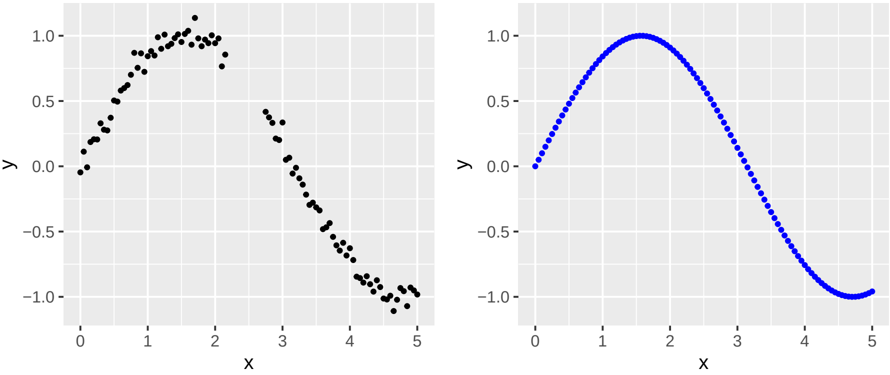
Data Preprocessing

Inês Dutra (with some material from Alípio Jorge)

October 2024

Major tasks

- Data cleaning
  - filling missing values
  - smoothing noise
  - removing outliers
  - resolving inconsistencies



Data integration

- You want to predict your customers preferences
  - customer data
  - products data
  - sales data
  - reviews
  - images from the products
  - posts on facebook
  - weather data

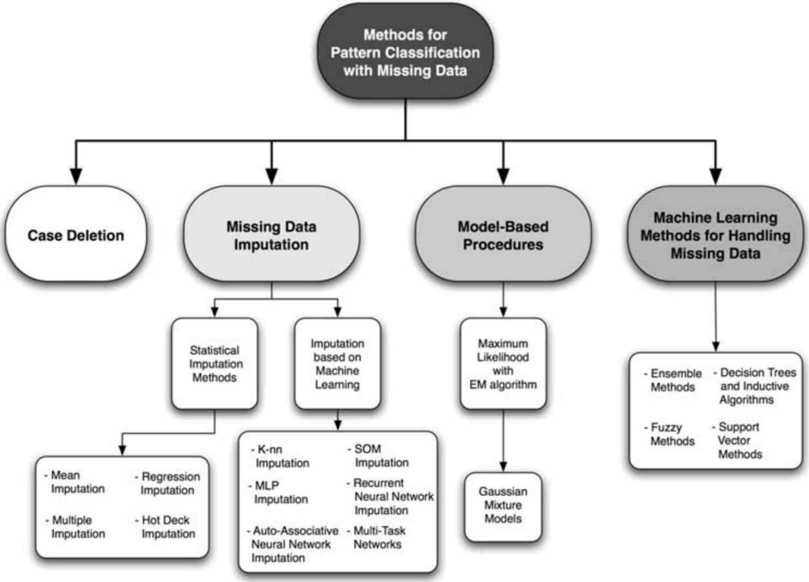
Missing Values

- **Missing Values** is perhaps the most common problem in RWD
- According to Rubin (1976), every data point has some likelihood of being missing
- The theory leads to the following categories:
  - MCAR: Missing Completey At Random, if the probability of missing is the same for each case
  - MAR: Missing At Random, if the probability of being missing is the same only within groups
  - MNAR (or NMAR): Missing Not At Random, if neither MCAR nor MAR holds

Missing Values

- What to do?
  - Nothing
  - Ignore the attribute
  - Ignore the tuple
  - Impute values (fill in)
    - (a lot to be said)

Missing Values



Missing Values

- **if** the method is robust to missing data and the amount of missing data is not too high
  - do Nothing
- **else**
  - **if** only a few cases have problems
    - ignore the cases
  - **if** the problem is on discardable attributes
    - ignore the attribute
  - **if** missing values persist
    - try value imputation

Always be **very careful** when you transform the data set

Data Imputation

Name	Age	Gender	Position	Salary
Manuel	25	M	assistant	23000
NA	36	M	manager	59000
Rui	27	M	NA	27000
Sofia	NA	F	manager	58000
Ana	48	F	CEO	77500

- Do Nothing
  - the Name column
  - Gender?

Data Imputation

Name	Age	Gender	Position	Salary
Manuel	25	M	assistant	23000
NA	36	M	manager	59000
Rui	27	M	NA	27000
Sofia	NA	F	manager	58000
Ana	48	F	CEO	77500

- Age?
- Use a global constant
  - **pro**: easy
  - **cons**: data bias, may affect inference

Data Imputation

Name	Age	Gender	Position	Salary
Manuel	25	M	assistant	23000
NA	36	M	manager	59000
Rui	27	M	NA	27000
Sofia	NA	F	manager	58000
Ana	48	F	CEO	77500

- Position, Age
- Use a measure of central tendency
  - *mean*, *median*, *mode*
  - **pros**: easy, gets the most likely value
  - **cons**: distorts the distribution
    - e.g.: average keeps average but affects variance

Data Imputation

Name	Age	Gender	Position	Salary
Manuel	25	M	assistant	23000
NA	36	M	manager	59000
Rui	27	M	NA	27000
Sofia	NA	F	manager	58000
Ana	48	F	CEO	77500

- Age, Position
- Use a measure of central tendency taken from **same group** or **same class**
  - **pros**: varied values imputed
  - **cons**: may still be too insensitive

Data Imputation

Name	Age	Gender	Position	Salary
Manuel	25	M	assistant	23000
NA	36	M	manager	59000
Rui	27	M	NA	27000
Sofia	NA	F	manager	58000
Ana	48	F	CEO	77500

- Age, Position
- Use a measure of central tendency using the **most likely** value for that case
  - e.g.: from *neighbours*, or using **linear regression**
  - **pros**: varied values imputed,
  - **cons**: needs processing, depends on distance measure and parameters

Data Imputation

Missingness Indicator Variable

One simple way to handle missingness in a variable, X_j, is to impute a value (like 0 or $\overline{X}_j$), then create a new variable, $X_{j,miss}$, that indicates this observation had a missing value. If X_j is categorical then just impute 0.

Then include both $X_{j,miss}$ and X_j as predictors in any model.

Illustration is to the right.

X_1	X_2	X_1^*	X_2^*	$X_{1,miss}$	$X_{2,miss}$
10	.	10	0	0	1
5	1	5	1	0	0
21	0	21	0	0	0
15	0	15	0	0	0
16	.	16	0	0	1
.	.	0	0	1	1
21	1	21	1	0	0
12	0	12	0	0	0
.	1	0	1	1	0

(source: https://harvard-iacs.github.io/2020-CS109A/lectures/lecture19/slides/Lecture19_Missingdata.pdf)

Noisy Data

- **Noise**
  - Random error or variance in a measured variable
- **Smoothing**
  - assume a value is always similar to neighbors
  - you **replace** values (stronger than imputation)
- **Outliers**
  - can be smoothed away if we assume they are noise
- Be very **careful**
  - do not smooth **legitimate** data (unless it helps)

Smoothing

Name	Age	Gender	Position	Salary
Manuel	25	M	assistant	23000
José	36	M	manager	59000
Rui	41	M	manager	57000
Sofia	105	F	manager	58000
Ana	28	F	assistant	28500

- Age=105 is an **outlier**
  - Binning: replace each value in group by the group mean
  - average of Age for each Position

Smoothing

Name	Age	Gender	Position	Salary
Manuel	25	M	assistant	23000
José	36	M	manager	59000
Rui	41	M	manager	57000
Sofia	105	F	manager	58000
Ana	28	F	assistant	28500

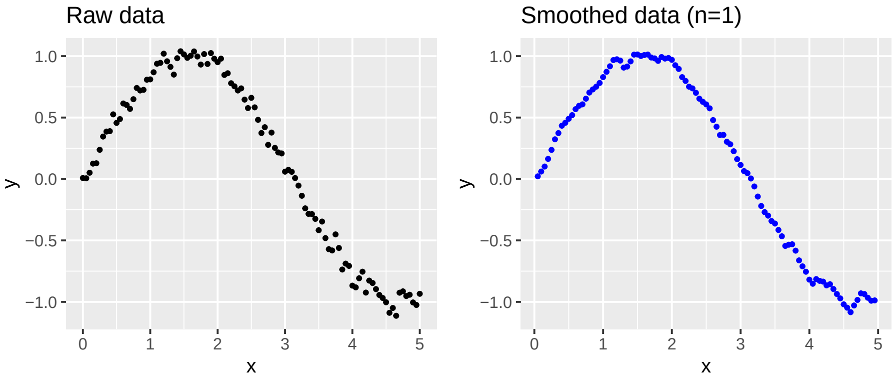
- Regression
  - try to predict 'Age' from the other attributes
  - replace the original values with the predicted ones
  - may lose **too much** information

Smoothing

Name	Age	Gender	Position	Salary
Manuel	25	M	assistant	23000
José	36	M	manager	59000
Rui	41	M	manager	57000
Sofia	105	F	manager	58000
Ana	28	F	assistant	28500

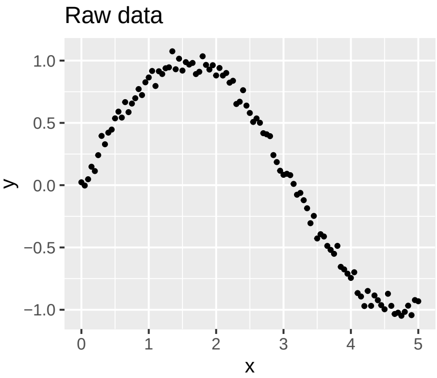
- Age=105 is an **outlier**
  - can be detected with clustering or using the IQR rule
  - can be replaced by the mean age of *manager*
  - i.e., detect outliers and replace them by a sensible mean

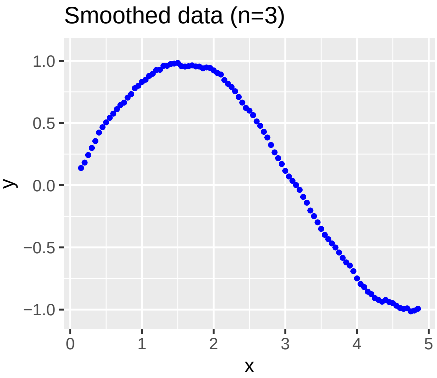
Smoothing



- Smoothing with **moving average**
  - replace each value y_i with $average(y_j), j = i - n, \ldots, i$
  - the larger the n, the smoother the line
  - Above $n = 1$

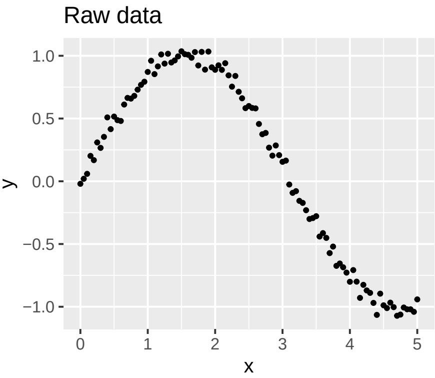
Smoothing

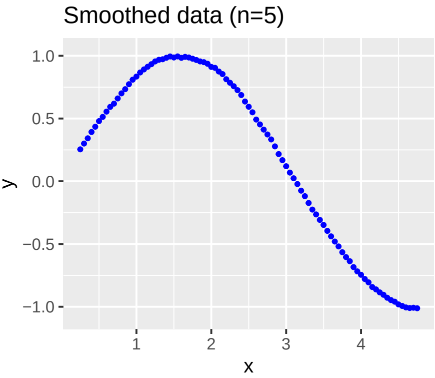




- Smoothing with **moving average**
  - replace each value y_i with $average(y_j), j = i - n, \ldots, i$
  - the larger the n, the smoother the line
  - Above $n = 3$

Smoothing





- Smoothing with **moving average**
  - replace each value y_i with $average(y_j), j = i - n, \ldots, i$
  - the larger the n, the smoother the line
  - Above $n = 5$

Data integration

- The same object can have different representations
  - customer in social network an in sales data
  - two companies merging
  - **entity identification problem**
- There may be **redundant variables**
  - **detect** redundancy
  - **remove** redundant variables

Redundancy analysis

- We can measure the "similarity" of two variables
  - Nominal: χ^2 statistical test
    - if the null hypothesis (variables are correlated) is accepted, one of the variables is redundant
    - if rejected, the variables are independent

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- o are the observed frequencies, e are the expected - high values of the χ_2 statistic mean **independence**

$$e_{ij} = \frac{\#(A = a_i) \times \#(B = b_j)}{n}$$

Example of step-by-step calculation

Redundancy analysis

- We can measure the "similarity" of two variables
  - Numerical: Pearson **correlation**
  - correlation is between -1 and 1
  - if correlation is close to zero, the variables are independent
- Example: *taxes payed* and *spending*
  - highly correlated
  - we cannot infer **causality**
    - A pays a lot of taxes because she spends a lot (**false**)
    - A spends a lot because she pays a lot of taxes (**false**)
- What is the relation between correlation and **Covariance**?

Other operations in data integration

- Eliminate **duplicate tuples**
  - the same customer appears in the DB (from two different sources)
- Detect **conflicting values**
  - different representations, units, encondings
  - e.g. sales in Euros and in Dollars
  - e.g. sales per day and sales per week

Data reduction: Dimensionality reduction

- reduce the number of variables
- **Principal Components Analysis** (PCA)
  - finds new variables that
    - are much **fewer** than original ones
    - each is a **linear combination** of the original ones
    - explain *most* but not all of what is observed
  - **cons**: new variables may not be interpretable

Data Reduction: Dimensionality reduction

- reduce the number of variables
- **Feature selection**
  - e.g. we want to predict if a customer is leaving a mobile operator (churn)
  - not all features are relevant for **this problem**
  - a **good feature** is correlated with the target variable
- **Techniques**
  - Eliminate features with low correlation
    - does not consider joint effects of variables
  - **Stepwise forward selection**
    - start with zero features, add the best feature, keep adding
    - stop when improvement stops
  - **Stepwise backward elimination**
    - start with all the features, ...
  - (among others)

References

- Han, Kamber & Pei, Data Mining Concepts and Techniques, Morgan Kaufman.
- García-Laencina, P.J., Sancho-Gómez, J. & Figueiras-Vidal, A.R. Pattern classification with missing data: a review. Neural Comput & Applic 19, 263–282 (2010).
- Pavel Horbonos, A brief guide to data imputation with Python and R: Make the data clean, Towards Data Science (2020)
- Stef van Buuren, Flexible Imputation of Missing Data
- Excel errors in scientific papers