LINEAR REGRESSION

Lê Bích Phượng

Trường Đại học Mỏ - Địa chất

lebichphuong@humg.edu.vn

Tháng 1 năm 2020

Mục lục

Lời mở đầu

Hồi quy là một vấn đề cơ bản trong machine learning, các vấn đề hồi quy xuất hiện trong một loạt các lĩnh vực nghiên cứu và ứng dụng, bao gồm phân tích chuỗi thời gian (ví dụ: nhận dạng hệ thống), tối ưu hóa, và các ứng dụng deeplearning (ví dụ: trò chơi máy tính, nhận dạng hình ảnh, chú thích video tự động). Hồi quy cũng là một thành phần then chốt của các thuật toán phân loại. Việc tìm một hàm hồi quy đòi hỏi phải giải quyết nhiều vấn đề khác nhau, bao gồm:

1. **Lựa chọn loại mô hình và tham số** của hàm hồi quy. Khi cho một tập dữ liệu, thì các lớp hàm nào (ví dụ: đa thức, ...)và các tham số nào (ví dụ: bậc của đa thức) ta nên chọn để mô hình hóa dữ liệu.
2. **Tìm các tham số tốt**. Khi đã chọn được mô hình của hàm hồi quy, ta cần tìm các tham số tốt cho mô hình. Ở đây, ta cần chọn tham số để cực tiểu hóa hàm loss.
3. **Overfitting và lựa chọn mô hình**. Overfitting là vấn đề mà hàm hồi quy phù hợp với dữ liệu huấn luyện *quá tốt* nhưng lại không có tính khái quát đối với dữ liệu không nhìn thấy.

4. **Mối quan hệ giữa hàm loss và các tham số priors**. Hàm loss (mục tiêu tối ưu) thường được tạo ra bởi một mô hình xác suất.
Chúng ta sẽ xem xét sự liên quan giữa hàm loss và các giả thuyết cơ bản trước đó gây ra những mất mát này.
5. **Mô hình không chắc chắn**. Trong bất kỳ cài đặt thực tế nào, chúng ta chỉ có quyền truy cập vào một lượng dữ liệu (huấn luyện) hữu hạn, có khả năng lớn để chọn lớp mô hình và các tham số tương ứng. Điều này có nghĩa là lượng dữ liệu huấn luyện hữa hạn thì không bao gồm tất cả các tình huống có thể xảy ra, chúng ta có thể muốn mô tả các tham số không chắc chắn còn lại để đạt được mức độ tin cậy của mô hình dự đoán tại thời điểm thử nghiệm; tập huấn luyện càng nhỏ thì mô hình không chắc chắn càng quan trọng.

Bài toán hồi quy tuyến tính

Ta xét bài toán hồi quy với hàm likelihood như sau:

$$p(y \mid \boldsymbol{x}) = \mathcal{N}(y \mid f(\boldsymbol{x}), \sigma^2). \tag{1}$$

Trong đó $\boldsymbol{x} \in \mathbb{R}^D$ là các biến đầu vào và $y \in \mathbb{R}$ là các giá trị hàm mục tiêu. Trong công thức (1),mối quan hệ hàm giữa $\boldsymbol{x}$ và y được cho bởi

$$y = f(\boldsymbol{x}) + \epsilon, \tag{2}$$

Trong đó $\epsilon = \mathcal{N}(0, \sigma^2)$ là phân phối của biến ngẫu nhiên độc lập, có phân phối Gausian với giá trị trung bình bằng 0 và phương sai là σ^2. Mục tiêu của ta là đi tìm một hàm tương tự với hàm f chưa biết.
Ta chọn một hàm được tham số hóa và tìm các tham số $\boldsymbol{\theta}$ mà nó "làm tốt" đối với mô hình dữ liệu.
Giả sử rằng phương sai (the noise variance) σ^2 đã biết và tập trung vào việc nghiên cứu các tham số mô hình $\boldsymbol{\theta}$. Trong hồi quy tuyến tính, ta xét trường hợp đặc biệt là các tham số $\boldsymbol{\theta}$ xuất hiện tuyến tính trong mô hình.

Bài toán hồi quy tuyến tính

Một ví dụ về hồi quy tuyến tính được đưa cho bởi

$$p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}(y \mid \boldsymbol{x}^\top \boldsymbol{\theta}, \sigma^2) \tag{3}$$

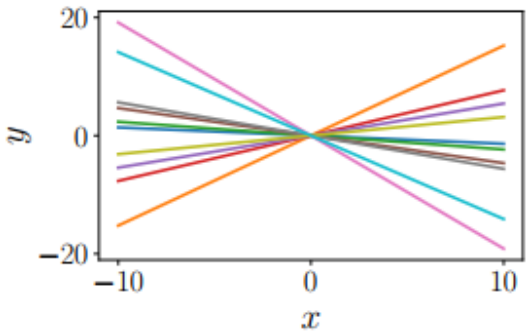
$$\iff y = \boldsymbol{x}^\top \boldsymbol{\theta} + \epsilon, \quad \epsilon = \mathcal{N}(0, \sigma^2), \tag{4}$$

Trong đó $\boldsymbol{\theta} \in \mathbb{R}^D$ là các tham số cần tìm. Lớp các hàm được mô tả bởi công thức (4) là các đường thẳng đi qua gốc tọa độ. Trong (4), ta đã chọn tham số hóa $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\theta}$.
Hàm likelihood trong (3) là hàm mật độ xác suất của y được tính tại $\boldsymbol{x}^\top \boldsymbol{\theta}$.

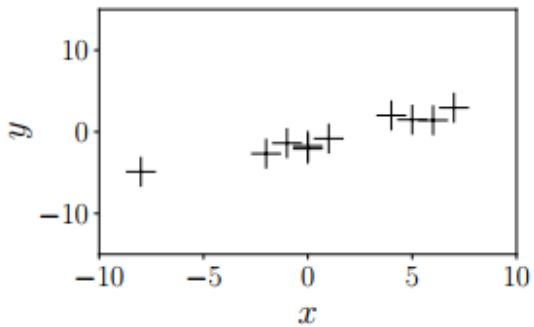
Ví dụ

Example

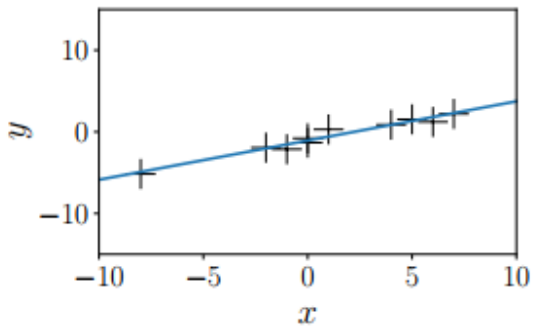
Với $x, \theta \in \mathbb{R}$, mô hình hồi quy tuyến tính trong hình sau mô tả các đường thẳng (các hàm tuyến tính) và tham số θ là hệ số góc của đường thẳng.



(a) Example functions (straight lines) that can be described using the linear model in (9.4).



(b) Training set.

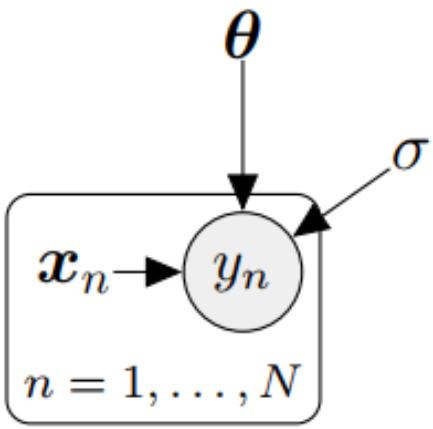


(c) Maximum likelihood estimate.

Hình: Linear regression example. (a) Example functions that fall into this category; (b) training set; (c) maximum likelihood estimate

Ước lượng tham số

Xét cài đặt hồi quy tuyến tính như trong (4) và giả sử cho trước một *tập huấn luyện* (training set) $\mathcal{D} := \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ bao gồm N đầu vào $\boldsymbol{x}_n \in \mathbb{R}^D$ và các mục tiêu tương ứng $y_n \in \mathbb{R}$, $n = 1, \ldots, N$. Mô hình đồ họa tương ứng được thể hiện trong hình sau.



Hình: Probabilistic graphical model for linear regression. Observed random variables are shaded, deterministic/ known values are without circles

Ước lượng tham số

Lưu ý rằng y_i và y_j độc lập có điều kiện với các đầu vào tương ứng $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ cho nên

$$p(\mathcal{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = p(y_1, \ldots, y_N \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{\theta}) \tag{5a}$$

$$= \prod_{n=1}^{N} p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}(y_n \mid \boldsymbol{x}_n^\top \boldsymbol{\theta}, \sigma^2), \tag{5b}$$

trong đó ta định nghĩa $\boldsymbol{X} := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ và $\mathcal{Y} := \{y_1, \ldots, y_N\}$ là các tập huấn luyện đầu vào và các mục tiêu tương ứng. The likelihood và các thành phần $p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})$ theo phân phối Gausian.; see (3).

Sau đây, ta sẽ thảo luận về cách tìm ra các tham số tối ưu $\boldsymbol{\theta}^* \in \mathbb{R}^D$ cho mô hình hồi quy tuyến tính (4). Khi các tham số $\boldsymbol{\theta}^*$ được tìm ra, ta có thể dự đoán các giá trị hàm theo như công thức (4), tại một giá trị $\boldsymbol{x}_*$ tùy ý, phân phối của mục tiêu y_* tương ứng là

$$p(y_* \mid \boldsymbol{x}_*, \boldsymbol{\theta}^*) = \mathcal{N}(y_* \mid \boldsymbol{x}_*^\top \boldsymbol{\theta}^*, \sigma^2). \tag{6}$$

Tiếp theo, ta sẽ xem xét việc đánh giá tham số bằng cách cực đại hóa hàm likelihood.

Ước lượng hợp lý cực đại

Để tìm ra các tham số tối ưu $\boldsymbol{\theta}_{\mathrm{ML}}$ của bài toán hồi quy tuyến tính, chúng ta cực tiểu hàm đối của log hàm hợp lý

$$-\log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) = -\log \prod_{n=1}^{N} p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) = -\sum_{n=1}^{N} \log p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}), \quad (7)$$

Trong mô hình hồi quy tuyến tính (4), hàm hợp lý, the likelihood, là Gaussian, do đó ta có

$$\log p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) = -\frac{1}{2\sigma^2}(y_n - \boldsymbol{x}_n^\top \boldsymbol{\theta})^2 + \text{const}, \quad (8)$$

trong đó hằng số bao gồm tất cả các thành phần độc lập với $\boldsymbol{\theta}$.

$$\mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \boldsymbol{x}_n^\top\boldsymbol{\theta})^2 \tag{9a}$$

$$= \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) = \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2, \tag{9b}$$

- $\boldsymbol{X} := [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]^\top \in \mathbb{R}^{N\times D}$ như bộ các đầu vào huấn luyện;
- $\boldsymbol{y} := [y_1, ..., y_N]^\top \in \mathbb{R}^N$ là vector tất các các mục tiêu huấn luyện

Chú ý rằng hàng thứ n trong ma trận thiết kế, the design matrix, tương ứng với $\boldsymbol{x}_n$ của đầu vào huấn luyện . Trong (9b), ta đã sử dụng tổng bình phương các lỗi giữa quan sát y_n và giá trị mà mô hình dự đoán tương ứng $\boldsymbol{x}_n^\top\boldsymbol{\theta}$ bằng bình phương khoảng cách giữa $\boldsymbol{y}$ và $X\boldsymbol{\theta}$.

Ta tính gradient của $\mathcal{L}$ theo tham số $\boldsymbol{\theta}$

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\boldsymbol{\theta}} = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\left(\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})\right) \tag{10a}$$

$$= \frac{1}{2\sigma^2}\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\left(\boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\theta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\theta}\right) \tag{10b}$$

$$= \frac{1}{\sigma^2}(-\boldsymbol{y}^\top\boldsymbol{X} + \boldsymbol{\theta}^\top\boldsymbol{X}^\top\boldsymbol{X}) \in \mathbb{R}^{1\times D}. \tag{10c}$$

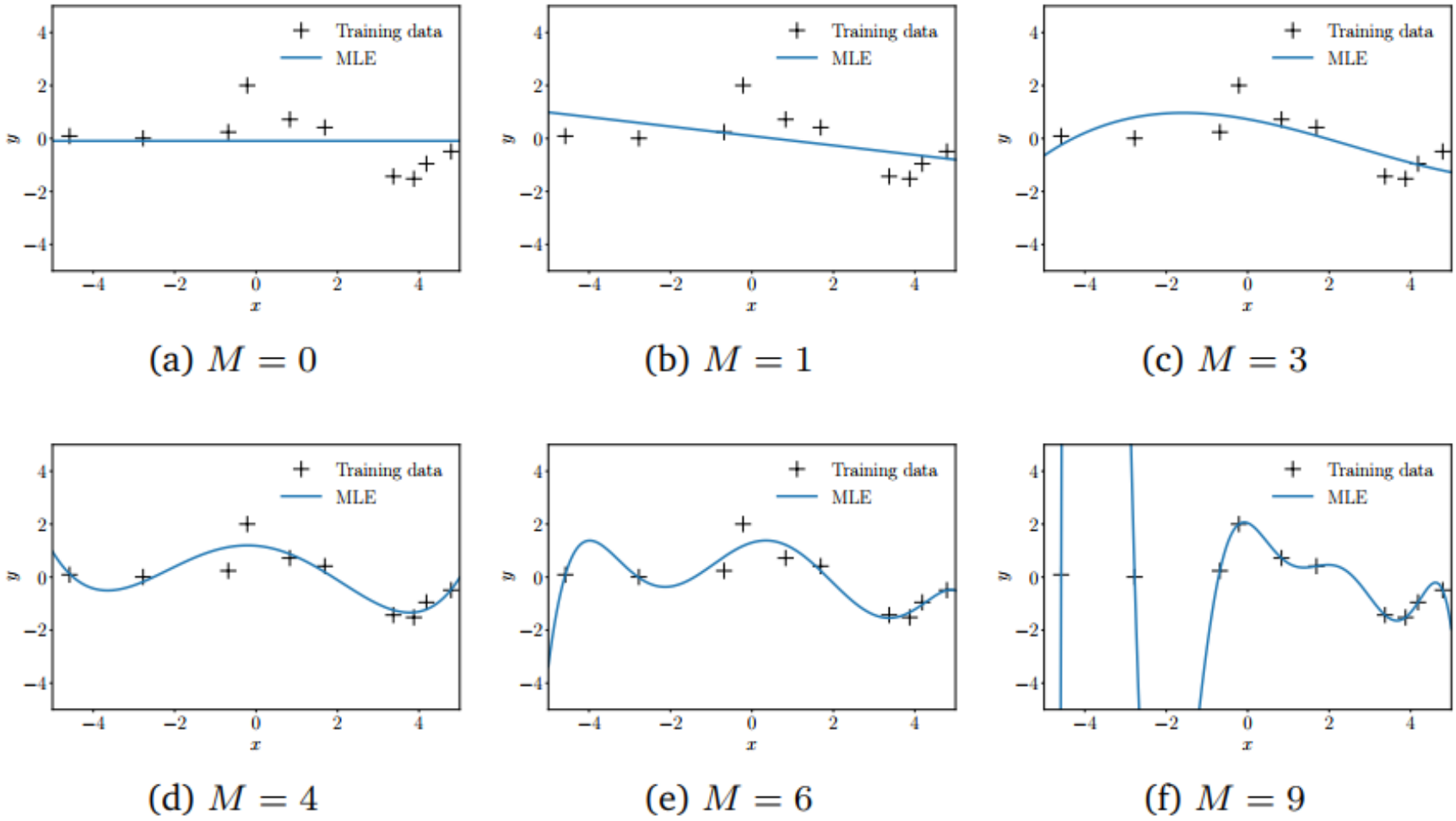
$\boldsymbol{\theta}_{\mathrm{ML}}$ là nghiệm của phương trình $\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\boldsymbol{\theta}} = \boldsymbol{0}^\top$

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\boldsymbol{\theta}} = \boldsymbol{0}^\top \overset{(10c)}{\iff} \boldsymbol{\theta}_{\mathrm{ML}}^\top\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{y}^\top\boldsymbol{X} \tag{11a}$$

$$\iff \boldsymbol{\theta}_{\mathrm{ML}}^\top = \boldsymbol{y}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \tag{11b}$$

$$\iff \boldsymbol{\theta}_{\mathrm{ML}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}. \tag{11c}$$

Vấn đề quá mức trong hồi quy tuyến tính



(a) $M = 0$ (b) $M = 1$ (c) $M = 3$

(d) $M = 4$ (e) $M = 6$ (f) $M = 9$

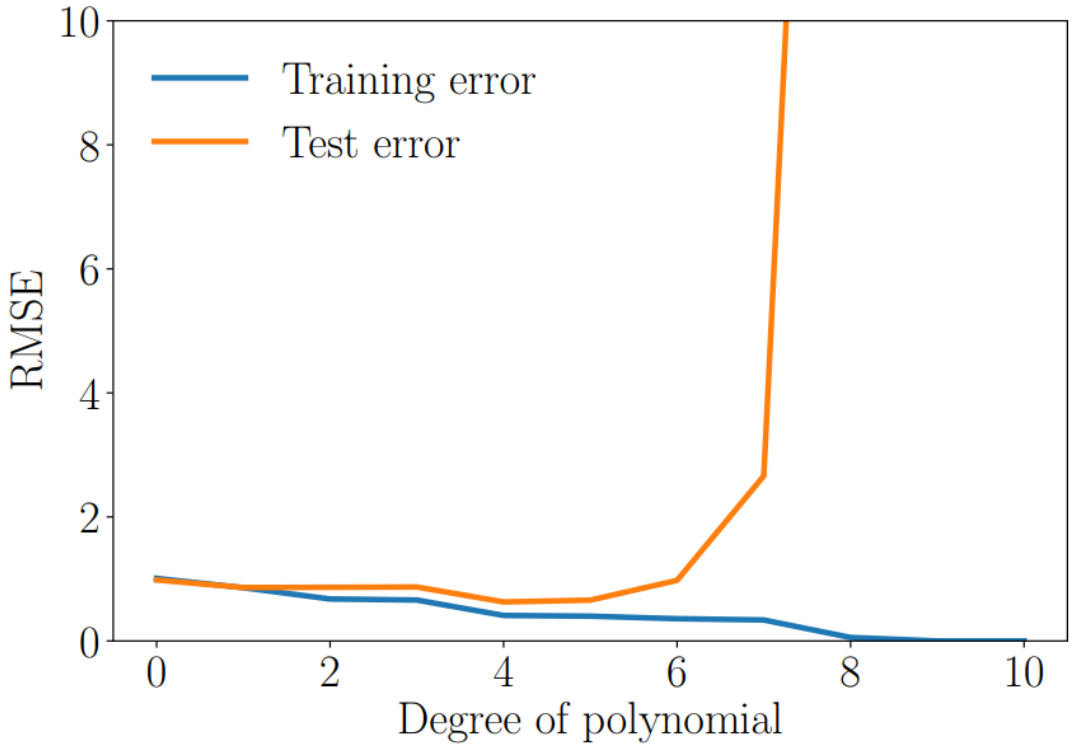
Hình: Maximum likelihood fits for different polynomial degrees M.

Vấn đề quá mức trong hồi quy tuyến tính

Hình trên minh họa một số đa thức ứng với $N = 10$ quan sát.

- Đa thức bậc thấp (hoặc hằng số khi $(M = 0)$ hoặc tuyến tính khi $(M = 1)$ biểu diễn kém,
- Với bậc $M = 3, ..., 5$ biểu diễn dữ liệu trơn chu hơn,
- $M = N - 1 = 9$, hàm sẽ đi qua mọi điểm của dữ liệu. Tuy nhiên các đa thức bậc cao này dao động mạnh và là sự đại diện kém cho hàm khởi tạo, do vậy ta đối mặt với hiện tượng *overfitting*.

Vấn đề quá mức trong hồi quy tuyến tính



Hình: Training and test error

Bây giờ nhìn vào the test error, ta thấy rằng ban đầu lỗi giảm sau đó tăng vọt lên, và bậc của đa thức bằng 4 là tốt nhất theo nghĩa lỗi của cả Training và Test đều thấp.

Ước lượng hậu nghiệm cực đại

Ước lượng hợp lý cực đại có hiện tượng overfitting. Để giảm thiểu ảnh hưởng của các giá trị tham số quá lớn, chúng ta có thể thay thế tham số $\boldsymbol{\theta}$ bởi prior distribution $p(\boldsymbol{\theta})$ trên các tham số.
Trên một tập dữ liệu $\mathcal{X}$, $\mathcal{Y}$ có sẵn, thay vì cực đại hóa hàm likelihood ta tìm các tham số mà cực đại hóa the posterior distribution $p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y})$.
Việc này được gọi là ước lượng *maximum a posteriori (MAP)*.
Ta bắt đầu với dạng log và tính the log-posterior như sau

$$\log p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y}) = \log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + \text{const}, \tag{12}$$

ở đó const là số hạng độc lập với $\boldsymbol{\theta}$.
Để tìm ước lượng MPA $\boldsymbol{\theta}_{\text{MAP}}$, ta đi cực tiểu hóa hàm đối của log của posterior distribution với biến số $\boldsymbol{\theta}$, nghĩa là ta giải

$$\boldsymbol{\theta}_{\text{MAP}} \in \arg\min_{\boldsymbol{\theta}} \{-\log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta})\}. \tag{13}$$

Ước lượng hậu nghiệm cực đại

The gradient của hàm đối của log posterior với biến $\boldsymbol{\theta}$ là

$$-\frac{\mathrm{d}\log p(\boldsymbol{\theta}\mid\mathcal{X},\mathcal{Y})}{\mathrm{d}\boldsymbol{\theta}} = -\frac{\mathrm{d}\log p(\mathcal{Y},\mathcal{X},\theta)}{\mathrm{d}\boldsymbol{\theta}} - \frac{\mathrm{d}\log p(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}, \tag{14}$$

$p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, b^2\boldsymbol{I})$

$$-\log p(\boldsymbol{\theta}\mid\mathcal{X},\mathcal{Y}) = \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{\Phi\theta})^\top(y-\boldsymbol{\Phi\theta}) + \frac{1}{2b^2}\boldsymbol{\theta}^\top\boldsymbol{\theta} + \text{const.} \tag{15}$$

$$-\frac{\mathrm{d}\log p(\boldsymbol{\theta}\mid\mathcal{X},\mathcal{Y})}{\mathrm{d}\boldsymbol{\theta}} = \frac{1}{\sigma^2}(\boldsymbol{\theta}^\top\boldsymbol{\Phi}^\top\Phi - \boldsymbol{y}^\top\Phi) + \frac{1}{b^2}\boldsymbol{\theta}^\top. \tag{16}$$

Ước lượng hậu nghiệm cực đại

$$\frac{1}{\sigma^2}(\boldsymbol{\theta}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - \boldsymbol{y}^\top\boldsymbol{\Phi}) + \frac{1}{b^2}\boldsymbol{\theta}^\top = \mathbf{0}^\top \tag{17a}$$

$$\iff \boldsymbol{\theta}^\top\left(\frac{1}{\sigma^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \frac{1}{b^2}\boldsymbol{I}\right) - \frac{1}{\sigma^2}\boldsymbol{y}^\top\boldsymbol{\Phi} = \mathbf{0}^\top \tag{17b}$$

$$\iff \boldsymbol{\theta}^\top\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \frac{\sigma^2}{b^2}\boldsymbol{I}\right) = \boldsymbol{y}^\top\boldsymbol{\Phi} \tag{17c}$$

$$\iff \boldsymbol{\theta}^\top = \boldsymbol{y}^\top\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \frac{\sigma^2}{b^2}\boldsymbol{I}\right)^{-1} \tag{17d}$$

vì vậy ướng lượng MAP là

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \frac{\sigma^2}{b^2}\boldsymbol{I}\right)^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y}. \tag{18}$$

Kết luận

Trong bài này, tôi trình bày về vấn đề hồi quy tuyến tính, một trong bốn trụ cột của học máy.

Tài liệu tham khảo

- C. Cortes, V. Vapnik, in *Machine Learning*, pp. 273–297 (1995).
- N. Cristianini, J. Shawe-Taylor, *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge University Press, New York, NY, USA (2000).
- J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA (2004).
- C. J. C. Burges, *Data Mining and Knowledge Discovery* **2**, 121 (1998).
- C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer (2006).
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*. 2016
- Hal Daumé III, *A Course In Machine Learning*, 2018
- Marc Peter Deisenroth, A.Aldo Faisal, Cheng Soon Ong, *Mathematics*