Fouille de données

▷ **Data**

Philippe Lenca et Romain Billot

philippe.lenca@telecom-bretagne.eu

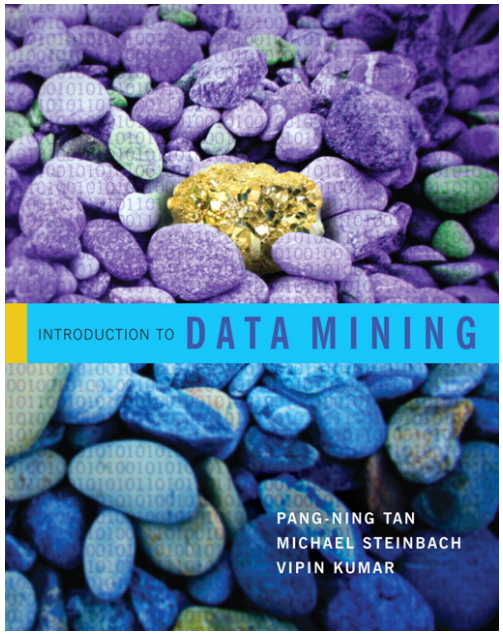
Telecom Bretagne

2015-2016

Outline

Outline

Introduction to Data Mining:
Pang-Ning Tan, Michael Steinbach, Vipin Kumar [**?**]



What is a Dataset?

A collection of data objects described by attributes

- Objects:
  the atomic elements from a dataset

  (examples, records, prototypes, objects, cases, points, samples, instances)
- Attributes:
  a property of an object

  (features, variables, fields, characteristics)

Attributes

Objects

TID	Marital status	Sex	Income	Age	Relation
id_1	single	M	100k€	22	▼
id_2	single	F	150k€	28	▲
id_3	married	F	120k€	32	▲
id_4	single	F	250k€	42	–
id_5	divorced	M	95k€	25	▼
id_6	single	M	120k€	55	▲
id_7	married	F	95k€	33	▲
id_8	divorced	M	150k€	47	▲
id_9	single	F	100k€	29	▲
id_{10}	single	M	80k€	26	–

Objects $\times$ **Attributes**

$\hookrightarrow$ Concepts: content inside the data that can be learned.

Outline

Attribute types

Attribute values are numbers or symbols assigned to an attribute

- An attribute can be mapped to different measurement scale

  Temperature can be measured in Celsius or Fahrenheit

  Age can be measured in month or year

  Height can be measured in feet or meter

- Different attributes can be mapped to the same scale

  ID and age could be integers

  Customer satisfaction on price and quality can be mapped on 'bad' 'reasonable' 'good'

$\hookrightarrow$ But properties of attribute values can be different.
Readings: [**?**], [**?**].

Attribute types

TID	Marital status	Sex	Income	Age	Relation
id_1	single	M	100k€	22	▼
id_2	single	F	150k€	28	▲
id_3	married	F	120k€	32	▲
id_4	single	F	250k€	42	–
id_5	divorced	M	95k€	25	▼
id_6	single	M	120k€	55	▲
id_7	married	F	95k€	33	▲
id_8	divorced	M	150k€	47	▲
id_9	single	F	100k€	29	▲
id_{10}	single	M	80k€	26	–

TID	Marital status	Sex	Income	Age	Relation
id_1	S	1	135k$	22	▼
id_2	S	2	202k$	28	▲
id_3	M	2	162k$	32	▲
id_4	S	2	337k$	42	–
id_5	D	1	128k$	25	▼
id_6	S	1	162k$	55	▲
id_7	M	2	128k$	33	▲
id_8	D	1	202k$	47	▲
id_9	S	2	135k$	29	▲
id_{10}	S	1	108k$	26	–

$\hookrightarrow$ Attribute types and scales of measurement have to be properly specified.

Levels of measurement

Four main levels of measurement

- nominal
- ordinal
- interval
- ratio

Levels of measurement

Nominal scale

Values are names or labels (identifiers):

- no ordering, no distance measure, no relation among values
- only equality tests and set membership can be performed
- categorical variables
- boolean as a special case

Examples

- ID numbers, ZIP codes
- color, marital status, sex

$\hookrightarrow$ Standard set structure (unordered).

Levels of measurement

Ordinal scale

Values with an order:

- no distance measure (no relative size or degree of difference)
- relation w.r.t. the order, addition does not make sense
- with a preference system one can transform an nominal scale toward an ordinal one

Examples

- rankings ('bad' $<$ 'medium' $<$ 'good')
- height ('short' $<$ 'medium' $<$ 'tall')

$\hookrightarrow$ Totally ordered set.

Levels of measurement

Interval scale

Values ordered and measured in fixed and equal units:

- zero point is not defined
- difference of two values makes sense
- sum or product does not make sense

Examples

- temperatures in Celsius or Fahrenheit
- calendar dates

$\hookrightarrow$ Affine line.

Levels of measurement

Ratio scale

Values ordered and measured on a scale with a zero point:

- estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind
- ratio quantities are treated as real numbers
- all mathematical operations are allowed

Examples

- temperatures in Kelvin
- distance, time

$\hookrightarrow$ One-dimensional vector space.

Levels of measurement

Brief synthesis

		Nominal	Ordinal	Interval	Ratio
distinctness	$= \neq$	▲	▲	▲	▲
order	$< >$		▲	▲	▲
addition	$+ -$			▲	▲
multiplication	$\times /$				▲

$\hookrightarrow$ Most schemes consider just two or three levels of measurement: nominal, ordinal and numerical. See also which scales are considered by the software.

Levels of measurement

Brief synthesis

Type	Description	Examples	Operations
Nominal	Distinction	ID, color, sex	mode, entropy, contingency correlation, χ^2 test
Ordinal	Order	grades	median, percentiles, rank correlation, run tests, sign tests
Interval	Differences	temperatures in C or F	mean, standard deviation, Pearson's correlation, t and F tests
Ratio	Differences and ratio	age	geometric mean, harmonic mean, percent variation

$\hookrightarrow$ Levels of measurement & legal operations.

Levels of measurement

Brief synthesis

Type	Transformation	Comments
Nominal	One to One	Any permutation of values
Ordinal	Monotonic increasing	$v_{new} = f(v_{old})$
Interval	Positive linear (affine)	$v_{new} = a \times v_{old} + b$
Ratio	Positive similarities	$v_{new} = a \times v_{old}$

$\hookrightarrow$ Levels of measurement & legal transformations.

Discrete and Continuous Attributes

Discrete Attribute

- finite or countably infinite set of values
- often represented as integer variables

Continuous Attribute

- infinite set of values
- real numbers
- in practice it can only be measured and represented using a finite number of digits (floating-point variables)

$\hookrightarrow$ Under software constraints.

Why Specifying Attribute Types?

Consistency, efficiency analysis

- check for valid values, missing values
- make adequate comparisons and legal operations
- express the best possible patterns

$\hookrightarrow$ Under software constraints.

Attribute roles

Main roles

- id
- predictive
- label, class
- descriptive

↪ Under software constraints.

Outline

Dataset types

Three main dataset types

- record e.g
  - data matrix
  - document data
  - transaction Data
- graph e.g.
  - World Wide Web
  - molecular structures
- ordered e.g.
  - spatial data
  - temporal data
  - sequential data
  - genetic Sequence data

Dataset types

Record data: tables, matrix.

A collection of records described with a fixed set of attributes.

TID	Marital status	Sex	Income	Age	Relation
id_1	single	M	100k€	22	▼
id_2	single	F	150k€	28	▲
id_3	married	F	120k€	32	▲
id_4	single	F	250k€	42	–
id_5	divorced	M	95k€	25	▼
id_6	single	M	120k€	55	▲
id_7	married	F	95k€	33	▲
id_8	divorced	M	150k€	47	▲
id_9	single	F	100k€	29	▲
id_{10}	single	M	80k€	26	–

Dataset types

Record data: document data

Each document becomes a 'term' vector:

- each term is a component (attribute) of the vector
- the value of each component is the number of times the corresponding term occurs in the document

ID	processus	décision	données
D_1	2	1	4
D_2	1	2	1
D_3	1	3	2

Dataset types

Record data: transaction data

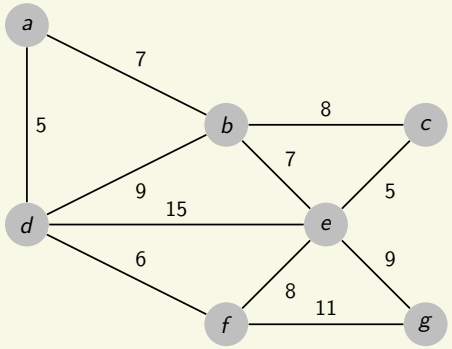
- each record (transaction) involves a set of items

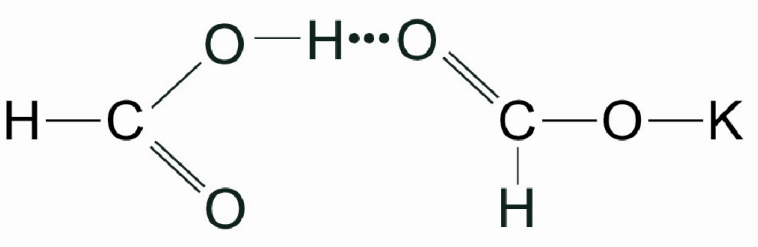
TID	Items
$\mathtt{id}_1$	Bread, Coke, Milk
$\mathtt{id}_2$	Bread, Coke
$\mathtt{id}_3$	Coke, Milk
$\mathtt{id}_4$	Beer, Bread, Diaper, Milk
...	...

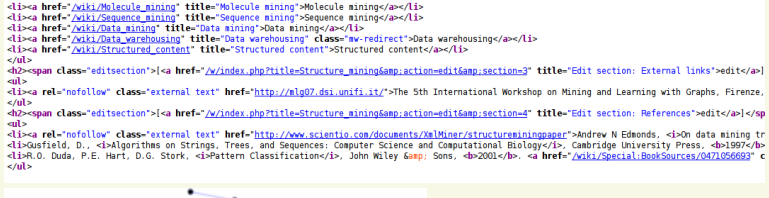
Dataset types

Graph data

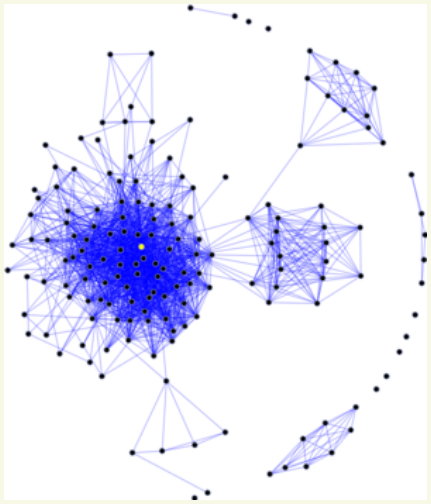
Generic graph.







```
<li><a href="/wiki/Molecule_mining" title="Molecule mining">Molecule mining</a></li>
<li><a href="/wiki/Sequence_mining" title="Sequence mining">Sequence mining</a></li>
<li><a href="/wiki/Data_mining" title="Data mining">Data mining</a></li>
<li><a href="/wiki/Data_warehousing" title="Data warehousing" class="mw-redirect">Data warehousing</a></li>
<li><a href="/wiki/Structured_content" title="Structured content">Structured content</a></li>
</ul>
<h2><span class="editsection">[<a href="/w/index.php?title=Structure_mining&amp;action=edit&amp;section=3" title="Edit section: External links">edit</a>]
<ul>
<li><a rel="nofollow" class="external text" href="http://mlg07.dsi.unifi.it/">The 5th International Workshop on Mining and Learning with Graphs, Firenze,
</ul>
<h2><span class="editsection">[<a href="/w/index.php?title=Structure_mining&amp;action=edit&amp;section=4" title="Edit section: References">edit</a>]</sp
<ul>
<li><a rel="nofollow" class="external text" href="http://www.scientio.com/documents/XmlMiner/structureminingpaper">Andrew N Edmonds, <i>On data mining tr
<li>Gusfield, D., <i>Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology</i>, Cambridge University Press, <b>1997</b>
<li>R.O. Duda, P.E. Hart, D.G. Stork, <i>Pattern Classification</i>, John Wiley &amp; Sons, <b>2001</b>. <a href="/wiki/Special:BookSources/0471056693" c
</ul>
```

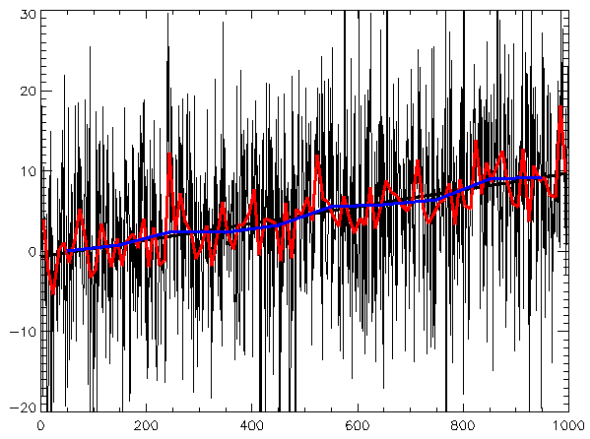


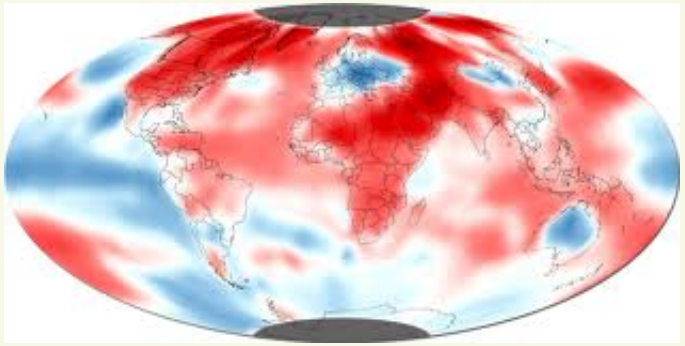
Dataset types

Ordered data

(CE) (A) (AE)
(BD) (C) (E)
(AB) (D) (CE)

AAATCGGACGCCGGGCTATA
CAACCGTACCCCGGGCTATA
GGTCCGTACCCCGGGCTAAA



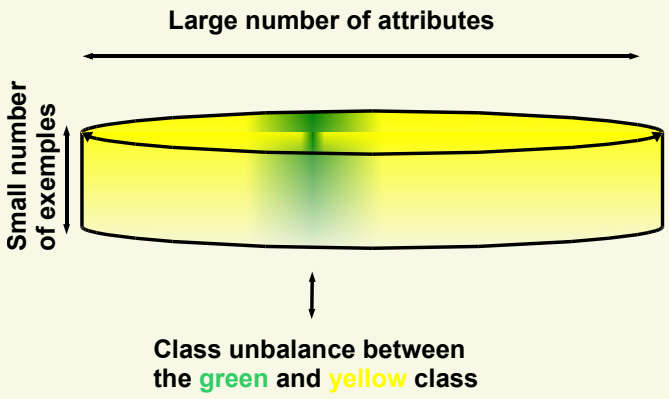


Outline

Dataset characteristics

Main characteristics

- dimensionality: number of attributes
- class label unbalance: prior class probabilities
- size: number of samples
- quality: missing values, outliers, precision
- sparsity: presence of attributes



Dataset characteristics

Main troubles

- dimensionality: curse of dimensionality, most existing algorithms only work well on data in large quantity, with a reasonable number of attributes.
- class label balance: high-unbalanced data (large differences in prior class probabilities) & bad performance of classifiers for the minority class.
- size: robustness, complexity issues,
- quality: quality of the outputs
- sparsity: presence of attributes, mining of associations among attribute sets only works if they frequently co-occur

$\hookrightarrow$ Challenges to the data mining community, especially when these characteristics are mixed together.

Dataset characteristics

Main characteristics

- dimensionality: number of attributes
- class label unbalance: prior class probabilities
- size: number of samples
- quality: missing values, outliers, precision
- sparsity: presence of attributes

$\hookrightarrow$ Challenges to the data mining community, especially when these characteristics are mixed together.
A first steep of the analysis is to take a 'picture' of the dataset.

Outline

Bibliographie