NeuMiss networks: differentiable programming for supervised learning with missing values

Marine Le Morvan, Julie Josse, Thomas Moreau, Erwan Scornet, Gaël Varoquaux

Objectives

Supervised learning with missing values:
- both in the training set and at prediction time in the test set,
- under possibly non-ignorable missingness (Missing Not At Random).

State-of-the-art and Challenges

A rich literature on statistical inference but few works on supervised learning with missing values.
- Most general methods for imputation are only valid if the missingness is ignorable.
- Samples are represented by varying subsets of input variables ⇒ learn compensation mechanisms.
- The total number of possible missing data patterns is exponential in the dimension (2^d) ⇒ keep the sample complexity polynomial.

Approach

Derive the analytical expression of the optimal predictor under various missing data mechanisms.
Propose a theoretically grounded neural network architecture (NeuMiss) designed to approximate these optimal predictors.

Notations and Assumptions

Random variables
- \( X \in \mathbb{R}^d \): complete data
- \( \tilde{X} \in \{X \cup \{\text{NA}\}\}^d \): incomplete data
- \( M \in \{0,1\}^d \): mask
- \( \text{obs}(M) \): indices of the observed entries

Ex. of realizations
- \( x = (1,1,2,3,1,8,5,27) \)
- \( \tilde{z} = (1,1,\text{NA},-3,1,8,\text{NA}) \)
- \( m = (0,1,0,0,1) \)
- \( \text{obs}(m) = (1,1,3,1,8) \)

Assumptions:
Linear model: \( Y = \beta_0^* + \sum_{j=1}^{d} \beta_j^* X_j + \epsilon \)
Gaussian data: \( X \sim \mathcal{N}(\mu, \Sigma) \)

Optimal (Bayes) predictor:
\[
f^* = \arg \min_{f(\mathcal{R}(\tilde{X})) \in \mathbb{R}} \mathbb{E} \left( (Y - f(X))^2 \right)
\]

The optimal predictors under various missing data mechanisms

Ignorable missing data mechanisms:
- Missing Completely At Random (MCAR): \( P(M = m) | X = x \) = \( P(M = m) \)
- Missing At Random (MAR): \( P(M = m) | X = x \) = \( P(M = m) | X_{\text{obs}(m)} \)

M(C)AR Bayes predictor
\[
f^*(X_{\text{obs}}, M) = \beta_0^* + (\beta_{\text{obs}}^* X_{\text{obs}}) + (\beta_{\text{mis}}^* \mu_{\text{mis}} + \Sigma_{\text{mis},\text{obs}}(\Sigma_{\text{obs}}^{-1})(X_{\text{obs}} - \mu_{\text{obs}}))
\]

Non-ignorable missing data mechanisms:

Gaussian self-masking (MNAR) Bayes predictor
\[
f^*(X_{\text{obs}}, M) = \beta_0^* + (\beta_{\text{obs}}^* X_{\text{obs}}) + (\beta_{\text{mis}}^* (\mu_{\text{mis}} + D_{\text{mis}} \Sigma_{\text{mis}_{\text{mis}}}^{-1})^{-1}
\times (\mu_{\text{mis}} + D_{\text{mis}} \Sigma_{\text{mis}_{\text{mis}}}^{-1} \Sigma_{\text{mis},\text{obs}}(\Sigma_{\text{obs}}^{-1})(X_{\text{obs}} - \mu_{\text{obs}})))
\]

Intuition:

The optimal predictors are linear per pattern

The slopes of the obs. variables depend on \( M \) to compensate for the missingness of other correlated variables:
\[
f^*(X_{\text{obs}}, M) = \beta(1 + \sum_{j=1}^{d} \beta_j^* M_j) X_j
\]

References


Approximating the Bayes predictors

Main difficulty: approx. of \( \Sigma_{\text{mis}}^{-1} \) for any obs, i.e., any missing data pattern!

Theoretical grounded architecture: In M(C)AR and under a simplifying assumption in Gaussian self-masking, NeuMiss with a depth \( \ell + 1 \) can exactly compute the order-\( \ell \) approximation of the Bayes predictor.

Experimental results

Simulated data
- Gaussian covariates
- Response is a linear model
- 50% missing values

Methods
- EM: Expectation Maximization.
- MLP [2]: feedforward neural network

Neumiss performances come close to the optimal performances.
Robustness to the missing data mechanism.
Suitable for medium-sized datasets thanks to weights sharing across mdp.