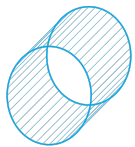
Xử lý dữ liệu khuyết

Nguyễn Trung Đức – 21110269

Giảng viên hướng dẫn: TS. Hoàng Văn Hà



Khoa Toán - Tin học, Trường Đại học Khoa học Tự nhiên

Ngày 20 tháng 1 năm 2025

Table of Contents

Giới thiệu về dữ liệu khuyết

Dữ liệu khuyết (Missing data) là một vấn đề phổ biến trong lĩnh vực Khoa học dữ liệu, đặc biệt với các bài toán học có giám sát (supervised learning).

Nguyên nhân: Do quá trình khảo sát không thu thập được dữ liệu, do lỗi ở phía thiết bị hoặc phần mềm thu thập dữ liệu, do quá trình xử lý dữ liệu,...

Do đó, các dữ liệu khuyết cũng nên được chia ra theo từng loại dựa trên nguyên nhân gây ra dữ liệu bị khuyết để có cách xử lý phù hợp.

Các cơ chế dữ liệu khuyết (Missing data mechanisms)

Cơ chế dữ liệu khuyết là các quy tắc hoặc lý do mà dữ liệu bị khuyết.

Theo Donald B. Rubin (1976) [1], có 3 cơ chế dữ liệu khuyết:

- **MCAR (Missing Completely At Random)**: Dữ liệu bị khuyết không phụ thuộc vào dữ liệu quan sát được hoặc không quan sát được.
- **MAR (Missing At Random)**: Dữ liệu bị khuyết chỉ phụ thuộc vào dữ liệu quan sát được và không phụ thuộc vào dữ liệu không quan sát được (hay chính nó).
- **MNAR (Missing Not At Random)**: Dữ liệu bị khuyết phụ thuộc vào chính giá trị bị khuyết hoặc các giá trị không quan sát được.

Xử lý dữ liệu khuyết

Thông thường, có 3 hướng tiếp cận:

- Loại bỏ toàn bộ các dữ liệu bị khuyết (Listwise deletion).
- Điền khuyết (Impute) dữ liệu bằng các phương pháp điền khuyết.
- Sử dụng các phương pháp có thể tự xử lý dữ liệu khuyết.

Giới thiệu bài toán

Bài toán hồi quy tuyến tính với dữ liệu khuyết.

- Trong thực tế, ta không thể biết được dữ liệu bị khuyết theo cơ chế nào nếu chỉ dựa vào dữ liệu.
- Đa số các hướng tiếp cận đều giả sử với cơ chế MCAR hay MAR.
- Ta muốn phương pháp phải mạnh (robust) với từng loại cơ chế khác nhau.

Giới thiệu bài toán

Bài toán hồi quy tuyến tính với dữ liệu khuyết.

- Trong thực tế, ta không thể biết được dữ liệu bị khuyết theo cơ chế nào nếu chỉ dựa vào dữ liệu.
- Đa số các hướng tiếp cận đều giả sử với cơ chế MCAR hay MAR.
- Ta muốn phương pháp phải mạnh (robust) với từng loại cơ chế khác nhau.

$\Rightarrow$ Bài báo [2] đề xuất phương pháp xấp xỉ bằng chuỗi Neumann cho dự đoán Bayes (dự đoán tối ưu nhất), dưới những giả định cơ chế dữ liệu khuyết khác nhau, cho ra kết quả chính xác, hiệu quả và ổn định hơn so với lại các phương pháp khác.

Một số ký hiệu

Ký hiệu:

- $X \in \mathbb{R}^d$: Dữ liệu đầy đủ
- $\widetilde{X} \in \{\mathbb{R} \cup \{\texttt{NA}\}\}^d$: Dữ liệu bị khuyết
- $M \in \{0,1\}^d$: Vector mask

Với mọi $1 \leqslant j \leqslant d$:

$$M_j = \begin{cases} 1, \text{ nếu } X_j \text{ bị khuyết} \\ 0, \text{ nếu } X_j \text{ không bị khuyết} \end{cases}$$

$$\widetilde{X}_j = \begin{cases} \texttt{NA}, \text{ nếu } M_j = 1 \\ X_j, \text{ nếu } M_j = 0 \end{cases}$$

Ví dụ:

$$\begin{aligned} x &= (1.1, 2.2, -3.5, 4, 5.6) \\ \widetilde{x} &= (1.1, \texttt{NA}, -3.5, 4, \texttt{NA}) \\ m &= (0, 1, 0, 0, 1) \\ x_{obs(m)} &= (1.1, -3.5, 4) \\ x_{mis(m)} &= (2.2, 5.6) \end{aligned}$$

Mô hình bài toán

Ta xét mô hình hồi quy tuyến tính tổng quát với các biến độc lập $X_1, X_2, \ldots, X_d \in \mathbb{R}$, biến phụ thuộc $Y \in \mathbb{R}$, hệ số hồi quy $\beta_0, \beta_1, \ldots, \beta_d$, và sai số ngẫu nhiên $\varepsilon \sim \mathcal{N}(0, \sigma^2)$:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d + \varepsilon \\ &= \beta_0 + \sum_{j=1}^{d} \beta_j X_j + \varepsilon \\ &= \beta_0 + \langle X, \beta \rangle + \varepsilon. \end{aligned}$$

với $\beta = (\beta_1, \beta_2, \ldots, \beta_d) \in \mathbb{R}^d$, $X = (X_1, X_2, \ldots, X_d) \in \mathbb{R}^d$, và $\langle \cdot, \cdot \rangle$ là tích vô hướng.

Dự đoán Bayes (Bayes predictor)

Giả sử quá trình sinh dữ liệu của Y được xác định bởi mô hình hồi quy tuyến tính cho dữ liệu đầy đủ X như sau:

$$Y = \beta_0^* + \langle X, \beta^* \rangle + \varepsilon,$$

với β_0^*, $\beta^* = (\beta_1^*, \beta_2^*, \ldots, \beta_d^*)$ là các hệ số chính xác (true coefficients) để xây dựng mô hình.

Dự đoán Bayes (Bayes predictor)

Giả sử quá trình sinh dữ liệu của Y được xác định bởi mô hình hồi quy tuyến tính cho dữ liệu đầy đủ X như sau:

$$Y = \beta_0^* + \langle X, \beta^* \rangle + \varepsilon,$$

với β_0^*, $\beta^* = (\beta_1^*, \beta_2^*, \ldots, \beta_d^*)$ là các hệ số chính xác (true coefficients) để xây dựng mô hình.

Sẽ khó để ước lượng các hệ số hồi quy khi dữ liệu bị khuyết, đặc biệt là khi d lớn, hay với các mẫu (pattern) dữ liệu khuyết phức tạp theo m.

Dự đoán Bayes (Bayes predictor)

Giả sử quá trình sinh dữ liệu của Y được xác định bởi mô hình hồi quy tuyến tính cho dữ liệu đầy đủ X như sau:

$$Y = \beta_0^* + \langle X, \beta^* \rangle + \varepsilon,$$

với β_0^*, $\beta^* = (\beta_1^*, \beta_2^*, \ldots, \beta_d^*)$ là các hệ số chính xác (true coefficients) để xây dựng mô hình.

Sẽ khó để ước lượng các hệ số hồi quy khi dữ liệu bị khuyết, đặc biệt là khi d lớn, hay với các mẫu (pattern) dữ liệu khuyết phức tạp theo m.

Thay vào đó, ta sẽ tìm một hàm f mà nó ánh xạ dữ liệu bị khuyết $\widetilde{X}$ thành giá trị Y, hay f là mô hình đưa ra dự đoán dựa trên dữ liệu bị khuyết $\widetilde{X}$.

Dự đoán Bayes (Bayes predictor)

Giả sử quá trình sinh dữ liệu của Y được xác định bởi mô hình hồi quy tuyến tính cho dữ liệu đầy đủ X như sau:

$$Y = \beta_0^* + \langle X, \beta^* \rangle + \varepsilon,$$

với β_0^*, $\beta^* = (\beta_1^*, \beta_2^*, \ldots, \beta_d^*)$ là các hệ số chính xác (true coefficients) để xây dựng mô hình.

Sẽ khó để ước lượng các hệ số hồi quy khi dữ liệu bị khuyết, đặc biệt là khi d lớn, hay với các mẫu (pattern) dữ liệu khuyết phức tạp theo m.

Thay vào đó, ta sẽ tìm một hàm f mà nó ánh xạ dữ liệu bị khuyết $\widetilde{X}$ thành giá trị Y, hay f là mô hình đưa ra dự đoán dựa trên dữ liệu bị khuyết $\widetilde{X}$.

Ta có dự đoán Bayes (dự đoán tối ưu):

$$f_{\widetilde{X}}^* \in \underset{f:\widetilde{\mathcal{X}} \to \mathbb{R}}{\arg\min}\, \mathbb{E}[(Y - f(\widetilde{X}))^2].$$

Dự đoán Bayes (Bayes predictor)

Do $\widetilde{X}$ tồn tại những phần tử `NA` đại diện cho dữ liệu khuyết, nên ta khó có thể tính toán f. Ta viết lại dự đoán Bayes dưới dạng một hàm của dữ liệu quan sát được $X_{\text{obs}(M)}$ và vector mask M:

$$
\begin{aligned}
f^*(X_{obs(M)}, M) &= \mathbb{E}[Y|M, X_{\text{obs}(M)}] \\
&= \mathbb{E}[\beta_0^* + \langle \beta^*, X\rangle | M, X_{obs(M)}] \\
&= \beta_0^* + \mathbb{E}[\langle \beta^*_{obs(M)}, X_{obs(M)}\rangle | M, X_{obs(M)}] + \mathbb{E}[\langle \beta^*_{mis(M)}, X_{mis(M)}\rangle | M, X_{obs(M)}] \\
&= \beta_0^* + \langle \beta^*_{obs(M)}, X_{obs(M)}\rangle + \langle \beta^*_{mis(M)}, \mathbb{E}[X_{mis(M)}|M, X_{obs(M)}]\rangle.
\end{aligned}
$$

với $\beta^*_{obs(M)}$, $\beta^*_{mis(M)}$ tương ứng với hệ số hồi quy của các phần tử không bị khuyết và bị khuyết.

Dự đoán Bayes (Bayes predictor)

Do $\widetilde{X}$ tồn tại những phần tử NA đại diện cho dữ liệu khuyết, nên ta khó có thể tính toán f. Ta viết lại dự đoán Bayes dưới dạng một hàm của dữ liệu quan sát được $X_{\text{obs}(M)}$ và vector mask M:

$$
\begin{aligned}
f^*(X_{obs(M)}, M) &= \mathbb{E}[Y|M, X_{\text{obs}(M)}] \\
&= \mathbb{E}[\beta_0^* + \langle \beta^*, X\rangle | M, X_{obs(M)}] \\
&= \beta_0^* + \mathbb{E}[\langle \beta^*_{obs(M)}, X_{obs(M)}\rangle | M, X_{obs(M)}] + \mathbb{E}[\langle \beta^*_{mis(M)}, X_{mis(M)}\rangle | M, X_{obs(M)}] \\
&= \beta_0^* + \langle \beta^*_{obs(M)}, X_{obs(M)}\rangle + \langle \beta^*_{mis(M)}, \mathbb{E}[X_{mis(M)}|M, X_{obs(M)}]\rangle.
\end{aligned}
$$

với $\beta^*_{obs(M)}$, $\beta^*_{mis(M)}$ tương ứng với hệ số hồi quy của các phần tử không bị khuyết và bị khuyết.

Khó để biểu diễn dự đoán Bayes dưới dạng đóng (closed-form) vì nó phụ thuộc vào phân phối của dữ liệu, cũng như đặc điểm riêng của từng cơ chế.

Dự đoán Bayes (Bayes predictor)

Do $\widetilde{X}$ tồn tại những phần tử NA đại diện cho dữ liệu khuyết, nên ta khó có thể tính toán f. Ta viết lại dự đoán Bayes dưới dạng một hàm của dữ liệu quan sát được $X_{\text{obs}(M)}$ và vector mask M:

$$\begin{aligned} f^*(X_{obs(M)}, M) &= \mathbb{E}[Y|M, X_{\text{obs}(M)}] \\ &= \mathbb{E}[\beta_0^* + \langle \beta^*, X\rangle | M, X_{obs(M)}] \\ &= \beta_0^* + \mathbb{E}[\langle \beta^*_{obs(M)}, X_{obs(M)}\rangle | M, X_{obs(M)}] + \mathbb{E}[\langle \beta^*_{mis(M)}, X_{mis(M)}\rangle | M, X_{obs(M)}] \\ &= \beta_0^* + \langle \beta^*_{obs(M)}, X_{obs(M)}\rangle + \langle \beta^*_{mis(M)}, \mathbb{E}[X_{mis(M)}|M, X_{obs(M)}]\rangle. \end{aligned}$$

với $\beta^*_{obs(M)}$, $\beta^*_{mis(M)}$ tương ứng với hệ số hồi quy của các phần tử không bị khuyết và bị khuyết.

Khó để biểu diễn dự đoán Bayes dưới dạng đóng (closed-form) vì nó phụ thuộc vào phân phối của dữ liệu, cũng như đặc điểm riêng của từng cơ chế.

Dẫu vậy, với dữ liệu X tuân theo phân phối chuẩn đa biến: $X \sim \mathcal{N}(\mu, \Sigma)$, ta vẫn có thể biểu diễn dự đoán Bayes dưới dạng đóng cho từng loại cơ chế cụ thể.

Dự đoán Bayes cho cơ chế M(C)AR

Dự đoán Bayes tổng quát có dạng:

$$f^*(X_{obs(M)}, M) = \beta_0^* + \langle \beta^*_{obs(M)}, X_{obs(M)} \rangle + \langle \beta^*_{mis(M)}, \mathbb{E}[X_{mis(M)} | M, X_{obs(M)}] \rangle.$$

Dự đoán Bayes cho cơ chế M(C)AR

$$f^*(X_{obs(M)}, M) = \beta_0^* + \langle \beta^*_{obs(M)}, X_{obs(M)} \rangle + \langle \beta^*_{mis(M)}, \mathbb{E}[X_{mis(M)}|M, X_{obs(M)}] \rangle.$$

Với mọi $m \in \{0,1\}^d$, ta có

Giả thiết cơ chế MCAR

$$P(M = m|X) = P(M = m).$$

Giả thiết cơ chế MAR:

$$P(M = m|X) = P(M = m|X_{obs(m)}).$$

Dự đoán Bayes cho cơ chế M(C)AR

$$f^*(X_{obs(M)}, M) = \beta_0^* + \langle \beta^*_{obs(M)}, X_{obs(M)} \rangle + \langle \beta^*_{mis(M)}, \mathbb{E}[X_{mis(M)} | M, X_{obs(M)}] \rangle.$$

Với mọi $m \in \{0,1\}^d$, ta có

Giả thiết cơ chế MCAR

$$P(M = m|X) = P(M = m).$$

Giả thiết cơ chế MAR:

$$P(M = m|X) = P(M = m|X_{obs(m)}).$$

Mệnh đề (Dự đoán Bayes với M(C)AR)

Giả sử dữ liệu được sinh ra từ mô hình tuyến tính và có phân phối chuẩn đa biến. Giả sử ta có giả thiết cơ chế MCAR hoặc MAR, thì dự đoán Bayes f^* có dạng:

$$f^*(X_{obs}, M) = \beta_0^* + \langle \beta^*_{obs}, X_{obs} \rangle + \langle \beta^*_{mis}, \mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}) \rangle.$$

Dự đoán Bayes cho cơ chế MNAR

Khó khăn: Với MNAR, dữ liệu bị khuyết phụ thuộc vào chính giá trị bị khuyết hoặc các giá trị không quan sát được, nên sẽ khó để mô hình hoá.

Dự đoán Bayes cho cơ chế MNAR

Khó khăn: Với MNAR, dữ liệu bị khuyết phụ thuộc vào chính giá trị bị khuyết hoặc các giá trị không quan sát được, nên sẽ khó để mô hình hoá.

Giả thiết Gaussian self-masking Cơ chế dữ liệu bị khuyết được gọi là self-masked với $P(M|X) = \prod_{k=1}^{d} P(M_k|X_k)$ và $\forall k \in [1, d]$,

$$P(M_k = 1|X_k) = K_k \exp\left(-\frac{1}{2}\frac{(X_k - \widetilde{\mu}_k)^2}{\widetilde{\sigma}_k^2}\right), \quad \text{với } 0 < K_k < 1.$$

- K_k là hằng số điều chỉnh xác suất X_k bị khuyết.
- Xác suất để X_k bị khuyết không phụ thuộc vào các giá trị khác.
- Xác suất để X_k bị khuyết tuân theo phân phối chuẩn $\mathcal{N}(\widetilde{\mu}_k, \widetilde{\sigma}_k^2)$.

Mệnh đề (Dự đoán Bayes với Gaussian self-masking)

Giả sử dữ liệu được sinh ra từ mô hình tuyến tính, tuân theo phân phối chuẩn đa biến và thoả giả thiết Gaussian self-masking. Cho $\Sigma_{mis|obs} = \Sigma_{mis,mis} - \Sigma_{mis,obs}(\Sigma_{obs})^{-1}\Sigma_{obs,mis}$, *và* D *là ma trận đường chéo sao cho* $\mathrm{diag}(D) = (\sigma_1^2, \ldots, \sigma_d^2)$. *Lúc này, dự đoán Bayes được viết dưới dạng:*

$$f^*(X_{obs}, M) = \beta_0^* + \langle \beta_{obs}^*, X_{obs} \rangle + \langle \beta_{mis}^*, (Id + D_{mis}\Sigma_{mis|obs}^{-1})^{-1} \times (\tilde{\mu}_{mis} + D_{mis}\Sigma_{mis|obs}^{-1}(\mu_{mis} + \Sigma_{mis,obs}\left(\Sigma_{obs}\right)^{-1}\left(X_{obs} - \mu_{obs}\right)))\rangle.$$

Dự đoán Bayes cho các cơ chế dữ liệu khuyết

Trường hợp M(C)AR:

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

Trường hợp MNAR:

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

Dự đoán Bayes cho các cơ chế dữ liệu khuyết

Trường hợp M(C)AR:

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

Trường hợp MNAR:

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

Khó khăn: Khi d lớn thì chi phí tính toán cho $(\Sigma_{obs})^{-1}$ sẽ rất lớn.

Xấp xỉ dự đoán Bayes bằng chuỗi Neumann

Chuỗi Neumann cho ma trận A được định nghĩa như sau:

$$\sum_{k=0}^{\infty} A^k = I + A + A^2 + \ldots,$$

và khi chuỗi hội tụ ($\|A\|_2 < 1$), tồn tại nghịch đảo của $(I - A)$ với:

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k.$$

Từ đây, chuỗi Neumann có thể được sử dụng để xấp xỉ nghịch đảo của một ma trận: Xét ma trận A khả nghịch, ta có:

$$A^{-1} = (I - (I - A))^{-1} = \sum_{k=0}^{\infty} (I - A)^k.$$

Xấp xỉ dự đoán Bayes bằng chuỗi Neumann

Xấp xỉ bằng chuỗi Neumann dưới dạng phép lặp:

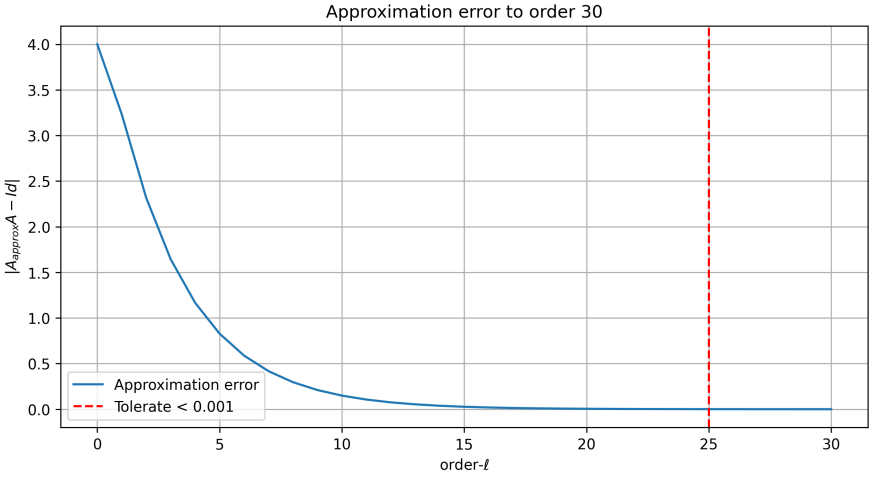
$$S^{(\ell)}_{obs(m)} = (Id - \Sigma_{obs(m)})S^{(\ell-1)}_{obs(m)} + Id.$$

Mệnh đề (Hội tụ tuyến tính của phép lặp Neumann)

Giả sử $\|\Sigma\|_2 < 1$. *Với mọi* $m \in \{0,1\}^d$, *phép lặp* $S^{(\ell)}_{obs(m)}$ *hội tụ tuyến tính về* $(\Sigma_{obs(m)})^{-1}$ *và thoả mãn với mọi* $\ell \geqslant 1$, *với* $\nu_{obs(m)}$ *là trị riêng nhỏ nhất của* $\Sigma_{obs(m)}$,

$$\|Id - \Sigma_{obs(m)}S^{\ell}_{obs(m)}\|_2 \leqslant (1 - \nu_{obs(m)})^{\ell}\|Id - \Sigma_{obs(m)}S^{(0)}_{obs(m)}\|_2.$$

Kết quả thực nghiệm: Xấp xỉ ma trận bằng chuỗi Neumann



Hình: Sai số của xấp xỉ ma trận tới bậc 30.

Xấp xỉ dự đoán Bayes bằng chuỗi Neumann

Ta có xấp xỉ bậc-ℓ của dự đoán Bayes trong trường hợp cơ chế M(C)AR dưới dạng:

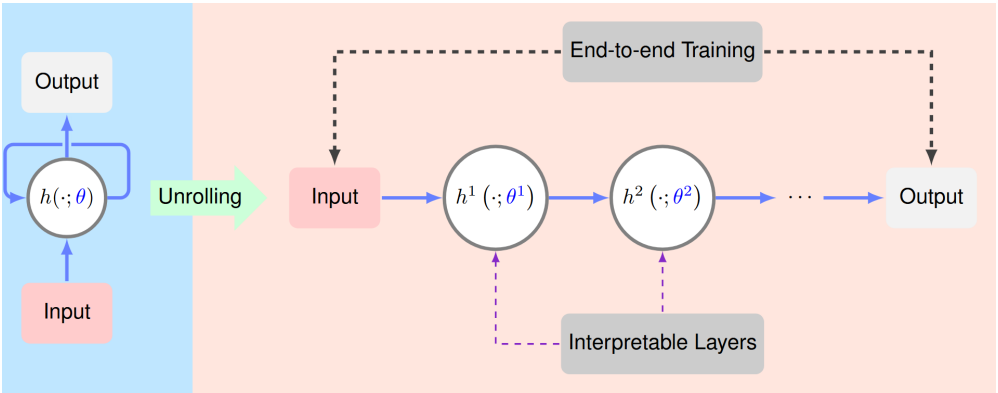
$$f^*_\ell(X_{obs}, M) = \langle \beta^*_{obs}, X_{obs} \rangle + \langle \beta^*_{mis}, \mu_{mis} + \Sigma_{mis,obs} S^{(\ell)}_{obs(m)} (X_{obs} - \mu_{obs}) \rangle.$$

Mệnh đề (Sai số giữa dự đoán Bayes và xấp xỉ bậc-ℓ của nó)

Cho ν là trị riêng nhỏ nhất của Σ. Giả sử dữ liệu được sinh ra qua mô hình tuyến tính, tuân theo phân phối chuẩn đa biến, giả thiết MCAR hay MAR thoả, và $\|\Sigma\|_2 < 1$. Thì với mọi $\ell \geqslant 1$,

$$\mathbb{E}\left[\left(f^*_\ell(X_{obs}, M) - f^*(X_{obs}, M)\right)^2\right] \leqslant \frac{(1-\nu)^{2\ell} \|\beta^*\|_2^2}{\nu} \, \mathbb{E}\left[\|Id - S^{(0)}_{obs(M)} \Sigma_{obs(M)}\|_2^2\right].$$

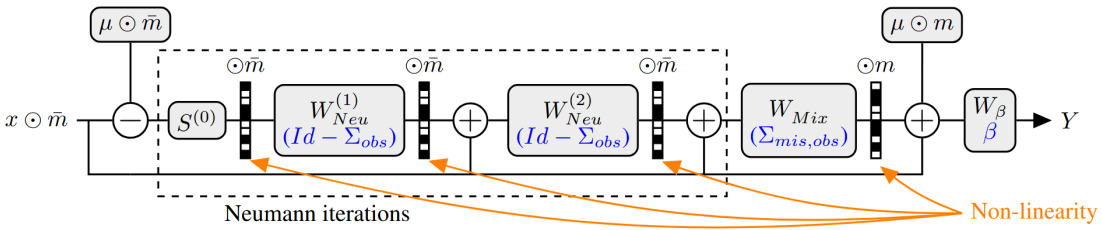
Algorithm Unrolling



Hình: Ý tưởng Algorithm Unrolling [3]: Thuật toán lặp trở thành Neural network (ảnh được trích từ [4])

Mạng NeuMiss

$$f^*_\ell(X_{obs}, M) = \langle \beta^*_{obs}, X_{obs} \rangle + \langle \beta^*_{mis}, \mu_{mis} + \Sigma_{mis,obs} S^{(\ell)}_{obs(m)} (X_{obs} - \mu_{obs}) \rangle.$$



Hình: Mạng NeuMiss độ sâu 4 – $\bar{m} = 1 - m$

Mạng NeuMiss: Xấp xỉ cho cơ chế MNAR

Dự đoán Bayes với cơ chế M(C)AR:

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

Giả sử $D_{mis}\Sigma_{mis|obs}^{-1} \approx Id$, thì dự đoán Bayes cho self-masking trở thành:

$$f^*(X_{obs}, M) \approx \beta_0^* + \langle \beta_{obs}^*, X_{obs} \rangle + \langle \beta_{mis}^*, \frac{1}{2}(\widetilde{\mu}_{mis} + \mu_{mis}) + \frac{1}{2}\Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}) \rangle,$$

hoặc khi $D_{mis}\Sigma_{mis|obs}^{-1} \approx \hat{D}_{mis}$ với $\hat{D}$ là một ma trận ma trận đường chéo:

$$f^*(X_{obs}, M) \approx \beta_0^* + \langle \beta_{obs}^*, X_{obs} \rangle + \langle \beta_{mis}^*, (Id + \hat{D}_{mis})^{-1}(\widetilde{\mu}_{mis} + \hat{D}_{mis}\mu_{mis}) + (Id + \hat{D}_{mis})^{-1}\hat{D}_{mis}\Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}) \rangle.$$

Kết quả thực nghiệm: Mạng NeuMiss

	Train Set		**Validation Set**		**Test Set**	
	R^2	MSE	R^2	MSE	R^2	MSE
NeuMiss depth-10	0.8072	0.208	0.7482	0.2808	0.7578	0.2674

Bảng: Hiệu suất của NeuMiss với độ sâu 10.

Với Bayes rate cho tập train là 0.8205 và tập test là 0.8078.

Kết quả thực nghiệm: Có và khi không có residual connection

	Train Set		Validation Set		Test Set	
	R^2	MSE	R^2	MSE	R^2	MSE
Có residual connection	0.8116	0.2033	0.7424	0.2873	0.7511	0.2747
Không có residual connection	0.8257	0.1880	0.7708	0.2556	0.7641	0.2604

Bảng: Hiệu suất của NeuMiss với độ sâu 10 khi có và không có residual connection.

Kết quả thực nghiệm: Với các độ sâu khác nhau

Độ sâu	Train Set		Validation Set		Test Set	
	R^2	MSE	R^2	MSE	R^2	MSE
1	0.7823	0.2349	0.7500	0.2788	0.7678	0.2563
5	0.7959	0.2202	0.7492	0.2796	0.7586	0.2664
10	0.8191	0.1951	0.7548	0.2735	0.7676	0.2565
15	0.8164	0.1980	0.7624	0.2650	0.7636	0.2610
20	0.8157	0.1988	0.7286	0.3027	0.7431	0.2836

Bảng: Hiệu suất của NeuMiss với các độ sâu khác nhau.

Kết quả thực nghiệm: Với các tỷ lệ dữ liệu khuyết khác nhau

Tỷ lệ khuyết	Train Set		Validation Set		Test Set	
	R^2	MSE	R^2	MSE	R^2	MSE
0.1%	0.8133	0.2014	0.7565	0.2716	0.7506	0.2752
0.2%	0.8145	0.2001	0.7533	0.2751	0.7517	0.2741
0.5%	0.8191	0.1952	0.7700	0.2565	0.7682	0.2558
0.8%	0.7973	0.2187	0.7505	0.2782	0.7572	0.2680

Bảng: Hiệu suất của NeuMiss độ sâu 10 với các tỷ lệ khuyết khác nhau.

Kết quả thực nghiệm: Với các tuỳ chỉnh số lượng mẫu và đặc trưng khác nhau

Mẫu	Đặc trưng	Train Set		Validation Set		Test Set	
		R^2	MSE	R^2	MSE	R^2	MSE
100	10	0.8609	0.1581	-1.6307	2.0416	-12.7133	7.1884
	20	0.9450	0.0596	-324.8323	205.5099	-418.3801	224.8063
1000	10	0.5976	0.4539	-1.5842	2.8056	-0.4908	1.6403
	20	0.7279	0.2953	-42.1458	41.3359	-64.2313	43.3147
5000	10	0.8186	0.2037	0.7657	0.2481	0.7715	0.2491
	20	0.8929	0.1189	-0.2029	1.3849	-0.1616	1.3378
10000	10	0.8138	0.2008	0.7489	0.2800	0.7568	0.2684
	20	0.8783	0.1336	0.1887	0.9035	0.2765	0.8806
50000	10	0.8249	0.1933	0.8124	0.2071	0.8001	0.2232
	20	0.8056	0.2153	0.7145	0.3151	0.7282	0.3021

Kết quả thực nghiệm: Với các phương pháp khác

	Train Set	**Test Set**
NeuMiss depth-10	**0.8072**	**0.7578**
KNN imputer + LR	0.6929	0.7018
Simple imputer + LR	0.6574	0.6478
SoftImpute + LR	0.7604	0.7524
MissForest + LR	0.7708	0.7442
MICE + LR	0.7496	0.7313
Simple imputer + MLP regressor	0.8022	0.7330

Bảng: Hiệu suất của NeuMiss độ sâu 10 với các phương pháp khác.

References I

[1] Donald B Rubin. "Inference and missing data". In: *Biometrika* 63.3 (1976), pp. 581–592.

[2] Marine Le Morvan et al. "NeuMiss networks: differentiable programming for supervised learning with missing values.". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 5980–5990.

[3] Karol Gregor and Yann LeCun. "Learning fast approximations of sparse coding". In: *Proceedings of the 27th international conference on international conference on machine learning*. 2010, pp. 399–406.

[4] Vishal Monga, Yuelong Li, and Yonina C Eldar. "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing". In: *IEEE Signal Processing Magazine* 38.2 (2021), pp. 18–44.

Thank You