L’art de naviguer en forêts malgré le manque d’information

Julie Josse, Nicolas Prost, Erwan Scornet, Gaël Varoquaux

CMAP, INRIA PARIETAL

Journée de rentrée du CMAP
Outline

1. Missing data: a medical issue
2. Missing value mechanisms
3. Decision Tree
4. Impact of missing values on a stump
Context: big data in health and social sciences

- More and more missing data due to:
  - high dimensionality (one feature may be missing)
  - difficulty of fine control on the acquisition process

- Causal conclusions from analysis challenging:
  - observational data (as opposed to experiments)
  - missing data induces selection biases

New data sources challenge missing-data methodology:

- high-dimensional
- observational
- uncontrolled confounds
Motivating data in health

**Traumabase**: 15,000 patients/250 var/15 hospitals

<table>
<thead>
<tr>
<th>Center</th>
<th>Age</th>
<th>Sex</th>
<th>Weight</th>
<th>Height</th>
<th>BMI</th>
<th>T°</th>
<th>Lactates</th>
<th>Glasgow</th>
<th>Transfusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beaujon</td>
<td>54</td>
<td>m</td>
<td>85</td>
<td>NR</td>
<td>NR</td>
<td>35.6</td>
<td>NA</td>
<td>12</td>
<td>yes</td>
</tr>
<tr>
<td>Lille</td>
<td>33</td>
<td>m</td>
<td>80</td>
<td>1.8</td>
<td>24.69</td>
<td>36.5</td>
<td>4.8</td>
<td>15</td>
<td>no</td>
</tr>
<tr>
<td>Pitie</td>
<td>26</td>
<td>m</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>36</td>
<td>3.9</td>
<td>3</td>
<td>no</td>
</tr>
<tr>
<td>Beaujon</td>
<td>63</td>
<td>m</td>
<td>80</td>
<td>1.8</td>
<td>24.69</td>
<td>36.7</td>
<td>1.66</td>
<td>15</td>
<td>yes</td>
</tr>
<tr>
<td>Pitie</td>
<td>30</td>
<td>w</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>36.6</td>
<td>NM</td>
<td>15</td>
<td>yes</td>
</tr>
</tbody>
</table>

- missing: Not Recorded, Made, Applicable, etc.

- predict the Glasgow score, start of a transfusion

- study the effect of a treatment on survival
Motivating data in health

*Traumabase*: 15 000 patients/ 250 var/ 15 hospitals

<table>
<thead>
<tr>
<th>Center</th>
<th>Age</th>
<th>Sex</th>
<th>Weight</th>
<th>Height</th>
<th>BMI</th>
<th>T°</th>
<th>Lactates</th>
<th>Glasgow</th>
<th>Transfusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beaujon</td>
<td>54</td>
<td>m</td>
<td>85</td>
<td>NR</td>
<td>NR</td>
<td>35.6</td>
<td>NA</td>
<td>12</td>
<td>yes</td>
</tr>
<tr>
<td>Lille</td>
<td>33</td>
<td>m</td>
<td>80</td>
<td>1.8</td>
<td>24.69</td>
<td>36.5</td>
<td>4.8</td>
<td>15</td>
<td>no</td>
</tr>
<tr>
<td>Pitie</td>
<td>26</td>
<td>m</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>36</td>
<td>3.9</td>
<td>3</td>
<td>no</td>
</tr>
<tr>
<td>Beaujon</td>
<td>63</td>
<td>m</td>
<td>80</td>
<td>1.8</td>
<td>24.69</td>
<td>36.7</td>
<td>1.66</td>
<td>15</td>
<td>yes</td>
</tr>
<tr>
<td>Pitie</td>
<td>30</td>
<td>w</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>36.6</td>
<td>NM</td>
<td>15</td>
<td>yes</td>
</tr>
</tbody>
</table>

- missing: Not Recorded, Made, Applicable, etc.

- predict the Glasgow score, start of a transfusion

- study the effect of a treatment on survival

**UK Biobank**: prospective epidemiology

- 1 Million patients of a normal aging population
- 10% have medical imaging data
- Observational data to study risk factors
A first look at data

**Ongoing work:** predict hemorrhagic shock automatically during the prehospital phase.

Database composed of 5153 patients with 5% of missing entries.
A first look at data

**Ongoing work:** predict hemorrhagic shock automatically during the prehospital phase.

Database composed of 5153 patients with 5% of missing entries. Among those

<table>
<thead>
<tr>
<th></th>
<th>Amount</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Age</td>
<td>7</td>
<td>0.1%</td>
</tr>
<tr>
<td>BMI</td>
<td>778</td>
<td>15%</td>
</tr>
<tr>
<td>GCS</td>
<td>18</td>
<td>0.4%</td>
</tr>
<tr>
<td>Heart rate</td>
<td>110</td>
<td>2.1%</td>
</tr>
<tr>
<td>Pulse pressure</td>
<td>126</td>
<td>2.5%</td>
</tr>
<tr>
<td>Hb level</td>
<td>301</td>
<td>5.8%</td>
</tr>
<tr>
<td>O2 saturation</td>
<td>171</td>
<td>3.3%</td>
</tr>
<tr>
<td>Expander</td>
<td>795</td>
<td>15.4%</td>
</tr>
</tbody>
</table>

All in all, 33% of patients have at least one missing value!
How to solve this problem?

1. Delete all missing values → very Very Very bad idea!
How to solve this problem?

1. Delete all missing values → very Very Very bad idea!

2. Impute data with your favorite imputation method
   - Replace all missing values by the mean/median/mode of the corresponding variable.
     - Good point: the mean/median/mode is unchanged!
     - Bad point 1: the variance of the imputed data is lower than reality
     - Bad point 2: structure of dependence between variable is destroyed.
   - Multiple imputation: very good idea but not the focus of this presentation.
How to solve this problem?

1. Delete all missing values $\rightarrow$ very Very Very bad idea!

2. Impute data with your favorite imputation method
   - Replace all missing values by the mean/median/mode of the corresponding variable.
     - Good point: the mean/median/mode is unchanged!
     - Bad point 1: the variance of the imputed data is lower than reality
     - Bad point 2: structure of dependence between variable is destroyed.
   - Multiple imputation: very good idea but not the focus of this presentation.

3. Design methods that handle missing values: very good idea and this is the focus of this presentation!
Outline

1. Missing data: a medical issue
2. Missing value mechanisms
3. Decision Tree
4. Impact of missing values on a stump
Available observations

We assume to be given an observed sample,

\[ \mathcal{D}_n^* = ((X_i^*, Y_i))_{1 \leq i \leq n} \sim (X^*, Y) \text{ i.i.d.}, \]

where \( X^* \in (\mathbb{R} \times \{\text{NA}\})^d \), resulting from the (unobserved) complete sample

\[ \mathcal{D}_n = ((X_i, M_i, Y_i))_{1 \leq i \leq n} \sim (X, M, Y) \text{ i.i.d.}. \]

For each variable index \( j \) and each realization \( i \),

- \( M_{i,j} = 0 \) means that \( X_{i,j} \) is observed,
- \( M_{i,j} = 1 \) means that \( X_{i,j} \) is missing.

In a nutshell,

\[ X^* := X \otimes (1 - M) + \text{NA} \cdot M, \]

where \( \otimes \) is the term-by-term product, \( \text{NA} \times 1 := \text{NA}, \text{NA} \times 0 := 0. \]
The most simple NAs

For a realization of the complete sample,

$$d_n = \begin{bmatrix}
2 & 3 & 1 & 0 & 0 & 0 & 1 & 0 & 15 \\
1 & 0 & 3 & 5 & 0 & 1 & 0 & 0 & 13 \\
9 & 4 & 2 & 5 & 0 & 0 & 0 & 1 & 18 \\
7 & 6 & 3 & 2 & 0 & 0 & 1 & 1 & 10
\end{bmatrix},$$

the observed sample is

$$d_n^\star = \begin{bmatrix}
2 & 3 & \text{NA} & 0 & 15 \\
1 & \text{NA} & 3 & 5 & 13 \\
9 & 4 & 2 & \text{NA} & 18 \\
7 & 6 & \text{NA} & \text{NA} & 10
\end{bmatrix}.$$
Rubin-MAR / realized MAR

Let $\mathbf{m}$ be the realization of $\mathbf{M}$. We separate the observed values and the missing values of $\mathbf{x}$: $\mathbf{x} = (\mathbf{x}_{\text{obs}}, \mathbf{x}_{\text{miss}})$. The missing data are said to be missing at random if $g(\mathbf{m}|\mathbf{x})$ is the same for all values of $\mathbf{x}_{\text{miss}}$, i.e.

$$\forall \mathbf{x}, \mathbf{x}', \mathbf{x}_{\text{obs}} = \mathbf{x}'_{\text{obs}} \Rightarrow g(\mathbf{m}|\mathbf{x}) = g(\mathbf{m}|\mathbf{x}').$$

Realized MAR is the minimal property to access the likelihood of missing data. But...

- difficult to comprehend!
- Not useful to generate data.
(Everywhere) MAR

The missing data are said to be everywhere missing at random if for all realizations $\mathbf{m}$ of $\mathbf{M}$, $g(\mathbf{m}|\mathbf{x})$ is the same for all values of $\mathbf{x}_{miss}$, i.e.

$$\forall \mathbf{m}, \forall \mathbf{x}, \mathbf{x}', \mathbf{x}_{obs} = \mathbf{x}'_{obs} \Rightarrow g(\mathbf{m}|\mathbf{x}) = g(\mathbf{m}|\mathbf{x}').$$

It should NOT be interpreted as $\mathbf{M} \perp \mathbf{X}_{miss}|\mathbf{X}_{obs}$.

MNAR

The pattern of missing data is said to be Missing Non At Random (MNAR) if it is not MAR, i.e., the pattern of missingness depends on unobserved values.
Outline

1. Missing data: a medical issue
2. Missing value mechanisms
3. Decision Tree
4. Impact of missing values on a stump
Where are Random Forests?
No random forests!

Forests are overrated! Let us consider a single tree.
How to build a tree?

- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.
How to build a tree?

- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.
How to build a tree?

- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.
How to build a tree?

- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.
How to build a tree?

- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.
How to build a tree?

- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.
How to build a tree?

- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.
How to build a tree?

**CART** are defined by

1. **A splitting rule**: minimize the variance within the resulting cells.
2. **A stopping rule**: stop when each cell contains less than \( \text{nodesize} = 2 \) observations.
How to split? Two classic strategies

"Classic" CART

- Exhaustive search
- Impurity of a node:
  \[ I = \sum (Y_i - \bar{Y})^2 \]
- Splitting criterion:
  \[ C(X_j) = I - I_L^{\text{best}} - I_R^{\text{best}} \]

Conditional trees

- Variable choice:
  \[ T(X_j) = \sum X_i^j Y_i \]
- Threshold choice: impurity
- Splitting criterion:
  \[ C(X_j) \propto T(X_j) \]

With missing values: summing over available points.
Another procedure: CART Missing Attributes

- Compute the splitting criterion corresponding to separating missing values from observed values.
- For each possible split, compute the splitting criterion corresponding to
  - Putting all missing values on the left node
  - Putting all missing values on the right node
- Choose the best split, i.e., the one maximizing the CART splitting criterion.
Outline

1. Missing data: a medical issue
2. Missing value mechanisms
3. Decision Tree
4. Impact of missing values on a stump
Four different algorithms:

- CART - \textit{rpart}
- Conditional Tree - \textit{CTree}
- Randomized CART
- CART - Taking into account missing attributes
Experiments

All experiments follow the same model:

\[ Y = 0.25X_1 + \varepsilon, \]

where

\[ X_2 \parallel X_1, \quad X_2 \parallel Y, \quad \varepsilon \parallel (X_1, X_2, Y), \quad X_1, X_2, \varepsilon \sim \mathcal{N}(0, 1). \]

Six different patterns of missing data:

- **MCAR1**: each entry is missing with the same probability.
- **MCAR2**: each entry of \( X_1 \) is missing with the same probability.
- **MAR1**: missing value on \( X_1 \) depends on values of \( X_2 \).
- **MAR2**: missing value on \( X_1 \) depends on values of \( Y \).
- **MNAR1**: missing value on \( X_1 \) depends on values of \( X_1 \).
- **MNAR2**: missing value on \( X_1 \) depends on values of \((X_1, Y)\).
MCAR1

Each entry is missing with probability $p/2$. 
Each entry of the first variable is missing with probability $p$. 

MCAR2
Each entry of the first variable is missing with probability

$$\mathbb{P}[M_1 = 1|X_2] = \frac{1}{1 + e^{-(X_2 + \lambda)}}.$$
MAR2

Each entry of the first variable is missing with probability

$$\mathbb{P}[M_1 = 1 | Y] = \frac{1}{1 + e^{-(\lambda + \gamma)}}.$$
Each entry of the first variable is missing with probability

\[
P[M_1 = 1 | X_1] = \frac{1}{1 + e^{-(X_1 + \lambda)}}.
\]
Each entry of the first variable is missing with probability

\[ P[M_1 = 1|(X_1, Y)] = \frac{\lambda}{\lambda + |Y - 0.25X_1|}. \]
Causal conclusions:

\begin{align*}
Y \text{ outcome, } X \text{ covariates, } W \text{ treatment 0 or 1} \\
\text{Average Treatment Effect } \tau &= E[Y_i(1) - Y_i(0)] \\
- \text{ experimental design: } &\bar{Y}_1 - \bar{Y}_0 \\
- \text{ observational data: adjust for the covariate} \\
\text{Unconfoundness: } (Y_i \perp \perp W_i | X_i)
\end{align*}

Inverse probability weighting — “Doubly robusts”

Estimates weights: \( e(x) = P(W_i = 1 | X = x) \)

\begin{align*}
\text{Average Treatment Effect } \hat{\tau} &= \frac{1}{n} \sum_i \left( \frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i) Y_i}{1-\hat{e}(X_i)} \right)
\end{align*}

⇒ Random Forests with missing values
Thank you!

There are two types of people in this world:

Those who can extrapolate from incomplete data.