NeuMiss network classifiers: deep learning for classifying with missing values

Alexandre Perez-Lebel, Marine Le Morvan, and Gaël Varoquaux

ECML PKDD 2021
VIRTUAL
13-17 September

Many application settings, such as health or business, call for prediction on data with missing values. The optimal predictor then builds upon correlations between observed and missing covariates. In regression settings, NeuMiss networks have been introduced to approximate such Bayes predictor, providing the corresponding inductive bias in deep learning. Here, we extend these results to classification settings: we show that the same functional forms are Bayes consistent in classification settings, we adapt the neural architecture to classification, and we benchmark the resulting predictive pipeline showing that it predicts favorably compared to classical missing-values pipelines and is more robust to the missingness mechanism, including missing not at random settings.

Notations

- A pair of random variables (X, Y), where $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ or $\{-1, 1\}$.
- A random vector $M \in \{0,1\}^d$ acting as a mask on X: $M_j = 1 \Leftrightarrow X_j$ not observed.
- The incomplete feature vector $\tilde{X}$: $\tilde{X}_j = \texttt{NA}$ if $M_j = 1$ and $\tilde{X}_j = X_j$ if $M_j \neq 1$.
- We note *obs* (resp. *mis*) the indices of the zero (resp. non-zero) entries of M.
- We note X_{obs} (resp. X_{mis}) the observed (resp. missing) values of X.

Assumption: Gaussian data

The complete data is distributed as a multivariate Gaussian: $X \sim \mathcal{N}(\mu, \Sigma)$.

Assumption: Missing At Random (MAR)

For all $m \in \{0,1\}^d$, $\mathbb{P}[M = m \mid X] = \mathbb{P}[M = m \mid X_{obs}]$.

Missing Completely At Random (MCAR) is a special case of MAR where the probability of M does not depend on the covariates.

Assumption: Gaussian self-masking, instance of MNAR

The probability that a variable is missing depends on its own value through a Gaussian function:

$$\mathbb{P}[M \mid X] = \prod_{k=1}^{d} \mathbb{P}[M_k \mid X_k]$$

$$\forall\, 1 \leq k \leq d, \quad \mathbb{P}[M_k = 1 \mid X_k] = K_k \exp\left(-\frac{(X_k - \tilde{\mu}_k)^2}{2\,\tilde{\sigma}_k^2}\right)$$

where $0 < K_k < 1$, $\tilde{\mu}_k \in \mathbb{R}$ and $\tilde{\sigma}_k \in \mathbb{R}_+$.

NeuMiss in regression

Assumption: Linear model

The response $Y \in \mathbb{R}$ is linked to the complete data X through a linear model:

$$Y = \beta_0^\star, \sigma + \langle X, \beta^\star \rangle + \epsilon, \quad \text{where } \beta_0^\star \in \mathbb{R},\ \beta^\star \in \mathbb{R}^d,\ \ \epsilon \sim \mathcal{N}\left(0, \sigma^2\right).$$

Assumption: MAR Bayes predictor, Le Morvan et al. (2020)

For linear model, Gaussian data and in the MAR setting, the Bayes predictor reads:

$$f^\star(X_{obs}, M) = \beta_0^\star + \langle \beta_{obs}^\star, X_{obs} \rangle + \left\langle \beta_{mis}^\star, \mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}) \right\rangle \quad (1)$$

Assumption: GSM Bayes predictor, Le Morvan et al. (2020)

For linear model, Gaussian data and in the Gaussian self-masking setting:

$$f^\star(X_{obs}, M) = \beta_0^\star + \langle \beta_{obs}^\star, X_{obs} \rangle + \langle \beta_{mis}^\star, (Id + D_{mis}\Sigma^{-1}_{mis\,|\,obs}) \times (\tilde{\mu}_{mis} + D_{mis}\Sigma^{-1}_{mis\,|\,obs}(\mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs})))\rangle \quad (2)$$

with $\Sigma_{mis\,|\,obs} := \Sigma_{mis} - \Sigma_{mis,obs}\Sigma_{obs}^{-1}\Sigma_{obs,mis}$ and D diagonal of $(\tilde{\sigma}_1^2, ..., \tilde{\sigma}_d^2)$.

Approximation of $(\Sigma_{obs})^{-1}$: 2^d submatrices to approximate $((\Sigma_{obs})^{-1} \neq (\Sigma^{-1})_{obs})$. Neumann iterates. Order-l approximation $S_{obs}^{(l)}$ of $(\Sigma_{obs})^{-1}$ for all $l \geq 1$:

$$S_{obs}^{(l)} = (Id - \Sigma_{obs})S_{obs}^{(l-1)} + Id \quad \text{and} \quad S_{obs}^{(0)} = Id. \quad (3)$$

Practically, each layer i applies a $x \mapsto W^{(i)}x + x$ transformation followed by a new type of non-linearity, the elementwise multiplication by the mask

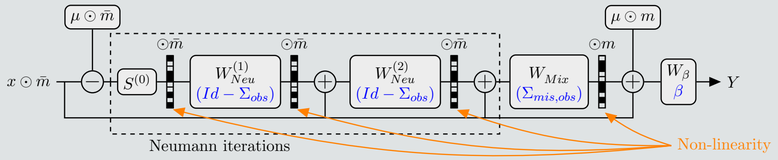


Figure 1: **NeuMiss network architecture with a depth of 4** — $\bar{m} = 1 - m$. Each weight matrix $W^{(k)}$ corresponds to a simple transformation of the covariance matrix indicated in blue. Figure taken from Le Morvan et al. (2020).

NeuMiss extended to classification

Assumption: Probit model

The response $Y \in \{-1, 1\}$ is linked to the complete data X through:

$$\mathbb{P}[Y = 1 \mid X, \beta_0^\star, \beta^\star] = \Phi(\langle X, \beta^\star \rangle + \beta_0^\star),$$

where $\Phi : \mathbb{R} \to [0, 1]$ is the CDF of the standard normal distribution.

Assumption: MAR Bayes predictor

For Gaussian data generated via the probit model in the MAR setting, then:

$$\mathbb{P}[Y = 1 \mid \tilde{X}] = \Phi\left(\frac{\nu}{(1 + \sigma^2)^{\frac{1}{2}}}\right), \qquad f_{\tilde{X}}^\star(\tilde{X}) = \text{sign}\,(\nu),$$

with: $\nu := (1)$

$$\sigma^2 := \beta_{mis}^{\star T}\Sigma_{mis}\beta_{mis}^\star - \beta_{mis}^{\star T}\Sigma_{mis,obs}(\Sigma_{obs})^{-1}\Sigma_{mis,obs}^T\beta_{mis}^\star$$

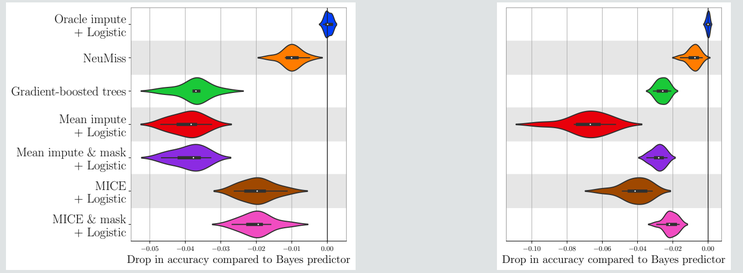
Assumption: GSM Bayes predictor

For Gaussian data generated via the probit model in the Gaussian self-masking setting, then:

$$\mathbb{P}[Y = 1 \mid \tilde{X}] = \Phi\left(\frac{\nu}{(1 + \sigma^2)^{\frac{1}{2}}}\right), \qquad f_{\tilde{X}}^\star(\tilde{X}) = \text{sign}\,(\nu),$$

with: $\nu := (2)$

$$\sigma^2 := \beta_{mis}^{\star T}\left(D_{mis}^{-1} + \Sigma_{mis|obs}^{-1}\right)^{-1}\beta_{mis}^\star$$



(a) MCAR (b) Gaussian self-masking

Figure 2: **Prediction benchmarks.** Accuracy of the benchmarked methods relative to the accuracy of the Bayes predictor, on 10 simulated datasets of 500 000 samples and 50 numerical features drawn from a multivariate gaussian $\mathcal{N}(\mu, \Sigma)$ for MCAR and GSM settings. About 50% values are missing in each.

alexandre.perez@inria.fr