PSI: Data and Data Preparation

Technological University Dublin City Campus

School of Computer Science

Defining the context

- What is the set of instances we are interested in?
- For example, if we want to analyse student attitudes, is it:
  - part-time students on City Campus
  - City Campus students
  - TU Dublin students
  - students in Ireland
  - students in Europe
- Data needs to be collected for the correct population or as an **unbiased sample** of the correct population
- If the data that is collected is wrong then all the analysis that follows will be invalid

Structured vs. unstructured data

- Data can be structured or unstructured
- For most methods, unstructured data must be transformed into structured data before analysis
- In this module, the term 'data' will be used as meaning 'structured data', unless indicated otherwise
- Structured data can be shown in a table

Generalized data table

	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	...	$\boldsymbol{x_p}$
i_1	x_{11}	x_{21}	...	x_{p1}
i_2	x_{12}	x_{22}	...	x_{p2}
.	.	.		.
.	.	.		.
.	.	.		.
i_n	x_{1n}	x_{2n}	...	x_{pn}

Data table: terminology in statistics (S), databases (DB) and machine learning (ML)

S: Variables ('independent', 'dependent')
DB: Columns, attributes
ML: Features (some of which are target ones)

S: Observations
DB: Rows,
records,
examples
ML: Instances

The terms are used across discipline boundaries as well!

	x_1	x_2	...	x_p
i_1	x_{11}	x_{21}	...	x_{p1}
i_2	x_{12}	x_{22}	...	x_{p2}
.	.	.		.
.	.	.		.
.	.	.		.
i_n	x_{1n}	x_{2n}	...	x_{pn}

Observed values, values

'Independent' vs. 'dependent': Let's say we are looking at happiness. Respondents are asked to record how they feel every day for a year, because we are studying seasonal variations and the effect of sunlight on the feeling of happiness. The independent variables here are 'amount of sunlight' (which we can obtain from meteorological records) and 'time of year'. The dependent variable is 'happiness'. However, the two 'independent' variables are not really independent of each other (in the summer there is more sunlight). *'Independent' in this context simply indicates the relationship with the 'dependent' variable*, which, by the same token, may be found not to depend on the 'independent' variables at all.

Variable types

- **discrete** (e.g. natural numbers) or **continuous variables** (e.g. real numbers)
- **categorical** or **numeric**
- classification according to scale:
  - **nominal scale**: categorical variable, no ordering among the possibe values (e.g. 'telecomms industry', 'pharmaceutical industry' etc.)
  - **ordinal scale**: categorical variable, values that can be ordered but without quantification of differences (e.g. low, medium and high)
  - **interval scale**: numeric variable, intervals between values can be compared but not ratios (e.g. temperature values $5°C$, $10°C$ and $15°C$)
  - **ratio scale**: numeric variable, both intervals and ratios between values can be calculated (e.g. prices €5, €10 and €15)
- dichotomous: the variable can take only two values and if the values are 0 and 1 the variable is referred to as *binary* (e.g. 'customer has defaulted on their mortgage' or 'has not defaulted')
- variables not used in analysis (e.g. customer id)

Data table: pets for adoption

Name	Species	Breed	Gender	Age	Vaccinated
Axel	Cat	Common	M	around 4	NO
Baz	Dog	Lurcher	M	2	YES
Clodagh	Cat	Persian	F	10	YES
Dan	Cockatoo	N/A	M	5	?
Emma	Dog	Staffie	F	1.5	YES

Data preparation

- Integration of data from multiple sources
  - removal of duplicate entries
  - unit conversion
  - normalisation
- Cleaning
  - resolution of ambiguities and errors (e.g abbreviations vs. full names or non-numeric values where numeric is expected, such as 'about 4')
  - removal or correction of invalid outliers, caused by:
    - erroneous entry (typing error)
    - a mistake in measurement (e.g. a traffic counter reports a flow of 30 cars per second, which indicates a fault)
    - mixing of different measurement units (e.g. an adult's weight recorded as 55lbs may in fact be in kilograms)
  - adjustments for time-dependent effects, such as inflation

- removal of columns that are not relevant to the analysis (e.g. customer id or calibration information)
- handling missing data:
  * removal of observations
  * imputation (ranging from simple replacement with the mean to complex methods such as multiple imputation)
- recording the steps of the cleaning process, so that if needed the information is available during future analyses

HOWTO

Min-max normalisation

Sometimes it is necessary to 'shift and rescale' a variable's data to a different range from the one they are originally provided in. This is typically done when a method requires normalised data, e.g. z-score for a hypothesis test, or when joining data sets that have the same variable specified on different scales (e.g. different units of measurement). The operation is called **min-max normalisation**. The formula to apply for each value is:

$$x_{i(NEW)} = \frac{x_{i(OLD)} - x_{min(OLD)}}{x_{max(OLD)} - x_{min(OLD)}}(x_{max(NEW)} - x_{min(NEW)}) + x_{min(NEW)}$$

Example:

Let's say we have the set of values 2, 4, and 7 ($x_1 = 2$, $x_2 = 4$ and $x_3 = 7$) and are required to normalise it into the range 0-1, the new values would be:

$$x_{1(NEW)} = \frac{2-2}{7-2}(1-0) + 0 = 0, \quad x_{2(NEW)} = \frac{4-2}{7-2}(1-0) + 0 = 0.4, \quad x_{3(NEW)} = \frac{7-2}{7-2}(1-0) + 0 = 1$$

The new set is 0, 0.4 and 1.

NOTES:

- When working with samples, the minimum and maximum values need to be those known for the population, rather than for the sample. Let's say we have data for exam marks given in percentages: 45, 67, 88 and 91. We know that the minimum and maximum in this case are 0 and 100 and we use those, rather than 45 and 91, for $x_{min(OLD)}$ and $x_{max(OLD)}$.
- The same technique can be used in any case where the scaling factor and a pair of corrsespondent source and target rangeg are known, even if no minimum and maximum don't exist. For example, converting temperatures from degrees celsius to farenheit involves a new 'minimum' of 32 and an old to new range mapping of 100 to 180: $temperature(°F) = \frac{temperature(°C)}{100} \times 180 + 32$

References The pictures in this presentation were taken from the following books. The source for each picture is cited beside it.

[DSB] *Data Science for Business: What you need to know about data mining and data-analytic thinking*, by Foster Provost and Tom Fawcett, O'Reilly Media, 2013.

[MSD] *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*, by Glenn J. Myatt and Wayne P. Johnson, John Wiley & Sons, 2014.

[US] Understanding Statistics, by Graham Upton and Ian Cook, Oxford University Press, 1996.