Data Science Portugal

June 12, 2019

Coimbra, Portugal

Handling Missing Data with Imputation

Ricardo Cardoso Pereira

Invited Assistant Teacher @ UC & ISEC

PhD Student @ University of Coimbra (CISUC)

Agenda

- Missing Data and its Mechanisms
- Methods to handle Missing Data
  - Case Deletion
  - Statistical Imputation
  - Machine Learning Imputation
- Issues with MNAR
- Open Challenges

Missing Data

- Problem often found in real-world contexts
- Occurs when values are missing for one or several features

23	87	?	12	?
8	?	0.3	?	5
43	0	0.4	56	?
93	?	0.2	9	2
4	99	0.5	?	1

→

0	0	1	0	1
0	1	0	1	0
0	0	0	0	1
0	1	0	0	0
0	0	0	1	0

Missing Data Mechanisms

- Describe how the missing values are related to the data
- Three different mechanisms exist
  - Missing **Completely** At **Random** (MCAR)
  - Missing At ~~**Random**~~ (MAR)
  - Missing **Not** At **Random** (MNAR)

	Number of cigarettes			
Age	**Complete**	**MCAR**	**MAR**	**MNAR**
15	2	2	**?**	2
15	9	**?**	**?**	**?**
15	4	**?**	**?**	4
16	2	2	**?**	2
16	2	2	**?**	2
16	7	7	**?**	**?**
16	3	3	**?**	3
17	9	**?**	9	**?**
17	6	6	6	**?**
17	4	**?**	4	4
17	5	5	5	5
17	5	5	5	5
18	7	**?**	7	**?**
18	6	6	6	**?**
18	7	**?**	7	**?**
19	3	3	3	3
19	8	**?**	8	**?**
19	3	**?**	3	3
20	9	9	9	**?**
20	2	2	2	2

Missing Completely At Random

- Occurs when the values are **randomly missing**
- The cause is not related to any observed or unobserved values

	Number of cigarettes	
Age	**Complete**	**MCAR**
15	2	2
15	9	?
15	4	?
16	2	2
16	2	2
16	7	7
16	3	3
17	9	?
17	6	6
17	4	?
17	5	5
17	5	5
18	7	?
18	6	6
18	7	?
19	3	3
19	8	?
19	3	?
20	9	9
20	2	2

Missing At Random

- Occurs when the missing values are **related to observed data**
- A strong correlation exists between the missing values and an observed feature

	Number of cigarettes	
Age	**Complete**	**MAR**
15	2	**?**
15	9	**?**
15	4	**?**
16	2	**?**
16	2	**?**
16	7	**?**
16	3	**?**
17	9	9
17	6	6
17	4	4
17	5	5
17	5	5
18	7	7
18	6	6
18	7	7
19	3	3
19	8	8
19	3	3
20	9	9
20	2	2

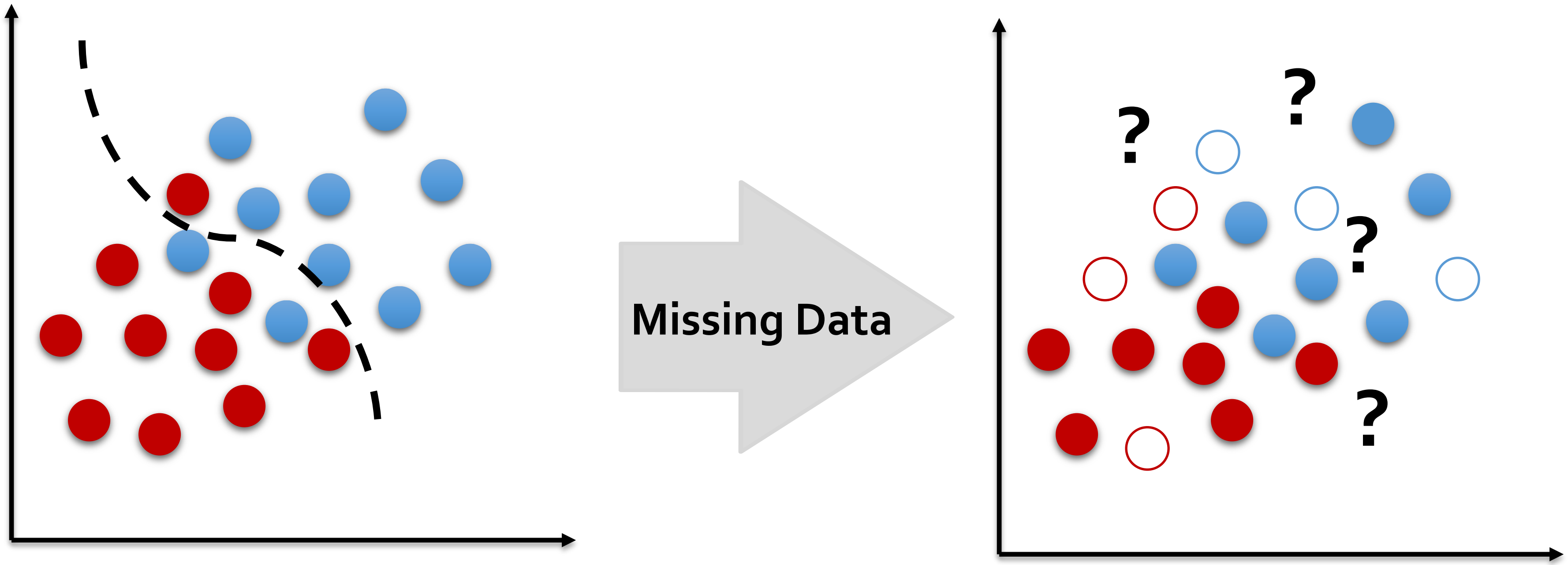
Missing Not At Random

- Occurs when the missing values are **related with themselves** or with **other unobserved values**
- Often called Non-Ignorable missing data

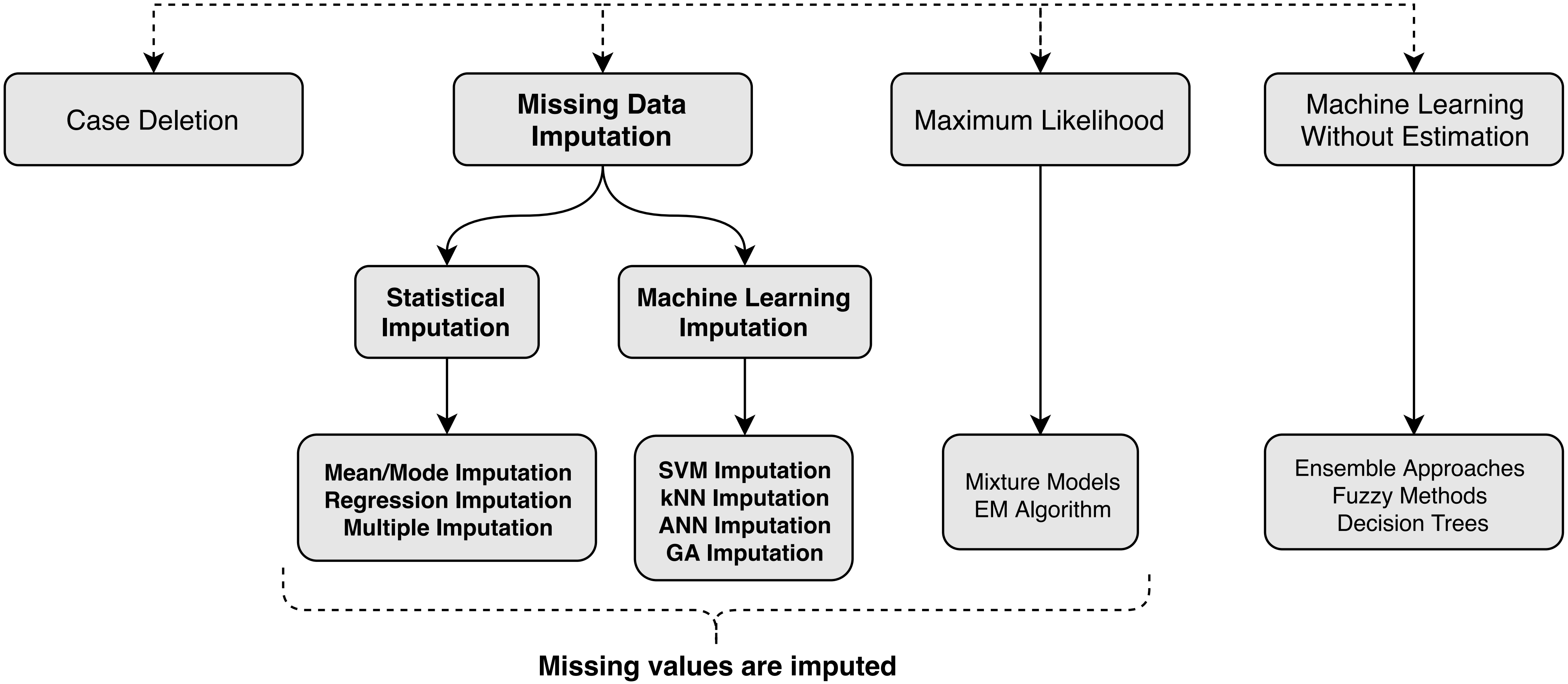
	Number of cigarettes	
Age	**Complete**	**MNAR**
15	2	2
15	9	?
15	4	4
16	2	2
16	2	2
16	7	?
16	3	3
17	9	?
17	6	?
17	4	4
17	5	5
17	5	5
18	7	?
18	6	?
18	7	?
19	3	3
19	8	?
19	3	3
20	9	?
20	2	2

Machine Learning with Missing Data

- Missing data can degrade the performance of ML models
- Some methods can cope with it (e.g., decision trees), but most don’t

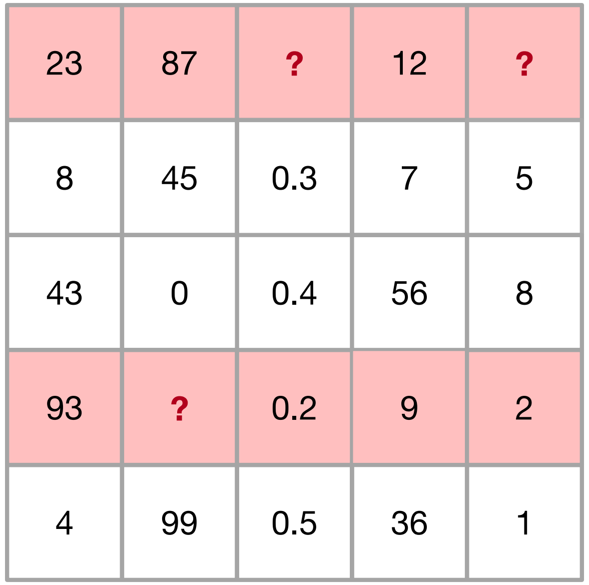


Methods to handle Missing Data



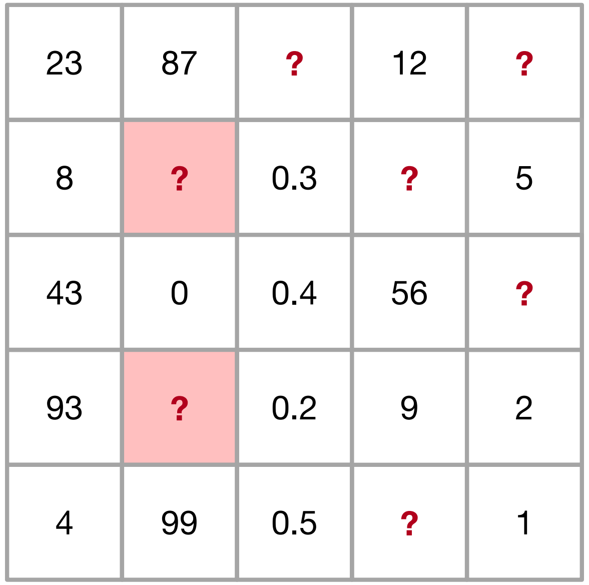
Case Deletion

- Records with missing values are… deleted
- In theory should only be applied with MCAR



23	87	?	12	?
8	45	0.3	7	5
43	0	0.4	56	8
93	?	0.2	9	2
4	99	0.5	36	1

Listwise Deletion



23	87	?	12	?
8	?	0.3	?	5
43	0	0.4	56	?
93	?	0.2	9	2
4	99	0.5	?	1

Pairwise Deletion

Mean/Mode Imputation

- Missing values are imputed with the mean
- The mode should be used for categorical data
- If the mechanism is not MCAR the imputation may be biased

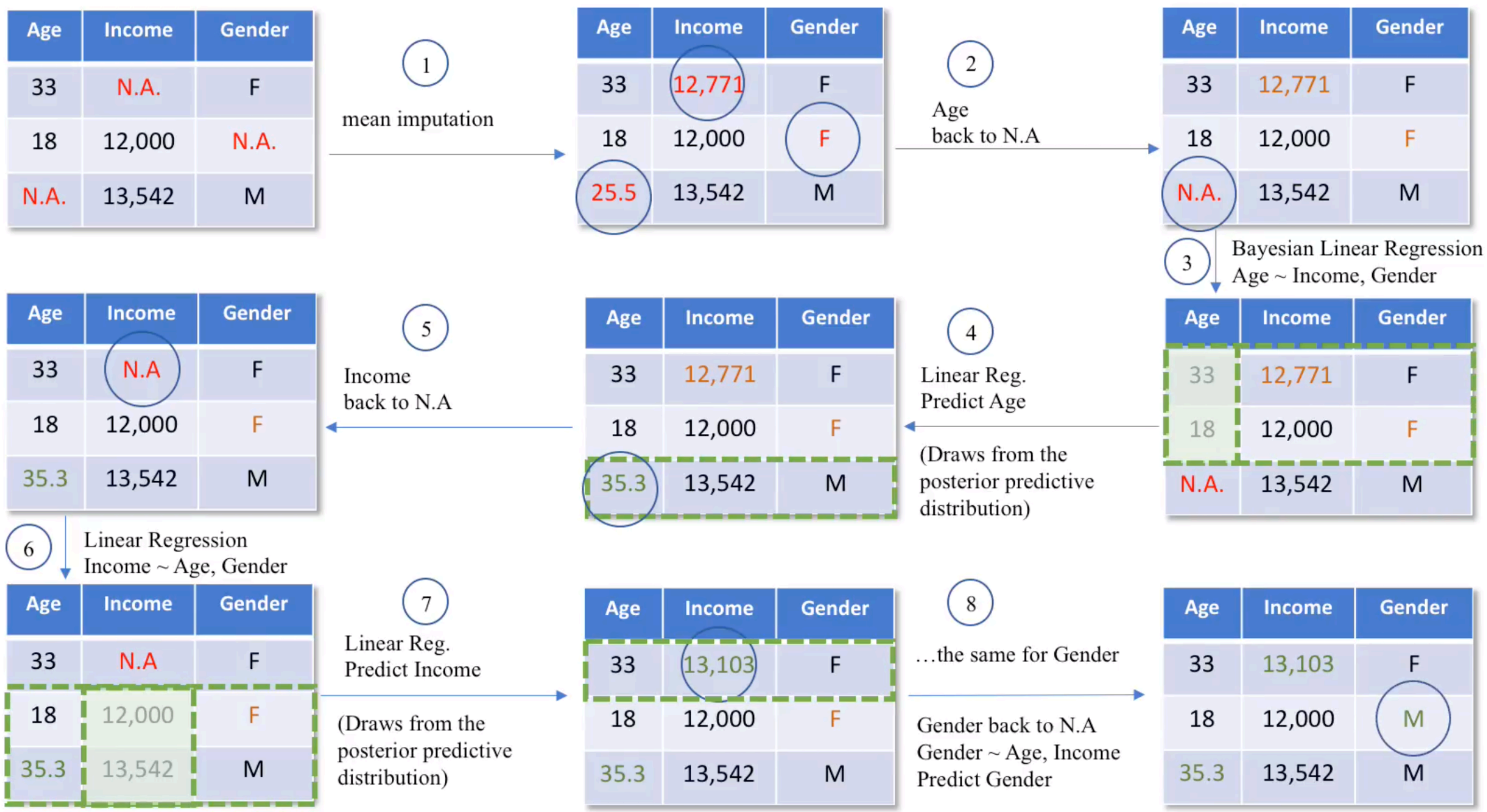
V_1	V_2	V_3	V_4	V_5
23	87	$E(V_3)$	12	$E(V_5)$
8	45	0.3	7	5
43	0	0.4	56	8
93	$E(V_2)$	0.2	9	2
4	99	0.5	36	1

Multiple Imputation by Chained Equations

- A series of regressions are modeled to each variable with missing data
- Each feature is modeled conditionally upon the other features
- In theory should only be applied with MAR

- The process is repeated multiple times to reduce bias
  - That’s why it’s called multiple imputation
  - But this concept can be applied with other methods

Multiple Imputation by Chained Equations



Age	Income	Gender
33	N.A.	F
18	12,000	N.A.
N.A.	13,542	M

1 mean imputation

Age	Income	Gender
33	12,771	F
18	12,000	F
25.5	13,542	M

2 Age back to N.A

Age	Income	Gender
33	12,771	F
18	12,000	F
N.A.	13,542	M

3 Bayesian Linear Regression Age ~ Income, Gender

Age	Income	Gender
33	12,771	F
18	12,000	F
N.A.	13,542	M

4 Linear Reg. Predict Age

(Draws from the posterior predictive distribution)

Age	Income	Gender
33	12,771	F
18	12,000	F
35.3	13,542	M

5 Income back to N.A

Age	Income	Gender
33	N.A	F
18	12,000	F
35.3	13,542	M

6 Linear Regression Income ~ Age, Gender

Age	Income	Gender
33	N.A	F
18	12,000	F
35.3	13,542	M

7 Linear Reg. Predict Income

(Draws from the posterior predictive distribution)

Age	Income	Gender
33	13,103	F
18	12,000	F
35.3	13,542	M

8 …the same for Gender

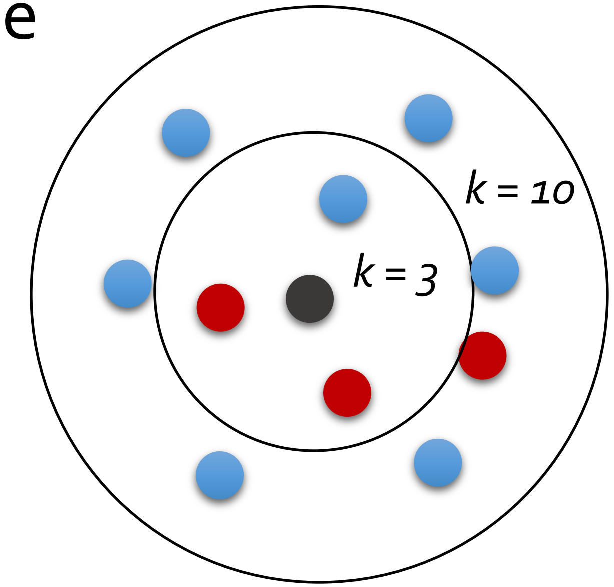
Gender back to N.A
Gender ~ Age, Income
Predict Gender

Age	Income	Gender
33	13,103	F
18	12,000	M
35.3	13,542	M

Ofir Shalev (@ofirdi) May 2018

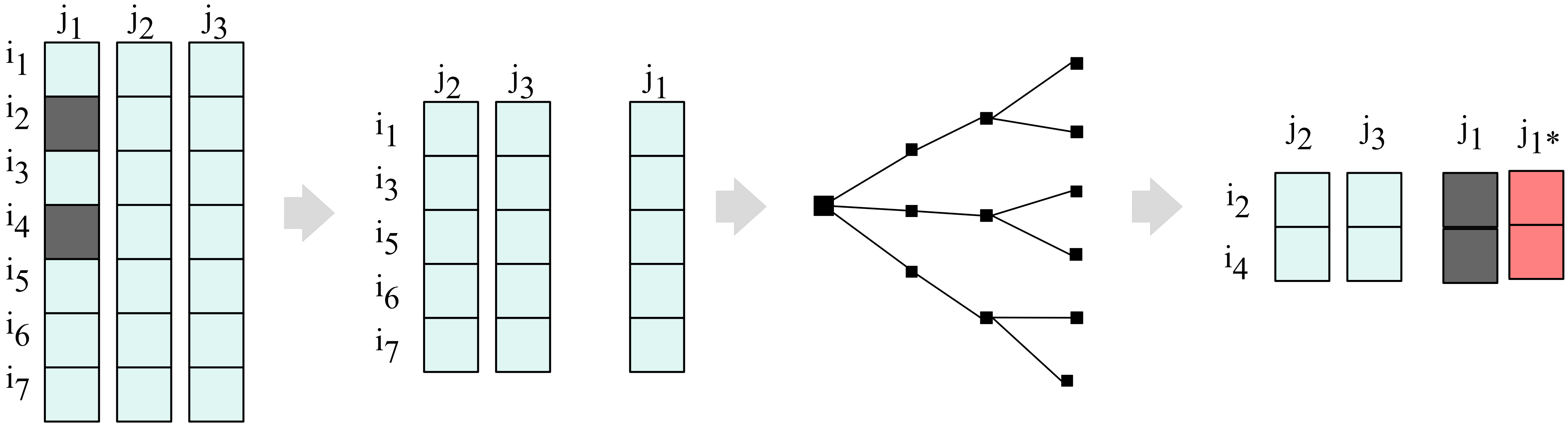
K-Nearest Neighbors Imputation

- Finds the *k* similar observations to the one that is being imputed
- Uses the values of the feature with missingness to generate the new value (mean, weighted mean, vote of majority, ...)
- The distance must be adjusted to the data type
  - Euclidean for numeric data
  - One-Hot Encoding to convert categorical data
  - Hamming distance for categorical data
- Is suitable for MAR and MCAR



Machine Learning-Based Imputation

- The pipeline used for most regression and classification algorithms can be adapted for missing values imputation (e.g., ANN, SVM, …)



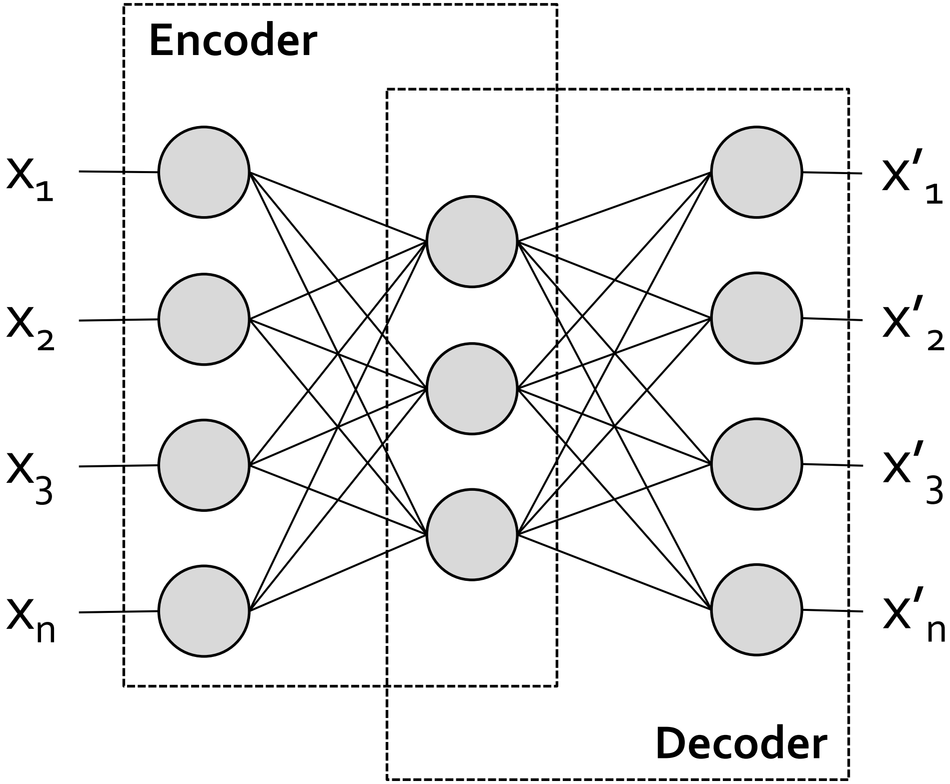
Missing Data

Model Training

Model Prediction

Stacked Denoising Autoencoders

- Special type of ANN that tries to reproduce the input at the output layer
- The Denoising variant learns from a corrupted version of the data
- Missing data is a type of corruption
- Is suitable for MAR and MCAR



What about MNAR?

- The described approaches are only valid for MAR and MCAR
- Imputation methods produce poor and biased results for MNAR
  - Expected since this mechanism is related to unobserved data
- Current solutions? --> **Sensitivity Analysis**
  - Try out different plausible MNAR models to see how consistent the results are
  - Multiple imputation strategies are often used
  - It’s just a test, not a solution…

What about MNAR?

- We could ignore it but MNAR is predominant in several contexts
- Example 1: IoT
  - Data collected from sensors is missing due to external factors
- Example 2: Clinical trials
  - Participants may be quitting a study for reasons related to the outcome that is being measured

Open Challenges

- New approaches to tackle the MNAR issues
- Identification of the missing mechanisms
- Use of generative models for imputation
    - Generative Adversarial Networks (GANs) are being used in very recent papers
- And many others…

Data Science Portugal

June 12, 2019

Coimbra, Portugal

Handling Missing Data with Imputation

Thank you! Questions?

rdpereira@dei.uc.pt

Ricardo Cardoso Pereira

Invited Assistant Teacher @ UC & ISEC

PhD Student @ University of Coimbra (CISUC)