ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN - TIN HỌC

Báo cáo Seminar Phương pháp toán

Xử lý dữ liệu khuyết

Sinh viên thực hiện: **Nguyễn Trung Đức**
Mã số sinh viên: **21110269**
Giảng viên hướng dẫn: **TS. Hoàng Văn Hà**

TP. Hồ Chí Minh - Ngày 20 tháng 1 năm 2025

Mục lục

Danh mục các thuật ngữ và từ viết tắt

Thuật ngữ	Ý nghĩa
Bayes predictor	Dự đoán Bayes
Imputation, Impute	Điền khuyết
Features	Đặc trưng
Mask	Che
Neural network	Mạng nơ-ron nhân tạo
Activation function	Hàm kích hoạt
Loss function	Hàm mất mát
Learning rate	Tốc độ học
Layer	Lớp (của Neural network)
Hidden layer	Lớp ẩn
Hidden unit	Đơn vị ẩn
Train set	Tập dùng để huấn luyện
Validation set	Tập dùng để đánh giá
Test set	Tập dùng để kiểm tra
Epoch	Một lần duyệt qua hết các dữ liệu trong train set
Batch size	Số lượng mẫu dữ liệu trong một lần huấn luyện

Từ viết tắt	Ý nghĩa
MCAR	Missing Completely At Random
MAR	Missing At Random
MNAR	Missing Not At Random
EM	Expectation Maximization
LR	Linear Regression
MLP	Multilayer perceptron
KNN	K-Nearest Neighbors
SVM	Support vector machine
PCA	Principal Component Analysis
MICE	Multiple Imputation by Chained Equations
GAN	Generative adversarial network
MSE	Mean Squared Error

Tóm tắt

Bài báo [1] xử lý vấn đề dữ liệu khuyết trong các bài toán học có giám sát, cụ thể là các bài toán hồi quy tuyến tính. Dựa trên việc xấp xỉ dự đoán Bayes (mô hình dự đoán tốt nhất theo lý thuyết) trên tập dữ liệu khuyết, bài báo đề xuất sử dụng chuỗi Neumann kết hợp với phương pháp Algorithm Unrolling [2] để xây dựng một kiến trúc neural network có tên là NeuMiss. Mạng NeuMiss sử dụng các hàm kích hoạt phi tuyến là các mask (chỉ số) của dữ liệu khuyết, giúp mạng hoạt động hiệu quả với các tập dữ liệu có độ lớn trung bình trở lên với các cơ chế dữ liệu khuyết khác nhau bao gồm cơ chế dữ liệu khuyết hoàn toàn ngẫu nhiên (MCAR), cơ chế dữ liệu khuyết ngẫu nhiên (MAR), và cơ chế dữ liệu khuyết không ngẫu nhiên (MNAR). Đặc biệt, mạng NeuMiss hoạt động hiệu quả trong những trường hợp mà các phương pháp truyền thống thường gặp khó khăn.

Mục đích của bài báo cáo này là đi trình bày lại các kiến thức nền tảng và kiến trúc của mạng NeuMiss, cũng như một số kết quả thực nghiệm cho các cài đặt khác nhau.

1 Kiến thức chuẩn bị

Trong mục này, chúng tôi trình bày một số kiến thức chuẩn bị cho các nội dung được nghiên cứu trong các phần sau.

1.1 Phân phối chuẩn đa biến

Phân phối chuẩn đa biến $X \sim \mathcal{N}(\mu, \Sigma)$ với vector trung bình là $\mu \in \mathbb{R}^d$ và ma trận hiệp phương sai đối xứng, nửa xác định dương $\Sigma \in \mathbb{R}^{d\times d}$, có hàm mật độ xác suất như sau:

$$f(x|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right),\ x \in \mathbb{R}^d,$$

với $|\Sigma|$ là định thức của Σ.

Lý do sử dụng phân phối chuẩn đa biến cho bài toán xử lý dữ liệu khuyết?

- Nhiều phương pháp thống kê dựa trên giả định dữ liệu có phân phối chuẩn, dễ làm việc hơn vì được nghiên cứu lâu dài, cùng với Định lý giới hạn trung tâm khiến cho phân phối chuẩn được sử dụng rộng rãi.
- Dữ liệu thực tế thường có nhiều cột đặc trưng (features), dẫn đến dữ liệu có số chiều cao, nên ta cần một phân phối có khả năng xử lý dữ liệu nhiều chiều và cho ta thấy được mối tương quan giữa các đặc trưng với nhau.

- Phân phối chuẩn đa biến cho phép mô hình hoá các phân phối xác suất đồng thời của các đặc trưng khác nhau.

Trong khuôn khổ bài báo cáo này, từ giờ về sau, hàm mật độ xác suất của phân phối chuẩn đa biến sẽ được gọi là hàm Gauss để cho thuận tiện.

Bổ đề 1.1 *(Tích của hai hàm Gauss là một hàm Gauss)*

Cho $f(x) = \exp\left(-\frac{1}{2}(x-a)^\top A^{-1}(x-a)\right)$ *và* $g(x) = \exp\left(-\frac{1}{2}(x-b)^\top B^{-1}(x-b)\right)$ *là hai hàm Gauss, với* A *và* B *là ma trận nửa xác định dương, ta có:*

$$f(x)g(x) = \exp\left(-\frac{1}{2}(a-b)^\top (A+B)^{-1}(a-b)\right)\exp\left(-\frac{1}{2}(x-\mu_p)^\top \Sigma_p^{-1}(x-\mu_p)\right),$$

với μ_p *và* Σ_p *lần lượt là vector trung bình và ma trận hiệp phương sai của tích hai hàm Gauss, phụ thuộc vào* a*,* A*,* b*, và* B*.*

Chứng minh: Xem mục A.2 trong bài báo [1].

1.2 Phân phối chuẩn có điều kiện

Một tính chất quan trọng của phân phối chuẩn đa biến đó là nếu 2 tập hợp biến ngẫu nhiên có phân phối chuẩn đa biến đồng thời (Joint Gaussian), thì phân phối có điều kiện của một tập khi biết tập còn lại cũng là một phân phối chuẩn đa biến, tức

$$f(x_a|x_b) = \mathcal{N}(x_a|\mu_{a|b}, \Sigma_{a|b}).$$

Giả sử x là vector d chiều có phân phối chuẩn đa biến $\mathcal{N}(x|\mu, \Sigma)$. Ta chia x thành 2 tập con riêng biệt x_a và x_b được cho bởi

$$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix}.$$

Lúc này x cũng được gọi là phân phối chuẩn đa biến bị chia (Partitioned Multivariate Gaussian). Ta cũng định nghĩa vector trung bình

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix},$$

và ma trận hiệp phương sai

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}.$$

Ta có:

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b),$$
$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}.$$

Chứng minh: Xem mục 2.3.1 của [3] hoặc [4]. ■

Ngoài ra, nếu ma trận hiệp phương sai $\Sigma_{ab} = 0$, ta có $\mu_{a|b} = \mu_a$, $\Sigma_{a|b} = \Sigma_{aa}$, tức x_a và x_b độc lập với nhau, với $P(x_a|x_b) = P(x_a)$. Tức là khi ta lấy tích của các biến ngẫu nhiên độc lập với nhau, ta chỉ cần lấy các phần tử đường chéo của ma trận hiệp phương sai Σ.

1.3 Chuẩn ma trận (Induced vector norm)

Với chuẩn 2 của vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ có dạng:

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{x^\top x} = \sqrt{\langle x, x \rangle},$$

ta xét chuẩn 2 (induced-2 norm) cho ma trận A, hay còn được biết đến là chuẩn phổ (spectral norm):

$$\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{\substack{x \neq 0 \\ \|x\|_2 = 1}} \|Ax\|_2 = \sigma_{max}(A),$$

với $\sigma_{max}(A)$ là giá trị suy biến (singular value) lớn nhất của A. Và khi A đối xứng, xác định dương, thì giá trị suy biến tương ứng với lại trị riêng của ma trận, tức $\|A\|_2 = \lambda_{\max}$.

1.4 Chuỗi Neumann

Chuỗi Neumann thường làm việc với các toán tử trong các không gian hàm. Nhưng trong khuôn khổ bài báo cáo, ta chỉ cần làm việc với ma trận – toán tử biến đổi tuyến tính.

Với ma trận $A \in \mathbb{R}^{d \times d}$, chuỗi Neumann của A được biểu diễn dưới dạng

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

với I là ma trận đơn vị. Chuỗi Neumann hội tụ với điều kiện chuẩn phổ của ma trận A bé hơn 1 ($\|A\|_2 < 1$). Và khi chuỗi hội tụ, tồn tại nghịch đảo của $(I - A)$ với:

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

Để chứng minh điều trên, ta xét:

$$S_n = \sum_{k=0}^{n} A^k.$$

Ta có:

$$\lim_{n\to\infty}(I-A)S_n = \lim_{n\to\infty}\left(\sum_{k=0}^{n} A^k - \sum_{k=0}^{n} A^{k+1}\right) = \lim_{n\to\infty}\left(I - A^{n+1}\right) = I.$$

Vậy nghịch đảo của $I - A$ là chuỗi Neumann của A.

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:

$$\begin{aligned} A^{-1} &= (I - I + A)^{-1} \\ &= (I - (I - A))^{-1} \\ &= (I - T)^{-1}, \end{aligned}$$

với $T = I - A$. Nếu T thoả điều kiện $\|T\|_2 < 1$, thì ta có theo chuỗi Neumann:

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

Và do đó:

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

Vì với điều kiện $\|I - A\|_2 < 1$, $I - A$ dần hội tụ khi $k \to \infty$, nên chuỗi chặt cụt (truncated) tại l hữu hạn có thể cho ta 1 xấp xỉ nghịch đảo ma trận tốt.

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

1.5 Bán kính phổ

Cho ma trận $A \in \mathbb{C}^{d\times d}$ với các trị riêng $\lambda_1, \ldots, \lambda_d$, bán kính phổ (spectral radius) $\rho(A)$ của A là giá trị tuyệt đối trị riêng lớn nhất của ma trận A:

$$\rho(A) = \max\{|\lambda_1|, \ldots, |\lambda_d|\}.$$

Khi A là ma trận thực đối xứng, bán kính phổ của A bằng với chuẩn phổ của A:

$$\rho(A) = \|A\|_2 = \lambda_{\max}.$$

1.6 Tỷ số Rayleigh (Rayleigh quotient)

Tỷ số Rayleigh thường được sử dụng để xấp xỉ trị riêng của ma trận thực đối xứng A dưới dạng:

$$r(x) = \frac{x^\top Ax}{x^\top x},$$

với $x \neq 0$.

Để cho đơn giản, ta cho $\|x\| = 1$, lúc này, việc tìm trị riêng lớn nhất của A có dạng:

$$\lambda_{\max}(A) = \rho(A) = \max_{\|x\|=1} x^\top Ax.$$

Người đọc có thể đọc thêm về phần kiến thức này ở phần V trong [5].

1.7 Phép nhân Hadamard (Hadamard product)

Trong các phép toán ma trận, phép nhân Hadamard, ký hiệu là $\odot$, là phép nhân từng phần tử (element-wise) của 2 ma trận với nhau.

Xét $A \in \mathbb{R}^{m\times n}$ và $B \in \mathbb{R}^{m\times n}$, với lần lượt a_{ij}, b_{ij} nằm ở dòng thứ i và cột thứ j của ma trận A, B. Phép nhân Hadamard của ma trận A và B là:

$$\mathbf{A} \odot \mathbf{B} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & \dots & a_{1n}b_{1n} \\ a_{21}b_{21} & a_{22}b_{22} & \dots & a_{2n}b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}b_{m1} & a_{m2}b_{m2} & \dots & a_{mn}b_{mn} \end{bmatrix}.$$

1.8 Điểm bất động

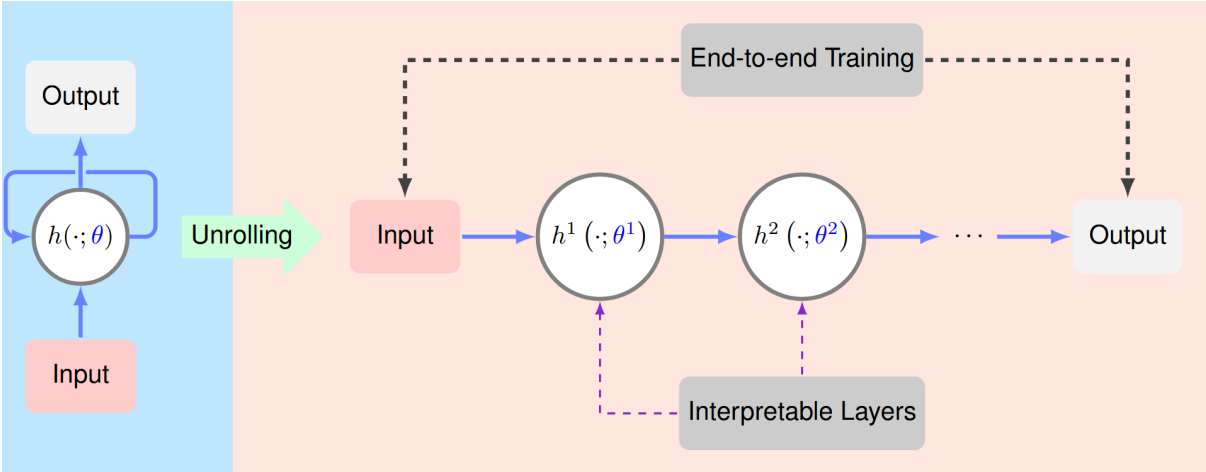
Điểm bất động (fixed point) x^* là điểm thoả

$$f(x^*) = x^*,$$

với hàm f cố định. Nói cách khác, điểm bất động là điểm không thay đổi với 1 phép biến đổi cho trước.

1.9 Algorithm Unrolling

Ý tưởng của Algorithm Unrolling được giới thiệu lần đầu trong công trình của Gregor và LeCun [2], nhằm liên kết các thuật toán lặp (Iterative algorithms) với các neural networks. Cụ thể, mỗi vòng lặp trong thuật toán được mô hình hóa như một lớp (layer) trong mạng, qua đó thuật toán lặp được biểu diễn như một chuỗi các lớp liên tiếp. Việc truyền qua mạng tương đương với việc thực thi thuật toán lặp một số lần hữu hạn.



Hình 1.1: **Tổng quan Algorithm unrolling:** cho một thuật toán lặp (bên trái), một neural network tương ứng (bên phải) có thể được tạo bằng cách "trải" các phép lặp h ra. Phép lặp h được thực thi với số lần hữu hạn, tương ứng với các lớp $h^1, h^2, \ldots$ của mạng. Mỗi phép lặp h phụ thuộc vào tham số θ, và tham số này được biểu diễn bởi tham số của mạng dưới dạng $\theta^1, \theta^2, \ldots$. Thay vì phải xác định các tham số một cách thủ công trong thuật toán lặp, ta học $\theta^1, \theta^2, \ldots$ trực tiếp từ quá trình huấn luyện mô hình. Bằng cách này, mạng có thể đạt được hiệu suất tốt hơn so với thuật toán lặp ban đầu. Ngoài ra, các lớp của mạng được kế thừa tính diễn giải (interpretability) một cách tự nhiên từ quá trình lặp. Các tham số có thể học được có màu xanh. (được trích dẫn từ [6])

2 Giới thiệu về xử lý dữ liệu khuyết

Dữ liệu bị khuyết 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). Mục tiêu chính của các bài toán này là dự đoán các giá trị hoặc phân loại các dữ liệu mới dựa trên các đặc trưng từ dữ liệu cũ. Phần lớn các phương pháp nhằm giải quyết các bài toán này đều được thiết kế để hoạt động trên tập dữ liệu hoàn chỉnh. Do đó, khi dữ liệu đầu vào bị khuyết, mô hình không thể học đầy đủ các quy luật cần thiết để có thể hoạt động tốt, làm giảm chất lượng của mô hình. Vậy nên ta cần tiền xử lý các giá trị bị khuyết, hoặc sử dụng các mô hình có thể tự xử lý dữ liệu khuyết trên các tập dữ liệu không hoàn chỉnh.

2.1 Các cơ chế dữ liệu khuyết

Dữ liệu bị khuyết có thể đến từ nhiều nguyên nhân khác nhau, ví dụ như: do người trả lời không cung cấp thông tin, do quá trình khảo sát không thu thập đủ 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. Việc sử dụng những phương pháp xử lý không phù hợp với từng loại dữ liệu khuyết có thể dẫn tới kết luận sai lệch từ dữ liệu.

Theo Donald B. Rubin, dữ liệu khuyết được chia thành 3 loại chính: MCAR, MAR, và MNAR [7].

- **Dữ liệu khuyết hoàn toàn ngẫu nhiên (Missing Completely At Random – MCAR)** là dữ liệu bị khuyết hoàn toàn không phụ thuộc vào dữ liệu quan sát được hoặc không quan sát được. Hay nói cách khác, việc dữ liệu bị khuyết là "hoàn toàn ngẫu nhiên".

  Ví dụ: Trong một khảo sát nghiên cứu, một số phiếu trả lời bị thất lạc ngẫu nhiên do quá trình vận chuyển của bưu điện. Việc thất lạc phiếu khảo sát là hoàn toàn ngẫu nhiên và không liên quan đến nội dung câu trả lời, người tham gia, hoặc bất kỳ yếu tố nào khác trong khảo sát.

- **Dữ liệu khuyết ngẫu nhiên (Missing At Random – MAR)** là dữ liệu bị khuyết chỉ phụ thuộc vào dữ liệu quan sát được mà không phụ thuộc vào dữ liệu không quan sát được (hay chính nó).

  Ví dụ: Trong một khảo sát có các câu hỏi về chủ đề nhạy cảm, phụ nữ thường ngại và không trả lời các câu hỏi nhạy cảm, có thể là do họ không thoải mái với câu hỏi đó. Tuy nhiên, đàn ông thì không có xu hướng như vậy và sẵn lòng trả lời. Như vậy, dữ liệu bị khuyết theo cơ chế MAR vì việc thiếu dữ liệu về câu trả lời của các câu hỏi nhạy cảm phụ thuộc vào giới tính của người được khảo sát, chứ không phải do giá trị của chính các câu hỏi nhạy cảm.

- **Dữ liệu khuyết không ngẫu nhiên (Missing Not At Random – MNAR)** là dữ liệu bị khuyết không thuộc một trong hai cơ chế dữ liệu khuyết bên trên, tức dữ liệu bị khuyết phụ thuộc vào dữ liệu quan sát được và không quan sát được, hoặc chỉ phụ thuộc vào dữ liệu dữ liệu không quan sát được, hay chính nó.

  MNAR thường xảy ra do người cung cấp thông tin từ chối tiết lộ thông tin cá nhân hay thông tin nhạy cảm.

  Ví dụ: Trong một cuộc khảo sát về sức khỏe tâm lý, người tham gia được yêu cầu trả lời câu hỏi về mức độ trầm cảm của họ. Tuy nhiên, những người có mức độ trầm cảm cao thường có xu hướng không trả lời hoặc từ chối cung cấp thông tin

về tình trạng của mình do cảm giác lo ngại, xấu hổ, hoặc sợ bị đánh giá. Tức dữ liệu bị khuyết phụ thuộc trực tiếp vào chính giá trị của mức độ trầm cảm.

Với MNAR, ta cần phải mô hình cơ chế dữ liệu khuyết vì sự khuyết dữ liệu có thể cung cấp nhiều thông tin hữu ích. MNAR thường phức tạp hơn MCAR hay MAR bởi dữ liệu bị khuyết phụ thuộc vào chính nó, không phải do ngẫu nhiên. Chính vì vậy, cơ chế MNAR thường xuất hiện ở trong thực tế, nhưng do nó khó xử lý nên phần lớn các phương pháp xử lý dữ liệu khuyết đều giả định dữ liệu bị khuyết với cơ chế MCAR và MAR, và các phương pháp này thường không thể áp dụng cho MNAR.

2.2 Phương pháp xử lý dữ liệu khuyết

Xử lý dữ liệu khuyết là một vấn đề đã được nghiên cứu trong khoảng thời gian dài. Các nhà nghiên cứu đã đề xuất nhiều phương pháp khác nhau, và thường sẽ có 3 hướng xử lý dữ liệu khuyết chính:

- 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 hay không bị ảnh hưởng bởi dữ liệu khuyết.

Thông thường, cách đơn giản để xử lý dữ liệu khuyết là xoá tất cả các dòng có chứa những dữ liệu khuyết để tập dữ liệu khuyết trở thành tập dữ liệu đầy đủ. Nhưng việc xoá tất cả dữ liệu khuyết sẽ khiến tập dữ liệu nhỏ đi, từ đó dữ liệu sẽ không còn nhiều thông tin hữu ích, chưa kể có một số loại dữ liệu khuyết phụ thuộc vào dữ liệu đã quan sát được, hay dữ liệu bị khuyết không hoàn toàn ngẫu nhiên, sẽ cung cấp nhiều thông tin hữu ích như mục 2.1 có đề cập.

Từ đó, việc xoá tất cả dữ liệu khuyết không phải là phương pháp tốt trong nhiều trường hợp vì nó có thể bỏ qua những dữ liệu có ích. Thay vào đó, ta sẽ muốn chèn dữ liệu vào những chỗ mà dữ liệu bị khuyết, hay còn được gọi là phương pháp điền khuyết (Imputation).

Điền khuyết là phương pháp điền dữ liệu khuyết bằng các giá trị cố định. Có nhiều cách để điền khuyết dữ liệu như: Simple Impute (Điền khuyết đơn giản), Interactive Impute (Điền khuyết tuần tự), Multiple Impute (Đa điền khuyết), MICE (Multiple Imputation by Chained Equations), PCA (Principal Component Analysis), hay sử dụng các thuật toán Machine learning như: Random Forest, KNN, Decision Tree, SVM,... không cần giả định về phân phối của dữ liệu do chúng đều là các phương pháp phi tham số.

Tuy nhiên, điền khuyết trong một số trường hợp cũng không phải lựa chọn tốt, khi mà nó có thể thay đổi phân phối của dữ liệu, làm mất tính liên kết giữa các đặc trưng với nhau.

Cùng với sự phát triển của các neural network, một số công trình nghiên cứu đã tận dụng sức mạnh của các kiến trúc mạng khác nhau để xử lý các vấn đề với dữ liệu khuyết,... ví dụ như: MLP (Multilayer perceptron), GAIN hay MisGAN (sử dụng kiến trúc GAN để điền khuyết dữ liệu), MIWAE (dựa trên Variational Autoencoder),...

Nhưng những phương pháp này cần có nhiều dữ liệu đầu vào hơn so với các phương pháp điền khuyết trên, cũng như các cơ chế dữ liệu khác nhau thường đòi hỏi các cách xử lý khác nhau.

Trong thực tế, các nhà phân tích dữ liệu thường phải giả sử 1 trong 3 loại cơ chế dữ liệu khuyết vì họ không biết rằng dữ liệu bị khuyết thực tế sẽ thuộc vào kiểu cơ chế nào. Bài báo [1] đề xuất một neural network có thể xử lý đồng thời cả 3 loại cơ chế trong bài toán hồi quy tuyến tính, mà không cần các bước điền khuyết truyền thống.

3 Tổng quan các mô hình nghiên cứu với dữ liệu khuyết

Để mô hình hoá vấn đề dữ liệu khuyết trong các bài toán hồi quy tuyến tính, ta định nghĩa một số ký hiệu:

Ta xét tập dữ liệu $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ gồm các cặp (X_i, Y_i) độc lập với nhau, được phân phối theo cặp tổng quát (X, Y), trong đó $X \in \mathbb{R}^d$ đại diện cho vector đặc trưng với d chiều, gồm các biến độc lập, và biến phụ thuộc $Y \in \mathbb{R}$.

Ta định nghĩa vector chỉ mục (indicator vector) $M \in \{0, 1\}^d$, sao cho với mọi $1 \leqslant j \leqslant d$:

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

Lúc này, vector ngẫu nhiên M đóng vai trò như một mask cho các giá trị dữ liệu bị khuyết trên X.

Ta kí hiệu $\mathcal{X} \in \mathbb{R}^d$ là không gian đầu vào, $\widetilde{\mathcal{X}}$ là không gian tương tự như $\mathcal{X} = (\mathbb{R} \cup \{\texttt{NA}\})^d$ nhưng có chứa thêm phần tử `NA` tượng trưng cho dữ liệu bị khuyết.

Ta kí hiệu vector đặc trưng bị khuyết $\widetilde{X} \in \widetilde{\mathcal{X}}$, sao cho với mọi $1 \leqslant j \leqslant d$:

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

Với giá trị hiện thực (realization), ký hiệu là chữ cái in thường: m của M, ta ký hiệu:

$obs(m)$ là các chỉ số của các giá trị 0 của m (các giá trị không bị khuyết),
$mis(m)$ là các chỉ số của các giá trị khác 0 của m (các giá trị bị khuyết).

Trong ngữ cảnh dữ liệu bị khuyết:

$X_{obs(M)}$ là các giá trị quan sát được trong X,
$X_{mis(M)}$ là các giá trị không quan sát được trong X.

Ví dụ 3.1 *Giả sử với giá trị hiện thực:* $x = (1.1, 2.2, -3.5, 4, 5.6) \in X$*, ta có:*

$$\begin{aligned} \widetilde{x} &= (1.1, \texttt{NA}, -3.5, 4, \texttt{NA}) \\ m &= (0, 1, 0, 0, 1) \\ obs(m) &= (0, 2, 3) \\ x_{obs(m)} &= (1.1, -3.5, 4) \\ mis(m) &= (1, 4) \\ x_{mis(m)} &= (2.2, 5.6). \end{aligned}$$

Khi không có sự nhầm lẫn nào, ta bỏ phần phụ thuộc m để tiện ký hiệu, ví dụ như X_{obs} thay vì $X_{obs(m)}$.

3.1 Hồi quy tuyến tính với dữ liệu khuyết

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 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d + \varepsilon,$$

hay ta có thể viết dưới dạng

$$Y = \beta_0 + \sum_{j=1}^{d} \beta_j X_j + \varepsilon = \beta_0 + \langle X, \beta \rangle + \varepsilon,$$

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.

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:

$$Y = \beta_0^* + \langle X, \beta^* \rangle + \varepsilon, \tag{3.1}$$

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

Do dữ liệu X có thể bị khuyết, nên việc ước lượng các hệ số hồi quy của mô hình có thể trở nên khó khăn. Đặc biệt khi với số chiều 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, tức f là mô hình đưa ra dự đoán dựa trên dữ liệu bị khuyết $\widetilde{X}$.

Vậy làm sao để biết f sẽ cho ra kết quả dự đoán tốt nhất? Kết quả dự đoán tốt nhất khi sai số của dữ liệu được dự đoán $f(\widetilde{X})$ và dữ liệu quan sát Y là nhỏ nhất. Ở đây, ta sử dụng hàm mất mát (loss function) bình phương sai số (squared error) có dạng:

$$L(Y, f(\widetilde{X})) = (Y - f(\widetilde{X}))^2.$$

Tiếp theo, ta muốn f hoạt động tốt trên toàn bộ tập dữ liệu D_n, bao gồm cả tập test, chứ không phải chỉ riêng 1 điểm dữ liệu. Do hàm mất mát chỉ dùng để đo sai số của 1 giá trị dự đoán, nên ta lấy kỳ vọng (trung bình) của hàm mất mát để giảm sai số dự đoán trung bình trên các điểm dữ liệu để đánh giá độ tốt của mô hình:

$$\mathbb{E}[L(Y, f(\widetilde{X}))] = \mathbb{E}[(Y - f(\widetilde{X}))^2],$$

Kỳ vọng của hàm mất mát ngoài ra còn có một tên gọi khác là hàm rủi ro (risk). Vì ta muốn mô hình f đưa ra dự đoán tốt nhất dựa trên dữ liệu khuyết $\tilde{X}$, tức mô hình có hàm rủi ro đạt giá trị thấp nhất (còn được gọi là Bayes risk), nên ta đi giải quyết bài toán tối ưu:

$$f_{\tilde{X}}^* \in \underset{f:\widetilde{\mathcal{X}} \to \mathbb{R}}{\arg\min}\, \mathbb{E}[(Y - f(\widetilde{X}))^2]. \tag{3.2}$$

Phương trình này nhắm tới việc tìm $f_{\tilde{X}}^*$ sao cho khi áp dụng vào dữ liệu bị khuyết $\tilde{X}$, nó sẽ đưa ra dự đoán trung bình gần nhất có thể với giá trị thực của Y. Lúc này, $f_{\tilde{X}}^*$ còn được gọi là dự đoán Bayes (Bayes predictor). Hay có thể nói, dự đoán Bayes $f_{\tilde{X}}^*$ là mô hình dự đoán Y tốt nhất dựa trên dữ liệu khuyết $\widetilde{X}$.

Do $\widetilde{\mathcal{X}}$ tồn tại những phần tử `NA` đại diện cho dữ liệu khuyết, nên ta khó có thể tính toán f dựa trên $\widetilde{X}$. Từ đây, ta viết lại phương trình của dự đoán Bayes (3.2) dưới dạng kỳ vọng

có điều kiện của giá trị Y với dữ liệu bị khuyết $\widetilde{X}$ để dễ tính toán:

$$\begin{aligned} f^*_{\widetilde{X}}(\widetilde{X}) &= \mathbb{E}[Y|\widetilde{X}] \\ &= \mathbb{E}\left[Y|M, X_{obs(M)}\right] \\ &= \sum_{m\in\{0,1\}^d} \mathbb{E}\left[Y|X_{obs(m)}, M=m\right] \mathbb{1}_{M=m}. \end{aligned} \tag{3.3}$$

Tức là thay vì ta đi tính toán dựa trên vector đặc trưng bị khuyết $\widetilde{X}$, ta tính toán dựa trên vector mask M và dữ liệu quan sát được theo vector mask $X_{obs(M)}$, với $\mathbb{1}_{M=m}$ là hàm chỉ thị (Indicator function):

$$\mathbb{1}_{M=m} = \begin{cases} 1 & \text{nếu } M = m, \\ 0 & \text{nếu } M \neq m. \end{cases}$$

Việc viết dự đoán Bayes dưới dạng tổng của các kỳ vọng có điều kiện trên tất cả giá trị hiện thực m như phương trình (3.3) cho ta thấy được độ phức tạp khi ta cần phải có 2^d mô hình để có thể học từng mẫu dữ liệu khuyết có thể xảy ra, và việc tính toán sẽ trở nên khó khăn khi số chiều d lớn. Ngoài ra, ta còn thấy được dự đoán Bayes là một tổ hợp các mô hình con để mô hình cho các mẫu dữ liệu khuyết, và nó dự đoán tốt hơn do ta mô hình dựa trên các mẫu dữ liệu khuyết khác nhau.

Ta viết lại dự đoán Bayes f^* dưới dạng một hàm của dữ liệu quan sát được $X_{\text{obs}(M)}$ và vector mask M:

$$f^*(X_{obs(M)}, M) = \mathbb{E}[Y|X_{obs(M)}, M].$$

Từ đây, ta có được dạng tổng quát của dự đoán Bayes:

$$f^*(X_{obs(M)}, M) = \beta^*_0 + \langle \beta^*_{obs(M)}, X_{obs(M)} \rangle + \langle \beta^*_{mis(M)}, \mathbb{E}[X_{mis(M)}|M, X_{obs(M)}] \rangle, \tag{3.4}$$

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

Chứng minh: Ta có:

$$\begin{aligned} f^*(X_{obs(M)}, M) &= \mathbb{E}[Y|M, X_{\text{obs}(M)}] \\ &= \mathbb{E}[\beta^*_0 + \langle \beta^*, X \rangle | M, X_{obs(M)}] \\ &= \mathbb{E}[\beta^*_0 | M, X_{obs(M)}] + \mathbb{E}[\langle \beta^*, X \rangle | M, X_{obs(M)}] \\ &= \beta^*_0 + \mathbb{E}[\langle \beta^*_{obs(M)}, X_{obs(M)} \rangle + \langle \beta^*_{mis(M)}, X_{mis(M)} \rangle | M, X_{obs(M)}] \\ &= \beta^*_0 + \mathbb{E}[\langle \beta^*_{obs(M)}, X_{obs(M)} \rangle | M, X_{obs(M)}] + \mathbb{E}[\langle \beta^*_{mis(M)}, X_{mis(M)} \rangle | M, X_{obs(M)}] \\ &= \beta^*_0 + \langle \beta^*_{obs(M)}, X_{obs(M)} \rangle + \langle \beta^*_{mis(M)}, \mathbb{E}[X_{mis(M)}|M, X_{obs(M)}] \rangle. \end{aligned}$$

Vậy ta có điều phải chứng minh. ■

3.2 Dự đoán Bayes (Bayes predictor)

Do có nhiều cơ chế dữ liệu khác nhau nên ta mong muốn phương pháp xử lý dữ liệu khuyết sẽ mạnh (robust) với từng cơ chế khác nhau, tức ta muốn dự đoán Bayes có thể được biểu diễn dưới dạng đóng (closed-form) tổng quát cho cả 3 loại cơ chế. Tuy nhiên, điều này gặp khó khăn vì dự đoán Bayes phụ thuộc vào phân phối của dữ liệu, cũng như đặc điểm riêng của từng cơ chế.

Dẫu vậy, với dữ liệu tuân theo phân phối chuẩn đa biến, ta vẫn có thể biểu diễn dự đoán Bayes dưới dạng đóng cho từng loại cơ chế cụ thể.

3.2.1 Mô hình cho cơ chế M(C)AR

Ta lần lượt đưa ra giả thiết về cơ chế MCAR và MAR dưới ngôn ngữ của xác suất:

Giả thiết 3.1 *(Cơ chế MCAR) Với mọi* $m \in \{0,1\}^d$*,*

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

Điều này cho thấy rằng, với cơ chế MCAR, xác suất để dữ liệu X bị khuyết theo vector mask M không phụ thuộc vào dữ liệu X.

Giả thiết 3.2 *(Cơ chế MAR) Với mọi* $m \in \{0,1\}^d$*,*

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

Với cơ chế MAR, xác suất để dữ liệu X bị khuyết theo vector mask M chỉ phụ thuộc vào dữ liệu quan sát được $X_{obs(m)}$ mà không phụ thuộc vào dữ liệu không quan sát được $X_{mis(m)}$ trong điểm dữ liệu X.

Và với 2 giả thiết trên, ta biểu diễn được dự đoán Bayes cho dữ liệu khuyết với cơ chế MCAR và MAR.

Mệnh đề 3.1 *(Dự đoán Bayes với M(C)AR) Giả sử dữ liệu được sinh ra từ mô hình tuyến tính được định nghĩa ở phương trình* (3.1) *và có phân phối chuẩn đa biến. Giả sử ta có giả thiết 3.1 hoặc 3.2, thì dự đoán Bayes* f^* *có dạng:*

$$f^*(X_{obs}, M) = \beta_0^* + \langle \beta_{obs}^*, X_{obs} \rangle + \langle \beta_{mis}^*, \mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}) \rangle, \quad (3.5)$$

Với obs (tương tự mis) thay vì obs(M) (tương tự mis(M)) nếu không có thêm trường hợp M nào khác.

Chứng minh: Từ phương trình (3.4), ta có dạng tổng quát của dự đoán Bayes:

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

Để tính dự đoán Bayes, ta cần quan tâm tới việc tính $\mathbb{E}[X_{mis(M)}|M, X_{obs(M)}]$. Hay nói cách khác, ta cần tính $\mathbb{E}[X_j|M, X_{obs}]$, với mọi $j \in mis$. Để làm được điều này, ta sẽ xét $P(X_j|M, X_{obs})$, với mọi $j \in mis$.

Cho $mis'(M, j) = mis(M) \backslash \{j\}$. Nếu không có gì nhầm lẫn, ta rút gọn kí hiệu thành mis'. Ta có:

$$\begin{aligned} P(X_j|M, X_{obs}) &= \frac{P(M, X_j, X_{obs})}{P(M, X_{obs})} \\ &= \frac{\int P(M, X_j, X_{obs}, X_{mis'}) \mathrm{d}X_{mis'}}{\int \int P(M, X_j, X_{obs}, X_{mis'}) \mathrm{d}X_{mis'} \mathrm{d}X_j} \\ &= \frac{\int P(M|X_{obs}, X_j, X_{mis'}) P(X_{obs}, X_j, X_{mis'}) \mathrm{d}X_{mis'}}{\int \int P(M|X_{obs}, X_j, X_{mis'}) P(X_{obs}, X_j, X_{mis'}) \mathrm{d}X_{mis'} \mathrm{d}X_j}. \end{aligned} \tag{3.6}$$

Trong trường hợp MCAR, với mọi $m \in \{0, 1\}^d$, $P(M = m|X) = P(M = m)$, ta có:

$$\begin{aligned} P(X_j|M, X_{obs}) &= \frac{P(M) \int P(X_{obs}, X_j, X_{mis'}) \mathrm{d}X_{mis'}}{P(M) \int \int P(X_{obs}, X_j, X_{mis'}) \mathrm{d}X_{mis'} \mathrm{d}X_j} \\ &= \frac{P(X_{obs}, X_j)}{P(X_{obs})} \\ &= P(X_j|X_{obs}). \end{aligned}$$

Còn đối với trường hợp MAR, với mọi $m \in \{0, 1\}^d$, $P(M = m|X) = P(M = m|X_{obs(m)})$, ta có:

$$\begin{aligned} P(X_j|M, X_{obs}) &= \frac{P(M|X_{obs}) \int P(X_{obs}, X_j, X_{mis'}) \mathrm{d}X_{mis'}}{P(M|X_{obs}) \int \int P(X_{obs}, X_j, X_{mis'}) \mathrm{d}X_{mis'} \mathrm{d}X_j} \\ &= \frac{P(X_{obs}, X_j)}{P(X_{obs})} \\ &= P(X_j|X_{obs}). \end{aligned}$$

Vì vậy, nếu cơ chế dữ liệu khuyết là MCAR hay MAR thì ta có:

$$P(X_j|M, X_{obs}) = P(X_j|X_{obs}),$$

tức:

$$\mathbb{E}[X_{mis(M)}|M, X_{obs(M)}] = \mathbb{E}[X_{mis(M)}|X_{obs(M)}].$$

Thế phương trình trên vào phương trình dự đoán Bayes tổng quát (3.4), ta được:

$$f^*(X_{obs}, M) = \beta_0^* + \langle \beta^*_{obs(M)}, X_{obs(M)} \rangle + \langle \beta^*_{mis(M)}, \mathbb{E}[X_{mis(M)}|X_{obs(M)}] \rangle. \tag{3.7}$$

Nếu không có gì nhầm lẫn cho trường hợp M nào khác, ta kí hiệu obs (tương tự mis) thay vì $obs(M)$ (tương tự $mis(M)$).

Vì dữ liệu X là vector có phân phối chuẩn đa biến $\mathcal{N}(\mu, \Sigma)$ theo giả thiết, nên ta xét phân phối chuẩn bị chia, tách riêng biệt phần dữ liệu quan sát được và không quan sát được, với

$$X = \begin{pmatrix} X_{mis} \\ X_{obs} \end{pmatrix}, \ \mu = \begin{pmatrix} \mu_{mis} \\ \mu_{obs} \end{pmatrix}, \ \Sigma = \begin{pmatrix} \Sigma_{mis,mis} & \Sigma_{mis,obs} \\ \Sigma_{obs,mis} & \Sigma_{obs,obs} \end{pmatrix},$$

ta có:

$$X_{mis}|X_{obs} \sim \mathcal{N}\left(\mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}), \Sigma_{mis,mis} - \Sigma_{mis,obs}(\Sigma_{obs})^{-1}\Sigma_{obs,mis}\right),$$

tức

$$\begin{aligned} \mu_{mis|obs} &= \mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}), \\ \Sigma_{mis|obs} &= \Sigma_{mis,mis} - \Sigma_{mis,obs}(\Sigma_{obs})^{-1}\Sigma_{obs,mis}. \end{aligned}$$

Từ đây, ta được:

$$\mathbb{E}[X_{mis}|X_{obs}] = \mu_{mis} + \Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs}).$$

Kết hợp với (3.7), ta có:

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

Vậy ta có điều phải chứng minh. ■

Qua mệnh đề trên, ta thấy rằng dự đoán Bayes trong trường hợp cơ chế MCAR và MAR là giống nhau, với giả thiết dữ liệu tuân theo phân phối chuẩn đa biến. Từ đây, ta có thể tính được dự đoán Bayes cho cả 2 cơ chế khác nhau, chỉ với phương trình (3.5). Tiếp theo, ta đi biểu diễn dự đoán Bayes cho cơ chế dữ liệu khuyết MNAR.

3.2.2 Mô hình cho cơ chế MNAR

Cơ chế MNAR phức tạp hơn so với các cơ chế còn lại do xác suất để dữ liệu bị khuyết phụ thuộc vào chính nó, nên dự đoán Bayes khó có thể được biểu diễn dưới dạng tổng quát cho cơ chế MNAR. Tuy nhiên, ta vẫn có thể biểu diễn được dưới giả thiết cơ chế dữ liệu khuyết là Gaussian self-masking – một dạng cơ chế MNAR, với xác suất để dữ liệu bị khuyết tuân theo phân phối chuẩn.

Giả thiết 3.3 *(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* $\forall k \in [1, d]$,

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

Ở đây, K_k là hằng số điều chỉnh xác suất X_k bị khuyết: K_k càng gần 1 thì xác suất X_k bị khuyết càng cao. Xác suất để X_k bị khuyết tuân theo phân phối chuẩn $\mathcal{N}(\widetilde{\mu}_k, \widetilde{\sigma}_k^2)$, với trung bình $\widetilde{\mu}_k$ và phương sai $\widetilde{\sigma}_k^2$. Nghĩa là xác suất này cao khi X_k gần với $\widetilde{\mu}_k$ và giảm dần khi X_k càng xa $\widetilde{\mu}_k$. Độ lớn của $\widetilde{\sigma}_k^2$ điều khiển mức độ giảm của xác suất khuyết theo khoảng cách từ X_k đến $\widetilde{\mu}_k$.

Tóm lại, Gaussian self-masking cho ta các giả thiết:

- Xác suất để X_k bị khuyết không phụ thuộc vào các giá trị khác trong cùng 1 điểm dữ liệu,
- Xác suất để X_k bị khuyết ($M_k = 1$) phụ thuộc vào chính X_k, và tuân theo phân phối chuẩn.

Sau khi đã có giả thiết cơ chế Gaussian self-masking, ta đi biểu diễn dự đoán Bayes cho cơ chế này.

Mệnh đề 3.2 *(Dự đoán Bayes với Gaussian self-masking) Giả sử dữ liệu được sinh ra từ mô hình tuyến tính được định nghĩa ở phương trình* (3.1) *tuân theo phân phối chuẩn đa biến và thoả giả thiết 3.3. 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:*

$$f^*(X_{obs}, M) = \beta_0^* + \langle \beta_{obs}^*, X_{obs} \rangle + \langle \beta_{mis}^*, (Id + D_{mis}\Sigma_{mis|obs}^{-1})^{-1} \times (\tilde{\mu}_{mis} + D_{mis}\Sigma_{mis|obs}^{-1}(\mu_{mis} + \Sigma_{mis,obs}\left(\Sigma_{obs}\right)^{-1}\left(X_{obs} - \mu_{obs}\right)))\rangle. \quad (3.8)$$

Chứng minh: Với giả thiết cơ chế dữ liệu bị khuyết Gaussian self-masking, ta có xác suất dữ liệu bị khuyết trên chính tập dữ liệu là:

$$P(M = m|X) = P(M_{mis(m)} = 1|X_{mis(m)})P(M_{obs(m)} = 0|X_{obs(m)}).$$

Từ đây, phương trình (3.6) có thể được viết lại thành:

$$\begin{aligned} P(X_j|M, X_{obs}) &= \frac{P(M_{obs} = 0|X_{obs}) \int P(M_{mis} = 1|X_{mis})P(X_{obs}, X_j, X_{mis'})\mathrm{d}X_{mis'}}{P(M_{obs} = 0|X_{obs}) \int\int P(M_{mis} = 1|X_{mis})P(X_{obs}, X_j, X_{mis'})\mathrm{d}X_{mis'}\mathrm{d}X_j} \\ &= \frac{\int P(M_{mis} = 1|X_{mis})P(X_{mis}|X_{obs})P(X_{obs})\mathrm{d}X_{mis'}}{\int\int P(M_{mis} = 1|X_{mis})P(X_{mis}|X_{obs})P(X_{obs})\mathrm{d}X_{mis'}\mathrm{d}X_j} \\ &= \frac{\int P(M_{mis} = 1|X_{mis})P(X_{mis}|X_{obs})\mathrm{d}X_{mis'}}{\int\int P(M_{mis} = 1|X_{mis})P(X_{mis}|X_{obs})\mathrm{d}X_{mis'}\mathrm{d}X_j}. \end{aligned} \tag{3.9}$$

Giờ đây, ta muốn tính $P(M_{mis} = 1|X_{mis})P(X_{mis}|X_{obs})$. Đặt D là ma trận đường chéo sao cho $\mathrm{diag}(D) = \widetilde{\sigma}^2$, với $\widetilde{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_d^2)$ được định nghĩa ở trong 3.3. Theo giả thiết, ta có $\forall k \in [1, d]$:

$$\begin{aligned} P(M|X) &= \prod_{k=1}^{d} P(M_k|X_k) \\ \Rightarrow P(M_{mis} = 1|X_{mis}) &= \prod_{k\in mis}^{d} P(M_k = 1|X_k). \end{aligned}$$

Kết hợp với định nghĩa trong giả thiết 3.3, ta có:

$$\begin{aligned} P(M_{mis} = 1|X_{mis}) &= \prod_{k\in mis}^{d} \left(K_k \exp\left(-\frac{1}{2}\frac{(X_k - \widetilde{\mu}_k)^2}{\widetilde{\sigma}_k^2} \right) \right) \\ &= \prod_{k\in mis}^{d} K_k \prod_{k\in mis}^{d} \exp\left(-\frac{1}{2}\frac{(X_k - \widetilde{\mu}_k)^2}{\widetilde{\sigma}_k^2} \right) \\ &= \prod_{k\in mis}^{d} K_k \ \exp\left(-\frac{1}{2}\sum_{k\in mis}^{d} \frac{(X_k - \widetilde{\mu}_k)^2}{\widetilde{\sigma}_k^2} \right). \end{aligned}$$

Qua các kí hiệu bằng ma trận, ta có xác suất để dữ liệu bị khuyết ($M_{mis} = 1$) khi biết các giá trị không quan sát được X_{mis} dưới dạng:

$$P(M_{mis} = 1|X_{mis}) = \prod_{k\in mis}^{d} K_k \ \exp\left(-\frac{1}{2}(X_{mis} - \widetilde{\mu}_{mis})^\top (D_{mis,mis})^{-1}(X_{mis} - \widetilde{\mu}_{mis}) \right). \tag{3.10}$$

Do X_{mis}, X_{obs} đều là các phân phối chuẩn đa biến, nên ta sử dụng phân phối chuẩn có

điều kiện, với $X_{mis}|X_{obs} \sim \mathcal{N}(X_{mis}|\mu_{mis|obs}, \Sigma_{mis|obs})$, ta có:

$$\begin{aligned} \mu_{mis|obs} &= \mu_{mis} + \Sigma_{mis,obs}\Sigma_{obs,obs}^{-1}(X_{obs} - \mu_{obs}), \\ \Sigma_{mis|obs} &= \Sigma_{mis,mis} - \Sigma_{mis,obs}\Sigma_{obs,obs}^{-1}\Sigma_{obs,mis}, \end{aligned} \tag{3.11}$$

và

$$P(X_{mis}|X_{obs}) = \frac{1}{\sqrt{(2\pi)^{|mis|}|\Sigma_{mis|obs}|}} \exp\left(-\frac{1}{2}(X_{mis} - \mu_{mis|obs})^\top \Sigma_{mis|obs}^{-1}(X_{mis} - \mu_{mis|obs})\right).$$

Từ phương trình (3.10), $P(M_{mis} = 1|X_{mis})$ và $P(X_{mis}|X_{obs})$ đều là hàm Gauss của X_{mis}. Áp dụng bổ đề 1.1, tích của chúng cũng là 1 hàm Gauss dưới dạng:

$$P(M_{mis} = 1|X_{mis})P(X_{mis}|X_{obs}) = K \exp\left(-\frac{1}{2}(X_{mis} - a_M)^\top (A_M)^{-1} (X_{mis} - a_M)\right),$$

với a_M và A_M phụ thuộc vào mẫu dữ liệu khuyết:

$$(A_M)^{-1} = D_{mis,mis}^{-1} + \Sigma_{mis|obs}^{-1}, \tag{3.12}$$

$$(A_M)^{-1}a_M = D_{mis,mis}^{-1}\tilde{\mu}_{mis} + \Sigma_{mis|obs}^{-1}\mu_{mis|obs}, \tag{3.13}$$

và:

$$K = \frac{\prod_{k\in mis}^{d} K_k}{\sqrt{(2\pi)^{|mis|}|\Sigma_{mis|obs}|}} \exp\left(-\frac{1}{2}(\tilde{\mu}_{mis} - \mu_{mis|obs})^\top (\Sigma_{mis|obs} + D_{mis|obs})^{-1}(\tilde{\mu}_{mis} - \mu_{mis|obs})\right).$$

Bởi vì K không phụ thuộc vào X_{mis}, nên đơn giản phương trình (3.9), ta được:

$$\begin{aligned} P(X_j|M, X_{obs}) &= \frac{\int \mathcal{N}(X_{mis}|a_M, A_M)\mathrm{d}X_{mis'}}{\int\int \mathcal{N}(X_{mis}|a_M, A_M)\mathrm{d}X_{mis'}\mathrm{d}X_j} \\ &= \int \mathcal{N}(X_{mis}|a_M, A_M)\mathrm{d}X_{mis'}. \end{aligned}$$

Phần mẫu số $\int\int \mathcal{N}(X_{mis}|a_M, A_M)\mathrm{d}X_{mis'}\mathrm{d}X_j = 1$ do tích phân của hàm mật độ xác suất trên miền xác định luôn bằng 1. Nó còn được coi là hằng số chuẩn hoá (Normalization constant) cho phần tử số. Qua đây, ta thấy được mật độ $P(X_j|M, X_{obs})$ chính là phân phối biên (marginal distribution) của X_j, được tính bằng cách lấy tích phân các biến còn lại trong $X_{mis'}$.

Do phân phối biên của một phân phối chuẩn đa biến là một phân phối chuẩn (trong

trường hợp này là phân phối chuẩn cho X_j), nên

$$P(X_j|M, X_{obs}) = \mathcal{N}\left(X_j|(a_M)_j, (A_M)_{j,j}\right),$$

với $(a_M)_j$ là phần tử thứ j của vector a_M, $(A_M)_{j,j}$ là phần tử thứ j nằm trên đường chéo của ma trận hiệp phương sai A_M. Và do đó:

$$\mathbb{E}[X_{mis}|M, X_{obs}] = (a_M)_{mis}. \tag{3.14}$$

Từ (3.12) ta có:

$$A_M = (D^{-1}_{mis,mis} + \Sigma^{-1}_{mis|obs})^{-1},$$

nhân cả 2 vế cho phương trình (3.13), ta được:

$$\begin{aligned}
a_M &= (D^{-1}_{mis,mis} + \Sigma^{-1}_{mis|obs})^{-1}(D^{-1}_{mis,mis}\tilde{\mu}_{mis} + \Sigma^{-1}_{mis|obs}\mu_{mis|obs}) \\
&= (D^{-1}_{mis,mis} + \Sigma^{-1}_{mis|obs})^{-1}D^{-1}_{mis,mis}(\tilde{\mu}_{mis} + D_{mis,mis}\Sigma^{-1}_{mis|obs}\mu_{mis|obs}) \\
&= (D_{mis,mis}(D^{-1}_{mis,mis} + \Sigma^{-1}_{mis|obs}))^{-1}(\tilde{\mu}_{mis} + D_{mis,mis}\Sigma^{-1}_{mis|obs}\mu_{mis|obs}) \\
&= (Id + D_{mis,mis}\Sigma^{-1}_{mis|obs})^{-1}(\tilde{\mu}_{mis} + D_{mis,mis}\Sigma^{-1}_{mis|obs}\mu_{mis|obs}).
\end{aligned}$$

Từ kết quả trên, kết hợp với (3.14) và (3.11), ta được:

$$\begin{aligned}
\mathbb{E}[X_{mis}|M, X_{obs}] = (Id + D_{mis,mis}\Sigma^{-1}_{mis|obs})^{-1} \\
\times (\tilde{\mu}_{mis} + D_{mis,mis}\Sigma^{-1}_{mis|obs}(\mu_{mis} + \Sigma_{mis,obs}\Sigma^{-1}_{obs,obs}(X_{obs} - \mu_{obs}))).
\end{aligned}$$

Kết hợp với phương trình của dự đoán Bayes dạng tổng quát (3.4), ta có:

$$\begin{aligned}
f^*(X_{obs}, M) = \beta^*_0 + \langle\beta^*_{obs}, X_{obs}\rangle + \langle\beta^*_{mis}, (Id + D_{mis}\Sigma^{-1}_{mis|obs})^{-1} \\
\times (\tilde{\mu}_{mis} + D_{mis}\Sigma^{-1}_{mis|obs}(\mu_{mis} + \Sigma_{mis,obs}\left(\Sigma_{obs}\right)^{-1}\left(X_{obs} - \mu_{obs}\right)))\rangle.
\end{aligned}$$

Vậy ta có điều phải chứng minh. ■

Qua mệnh đề này, dự đoán Bayes có thể được biểu diễn cho cơ chế MNAR, cụ thể là cơ chế Gaussian self-masking.

Từ 2 mệnh đề 3.1 và 3.2, mỗi mô hình trong 2^d mô hình ở phương trình (3.3) có thể được biểu diễn dưới dạng 1 hàm tuyến tính của dữ liệu quan sát được X_{obs}. Hay nói cách khác, dự đoán Bayes là tuyến tính theo từng mẫu dữ liệu khuyết với các giả thiết được sử dụng. Ngoài ra, ta còn thấy được phương trình dự đoán Bayes không phụ thuộc vào giá trị của dữ liệu bị khuyết X_{mis} cho cả 3 cơ chế.

4 Mạng NeuMiss

Để tính dự đoán Bayes trong phương trình (3.5) và (3.8), ta cần tính nghịch đảo của mỗi ma trận hiệp phương sai con $\Sigma_{obs(m)}$ với mỗi mẫu dữ liệu khuyết $m \in \{0,1\}^d$. Điều này tương đương với việc xây dựng một mô hình riêng biệt cho từng mẫu dữ liệu khuyết.

Với số lượng đơn vị ẩn (hidden units) tỷ lệ với 2^d mẫu dữ liệu khuyết, một mạng MLP với hàm kích hoạt (activation function) ReLU có thể học được các mô hình độc lập cho từng trường hợp. Tuy nhiên, khi số chiều d lớn thì chi phí tính toán cũng trở nên lớn hơn. Các kiến trúc như MLP sẽ tạo ra số lượng tham số rất lớn vì chúng không chia sẻ thông tin giữa những mẫu dữ liệu khuyết tương tự nhau, nên chúng không tận dụng được mối quan hệ giữa các mẫu dữ liệu có đặc điểm khuyết giống nhau.

Một hướng tiếp cận khác là ước lượng vector μ và ma trận hiệp phương sai Σ bằng thuật toán EM (Expectation Maximization), sau đó tính nghịch đảo của Σ_{obs}. Dẫu vậy, cách tiếp cận này cũng không hiệu quả do độ phức tạp tính toán tăng cao khi số chiều d lớn.

Do đó, bài báo [1] đề xuất một giải pháp trung hoà giữa 2 hướng trên: mô hình mối quan hệ giữa các hệ số cho các mẫu dữ liệu khuyết khác nhau thay vì ước lượng trực tiếp ma trận hiệp phương sai. Hiểu một cách trực quan, các dữ liệu quan sát được trong một mẫu sẽ được dùng để ước lượng tham số hồi quy cho các mẫu dữ liệu khác, thông qua việc chia sẻ tham số giữa các mẫu dữ liệu có đặc điểm chung.

Trong mục này, ta sẽ xấp xỉ nghịch đảo của ma trận hiệp phương sai Σ_{obs}, tức xấp xỉ dự đoán Bayes bằng chuỗi Neumann, cũng như xét sự hội tụ của chuỗi và dự đoán Bayes. Với phép lặp xấp xỉ này, kết hợp với phương pháp Algorithm Unrolling [2], ta sẽ có được một neural network mang tên NeuMiss (tiền tố Neu viết tắt cho Neumann và Neural network, Miss cho Missing). Mạng NeuMiss có thể tận dụng các tham số có thể học được trong mạng để xấp xỉ dự đoán Bayes, với các tham số được chia sẻ với nhau thông qua các mẫu dữ liệu khuyết, giúp giảm độ phức tạp tính toán và hiệu quả hơn.

4.1 Xấp xỉ ma trận bằng chuỗi Neumann

Thử thách lớn nhất của phương trình (3.5) và (3.8) là việc tính nghịch đảo của ma trận $\Sigma_{obs(m)}$ với mọi $m \in \{0,1\}^d$. Khi d lớn thì chi phí tính toán sẽ rất lớn. Nên thay vì đi tính trực tiếp, bài báo đề xuất việc xấp xỉ nghịch đảo của $\Sigma_{obs(m)}$ một cách đệ quy bằng cách sử dụng chuỗi Neumann.

Đầu tiên, ta chọn 1 ma trận khởi tạo $S^{(0)}$ với $d \times d$ chiều. $S^{(0)}_{obs(m)}$ được định nghĩa là ma trận con của $S^{(0)}$, được tạo ra bằng cách chọn các cột và hàng sao cho dữ liệu không bị khuyết (các giá trị mà tại đó, $m = 0$) và là xấp xỉ bậc-0 của $(\Sigma_{obs(m)})^{-1}$. Sau đó, với mọi

$m \in \{0,1\}^d$, ta định nghĩa xấp xỉ bậc-ℓ của $(\Sigma_{obs(m)})^{-1}$ qua phép lặp sau: Với mọi $\ell \geqslant 1$,

$$S^{(\ell)}_{obs(m)} = (Id - \Sigma_{obs(m)})S^{(\ell-1)}_{obs(m)} + Id. \tag{4.1}$$

Phép lặp $S^{(\ell)}_{obs(m)}$ hội tụ tuyến tính về $(\Sigma_{obs(m)})^{-1}$, và cũng là chuỗi Neumann chặt cụt tới ℓ nếu $S^{(0)} = Id$.

Chứng minh: Ta xấp xỉ nghịch đảo ma trận $\Sigma_{obs(m)}$ bằng chuỗi Neumann:

$$(\Sigma_{obs(m)})^{-1} = \sum_{k=0}^{\infty}(Id - \Sigma_{obs(m)})^k,$$

tức

$$S_{obs(m)} = \sum_{k=0}^{\infty}(Id - \Sigma_{obs(m)})^k.$$

Chuỗi hội tụ khi $\|Id - \Sigma_{obs(m)}\|_2 < 1$. Vì bán kính phổ của Σ nhỏ hơn 1, nên bán kính phổ của mỗi ma trận con $\Sigma_{obs(m)}$ của Σ cũng nhỏ hơn 1, theo định lý Cauchy đan nhau (Cauchy Interlace Theorem) [8], hoặc qua định nghĩa của các trị riêng dưới dạng tỷ số Rayleigh:

$$\rho(\Sigma_{obs(m)}) = \max_{u \in \mathbb{R}^{|obs(m)|}} u^\top \Sigma_{obs(m)} u = \max_{\substack{x \in \mathbb{R}^d \\ x_{mis}=0}} x^\top \Sigma x \leqslant \max_{x \in \mathbb{R}^d} x^\top \Sigma x = \rho(\Sigma).$$

Thay vì viết và tính toán dưới dạng chuỗi vô hạn, ta chỉ cần chuỗi xấp xỉ tới bậc-ℓ:

$$S^{(\ell)}_{obs(m)} = \sum_{k=0}^{\ell}(Id - \Sigma_{obs(m)})^k = \sum_{k=0}^{\ell-1}(Id - \Sigma_{obs(m)})^k + (Id - \Sigma_{obs(m)})^\ell S^{(0)}_{obs(m)}.$$

Với ma trận khởi tạo $S^{(0)}_{obs(m)}$, ta có thể định nghĩa một cách đệ quy thông qua phép lặp:

$$S^{(\ell)}_{obs(m)} = (Id - \Sigma_{obs(m)})S^{(\ell-1)}_{obs(m)} + Id. \tag{4.1}$$

Để rõ hơn, ta khai triển phương trình đệ quy trên:

$$\begin{aligned}
S^{(\ell)}_{obs(m)} &= (Id - \Sigma_{obs(m)})S^{(\ell-1)}_{obs(m)} + Id \\
&= (Id - \Sigma_{obs(m)})((Id - \Sigma_{obs(m)})S^{(\ell-2)}_{obs(m)} + Id) + Id \\
&= (Id - \Sigma_{obs(m)})^2 S^{(\ell-2)}_{obs(m)} + (Id - \Sigma_{obs(m)}) + Id \\
&= (Id - \Sigma_{obs(m)})^3 S^{(\ell-3)}_{obs(m)} + (Id - \Sigma_{obs(m)})^2 + (Id - \Sigma_{obs(m)}) + Id \\
&\vdots \\
&= (Id - \Sigma_{obs(m)})^\ell S^{(0)}_{obs(m)} + \sum_{k=0}^{\ell-1} (Id - \Sigma_{obs(m)})^k.
\end{aligned}$$

Ở (4.1), $(\Sigma_{obs(m)})^{-1}$ là điểm bất động của phương trình. Thật vậy:

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

Lấy phương trình này trừ đi (4.1), ta có:

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

Nhân 2 vế phương trình trên cho $\Sigma_{obs(m)}$, ta được:

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

Lấy chuẩn l_2 (chuẩn phổ) cho cả 2 vế và sử dụng bất đẳng thức Cauchy-Schwartz, ta được:

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

Cho $\nu_{obs(m)}$ là trị riêng nhỏ nhất của $\Sigma_{obs(m)}$, với giá trị dương do Σ khả nghịch. Vì trị riêng của $\Sigma_{obs(m)}$ có chặn trên là 1, ta có $\|Id - \Sigma_{obs(m)}\|_2 = (1 - \nu_{obs(m)})$. Bằng phép lặp đệ quy, ta nhận được:

$$\|Id - \Sigma_{obs(m)} S^{\ell}_{obs(m)}\|_2 \leqslant (1 - \nu_{obs(m)})^\ell \|Id - \Sigma_{obs(m)} S^{(0)}_{obs(m)}\|_2. \tag{4.2}$$

Vậy $S^{(\ell)}_{obs(m)}$ hội tụ tuyến tính về $(\Sigma_{obs(m)})^{-1}$, và tốc độ hội tụ được xác định bởi trị riêng nhỏ nhất của $\Sigma_{obs(m)}$. ■

Giờ ta đi định nghĩa xấp xỉ bậc-ℓ của dự đoán Bayes trong trường hợp cơ chế MAR (phương trình (3.5)) dưới dạng:

$$f^*_\ell(X_{obs}, M) = \langle \beta^*_{obs}, X_{obs} \rangle + \langle \beta^*_{mis}, \mu_{mis} + \Sigma_{mis,obs} S^{(\ell)}_{obs(m)} (X_{obs} - \mu_{obs}) \rangle, \tag{4.3}$$

Sai số giữa dự đoán Bayes và xấp xỉ bậc-ℓ của nó được cho bởi mệnh đề sau:

Mệnh đề 4.1 *Cho ν là trị riêng nhỏ nhất của Σ. Giả sử dữ liệu được sinh ra qua mô hình tuyến tính* (3.1) *và tuân theo phân phối chuẩn đa biến. Giả sử giả thiết 3.1 hay 3.2 thoả và bán kính phổ của Σ nhỏ hơn 1. Thì với mọi $\ell \geqslant 1$,*

$$\mathbb{E}\left[\left(f_\ell^*(X_{obs}, M) - f^*(X_{obs}, M)\right)^2\right] \leqslant \frac{(1-\nu)^{2\ell}\|\beta^*\|_2^2}{\nu}\ \mathbb{E}\left[\|Id - S^{(0)}_{obs(M)}\Sigma_{obs(M)}\|_2^2\right]. \quad (4.4)$$

Chứng minh: Với mệnh đề 3.1 và xấp xỉ bậc-ℓ cho dự đoán Bayes trong phương trình (4.3), ta có:

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

Do ta chỉ quan tâm tới phần xấp xỉ nên ta tạm thời bỏ qua β_0^*. Xét:

$$\begin{aligned}
&\mathbb{E}[(f^*_{\widetilde{X},\ell}(\widetilde{X}) - f^*_{\widetilde{X}}(\widetilde{X}))^2]\\
&= \mathbb{E}\left[\langle \beta^*_{mis}, \Sigma_{mis,obs}(S^{(\ell)}_{obs} - \Sigma^{-1}_{obs})(X_{obs} - \mu_{obs})\rangle^2\right]\\
&= \mathbb{E}\left[(\beta^*_{mis})^\top\left(\Sigma_{mis,obs}(S^{(\ell)}_{obs} - \Sigma^{-1}_{obs})(X_{obs} - \mu_{obs})\right)\left(\Sigma_{mis,obs}(S^{(\ell)}_{obs} - \Sigma^{-1}_{obs})(X_{obs} - \mu_{obs})\right)^\top \beta^*_{mis}\right]\\
&= \mathbb{E}\left[(\beta^*_{mis})^\top\Sigma_{mis,obs}(S^{(\ell)}_{obs} - \Sigma^{-1}_{obs})(X_{obs} - \mu_{obs})(X_{obs} - \mu_{obs})^\top(S^{(\ell)}_{obs} - \Sigma^{-1}_{obs})\Sigma_{obs,mis}\beta^*_{mis}\right]\\
&= \mathbb{E}\left[(\beta^*_{mis})^\top\Sigma_{mis,obs}(S^{(\ell)}_{obs} - \Sigma^{-1}_{obs})\mathbb{E}[(X_{obs} - \mu_{obs})(X_{obs} - \mu_{obs})^\top|M](S^{(\ell)}_{obs} - \Sigma^{-1}_{obs})\Sigma_{obs,mis}\beta^*_{mis}\right]\\
&= \mathbb{E}\left[(\beta^*_{mis})^\top\Sigma_{mis,obs}(S^{(\ell)}_{obs} - \Sigma^{-1}_{obs})\Sigma_{obs}(S^{(\ell)}_{obs} - \Sigma^{-1}_{obs})\Sigma_{obs,mis}\beta^*_{mis}\right]\\
&= \mathbb{E}\left[(\beta^*_{mis})^\top\Sigma_{mis,obs}(\Sigma_{obs}S^{(\ell)}_{obs} - Id_{obs})\Sigma^{-1}_{obs}\Sigma^{\frac{1}{2}}_{obs}\Sigma^{\frac{1}{2}}_{obs}\Sigma^{-1}_{obs}(\Sigma_{obs}S^{(\ell)}_{obs} - Id_{obs})\Sigma_{obs,mis}\beta^*_{mis}\right]\\
&= \mathbb{E}\left[\left\|\Sigma^{\frac{1}{2}}_{obs}\Sigma^{-1}_{obs}(\Sigma_{obs}S^{(\ell)}_{obs} - Id_{obs})\Sigma_{obs,mis}\beta^*_{mis}\right\|_2^2\right]\\
&= \mathbb{E}\left[\left\|\Sigma^{-\frac{1}{2}}_{obs}(Id_{obs} - \Sigma_{obs}S^{(\ell)}_{obs})\Sigma_{obs,mis}\beta^*_{mis}\right\|_2^2\right].
\end{aligned}$$

Với $\|\Sigma_{obs}x\|_2^2 = \sum_{i\in obs}(\Sigma_i^\top x)^2 \leqslant \sum_{i=1}^d(\Sigma_i^\top x)^2 = \|\Sigma x\|_2^2$,, tức $\|\Sigma_{obs}x\|_2 \leqslant \|\Sigma x\|_2$, ta có:

$$\|\Sigma_{obs,mis}\|_2 = \max_{\|x_{mis}\|_2=1}\|\Sigma_{obs,mis}x_{mis}\|_2 \leqslant \max_{\substack{\|x\|_2=1\\ x_{obs}=0}}\|\Sigma_{obs}x\|_2 \leqslant \max_{\substack{\|x\|_2=1\\ x_{obs}=0}}\|\Sigma x\|_2 \leqslant \max_{\|x\|_2=1}\|\Sigma x\|_2 = \|\Sigma\|_2.$$

Với trị riêng nhỏ nhất của Σ, ta có:

$$\lambda_{\min}(\Sigma) = \min_{\|x\|_2=1} x^\top\Sigma x \leqslant \min_{\substack{\|x\|_2=1\\ x_{mis}=0}} x^\top\Sigma x = \min_{\|x_{obs}\|_2=1} x_{obs}^\top\Sigma_{obs}x_{obs} = \lambda_{\min}(\Sigma_{obs}),$$

mà $\|\Sigma^{-1}\|_2 = \dfrac{1}{\lambda_{\min}(\Sigma)}$ và $\|\Sigma^{-1}_{obs}\|_2 = \dfrac{1}{\lambda_{\min}(\Sigma_{obs})}$, nên $\|\Sigma^{-1}_{obs}\|_2 \leqslant \|\Sigma^{-1}\|_2$.

Do đó, ta được:

$$\begin{aligned}\mathbb{E}[(f^*_{\widetilde{X},\ell}(\widetilde{X}) - f^*_{\widetilde{X}}(\widetilde{X}))^2] &\leqslant \|\Sigma^{-1}\|_2\|\Sigma\|_2^2\|\beta^*\|_2^2\ \mathbb{E}\big[\|Id_{obs} - \Sigma_{obs}S^{(\ell)}_{obs}\|_2^2\big] \\ &\leqslant \frac{1}{\nu}\|\beta^*\|_2^2\ \mathbb{E}\big[(1-\nu_{obs(m)})^{2\ell}\|Id_{obs} - \Sigma_{obs}S^{(0)}_{obs}\|_2^2\big] \qquad \text{(do (4.2))} \\ &= \frac{(1-\nu_{obs(m)})^{2\ell}\|\beta^*\|_2^2}{\nu}\ \mathbb{E}\big[\|Id_{obs} - \Sigma_{obs}S^{(0)}_{obs}\|_2^2\big].\end{aligned}$$

Vậy ta có điều phải chứng minh. ■

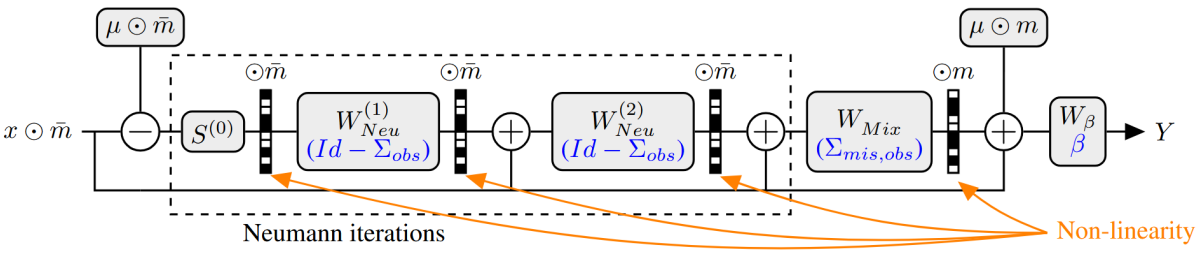
Mệnh đề 4.1 cho biết: Sai số của xấp xỉ bậc-ℓ phân rã (decay), hay tiến về 0 rất nhanh theo hàm mũ khi ℓ tăng. Quan trọng hơn, nếu ma trận con $S^{(0)}_{obs(m)}$ của $S^{(0)}$ là một xấp xỉ tốt của Σ^{-1}_{obs}, nghĩa là nếu ta chọn $S^{(0)}$ để tối thiểu hoá kỳ vọng ở vế bên phải của bất đẳng thức (4.4), thì mô hình của chúng ta sẽ cung cấp một xấp xỉ tốt cho dự đoán Bayes ngay cả khi bậc $\ell = 0$. Điều này đúng với ma trận hiệp phương sai dạng đường chéo, vì việc chọn $S^{(0)} = \Sigma^{-1}$ không có sai số xấp xỉ, do $(\Sigma^{-1})_{obs} = (\Sigma_{obs})^{-1}$.

4.2 Kiến trúc của mạng NeuMiss

Bài báo [1] đề xuất một kiến trúc neural network có tên là NeuMiss để xấp xỉ dự đoán Bayes, với nghịch đảo Σ^{-1}_{obs} được tính toán bằng một phiên bản được "unroll" của phép lặp Neumann. Hình 4.1 cho ta kiến trúc của neural network sử dụng xấp xỉ bậc-3, tương ứng với độ sâu 4. Với x là dữ liệu đầu vào, giá trị bị khuyết được thay bằng 0, và μ là tham số có thể huấn luyện (trainable) tương ứng với μ ở trong phương trình xấp xỉ bậc-ℓ của dự đoán Bayes (4.3):

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

Để giống với dự đoán Bayes (phương trình (4.3)), ma trận trọng số (weight matrix) W là một phép biến đổi đơn giản của ma trận hiệp phương sai như trong hình 4.1.



Hình 4.1: **Kiến trúc mạng NeuMisss với độ sâu** 4 — $\bar{m} = 1 - m$. Mỗi ma trận trọng số $W^{(k)}$ tương ứng với một phép biến đổi của ma trận hiệp phương sai.

Mỗi phép lặp Neumann đi qua một ma trận trọng số, qua tất cả $W^{(k)}_{Neu}$. Hay đúng hơn là

ta học mỗi lớp một cách độc lập, theo như bài báo về Algorithm Unrolling [2] đã đề cập, khi mà các trọng số của các lớp có thể cải thiện hiệu quả xấp xỉ của mô hình.

Các trị số cho các giá trị quan sát được thay đổi với mỗi mẫu dữ liệu, nên có thể dẫn tới khó khăn trong việc cài đặt. Ví dụ như dữ liệu bị khuyết với mẫu m, ma trận trọng số $S^{(0)}, W_{Neu}^{(1)}, W_{Neu}^{(2)}$ của hình 4.1 nên được masked sao cho các hàng và cột tương ứng với các trị số $mis(m)$ có giá trị bằng 0, và các hàng của W_{Mix} tương ứng với $obs(m)$ cũng như các cột của W_{Mix} tương ứng với $mis(m)$ có giá trị bằng 0.

Triển khai neural network với ma trận trọng số được masked theo các cách khác nhau với mỗi mẫu dữ liệu có thể phức tạp, nên bài báo đề xuất cách làm sau: Cho W là ma trận trọng số, v là một vector, và $\bar{m} = 1 - m$. Lúc này, $(W \odot \bar{m}\bar{m}^\top)v = (W(v \odot \bar{m})) \odot \bar{m}$, nghĩa là sử dụng ma trận trọng số được masked tương đương với việc masking vector đầu vào và đầu ra. Mạng NeuMiss có thể được xem là một neural network truyền thống với các hàm kích hoạt phi tuyến là **tích của các mask với nhau**. Cách tiếp cận này khiến việc cài đặt trở nên đơn giản hơn và cải thiện tốc độ tính toán của neural network, cũng như dễ dàng diễn giải hơn khi các mask điều khiển các thông tin của dữ liệu bị khuyết khi mạng được huấn luyện.

Theo như lý thuyết, chuỗi Neumann cần phải được lặp nhiều lần mới cho ra kết quả xấp xỉ nghịch đảo tốt, nên mạng NeuMiss cần phải sâu vì mỗi lớp tượng trưng cho 1 phép lặp. Do đó, mạng cần phải có residual connection, giúp model học được tốt hơn khi số lượng lớp nhiều. Mặc dù sau khi thực nghiệm, việc có hay không có residual connection không ảnh hưởng quá nhiều đến hiệu suất của mạng NeuMiss (xem 5.3).

Xấp xỉ trong trường hợp Gaussian self-masking

Mặc dù kiến trúc mạng NeuMiss được xây dựng dựa trên dự đoán Bayes trong trường hợp MCAR và MAR, nhưng nó cũng có thể được sử dụng cho cơ chế self-masking (3.8). Giả sử $D_{mis}\Sigma_{mis|obs}^{-1} \approx Id$, thì dự đoán Bayes cho self-masking trở thành:

$$\begin{aligned} f^*(X_{obs}, M) &\approx \beta_0^* + \langle \beta_{obs}^*, X_{obs} \rangle + \langle \beta_{mis}^*, (2Id)^{-1}(\tilde{\mu}_{mis} + Id(\mu_{mis} + \Sigma_{mis,obs}\left(\Sigma_{obs}\right)^{-1}\left(X_{obs} - \mu_{obs}\right)))\rangle \\ &\approx \beta_0^* + \langle \beta_{obs}^*, X_{obs} \rangle + \langle \beta_{mis}^*, \frac{1}{2}(\tilde{\mu}_{mis} + \mu_{mis} + \Sigma_{mis,obs}\left(\Sigma_{obs}\right)^{-1}\left(X_{obs} - \mu_{obs}\right))\rangle \\ &\approx \beta_0^* + \langle \beta_{obs}^*, X_{obs} \rangle + \langle \beta_{mis}^*, \frac{1}{2}(\widetilde{\mu}_{mis} + \mu_{mis}) + \frac{1}{2}\Sigma_{mis,obs}(\Sigma_{obs})^{-1}(X_{obs} - \mu_{obs})\rangle. \end{aligned}$$

Phương trình này giống như phương trình dự đoán Bayes cho cơ chế M(C)AR (3.5), nhưng với μ_{mis} được thay bởi $\frac{1}{2}(\widetilde{\mu}_{mis} + \mu_{mis})$ và $\Sigma_{mis,obs}$ trở thành $\frac{1}{2}\Sigma_{mis,obs}$. Dưới phép xấp xỉ này, dự đoán Bayes cho cơ chế self-masking có thể được mô hình bởi kiến trúc được đề xuất. Điểm khác biệt duy nhất là giá trị được quan tâm là tham số μ và W_{mix} của mạng.

Một phép xấp xỉ khác cũng phù hợp cho dự đoán Bayes: $D_{mis}\Sigma^{-1}_{mis|obs} \approx \hat{D}_{mis}$ khi $\hat{D}$ là một ma trận ma trận đường chéo. Trong trường hợp này, kiến trúc được đề xuất có thể mô hình dự đoán Bayes cho self-masking:

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

Ở đây, tham số μ của mạng nhắm tới $(Id + \hat{D}_{mis})^{-1}(\widetilde{\mu}_{mis} + \hat{D}_{mis}\mu_{mis})$ và W_{mix} nhắm tới $(Id + \hat{D}_{mis})^{-1}\hat{D}_{mis}\Sigma_{mis,obs}$ thay vì chỉ là $\Sigma_{mis,obs}$ trong trường hợp M(C)AR. Do đó, kiến trúc được đề xuất có thể xấp xỉ tốt dự đoán Bayes trong trường hợp self-masking bằng cách điều chỉnh giá trị tham số μ và W_{mix} học được nếu $D_{mis}\Sigma^{-1}_{mis|obs}$ gần giống với ma trận đường chéo.

5 Kết quả thực nghiệm

Sau khi đã đi qua nền tảng lý thuyết của mạng NeuMiss, ta đi cài đặt và kiểm chứng các kết quả có đúng với lý thuyết hay không, dựa trên code của tác giả bài báo [1].

Toàn bộ code dùng để thực nghiệm trong bài báo cáo này được chạy trên Google Colab[2], có tại: https://github.com/ngntrgduc/seminar.

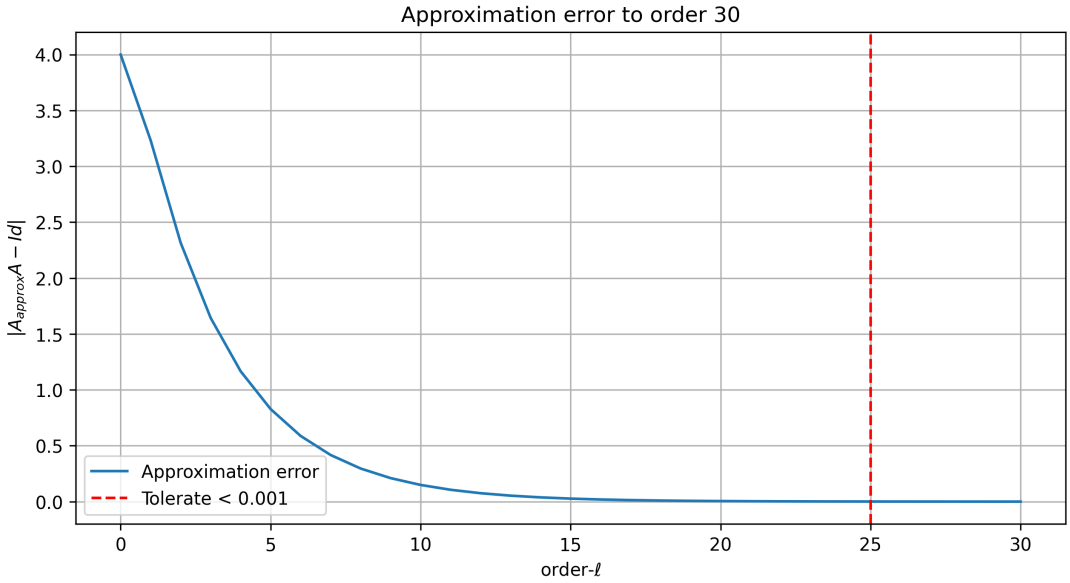
5.1 Xấp xỉ ma trận bằng chuỗi Neumann

Trước tiên, ta kiểm chứng độ hiệu quả cho việc xấp xỉ nghịch đảo ma trận nửa xác định dương, với bán kính phổ bé hơn 1 bằng chuỗi Neumann.

Sử dụng $seed = 0$ như trong cài đặt của bài báo, ta có được một xấp xỉ nghịch đảo của ma trận tại bậc 25 khá tốt khi mức sai số cho phép (tolerate) bé hơn 0.001, mặc dù từ bậc 15 thì sai số của xấp xỉ giảm không đáng kể. Sai số của xấp xỉ được tính bằng tổng sai khác của tích ma trận gốc và ma trận được xấp xỉ với lại ma trận đơn vị.

Với các trường hợp bán kính phổ của ma trận lớn hơn 1, chuỗi Neumann không hội tụ, nên không thể xấp xỉ ma trận được. Trong bài báo, dữ liệu được tạo ra với một ma trận hiệp phương sai với bán kính phổ lớn hơn 1. Sở dĩ bài báo không quan tâm tới bán kính phổ lớn hơn 1 là vì mạng NeuMiss sử dụng các ma trận trọng số để học và xấp xỉ sao cho các lớp ma trận trọng số sẽ tương ứng với xấp xỉ nghịch đảo Σ^{-1}_{obs}, không bắt buộc bán kính phổ ma trận hiệp phương sai nhỏ hơn 1 để hoạt động. Nên trong code tạo dữ

[1] Code của bài báo có tại https://github.com/marineLM/NeuMiss
[2] https://colab.research.google.com/



Hình 5.1: Sai số của xấp xỉ ma trận tới bậc 30.

liệu của tác giả, không có điều kiện kiểm tra bán kính phổ nhỏ hơn 1.

5.2 Mạng NeuMiss

Ta tiến hành kiểm tra độ hiệu quả của mạng NeuMiss trên tập dữ liệu được sinh ra từ phân phối chuẩn đa biến, với Y được tính bằng hàm tuyến tính của X như trong phương trình (3.1), cùng với 50% dữ liệu bị khuyết ngẫu nhiên (MCAR) ở mỗi đặc trưng trong tập dữ liệu và nhiễu ε tuân theo phân phối chuẩn với tỷ lệ signal-to-noise (SNR) bằng 10 (tỷ lệ nhiễu nhỏ hơn 10 lần so với "tín hiệu" Y).

Do dữ liệu được sinh ra từ mô hình tuyến tính, nên bài báo sử dụng metric R bình phương (R-squared – R^2), hay còn được gọi là hệ số xác định (coefficient of determination), để tính tỷ lệ của độ biến thiên cho biến phụ thuộc Y được giải thích bởi biến độc lập X, từ đó đánh giá độ hiệu quả của mô hình.

Ngoài ra, ta còn có các biến thể khác của R^2 như:

- R^2 hiệu chỉnh (Adjusted R-squared) cho ta biết được tỷ lệ biến thiên được giải thích chỉ bằng các biến độc lập thật sự ảnh hưởng tới biến phụ thuộc.
- R^2 cho dự đoán (Predicted R-Squared) cho biết độ chính xác của mô hình trên tập dữ liệu chưa biết.

Tuy nhiên, ta không quan tâm tới các mối quan hệ của các biến độc lập với biến phụ thuộc mà chỉ quan tâm tới độ tốt của mô hình trên tập dữ liệu đang xét, nên R^2 hiệu chỉnh hay R^2 cho dự đoán không được sử dụng trong trường hợp này.

Bayes rate (Tỷ lệ Bayes) là giá trị R^2 tốt nhất mà mô hình đạt được dựa trên lý thuyết, được ước lượng thông qua phương trình dự đoán Bayes (3.1). Khi hệ số xác định càng gần với Bayes rate, tức kết quả R^2 – Bayes rate càng tiến về 0 thì độ hiệu quả của mô hình càng tốt.

Sử dụng PyTorch để cài đặt mạng NeuMiss với độ sâu 10, hàm mất mát là MSE, train (huấn luyện) trên CPU, thời gian train tầm 2 phút cho 500 epochs, với batch size là 256, learning rate là 0.001. Mạng cho ra chỉ số R^2 cho tập train quanh ngưỡng 0.8, tập validation và test quanh ngưỡng 0.75.

	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 1: Hiệu suất của NeuMiss với độ sâu 10.

5.3 Một số kết quả khác

Sau đây là một số kết quả thực nghiệm khác cho mạng NeuMiss, được train với 500 epochs.

So sánh mạng khi có residual connection với khi không có residual connection

Với mạng NeuMiss độ sâu 10:

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

Khi không có residual connection, các chỉ số tốt hơn 1 chút so với khi có residual connection. Có thể do khi mạng có độ sâu lớn, việc có residual connection giúp mạng hoạt động hiệu quả hơn.

Qua đây, ta có thể thấy việc có residual connection không ảnh hưởng nhiều tới kết quả so với mạng không có residual connection, như trong bài báo [1] đã đề cập.

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

So sánh các mạng với độ sâu khác nhau

Mạng NeuMiss càng sâu thì hiện tượng hiệu suất giảm dần (diminishing returns) xuất hiện như trong bài báo [1] đã đề cập. Điều này cho thấy việc tăng độ sâu của mạng không đảm bảo sẽ cải thiện kết quả trên tập test.

So sánh các mạng với 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 4: Hiệu suất của NeuMiss độ sâu 10 với các tỷ lệ khuyết khác nhau.

Ta nhận thấy rằng dù tỷ lệ khuyết nhiều nhưng NeuMiss vẫn cho ra kết quả tốt.

So sánh các mạng với số lượng mẫu và đặc trưng khác nhau

Qua Bảng 5, ta thấy rằng mạng NeuMiss chỉ hoạt động tốt khi số lượng mẫu ở mức trung bình trở lên. Điều này cho thấy mạng NeuMiss phù hợp với các tập dữ liệu trung bình, như trong bài báo [1] đã đề cập.

So sánh với các phương pháp khác

Ta so sánh mạng NeuMiss với các phương pháp điền khuyết kết hợp việc sử dụng mô hình hồi quy tuyến tính để dự đoán. Các phương pháp này được cài đặt thông qua thư

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

Bảng 5: Hiệu suất của NeuMiss độ sâu 10 với các tuỳ chỉnh số lượng mẫu và đặc trưng khác nhau cho tập dữ liệu.

viện *fancyimpute*[3], *scikit-learn*[4], với kết quả tốt nhất được chọn từ nhiều tuỳ chỉnh khác nhau. Các kết quả được so sánh với hiệu suất của mạng NeuMiss ở bảng 1, sử dụng chỉ số R^2 để đánh giá trên tập train và tập test.

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

Qua các kết quả này, ta có thể thấy mạng NeuMiss tỏ ra vượt trội hơn so với các phương pháp điền khuyết truyền thống.

[3] https://pypi.org/project/fancyimpute/
[4] https://pypi.org/project/scikit-learn/

Tài liệu tham khảo

[1] 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.

[2] 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.

[3] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[4] Macro. *Deriving the conditional distributions of a multivariate normal distribution*. Cross Validated. Mar. 2023. URL: https://stats.stackexchange.com/q/30600.

[5] L.N. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, 1997.

[6] 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.

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

[8] Suk-Geun Hwang. "Cauchy's interlace theorem for eigenvalues of Hermitian matrices". In: *The American mathematical monthly* 111.2 (2004), pp. 157–159.