Lecture Note: Mathematics for Machine Learning

Nguyễn Quản Bá Hồng*

Ngày 6 tháng 2 năm 2025

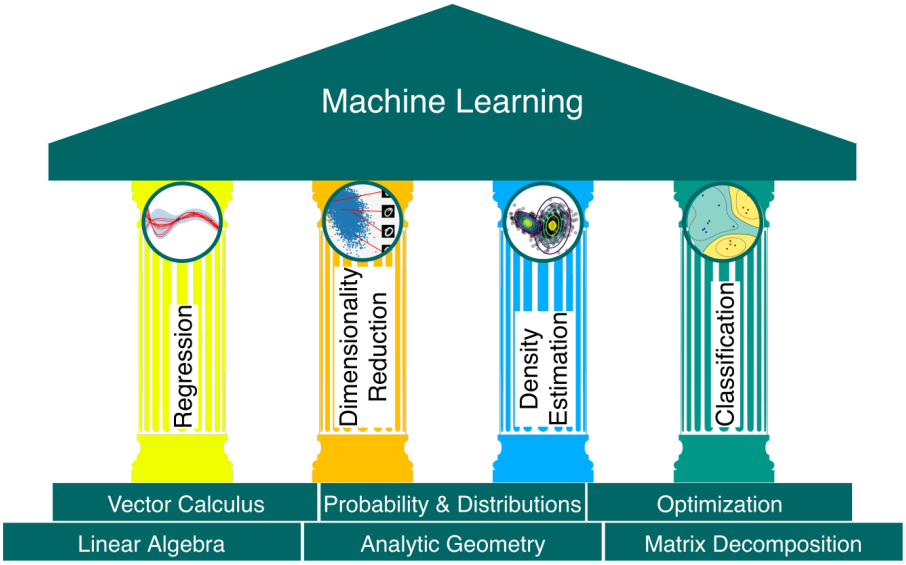
Tóm tắt nội dung

This text is a part of the series *Some Topics in Advanced STEM & Beyond*:
URL: https://nqbh.github.io/advanced_STEM/.
Latest version:

- *Lecture Note: Mathematics for Machine Learning.*
  PDF: URL: https://github.com/NQBH/advanced_STEM_beyond/blob/main/machine_learning/lecture/NQBH_mathematics_for_machine_learning_lecture.pdf.
  TEX: URL: https://github.com/NQBH/advanced_STEM_beyond/blob/main/machine_learning/lecture/NQBH_mathematics_for_machine_learning_lecture.tex.
- *Slide: Mathematics for Machine Learning – Toán Học cho Học Máy.*
  PDF: URL: https://github.com/NQBH/advanced_STEM_beyond/blob/main/machine_learning/slide/NQBH_mathematics_for_machine_learning_slide.pdf.
  TEX: URL: https://github.com/NQBH/advanced_STEM_beyond/blob/main/machine_learning/slide/NQBH_mathematics_for_machine_learning_slide.tex.
- *Personal Expository Notes on Machine Learning.*
  PDF: URL: https://github.com/NQBH/draft/blob/master/machine_learning/NQBH_machine_learning.pdf.
  TEX: URL: https://github.com/NQBH/draft/blob/master/machine_learning/NQBH_machine_learning.tex.
- Codes:
  - Python for Machine Learning: https://github.com/NQBH/advanced_STEM_beyond/tree/main/machine_learning/Python.
  - C/C++ for Machine Learning:

Mục lục

*A Scientist & Creative Artist Wannabe. E-mail: nguyenquanbahong@gmail.com. Bến Tre City, Việt Nam.



Hình 1: Foundations & 4 pillars of ML. Source: [DFO23, Fig. 1.1, p. 14].

1 Linear Regression – Hồi Quy Tuyến Tính

References – Tài nguyên.

1. [Bac24]. Francis Bach. *Learning Theory from First Principles*. Chap. 3: Linear Least-Squares Regression.
2. [DFO23]. Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong. *Mathematics for Machine Learning*. Chap. 9: Linear Regression.
3. Andrew Ng's Machine Learning Specialization slides: .
4. [Tiệ25]. Vũ Hữu Tiệp. *Machine Learning Cơ Bản*. Chap. 7: Hồi Quy Tuyến Tính.
5. Wikipedia: Wikipedia/linear function. Wikipedia/linear regression.

A general idea – 1 ý tưởng tổng quát. Ý tưởng này có vẻ được mượn từ Lý Thuyết Thống Kê Có Tham Số (Parametric Statistics Theory), see, e.g., Wikipedia/parametric statistics.

For $d \in \mathbb{N}^\star$ pairs $(\mathbf{x}_i, y_i)_{i=1}^d$ of input-output, we represent some (mathematically reasonable) hypotheses h about the data using the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_{\mathrm{p}})$ where $\mathrm{p} = \text{size}\,\boldsymbol{\theta}$ is the size/dimension (kích thước/số chiều) of the vector of parameters, i.e., the total number of parameters in the chosen hypotheses h. If the data is correctly predicted according to hypothesis $h_{\boldsymbol{\theta}}$, then $y \approx h_{\boldsymbol{\theta}}(\mathbf{x})$.

Example 1 (Some examples of hypothesis function $h_{\boldsymbol{\theta}}$ – Vài ví dụ về hàm giả thiết $h_{\boldsymbol{\theta}}$)**.**

(i) (Simplest: Linear function/affine mapping) *We assume $h_{\boldsymbol{\theta}}$ be a linear function, i.e., $h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x$ for $d = 1$ & $h_{\boldsymbol{\theta}}(\mathbf{x}) = \theta_0 + [\theta_1, \theta_2, \dots, \theta_d] \cdot \mathbf{x}$ for $d \geq 2$.*

(ii) (Polynomial)*: When $d = 1$, we can assume $h_{\boldsymbol{\theta}} \in \mathbb{R}[\mathbf{x}]$ or even $h_{\boldsymbol{\theta}} \in \mathbb{C}[\mathbf{x}]$, i.e., a polynomial with real/complex coefficients in the variable x:*

$$h_{\boldsymbol{\theta}}(x) = \sum_{i=0}^{\mathrm{p}} \theta_i x^i = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_{\mathrm{p}} x^{\mathrm{p}}.$$

(iii) (Hàm phân thức) *When $d = 1$, we can assume*

$$h_{\boldsymbol{\theta}}(x) = \frac{\sum_{i=0}^{m} \theta_i x^i}{\sum_{m+1}^{\mathrm{p}} \theta_i x^{i-m-1}} = \frac{\theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_m x^m}{\theta_{m+1} + \theta_{m+2} x + \cdots + \theta_{\mathrm{p}} x^{p-m-1}}.$$

(iv) (Hàm căn thức) *We can assume*

$$h_{\boldsymbol{\theta}}(x) = \sqrt[n]{\sum_{i=0}^{\mathrm{p}} \theta_i x^i} = \sqrt[n]{\theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_{\mathrm{p}} x^{\mathrm{p}}},$$

$$h_{\boldsymbol{\theta}}(x) = \sqrt[n]{\frac{\sum_{i=0}^{m} \theta_i x^i}{\sum_{m+1}^{\mathrm{p}} \theta_i x^{i-m-1}}} = \sqrt[n]{\frac{\theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_m x^m}{\theta_{m+1} + \theta_{m+2} x + \cdots + \theta_{\mathrm{p}} x^{\mathrm{p}-m-1}}},$$

for a given (fixed) $n \in \mathbb{R}$ or an adjustable n, i.e., n itself is also a component of $\boldsymbol{\theta}$.

(v) (Elementary function – Hàm sơ cấp) *We can assume* $h_{\boldsymbol{\theta}}(\mathbf{x})$ *to be an elementary function of the variable* $\mathbf{x}$. *For elementary functions, see, e.g., Wikipedia/ elementary function.*

A typical learning algorithm finds the best hypothesis $h_{\boldsymbol{\theta}}$ for the training set $(\mathbf{x}_i, y_i)_{i=1}^N$. We can then estimate the values of y for the test set $(\mathbf{x}_i, y_i)_{i=N+1}^d$ using the "optimal" hypothesis $h_{\boldsymbol{\theta}}(\mathbf{x})$ obtained from our learning algorithm. In particular, if $h_{\boldsymbol{\theta}}$ is a linear function of $x \in \mathbb{R}$, this procedure is called *linear regression.*

1.1 Introduction to linear regression – Giới thiệu về hồi quy tuyến tính

1st impression. Hồi quy tuyến tính (linear regression) là:

- 1 thuật toán hồi quy mà đầu ra là 1 hàm số tuyến tính (linear function, i.e., $y = ax + b$, $a, b \in \mathbb{R}$, $a \neq 0$) của đầu vào. A brief notation:
$$\text{output} = \begin{cases} \texttt{linear\_function}(\text{input}) = a\,\text{input} + b & \text{if } d = 1, \\ \texttt{linear\_function}(\mathbf{inputs}) = \mathbf{w} \cdot \mathbf{inputs} + b & \text{if } d \geq 2, \end{cases}$$
- Thuật toán đơn giản nhất trong nhóm các thuật toán học có giám sát (supervised learning algorithms).

Problem 1 (Housing prices – giá cả nhà đất/bất động sản)**.** *Let* $m \in \mathbb{N}^\star$. *Suppose we have a 2-column table whose the 1st column consists of sizes of* m *houses:* $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ *& the 2nd one consists of their corresponding housing prices* $\mathbf{y} = (y_1, y_2, \ldots, y_m)$. *Here* m *is the number of training examples,* $\mathbf{x}$ *is the vector of input variable or features (size), &* $\mathbf{y}$ *is the vector of output variables or target variables (price).*

Bài toán 1 (Housing price estimation – Ước lượng giá nhà, [Tiệ25], Sect. 7.1, p. 100)**.** *Xét bài toán ước lượng giá của 1 căn phòng rộng* x_1 m^2, *có* x_2 *phòng ngủ, & cách trung tâm thành phố* x_3 *km. Giả sử có 1 tập dữ diệu của* $N = 10000$ *căn nhà trong thành phố đó. Liệu có thể dự đoán được giá* y *của 1 căn nhà mới (i.e., khác* N *căn nhà đã có dữ liệu) thông qua 3 thông số về diện tích* $x_1 \in (0, \infty)$, *số phòng ngủ* $x_2 \in \mathbb{N}$, *& khoảng cách tới trung tâm thành phố* $x_3 \in [0, \infty)$*?*

Mathematical notations – Ký hiệu toán học. Đặt $\mathbf{x} = [x_1, x_2, x_3]^\top \in (0, \infty) \times \mathbb{N} \times [0, \infty)$ là 1 vector cột chứa dữ liệu đầu vào (inputs) gồm 3 thông số diện tích x_1, số phòng ngủ x_2, & khoảng cách đến trung tâm thành phố x_3; $\mathbf{x}$ được gọi là *vector đặc trưng*. Đặt giá nhà (đầu ra/output) $y \in (0, \infty)$.

Reasoning – Biện luận. Nếu diện tích nhà càng lớn thì giá nhà càng cao: $x_1 \uparrow \Rightarrow y \uparrow$, nếu nhà có càng nhiều phòng ngủ thì giá nhà càng cao: $x_2 \uparrow \Rightarrow y \uparrow$, & nếu nhà càng ở gần trung tâm thành phố thì giá nhà càng cao: $x_3 \downarrow 0 \Rightarrow y \uparrow$ (why? economics!). Có thể sử dụng mô hình đầu ra là 1 hàm tuyến tính đơn giản của đầu vào:

$$y \approx \hat{y} := f(\mathbf{x}) = w_1 x_1 + w_2 x_2 + w_3 x_3 = \sum_{i=1}^{3} w_i x_i = \mathbf{x}^\top \mathbf{w} = \mathbf{w}^\top \mathbf{x} = \mathbf{x} \cdot \mathbf{w} = \langle \mathbf{x}, \mathbf{w} \rangle, \tag{lnr}$$

với $\mathbf{w} = [w_1, w_2, w_2]^\top$ là *vector trọng số* (weight vector) với $w_1, w_2 \in (0, \infty), w_3 \in (-\infty, 0)$ (why?) cần tìm. Mối quan hệ cho bởi (lnr) là 1 mối quan hệ tuyến tính.

Question 1 (Regression factor)**.** *Why is it called* regression*?*

Bản chất. 2 bài toán dự đoán giá nhà trên là 2 bài toán dự đoán giá trị đầu ra dựa trên vector đặc trưng đầu vào. Ngoài ra, giá trị của đầu ra y có thể nhận rất nhiều giá trị thực dương khác nhau. Nên ta cần tính đi tính lại, tính kết quả đầu ra mới dựa trên các kết quả đầu ra cũ, để nhận được 1 kết quả được xem là tối ưu theo nghĩa nào đó. Vì vậy, đây cũng là 1 bài toán hồi quy. Kết hợp bản chất hồi quy với mối quan hê tuyến tính $\hat{y} = \mathbf{x}^\top \mathbf{w}$ cho ra tên gọi *hồi quy tuyến tính.* Về hàm tuyến tính, see, e.g., Wikipedia/linear function.

Some reasons to study linear least-squares regression. From [Bac24, Sect. 3.1, pp. 45–46].

- Linear least-squares regression already captures many of the concepts in learning theory, e.g. bias-variance trade-off, as well as dependence of generalization performance on the underlying dimension of the problem with no regularization, or on dimensionless quantities when regularization is added.

  – Hồi quy tuyến tính bình phương nhỏ nhất đã nắm bắt được nhiều khái niệm trong lý thuyết học tập, ví dụ như sự đánh đổi giữa độ lệch và phương sai, cũng như sự phụ thuộc của hiệu suất tổng quát vào chiều cơ bản của vấn đề không có chính quy hóa hoặc vào các đại lượng không có thứ nguyên khi chính quy hóa được thêm vào.

- Because of its simplicity, many results can be easily derived without the need for complicated mathematics, both in terms of algorithms & statistical analysis (simple linear algebra for the simplest results in the fixed design setting).

  – Do tính đơn giản của nó, nhiều kết quả có thể dễ dàng thu được mà không cần đến các phép toán phức tạp, cả về mặt thuật toán & phân tích thống kê (đại số tuyến tính đơn giản cho kết quả đơn giản nhất trong bối cảnh thiết kế cố định).

- Using nonlinear features, it can lead to arbitrary nonlinear predictions.

  – Sử dụng các tính năng phi tuyến tính, nó có thể dẫn đến những dự đoán phi tuyến tính tùy ý.

Question 2 (Optimal choice of parameters $\boldsymbol{\theta}$)**.** *How to choose parameters* $\boldsymbol{\theta}$ *"optimally"? "Optimal" in which sense?*
– Làm thế nào để chọn tham số $\boldsymbol{\theta}$ *"tối ưu"? "Tối ưu" theo nghĩa nào?*

1.2 A general mathematical setting

Tổng quát, nếu mỗi điểm dữ liệu được mô tả bởi 1 vector đặc trưng d chiều $x \in \mathbb{R}^d$, hàm dự đoán đầu ra được viết dưới dạng

$$y = \mathbf{x}^\top \mathbf{w} = \mathbf{x} \cdot \mathbf{w} = \sum_{i=1}^{d} w_i x_i = w_1 x_1 + \cdots + w_d x_d.$$

Nếu tính thêm bias (e.g., tiện đặt cọc, chi phí phát sinh, giảm giá, voucher discount, etc.) $b \in \mathbb{R}$ thì thêm vào thành:

$$y = \mathbf{x}^\top \mathbf{w} + b = \mathbf{x} \cdot \mathbf{w} + b = \sum_{i=1}^{d} w_i x_i + b = w_1 x_1 + \cdots + w_d x_d + b.$$

Với bài toán hồi quy nói chung, ta cần cực tiểu sự sai khác e (error) giữa đầu ra thực sự y & đầu ra dự đoán $\hat{y}$ với e thường được chọn khoảng cách giữa 2 điểm y & $\hat{y}$, e.g.:

$$\frac{1}{2}e^2 = \frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - \mathbf{x}^\top \mathbf{w} - b)^2.$$

(Hệ số $\frac{1}{2}$ đóng vai trò là scaling factor, giúp khi tính đạo hàm thì sẽ đơn giản được hệ số, thật vậy: vì $\left(\frac{1}{2}x^2\right)' = x$ nên hệ số $\frac{1}{2}$ sẽ bị triệ tiêu khi lấy đạo hàm của e theo tham số mô hình $\mathbf{w}$.) Các cách chọn sai số e khác:

1. $e = y - \hat{y}$: Khuyết điểm: $y - \hat{y}$ có thể âm, do đó việc cực tiểu hóa e không có ý nghĩa.
2. $e = |y - \hat{y}|$: Khuyết điểm: Hàm trị tuyệt đối (absolute value function) $f : \mathbb{R} \to [0, \infty)$, $f(x) = |x|$ tuy liên tục nhưng không khả vi tại gốc tọa độ (why?), không thuận tiện cho việc tối ưu.
3. $e = \frac{1}{p}|y - \hat{y}|^p$ với $p \in (1, \infty)$. Hàm số $f(x) = \frac{1}{p}x^p$ có đạo hàm $f'(x) = |x|^{p-1}$. Khuyết điểm: Cần nhiều tính toán hơn số với hàm bình phương, i.e., khi $p = 2$.

Hàm mất mát đóng vai trò là tiêu chí cho việc tối ưu mô hình: việc tìm mô hình tốt nhất/tối ưu nhất (best/optimal) tương đương với việc tìm $\mathbf{w}$ để cực tiểu hàm số:

$$L(\mathbf{w}) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - \mathbf{x}_i^\top \mathbf{w})^2.$$

Optimization problem – Bài toán tối ưu:

$$\min_{\mathbf{w}} L(\mathbf{w}).$$

Nếu bài toán có nghiệm duy nhất, ký hiệu:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} L(\mathbf{w}).$$

2 Overfitting – Quá Khớp

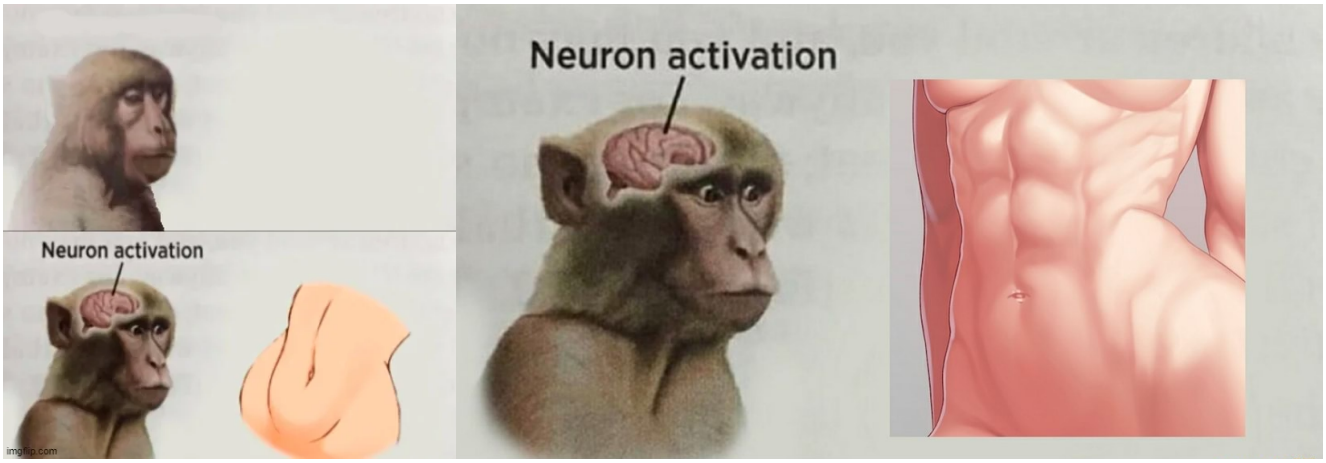
3 K Neighbors – K Lân Cận

4 Phân Cụm K-Means

5 Artificial Neural Networks (ANNs) – Mạng Neuron Nhân Tạo

Distinguish: Artificial Neural Network (ANN) vs. Biological Neural Network (BNN), see, e.g., Wikipedia/neural network (machine learning) vs. Wikipedia/neural network (biology), Psychology Wiki/biological neural network.

Definition 1 (Activation function $\sigma(\cdot)$)**.**



Hình 2: A typical monkey's (maybe man also?) biological neural networks (BNN) gets activated by 2D/anime girl's strong abs.

6 Perception Learning Algorithm – Thuật Toán Học Perceptron

7 Logistic Regression – Hồi Quy Logistic

8 Softmax Regression – Hồi Quy Softmax

9 Deep Neural Networks & Backpropagation – Mạng Neuron Đa Tầng & Lan Truyền Ngược

10 Miscellaneous

10.1 Contributors

Tài liệu

[Bac24] Francis Bach. "Learning Theory from First Principles". In: Adaptive Computation and Machine Learning series (2024), p. 496.

[DFO23] Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. *Mathematics for Machine Learning*. 1st edition. Cambridge University Press, 2023, p. 398.

[Tiệ25] Vũ Khắc Tiệp. *Machine Learning Cơ Bản*. 2025, p. 422.