation

Be careful: deciding whether differences are meaningful is not always easy.
E.g., if a text contains a word 1 or 0 times, is this the same as 101 or 100 times?

$x' = b \cdot x + c$ preserves the properties of this scale.

Numeric Attributes: Ratio Scale

Values where zero and multiples make sense.

Example: Weight, Height

Appropriate statistics (additionally):

- Geometric mean, harmonic mean
- Coefficient of variation

Be careful: $x' = b \cdot x + c$ *no longer* preserves these properties, except $c = 0$!

Discrete versus Continuous Attributes

Discrete Attribute

- finite or countably infinite set of values
- data type is typically `integer`
- examples: count, zipcode, customerID, ...

Continuous Attribute

- continuous values are real numbers
- often represented by finite number of digits
- data types: `float`, `double`, `decimal`
- examples: price, weight, temperature

Discrete versus Continuous Attributes /2

> *The Census report, like most such surveys, had cost an awful lot of money and didn't tell anybody anything they didn't already know – except that every single person in the Galaxy had 2.4 legs and owned a hyena.*
>
> *Since this was clearly not true the whole thing eventually had to be scrapped.*
>
> — So Long, and Thanks For All the Fish, Douglas Adams

Record-oriented Data

Set of records with a fixed set of attributes

CustomerID	Debt	Income	Type of employment	Credit rating
1	High	High	Self-employed	poor
2	High	High	Employee	poor
3	High	Low	Employee	poor
4	Low	Low	Employee	good
5	Low	Low	Self-employed	poor
6	Low	High	Self-employed	good
7	Low	High	Employee	good

Text Data

Documents can be represented as vectors (bag-of-words model)

Term (word) represents element in vector

- 1 if term occurs in document, 0 if term does not occur in document
- Frequency of term occurrence

	DBMS	KDD	Mining	Data	Web
Document #1	3	2	0	1	0
Document #2	0	2	1	1	4
Document #3	0	0	2	4	3

Transactional Data

Special case of record-oriented data

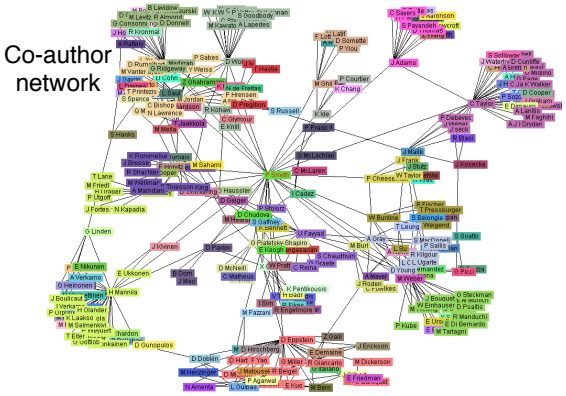
Data record corresponds to a transaction
consisting of a transaction ID and a set of elements (items)

Transaction ID	Items
1	milk, butter
2	milk, honey, butter
3	milk, bread, butter
4	milk, bread, honey

Graph Data

Data is a single graph or a collection of graphs (labeled, directed, multirelational, …)

Examples: social networks, Web link structure, molecules, …



Sequence-based Data

Ordered sequence of elements, e.g., alerts, Web log entries, gen sequence, spatio-temporal data (location of a moving object over time), …

```
194.145.89.65 - - [3/Oct/2015:14:57:21 +0200] "GET ... HTTP/1.0"
194.145.89.65 - - [3/Oct/2015:14:57:21 +0200] "GET ... HTTP/1.0"
195.36.75.26 - - [3/Oct/2015:14:58:54 +0200] "GET ... HTTP/1.0"
195.37.152.250 - - [3/Oct/2015:15:02:55 +0200] "GET ... HTTP/1.1"
195.37.152.250 - - [3/Oct/2015:15:02:55 +0200] "GET ... HTTP/1.1"
193.51.91.2 - - [3/Oct/2015:15:06:20 +0200] "GET ... HTTP/1.0"
65.54.188.64 - - [3/Oct/2015:15:07:13 +0200] "GET ... HTTP/1.0"
84.168.66.17 - - [3/Oct/2015:15:12:02 +0200] "GET ... HTTP/1.1"
84.168.66.17 - - [3/Oct/2015:15:12:08 +0200] "GET ... HTTP/1.1"
68.142.251.148 - - [3/Oct/2015:15:22:14 +0200] "GET ... HTTP/1.0"
68.142.250.20 - - [3/Oct/2015:15:22:14 +0200] "GET ... HTTP/1.0"
```

```
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
```

Data Quality

Data quality: "fitness for use"

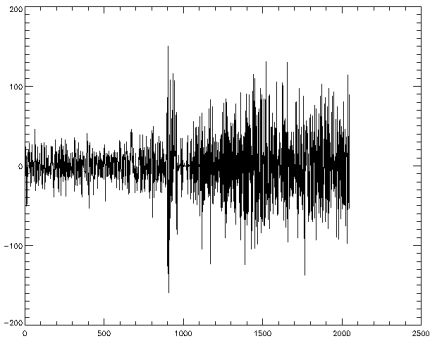
- Subjective: dependent on the context, user, ...
- Multi-dimensional: different dimensions, data characteristics, ...
- Characteristics describe data quality problems

Tasks:

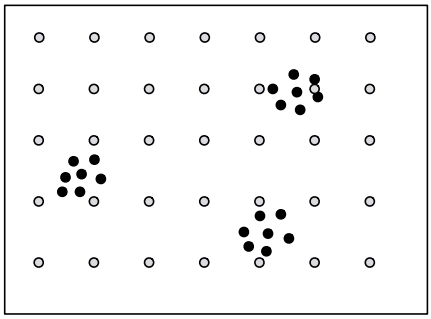
- Measuring data quality (assessment): estimate quality; improvement necessary? Cost-benefit analysis after improvements
- Data Cleaning: detection and removal of inconsistencies, contradictions, and errors in data with the goal to improve data quality

Data Quality Problems: Noise

Noise: random error or variance in measured variable



Sensor data



Charge-Coupled Device

Data Quality Problems: Missing Values

Missing data at different levels of object description

- Instance level: values, data record, part of a relation, ...
- Schema level: attribute, ...

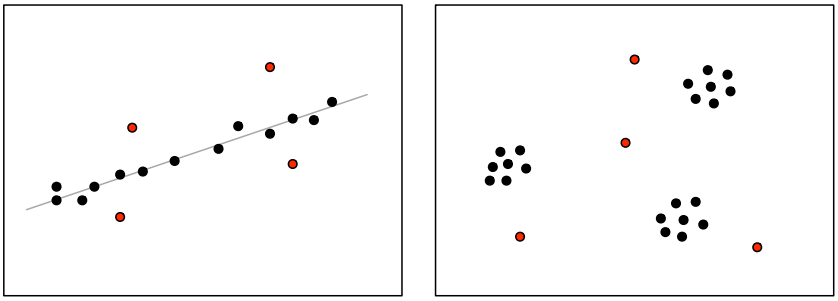
Problems specific to the instance level:

- Handling of null values: missing values, default value, value not applicable/meaningful?
- Refuse or replace (while maintaining distribution of data values)

CustomerID	Name	Email	
123	Leo Pren	$\perp$	e.g., no email address
125	Ann Joy	$\perp$	e.g., email address not known
126	Just Vorfan	$\perp$	...

Data Quality Problems: Outliers

Outlier: value (observation) that is numerically distant from the rest of the data



Issues:

- Detection: distribution, "geometry", time series
- Interpretation: error in measurement or real / valid observation?

Data Quality Problems: Duplicates

Duplicate: data objects that represent the same real-world object

- exact duplicate: trivial to determine using SQL

```
select distinct ... from ...
```

- but ...

CustomerID	Name	Address
3346	Just Vorfan	Hafenstraße 12
3346	Justin Forfun	Hafenstr. 12
5252	Lilo Pause	Kuhweg 42
5268	Lisa Pause	Kuhweg 42
$\perp$	Ann Joy	Domplatz 2a
$\perp$	Anne Scheu	Domplatz 28

Summary Sections 1 and 2

- Get a good understanding of the type of data (record-oriented, text, transactional, ...) you want to analyse, mine, and explore
- Investigate the attribute types of data objects and appropriate transformations
- Be aware of data quality problems. Always investigate potential data quality problems!
- All of the above are tasks to accomplish before any data mining task