The industrial challenge of missing data

Reza Sahraeian

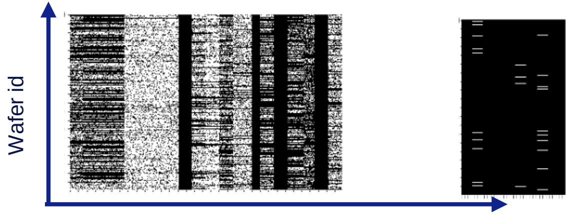
Data Scientist @ASML

October 2020

PyData Eindhoven

Why am I here

- At ASML we face different kinds of missingness.



- As a data scientist/analyst/engineer it is important to know how to deal with missing data

- And as an implementer (and python user), you need to what's happening in your code with missing data.

With incomplete data:

sklearn.ensemble.RandomForestRegressor → returns ValueError
LightGBM → runs without error (by default replaces missing values by zero!)
XGboost → runs without error (find the right splitting for missing variables according to the training loss function)

tf.keras.layers.MaxPooling2D → ignores missing (NANs) for max operation
tf.keras.layers.AveragePooling2D → Considers missing (NANs) for averaging operation

Today

- Why missing data and why should we care?
  - Why is it important
  - The origin and types of missingness

- How to deal with missing data
  - Categorizing imputation approaches

- Some practical results

Missing data

- Different industrial data has different challenges:
  - Privacy
  - Expense
  - Machine's or human's mistakes
  - …
- Many AI/ML/Data Science methods are developed for complete data
- Inappropriate approach imposes noise or bias on data

- Types of missingness:
  I. Missing completely at random (MCAR)

Income (Gross)	gender
NaN	m
40.0	m
80.0	f
NaN	f
70.0	m
65.0	f

There is no relationship/dependency between missingness and observed values (causes and values of missing is uncorrelated to the data)
→ men or women are not more inclined to share their salary info
→ gender does not impact on salary

Missing data

- Different industrial data has different challenges:
  - Privacy
  - Expense
  - Machine’s or human’s mistakes
  - …
- Many AI/ML/Data Science methods are developed for complete data
- Inappropriate approach imposes noise or bias on data

- Types of missingness:
  I. Missing completely at random (MCAR)
  II. Missing at random (MAR)

There is a relationship/dependency between missing values and observed ones, but not the missing values.

Income (Gross)	gender	Experience (year)
NaN	m	10
40.0	m	2
80.0	f	7
NaN	f	6
70.0	m	6
65.0	f	5

Missing data

- Different industrial data has different challenges:
  - Privacy
  - Expense
  - Machine's or human's mistakes
  - …
- Many AI/ML/Data Science methods are developed for complete data
- Inappropriate approach imposes noise or bias on data

- Types of missingness:
  - I. Missing completely at random (MCAR)
  - II. Missing at random (MAR)
  - III. Missing not at random (MNAR)

Income (Gross)	gender	Position
NaN	m	CEO
40.0	m	junior
80.0	f	senior
NaN	f	CEO
70.0	m	junior
65.0	f	senior

The cause of missingness is not known and we cannot draw any conclusion from observed data!

Missing data

- Different industrial data has different challenges:
  - Privacy
  - Expense
  - Machine’s or human’s mistakes
  - …
- Many AI/ML/Data Science methods are developed for complete data
- Inappropriate approach imposes noise or bias on data

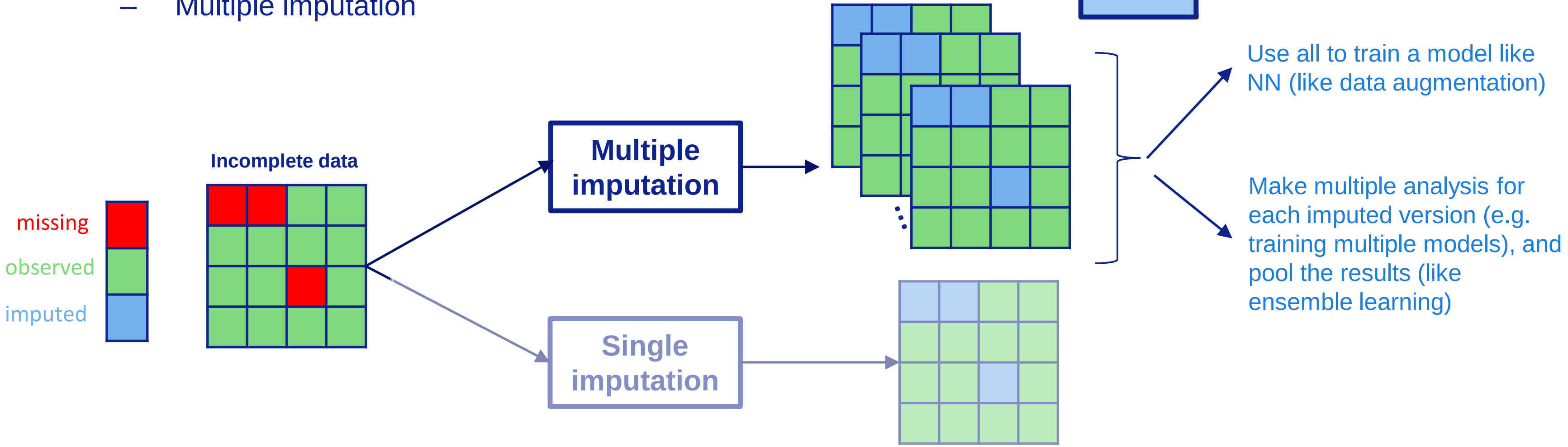
- Types of missingness:
  - I. Missing completely at random (MCAR)
  - II. Missing at random (MAR)
  - III. Missing not at random (MNAR)

Points:
- The size and balance of data must be considered before distinguishing the type.
- Finding the type of missingness is not easy and unfortunately sometimes impossible.
- We may face different missing types in one dataset.
- As a rule of thumb, we may assume missing at random type unless there is a good reason not to!

How to deal with missingness (if we should …)

- What to do depends on missing type…
- Know the art of dropping (ignoring) missing values
  - Sample size
  - Loss of information
  - Induced error of imputation

- Imputation techniques can be categorized as
  - Single imputation
  - Multiple imputation



How to deal with missingness (if we should …)

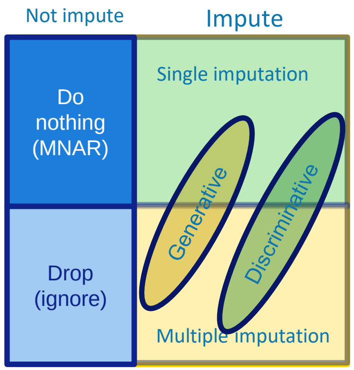
- What to do depends on missing type…
- Know the art of dropping (ignoring) missing values
  - Sample size
  - Loss of information
  - Induced error of imputation

- Imputation techniques can be categorized as
  - Single imputation
  - Multiple imputation

  Or
  - Discriminative models:

    $X_{miss} = f(X_{obs})$ (MAR)
  - Generative models :

    $P(X_{miss})$ (MCAR) and $P(X_{miss}|X_{obs})$ (MAR)



Single imputation

- One candidate is given to be treated as the true value

- Some popular methods:
  - **Mean (median) imputation (generative)**

Income (Gross)	gender	Experience (year)
NaN	m	10
40.0	m	2
80.0	f	7
NaN	f	6
70.0	m	6
65.0	f	5

mean

Income (Gross)	gender	Experience (year)
63.75	m	10
40.00	m	2
80.00	f	7
63.75	f	6
70.00	m	6
65.00	f	5

Single imputation

- One candidate is given to be treated as the true value
- Some popular methods:
  - Mean (median) imputation (generative)
  - **K-nearest neighbor (KNN)** (discriminative)

Number of neighbors = 2

Income (Gross)	gender	Experience (year)
NaN	m	10
40.0	m	2
80.0	f	7
NaN	f	6
70.0	m	6
65.0	f	5

1st neighbor

2nd neighbor

Income (Gross)	gender	Experience (year)
75.0	m	10
40.0	m	2
80.0	f	7
75.0	f	6
70.0	m	6
65.0	f	5

Single imputation

- One candidate is given to be treated as the true value

- Some popular methods:
  - Mean (median) imputation (generative)
  - K-nearest neighbor (KNN) (discriminative)
  - **Discriminative model training** (discriminative)
    - Linear regression
    - Neural nets
    - **Random forest**
    - …

Income (Gross)	gender	Experience (year)
76.300	m	10
40.000	m	2
80.000	f	7
68.825	f	6
70.000	m	6
65.000	f	5

Model

imputation

Income (Gross)	gender	Experience (year)
NaN	m	10
40.0	m	2
80.0	f	7
NaN	f	6
70.0	m	6
65.0	f	5

Y Model X

train

Single imputation

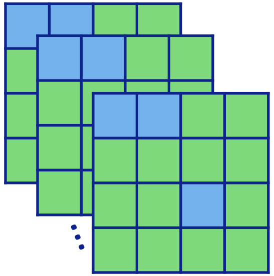
- One candidate is given to be treated as the true value

- Some popular methods:
  - Mean (median) imputation (generative)
  - K-nearest neighbor (KNN) (discriminative)
  - Discriminative model training (discriminative)
    - Linear regression
    - Neural nets
    - Random forest
    - …

- Other approaches: PCA, EM, …
- <u>Problem: we do not account for uncertainty</u>

Multiple Imputation

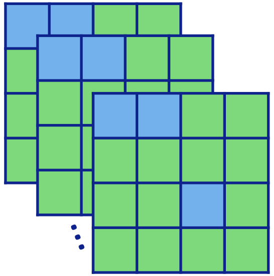
- With multiple imputation we account for uncertainty by creating multiple imputed version of data.



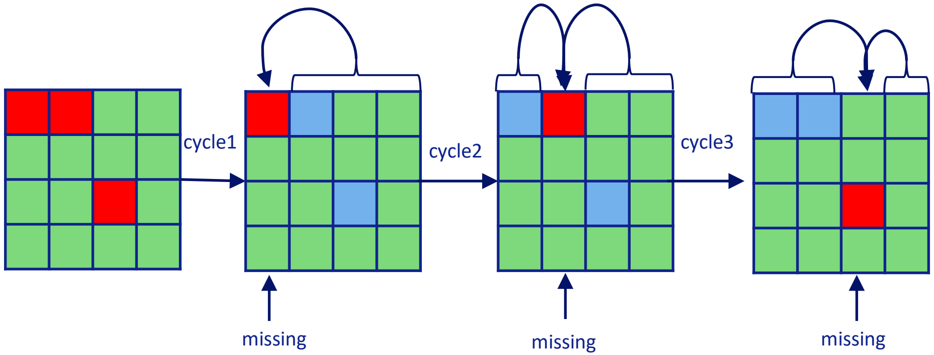
- How?
  - Bootstrapping
  - Generative models
  - Different imputation techniques
  - Others: MICE, denoising autoencoder, …

Multiple Imputation

- With multiple imputation we account for uncertainty by creating multiple imputed version of data.



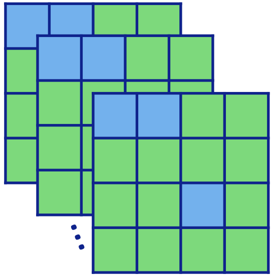
- How?
  - Bootstrapping
  - Generative models
  - Different imputation techniques
  - Others: **MICE**, denoising autoencoder, …



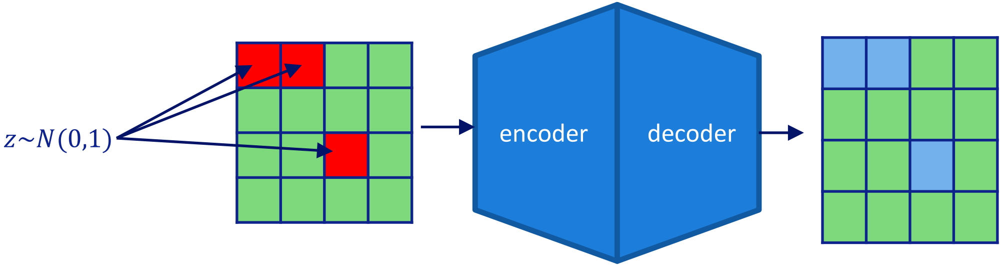
- It starts with initial imputation, e.g. mean imputation
- At each cycle only one variable is considered missing and is imputed via other variables.
- The whole process may be repeated.

Multiple Imputation

- With multiple imputation we account for uncertainty by creating multiple imputed version of data.



- How?
  - Bootstrapping
  - Generative models
  - Different imputation techniques
  - Others: MICE, **denoising autoencoder**, …



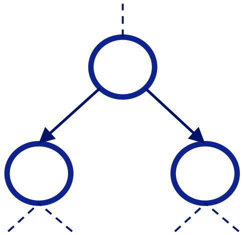
Missingness is assumed as infinite noise!

More on Imputation

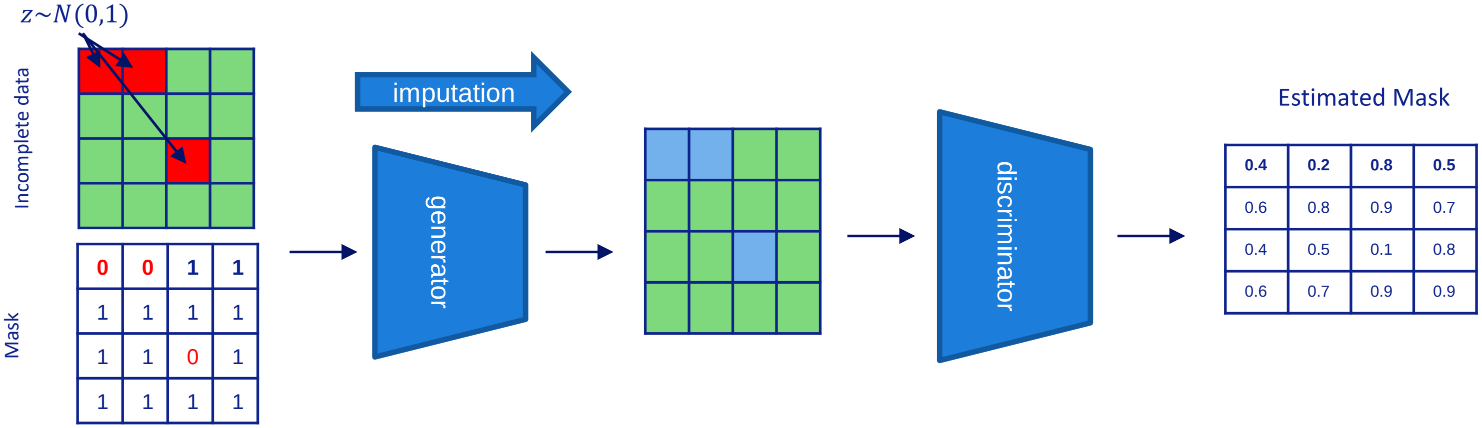
- **Exploiting target (output) for imputation:**
  - This is not straightforward and is usually done implicitly, For example in XGboost package:

During training model learns how to split the node for certain variable:

If var > threshold go left
If var < threshold go right
If var is missing go both sides and choose the side which has lower loss



- **Complete data is not available for training.**
  - For example in GAN based imputation (e.g. GAIN):

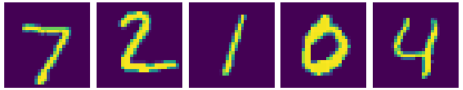


MNIST examples

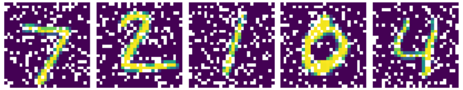
Sklearn.imputer

fancyimpute

Original Samples

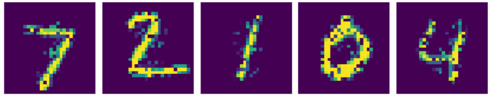


Samples with missing values (20% missing)

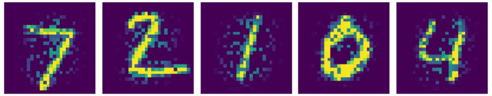


Single imputation

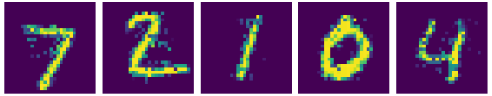
Median imputation



Mean imputation

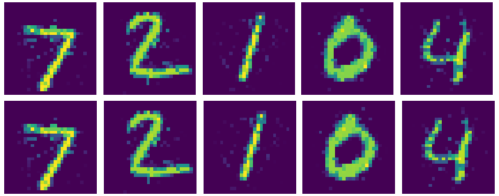


KNN imputation (n=5)

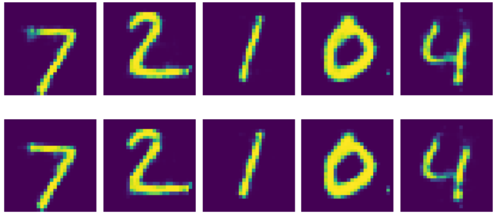


multiple imputation

MICE

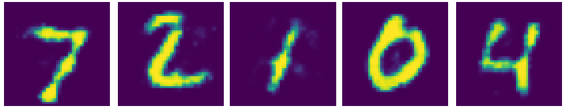


DAE



Generative adversarial imputation network

GAIN imputation



Takeaway

- Understand the missing type and data before doing anything (tips: missing rate, balance, correlation, data size, …)
- There is no single magical method to deal with all missingness, the right choice depends on your data.
- Benefit from multiple imputation to account for uncertainty.
- Be vigilant in using open source packages.
- Check literature for new methodologies.

ASML