NeuMiss network classifiers: deep learning for classifying with missing values

ADS 2021 Workshop

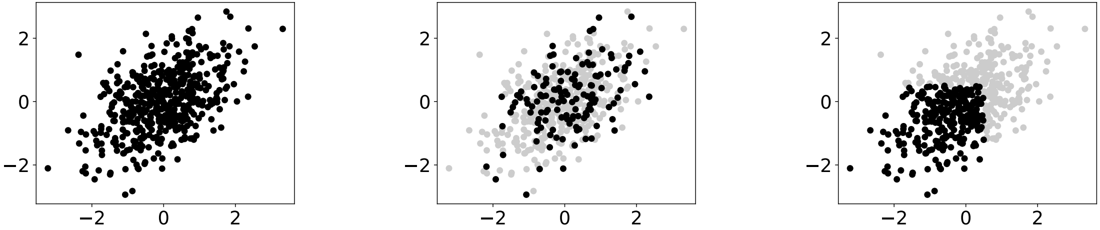
Alexandre Perez-Lebel
Marine Le Morvan
Gaël Varoquaux

Inria Saclay, France

September 17, 2021

Supervised learning with missing values

$X \in \mathbb{R}^d$:	Complete data (unavailable)
$\widetilde{X} \in (\mathbb{R} \cup \{\mathtt{NA}\})^d$:	Incomplete data (available)



Complete
X

Missing Completely At Random (MCAR)
$\tilde{X}$

Missing Not At Random (MNAR)
$\tilde{X}$

Black:	Fully observed samples.	*e.g* $x = (-1, 0.2)$.
Gray:	At least one coordinate missing.	*e.g* $\tilde{x} = (\mathtt{NA}, 0.2)$

Linear regression with missing values

- Gaussian data

$$X \sim \mathcal{N}(\mu, \Sigma)$$

- Linear model

$$Y = \langle X, \beta^\star \rangle + \epsilon, \qquad \epsilon \sim \mathcal{N}\left(0, \sigma^2\right).$$

- Optimal predictor

$$f^\star \in \underset{f:(\mathbb{R} \cup \{\texttt{NA}\})^d \mapsto \mathbb{R}}{\operatorname{argmin}} \mathbb{E}\left[\left(Y - f(\widetilde{X})\right)^2\right]$$

- Q: Expression and approximation of the optimal predictor with missing values $f^\star$?

NeuMiss networks for linear regression with missing values

Linear regression: notations and assumptions

Random variables:

$X \in \mathbb{R}^d$: complete data (unavailable).

$\tilde{X} \in (\mathbb{R} \cup \{\texttt{NA}\})^d$: incomplete data (available).

$M \in \{0,1\}^d$: mask.

$obs(M)$ indices of the observed entries.
$mis(M)$ indices of the missing entries.

Notation abuse: $A_{obs(m),obs(m)} = A_{obs(m)} = A_{obs}$.

Examples of realizations:

$x = (1, 2, 3, 8, 5)$

$\widetilde{x} = (1, \texttt{NA}, 3, 8, \texttt{NA})$

$m = (0, 1, 0, 0, 1)$

$x_{obs} = (1, 3, 8)$
$x_{mis} = (2, 5)$

Assumptions:

Linear model: $Y = \beta_0^\star + \langle \beta^\star, X \rangle + \epsilon,$

Gaussian data: $X \sim \mathcal{N}(\mu, \Sigma)$

Optimal predictor:

$$f^\star \in \underset{f:(\mathbb{R}\cup\{\texttt{NA}\})^d \mapsto \mathbb{R}}{\operatorname{argmin}} \mathbb{E}\left[\left(Y - f(\widetilde{X})\right)^2\right]$$

Linear regression: optimal predictor in MCAR and MAR

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): $\mathbb{P}[M = m \mid X] = \mathbb{P}[M = m]$.

Proposition (MAR optimal predictor, [Le Morvan et al., 2020])

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

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

Linear regression: the NeuMiss architecture

[Le Morvan et al., 2020] introduced NeuMiss, a neural network that approximates the above predictor.

Key point: approximation of $(\Sigma_{obs})^{-1}$ with Neumann iterates.
2^d submatrices to approximate: $(\Sigma_{obs})^{-1} \neq (\Sigma^{-1})_{obs}$.

Order-l approximation $S_{obs}^{(l)}$ of $(\Sigma_{obs})^{-1}$ is defined for all $l \geq 1$ with:

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

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

Linear regression: the NeuMiss architecture

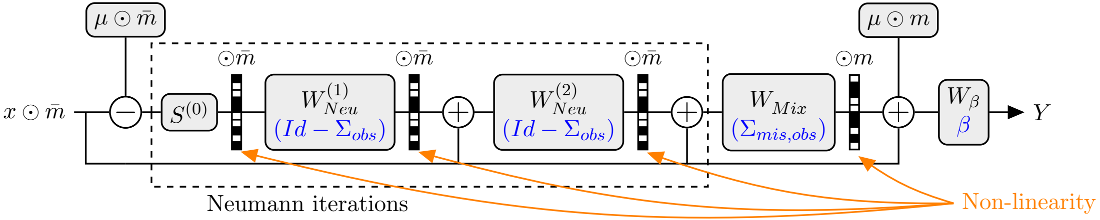


Figure: **NeuMiss network architecture with a depth of 4** — $\bar{m} = 1 - m$. Each weight matrix $W^{(k)}$ corresponds to a simple transformation of the covariance matrix indicated in blue. Figure taken from [Le Morvan et al., 2020].

Reminder of the MAR optimal predictor:

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

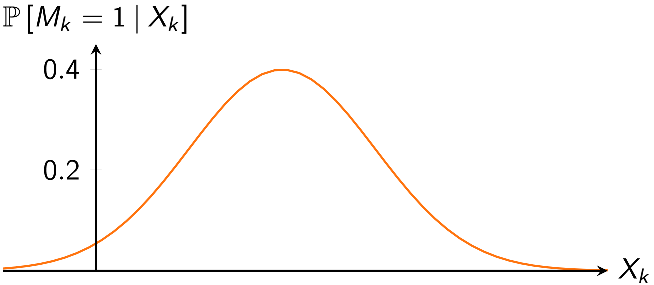
Linear regression: Gaussian self-masking

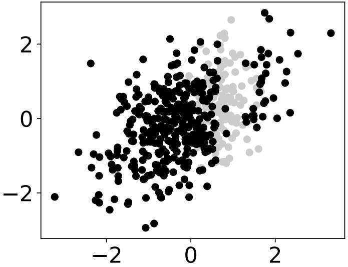
Assumption (Gaussian self-masking, instance of MNAR)

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

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

$$\forall\ 1 \leq k \leq d, \quad \mathbb{P}\left[M_k = 1 \mid X_k\right] = K_k \exp\left(-\frac{(X_k - \tilde{\mu}_k)^2}{2\,\tilde{\sigma}_k^2}\right) \quad \textit{where } 0 < K_k < 1.$$





Extension to specific classification settings

Classification: assumptions

Assumptions:

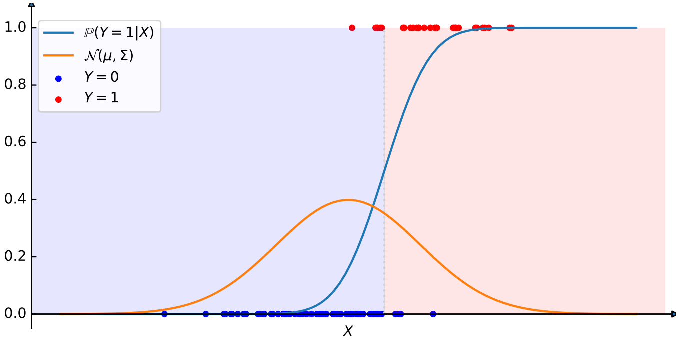
Binary classification: $Y \in \{0,1\}$,

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

Gaussian data: $X \sim \mathcal{N}(\mu, \Sigma)$

Optimal predictor:

$$f^\star \in \underset{f:(\mathbb{R}\cup\{\mathtt{NA}\})^d \mapsto \mathbb{R}}{\operatorname{argmin}} \mathbb{E}\left[\mathbb{1}_{f(\widetilde{X})\neq Y}\right]$$



Classification: optimal predictor in MAR

Proposition (MAR optimal predictor)

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

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

with:

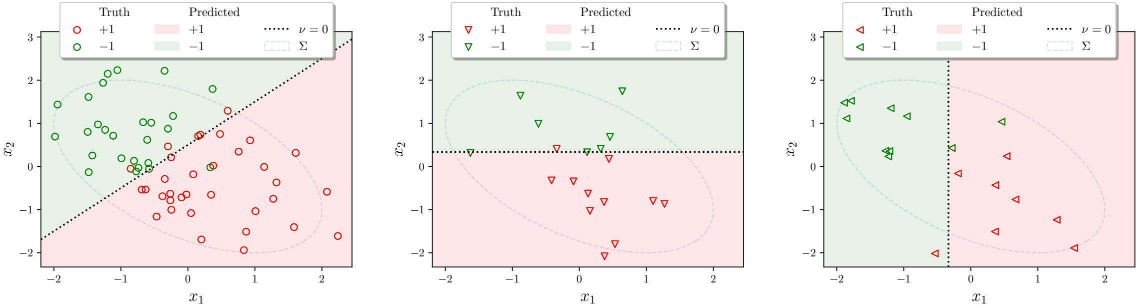
$$\nu := \beta_0^\star + \langle \beta_{obs}^\star, X_{obs} \rangle + \langle \beta_{mis}^\star, \mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}) \rangle$$

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

and the optimal predictor can be written:

$$f_{\tilde{X}}^\star(\tilde{X}) = \mathrm{sign}\,(\nu)$$

Classification: 2D example



(a) No missing values (b) x_1 missing (c) x_2 missing

Figure: **MCAR.** Binary classification using the MCAR optimal predictor on 100 two-dimensional samples drawn from $\mathcal{N}(0, \Sigma)$ with 25% MCAR missing values.

One boundary per missing values pattern: 2^d boundaries.

Classification: optimal predictor in Gaussian self-masking

Proposition (Gaussian self-masking optimal predictor)

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

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

with:

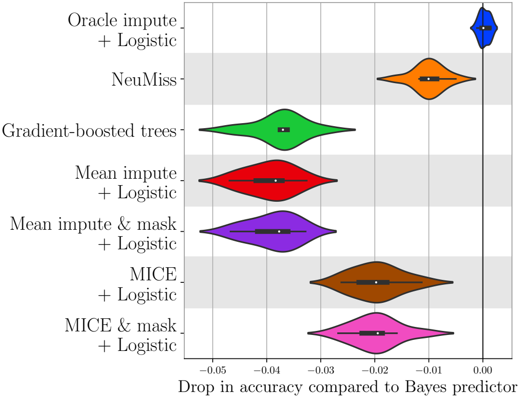
$$\begin{aligned} \nu &= \textit{Optimal predictor in regression (Gaussian self-masking)} \\ \sigma^2 &= \ldots \end{aligned}$$

and the optimal predictor can be written:

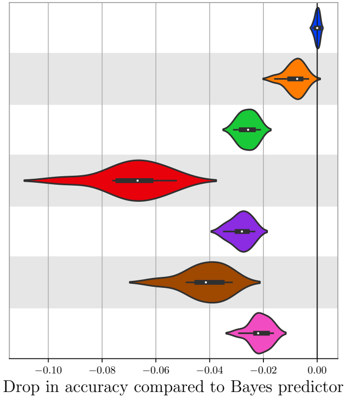
$$f^{\star}_{\tilde{X}}(\tilde{X}) = \mathrm{sign}\,(\nu) \tag{2}$$

Empirical study of NeuMiss in classification with missing values

Classification: empirical study of NeuMiss



(a) MCAR



(b) Gaussian self-masking

Figure: **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.

Conclusion

Conclusion

- Theoretically-grounded architecture and adaptation.
- Adaptation to binary classification with Gaussian-Probit model is easy.
- Still robust to the missing values mechanism.
- As in regression, better performance than imputation.

Future work: loosen assumptions (probit, gaussian, binary).

Thank you for you attention!

References

Le Morvan, M., Josse, J., Moreau, T., Scornet, E., and Varoquaux, G. (2020).
NeuMiss networks: differentiable programming for supervised learning with missing values.
Advances in Neural Information Processing Systems, 33:5980–5990.