xPhO Physics Club

Mô hình hóa và xử lý dữ liệu

Người trình bày: Nguyễn Thành Long

Mục lục

7 Đại lượng SI và 7 thứ nguyên

Đại lượng	Ký hiệu	Đơn vị
Chiều dài	L	Meter (m)
Khối lượng	M	Kilogram (kg)
Thời gian	T	Second (s)
Cường độ dòng điện	I	Ampere (A)
Nhiệt độ	Θ	Kelvin (K)
Lượng chất	N	Mol (mol)
Cường độ sáng	J	Candela (cd)

Bảng: 7 đại lượng cơ bản trong hệ SI và ký hiệu thứ nguyên tương ứng.

- $\Delta t^{\alpha} m^{\beta} R^{\gamma} g^{\delta}$ là đại lượng không thứ nguyên.

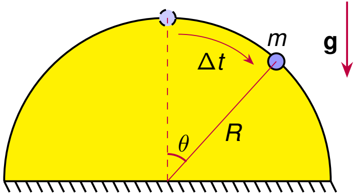
$$\left[\Delta t^{\alpha} m^{\beta} R^{\gamma} g^{\delta}\right] = T^{\alpha}(M)^{\beta}(L)^{\gamma+\delta}(T^{-2})^{\delta}. \qquad (1)$$

nên

$$\begin{cases} \alpha - 2\delta = 0 \\ \beta = 0 \\ \gamma + \delta = 0 \end{cases} \rightarrowtail \begin{cases} \alpha = 2\delta \\ \beta = 0 \\ \gamma = -\delta \end{cases} \qquad (2)$$

Chọn $\delta = 1$ ta có

$$\Delta t = \sqrt{\frac{R}{g}}\, f(\theta). \qquad (3)$$



Hình: Chất điểm trượt trên mặt tròn.

Khi số biến lớn hơn số bậc tự do?

Bài toán dao động con lắc lò xo [1]

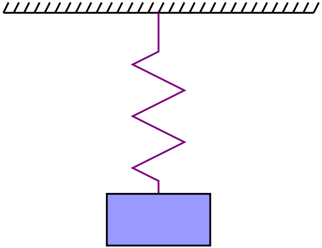
$$\left[\omega m^{\alpha} k^{\beta} \rho^{\gamma} V^{\delta} g^{\varepsilon}\right] = T^{-1-2\beta-2\varepsilon} M^{\alpha+\beta+\gamma} L^{3\delta-3\gamma-\varepsilon}. \tag{4}$$

Giải hệ phương trình và biểu diễn theo 2 biến tự do δ, ε:

$$\omega m^{\alpha} k^{\beta} \rho^{\gamma} V^{\delta} g^{\varepsilon} = \left(\frac{\omega m^{1/2}}{k^{1/2}}\right)\left(\frac{\rho V}{m}\right)^{\delta}\left(\frac{m^{4/3} g}{k \rho^{1/3}}\right)^{\varepsilon}. \tag{5}$$

nên

$$\omega = \sqrt{\frac{k}{m}}\, f\left(\frac{\rho V}{m}, \frac{m^{4/3} g}{k \rho^{1/3}}\right). \tag{6}$$



Hình: Con lắc lò xo.

- Khối lượng m, độ cứng lò xo k, khối lượng riêng của khí ρ, thể tích chất lỏng V, gia tốc trọng trường g.
- Tần số $\omega = f(m, k, \rho, V, g)$.

Mục lục

x	y	x	y
1.01	1.45	10.97	6.53
2.04	2.03	11.94	6.98
2.98	2.47	12.98	7.53
3.95	3.01	13.95	8.00
5.01	3.49	15.01	8.50
5.99	4.02	15.99	8.93
7.02	4.47	17.02	9.49
7.98	4.95	18.07	10.02
9.03	5.52	19.06	10.52
10.01	6.02	19.91	11.03

Bảng: Dữ liệu mẫu.

- Bài toán: Tìm hàm $f(x)$ sao cho $y \approx f(x)$.

Dự đoán mô hình: $f(x) = \beta_0 + \beta_1 x$.

- Tham số cần tìm: $\boldsymbol{\theta} = (\beta_0, \beta_1)$.
- Hàm mất mát (Mean Squared Error):

$$L(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i; \boldsymbol{\theta}))^2$$

  với N là số lượng mẫu dữ liệu.
- Mục tiêu: Tìm $\boldsymbol{\theta}$ sao cho $L(\boldsymbol{\theta})$ nhỏ nhất.

Nghiệm của hồi quy tuyến tính

- Điều kiện đủ để hàm mất mát đạt cực tiểu:

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \beta_0} = -\frac{2}{N}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i) = 0, \tag{7}$$

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \beta_1} = -\frac{2}{N}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)x_i = 0. \tag{8}$$

Có thể thử với máy tính cầm tay Casio!

- Giải hệ phương trình trên, ta được nghiệm:

$$\beta_1 = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2}, \tag{9}$$

$$\beta_0 = \frac{1}{N}\sum_{i=1}^{N} y_i - \beta_1 \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{10}$$

Vậy hàm hồi quy tuyến tính là:

$$f(x) = 0.961 + 0.499x.$$

- Mô hình hồi quy tuyến tính đa biến [2]:

$$f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = \boldsymbol{\theta}^T \mathbf{x}, \tag{11}$$

với $\mathbf{x} = (1, x_1, x_2, \ldots, x_p)$ là vector đặc trưng, và $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p)$ là vector tham số.

- Hàm mất mát (Mean Squared Error):

$$L(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \boldsymbol{\theta}^T \mathbf{x}_i)^2. \tag{12}$$

- Kết quả tính hồi quy:

$$\boldsymbol{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}, \tag{13}$$

Hồi quy đa biến và hồi quy đa thức

- Mở rộng với trường hợp $\boldsymbol{\theta}$ tuyến tính.

Ví dụ:

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \sin(x_1) + \beta_4 x_1 \cos(x_2) + \beta_5 x_2^2. \tag{14}$$

Dữ liệu mở rộng

$$\mathbf{x} = (1, x_1, x_2, \sin(x_1), x_1 \cos(x_2), x_2^2). \tag{15}$$

- Coi hồi quy đa thức là trường hợp đặc biệt của hồi quy đa biến.

$$\mathbf{x} = (1, x, x^2, x^3, \ldots, x^d). \tag{16}$$

Điểm yếu: **Nhạy cảm với nhiễu!**

Mục lục

Thuật toán Gradient Descent

- Tìm cực tiểu của hàm mất mát $L(\boldsymbol{\theta})$ mà không cần tính ma trận nghịch đảo.
- Cập nhật tham số: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla L(\boldsymbol{\theta})$, với η là tốc độ học (learning rate).
- Lặp lại quá trình cho đến khi hội tụ.

Ví dụ: $L(\theta) = 3 + (y - \theta x - 1)^2$ với bộ giá trị $(x, y) = (1, 2)$

- Tính đạo hàm: $\nabla L = 2(y - \theta x - 1)(-x)$
- Cập nhật tham số: $\theta_{n+1} = \theta_n - \eta \nabla L$.
- Dừng thuật toán khi $|L(\theta_{n+1}) - L(\theta_n)| < \epsilon$.

Chọn $\theta_0 = 0$, $\eta = 0.4$!

Bảng: Quá trình hội tụ của thuật toán Gradient Descent.

Bước	θ	$L(\theta)$
0	0.0	4.0
1	0.8	3.04
2	0.96	3.0016
3	0.992	3.00000012
4	0.99968	3.000000102
5	0.999936	3.000000004

- Tối ưu Newton:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \eta \mathbf{H}^{-1} \nabla L(\boldsymbol{\theta}_n), \tag{17}$$

với $\mathbf{H}$ là ma trận Hessian của L, tức là $\mathbf{H} = \nabla^2 L(\boldsymbol{\theta})$.

- Tối ưu Gauss-Newton:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \eta (\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T \mathbf{r}, \tag{18}$$

với $\mathbf{J}$ là ma trận Jacobian của vector sai số $\mathbf{r}$.

- Tối ưu Levenberg-Marquardt (Kết hợp giữa Gradient Descent và Gauss-Newton).

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \eta (\mathbf{J}^T \mathbf{J} + \lambda \mathbf{I})^{-1} \mathbf{J}^T \mathbf{r}, \tag{19}$$

với λ là tham số điều chỉnh.

Tài liệu tham khảo I

[1] D. S. Lemons, *A Student's Guide to Dimensional Analysis* (Student's Guides). Cambridge University Press, 2017.

[2] V. H. Tiệp, *Machine Learning cơ bản*. 2020. [Online]. Available: https://machinelearningcoban.com/ebook/.