Machine Learning and Data Mining

The Data

Claudio Sartori

DISI
Department of Computer Science and Engineering – University of Bologna, Italy
claudio.sartori@unibo.it

Issues on data

- Type
  - Quantitative, qualitative, structured, unstructured, ...
- Quality
  - Data are never perfect
    - Missing, inconsistent, duplicated, wrong
  - Outliers
    - Small amount of data which are different from the rest, due to anomalies or errors
  - Some ML techniques are more robust w.r.t. errors than others
  - Better data quality $\Rightarrow$ better results
    - trivial: garbage–in–garbage–out

- Pre-processing activities transform data to ease the ML activities

Some examples of datasets

UCI Machine Learning Repository

- Iris
- Adult
- Breast Cancer
- Wine Quality
- Car evaluation

Explore them by yourself

Data Types and numerical properties

[Tan et al.(2006)Tan, Steinbach, and Kumar, Stevens(1946)]

Data Type		*Description*	*Examples*	*Descriptive statistics allowed*
Categorical	Nominal	The values are a set of labels, the available information allows to distinguish a label from another Operators: $=$ and $\neq$	zip code, eye color, sex, ...	mode, entropy, contingency, correlation, χ^2 test
	Ordinal	The values provide enough information for a total ordering Operators: $<>\leqslant\geqslant$	hardness of minerals, non–numerical quality evaluations (bad, fair, good, excellent)	median, percentiles, rank correlations
Numerical	Interval	The difference is meaningful Operators: $+-$	Calendar dates, temperatures in centigrades and Fahrenheit	average, standard deviation, Pearson's correlation, F and t tests
	Ratio	Have a univocal definition of 0 Allow all the mathematic operations on numbers	Kelvin temperatures, masses, length, counts	geometric mean, harmonic mean, percentage variation

The "description" and "descriptive statistics" columns are *incremental*, i.e. the properties described in a row are added to the properties described in the rows above

Interval data

Interval data is commonly used in statistical research, academic assessment, scientific studies, and probability calculations. Here are some examples with thorough explanations to help you understand how they are used in various industries:

- Temperature
- Scores
- Time
- IQ Test
- CGPA

Focus on *interval data* I

- *Temperature* - The best examples of interval scales. Positive and negative readings are allowed. Zero is an arbitrary value, hnce, 0°C and 0°F indicate different temperatures. Given two temperatures $t_1 = 10^{\circ}C$ and $t_2 = 12^{\circ}C$ it does not have any sense to say that their relative difference is 20%, changing the scale the relative difference would change
- *Scores* - When grading test scores, such as the SAT, the digits 0 to 200 are not used when scaling the raw score to the section score. Absolute zero is not employed as a reference point in this scenario. As a result, it is interval data.
- *Time* - Every hour in a clock is separated by an equal distance of 60 minutes; not more, not less, perfectly equal, but the zero is arbitrary and changes with time zones. This is why clocks are referred to as interval scales since they are equidistant and measurable.

Interval vs ratio

Interval does not preserve relative values upon scale change

	2021	**2022**	**Var%**
Revenues $	204000.00	217000.00	6.37%
Revenues Euro	192452.83	204716.98	6.37%

1 Euro	$ 1.06

	2021	**2022**	**Var%**
AVG Temp C	16	17	6.25%
AVG Temp F	60.8	62.6	2.96%

Farenheit	Centigrades*9/5+32

Discuss the types of data in the columns

Patient	Treatment	Treatment Day	Temperature	Pain
XXXX	a	2	37	3
XXXX	a	3	37	2
XXXX	a	4	36.5	1
YYYY	b	1	38	3
YYYY	b	2	37.5	2

Example 1

Patient	Weight	BirtyYear	Age	Sex
XXXX	78	1970	50	M
YYYY	56	1980	40	F

Example 2

https://app.wooclap.com/AANACK

Allowed transformations

Data Type		*Transformation*	*Comment*	
Categorical	Nominal	Any one–to–one correspondence	the SSN can be arbitrarily reassigned (masking)	
	Ordinal	Any order preserving transformation new $\leftarrow f(\text{old})$ where f is a monotonic function	(bad, fair, good, excellent) can be substituted by (1,2,3,4)	
Numerical	Interval	Linear functions new $\leftarrow a + b * \text{old}$	centigrades and Fahrenheit temperatures can be converted either way	
	Ratio	Allow any mathematical function, *standardization*, variation in percentage	Kelvin temperatures, masses, length, counts	

The transformations above do not change the meaning of the attribute

- e.g. Kelvin temperature increments can be expressed with percentages, since they have a physical meaning related to energy
- Centigrade temperatures can be linearly transformed into Fahrenheit, but percentage increments do not have any physical meaning, since the zero level is arbitrary

Why should we transform them?

Number of values

- Discrete domains
  - allow a finite number of values (or infinitely countable)
    - codes, counts, ...
  - special case: binary attributes
  - special case: identifier
    - useful for data manipulation, not for analysis
- Continuous domains
  - floating point variables
- nominals and ordinals are discrete, possibly binary
- intervals and ratio are continuous (possibly with approximation)
- counts are discrete and ratio

Discuss the number of values in the columns

Patient	Treatment	Treatment Day	Temperature	Pain
XXXX	a	2	37	3
XXXX	a	3	37	2
XXXX	a	4	36.5	1
YYYY	b	1	38	3
YYYY	b	2	37.5	2

Example 1

Patient	Weight	BirtyYear	Age	Sex
XXXX	78	1970	50	M
YYYY	56	1980	40	F

Example 2

https://app.wooclap.com/FJJYMF

Asymmetric attributes

- Only presence is considered important (a non null value)
  - e.g. a student record with one attribute per offered exam
    - only passed exams are interesting, in general the exams not passed will be in much greater number, and do not carry much information
- In particular, binary asymmetric attributes are relevant in the discovery of association rules

General characteristics of data sets

- Dimensionality
  - the difference between having a small or a large (hundreds, thousands, ...) of attribute is also qualitative
    - see the curse of dimensionality, later
- Sparsity
  - when there are many zeros or nulls
- Beware the nulls in disguise
  - a widespread bad habit is to store zero or some special value when a piece information is not available
- Resolution
  - has a great influence on the results
    - the analysis of too detailed data can be affected by noise
    - the analysis of too general data can hide interesting patterns

Record data

- Tables
  - e.g. relational
- Transaction
  - a row is composed by: TID + set of Items
- Data matrix
  - numeric values of the same type
  - a row is a point in a vector space
- Sparse data matrix
  - asymmetric values of the same type

Relational table

The set of attributes is the same for all the records

Tid	*Refund*	*Marital Status*	*Taxable Income*	*Cheat*
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

https://wooclap.com/QZSXEY

Data matrix

- Numeric attributes
- Each row is a point in a vector space
- N rows and D dimensions (attributes, columns, properties)

Projection of x load	*Projection of y load*	*Distance*	*Load*	*Thickness*
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document representation

- Each row represents a document
- Each column represents a term
- Each cell contains the absolute frequency of the term in the document
  - the sequence of terms is lost

	team	coach	play	ball	score	game	won	lost	timeout	season
doc1	3	0	5	0	2	6	0	2	0	2
doc2	0	7	0	2	1	0	0	3	0	0
doc3	0	1	0	0	1	2	2	0	3	0

Transactional data

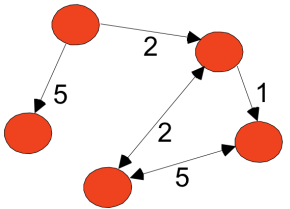
- Each record contains a set of objects
  - strictly speaking it isn't a relational table
- The reference example is the *market basket*
  - a commercial transaction

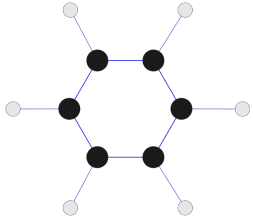
TID	*Items*
1	bread, coke, milk
2	beer, bread
3	beer, coke, diaper, milk
4	beer, bread, diaper, milk
5	coke, diaper, milk

Graph data

- Web pages
- Set of nodes and (oriented) arcs
- Molecular structures

```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```





Ordered data

- Spatial
- Temporal
- Sequence
  - of events, objects, ...
- Genetic bases

```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

Reality is complicated

- texts
- images
- moving images
- ...

Data Quality

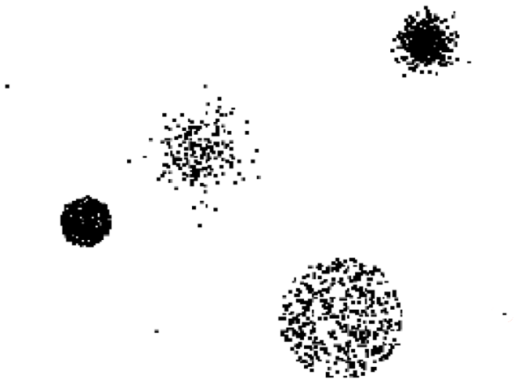
- Which are the problems?
- How can we detect the problems?
- What can we do about these problems?
- Examples
  - noise and outliers
  - missing values
  - duplicates
  - inconsistencies

Noise

- Modification of original values
- Uninteresting mixed to the interesting data
  - noise in transmission
  - web crawler accesses mixed to the human accesses

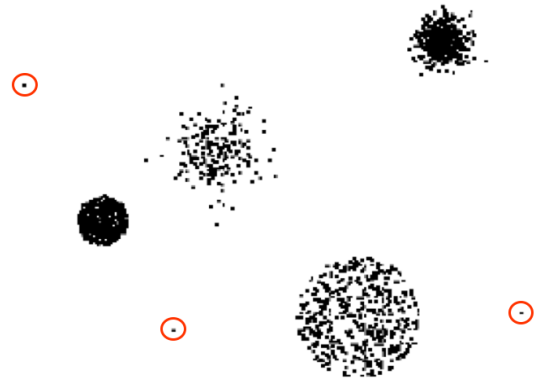
Outliers - I

- Data whose characteristics are considerably different from most of the data in the dataset
- Can be generated by
  - noise
  - rare events/processes



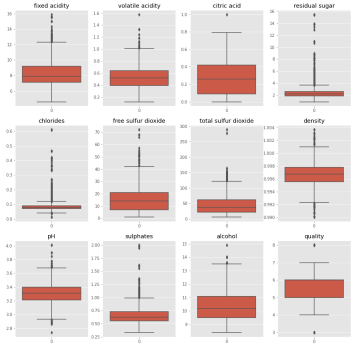
Outliers - II

- Data whose characteristics are considerably different from most of the data in the dataset
- Can be generated by
  - noise
  - rare events/processes



How to detect outliers: descriptive statistics

- IQR - InterQuartile Range
  - $Q1$: first quartile, $Q3$: third quartile, $IQR = Q3 - Q1$
- Lower boundary $= Q1 - IQR * 1.5$
- Upper boundary $= Q3 + IQR * 1.5$
- consider outlier the values out of the whiskers
  - the upper whisker will extend to last datum less than Q3 + 1.5*IQR



Boxplot of the Wine dataset

Missing values

- Reasons
  - data were not collected
    - e.g. persons are reluctant to communicate income or weight
  - the information is not applicable
    - e.g. children do not have a working annual income
- Management of missing values
  - do not consider objects with missing values
    - sometimes not a good idea
  - estimate/default
  - ignore
    - cannot be done for all the learning schemes
  - insert all the possible values, weighted with probabilities

Duplicated data

- Data objects that are duplicates, or almost duplicated
  - major issue when merging data from different sources
- Data cleaning
  - the (difficult) process of dealing with duplicated/inconsistent data

Bibliography I

- S. S. Stevens.
  On the theory of scales of measurement.
  Science, 103(2684):677–680, 1946.
  ISSN 0036-8075.
  doi: 10.1126/science.103.2684.677.
  URL http://science.sciencemag.org/content/103/2684/677.
- Pang-Nin Tan, Michael Steinbach, and Vipin Kumar.
  Data Mining.
  Addison Wesley, 2006.
  ISBN 0-321-32136-7.