Tuần 3 - Thống kê mô tả và Ứng dụng của nó trong Học Máy

Time-Series Team

Ngày 19 tháng 7 năm

Buổi học số 6 (Thứ 6, 19/07/2025) được chia thành 3 phần chính tập trung vào thống kê mô tả cơ bản:

- **Phần 1: Thống kê mô tả cơ bản**
- **Phần 2: Thống kê mô tả nâng cao**
- **Phần 3: Ứng dụng của Thống kê mô tả trong Học Máy**

Phần 1: Thống kê mô tả cơ bản

1.1 Trung bình tổng thể (Population Mean)

Trong thống kê, nếu ta có toàn bộ dữ liệu từ một quần thể, thì giá trị trung bình tổng thể (population mean) được tính bằng:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ví dụ: Khảo sát cân nặng của toàn bộ quần thể chuột gồm 200,000 con, trong đó một vài giá trị mẫu là: 100g, 20g, 30g, 40g. Như ta đã học ở buổi thứ 3 và thứ 4, việc tính toán trung bình tổng thể cho phép ta hiểu được một đặc trưng chính của toàn bộ quần thể.

Vấn đề thực tế

Trong thực tế, chúng ta **hiếm khi có đủ thời gian và nguồn lực** để thu thập toàn bộ dữ liệu từ quần thể.

→ Do đó, ta cần ước lượng trung bình tổng thể bằng cách sử dụng một mẫu nhỏ (sample).

1.2 Trung bình mẫu (Sample Mean)

Thay vì thu thập toàn bộ dữ liệu, ta chọn một mẫu đại diện. Trung bình mẫu (sample mean) được tính bằng:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ví dụ: Ta chọn 5 mẫu chuột có cân nặng lần lượt là 3g, 13g, 19g, 24g, 29g, thì:

$$\bar{x} = \frac{3 + 13 + 19 + 24 + 29}{5} = 17.6$$

Giá trị trung bình mẫu này là một **ước lượng** cho trung bình tổng thể μ.

Tuy nhiên, khi dùng trung bình mẫu để tính phương sai, một vấn đề nảy sinh:

Vấn đề thống kê

Nếu chia phương sai cho n (giống như cách tính phương sai tổng thể), ta sẽ **liên tục đánh giá thấp mức độ phân tán thực sự** quanh trung bình tổng thể.

Lý do là vì khi tính trung bình mẫu, chính mẫu đó đã được dùng để xác định trung bình — nên một bậc tự do đã bị mất. Điều này dẫn đến cần điều chỉnh công thức phương sai mẫu bằng cách chia cho $n-1$ thay vì n.

1.3 Phương sai và độ lệch chuẩn

Phương sai tổng thể:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Phương sai mẫu:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Độ lệch chuẩn:

$$\sigma = \sqrt{\sigma^2} \quad \text{hoặc} \quad s = \sqrt{s^2}$$

1.4 Tại sao chia cho $n-1$? (Degrees of Freedom)

Giả sử bạn biết trung bình mẫu là 17.6. Khi đó, nếu bạn đã biết 4 giá trị đầu tiên trong mẫu, thì giá trị thứ 5 không thể tự do thay đổi — nó bị ràng buộc bởi trung bình.

$\Rightarrow$Chỉ có $n-1$ **giá trị là tự do thay đổi**, gọi là **bậc tự do**.

Ví dụ trực quan:

- Tung 100 đồng xu: Chỉ cần biết số lần xuất hiện mặt ngửa là đủ để suy ra số lần mặt sấp $\Rightarrow$chỉ cần 1 thông tin $\Rightarrow$1 bậc tự do.
- Đèn giao thông có 3 màu. Nếu bạn biết nó không phải Đỏ và cũng không phải Vàng, thì bạn biết chắc nó là Xanh $\Rightarrow$2 bậc tự do.

Tóm tắt ý chính:

- Trung bình tổng thể μ phản ánh giá trị trung tâm của toàn bộ quần thể.
- Trung bình mẫu $\bar{x}$ được dùng để ước lượng μ khi ta chỉ có mẫu nhỏ.
- Khi tính phương sai từ mẫu, cần dùng $n-1$ để không đánh giá thấp độ phân tán thực sự.
- Khái niệm "bậc tự do" giải thích lý do tại sao cần điều chỉnh như vậy.

Phần 2: Thống kê mô tả nâng cao

2.1 Covariance (Hiệp phương sai)

Hiệp phương sai phản ánh chiều hướng biến thiên giữa hai biến ngẫu nhiên.

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_X)(y_i - \mu_Y)$$

Ví dụ minh họa: Số lượng táo xanh và táo đỏ được ghi nhận tại các siêu thị:

Thành phố	Táo xanh	Táo đỏ
Hà Nội	5	12
TP.HCM	15	10
Cần Thơ	18	28
Đà Nẵng	25	38
Bình Định	29	33

Trực quan:

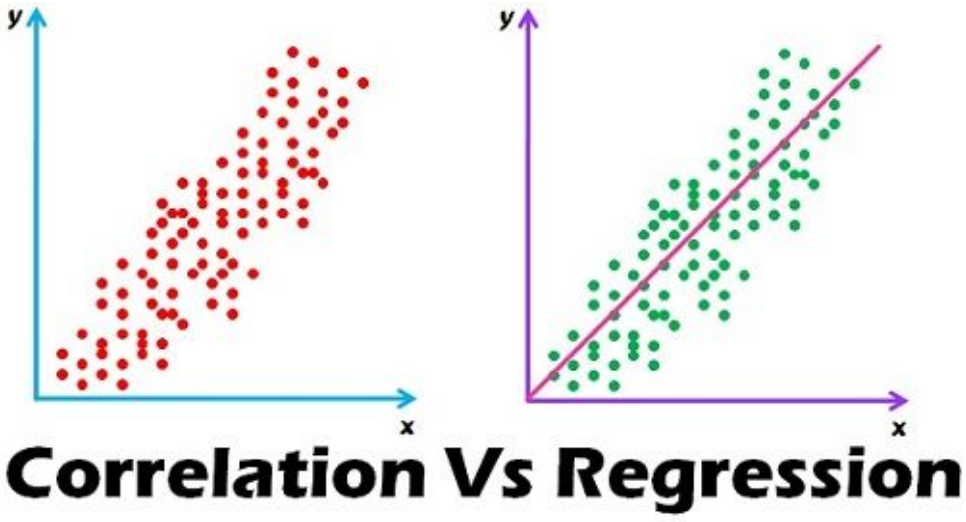
- Các siêu thị có lượng táo xanh cao thì cũng có lượng táo đỏ cao.
- Ngược lại, nơi ít táo xanh thì cũng ít táo đỏ.
- Covariance > 0, cho thấy xu hướng đồng biến.

Hạn chế của Covariance

- Không cho biết độ mạnh của mối quan hệ.
- Nhạy cảm với đơn vị đo (kg, triệu, mm...) ⇒khó so sánh giữa các tập dữ liệu khác nhau.

2.2 Correlation (Tương quan)

Trước khi đi sâu vào phép toán, có 2 khái niệm dễ gây hiểu lầm ta cần làm sáng tỏ, đó là tương quan khác biệt như thế nào với hồi quy.



Sự khác biệt cốt lõi giữa Correlation và Regression:

Sự khác biệt chính

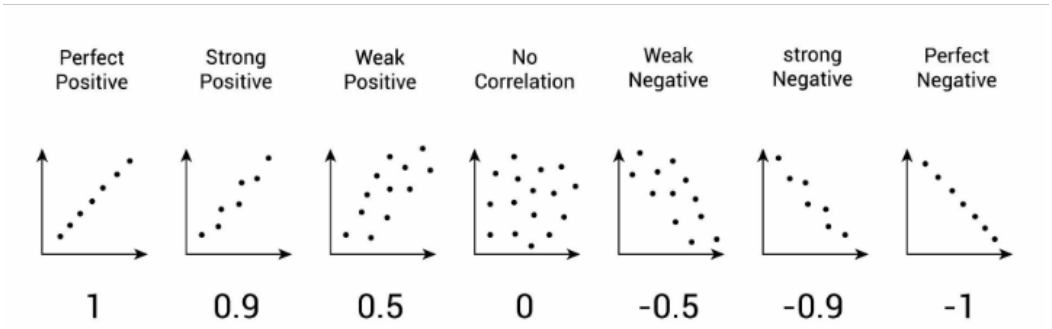
- **Correlation** đo lường **mức độ liên hệ** giữa hai biến độc lập (ví dụ: x và y). Nó không giả định rằng có một biến gây ảnh hưởng lên biến kia.
- **Regression** mô tả **mối quan hệ nhân quả**, trong đó một biến độc lập giải thích hoặc ảnh hưởng đến biến phụ thuộc.

Tóm tắt

- Correlation: "X và Y có liên hệ như thế nào?"
- Regression: "X thay đổi bao nhiêu thì Y thay đổi bao nhiêu?"

Tương quan là khái niệm mô tả xu hướng mối liên hệ giữa hai biến. Có 3 dạng chính:

- Tương quan dương: X tăng $\Rightarrow$Y tăng
- Tương quan âm: X tăng $\Rightarrow$Y giảm
- Không tương quan: X và Y không liên hệ



Hình 1: Pearson Correlation

Chú ý

Tương quan ≠ Nhân quả. Hai biến có tương quan không có nghĩa là một biến gây ra biến kia!

2.3 Correlation Coefficient (Hệ số tương quan)

Hệ số tương quan Pearson r là phiên bản chuẩn hóa của hiệp phương sai, giúp đánh giá mức độ và chiều hướng của mối quan hệ tuyến tính giữa hai biến ngẫu nhiên.

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Tính chất:

- Giá trị nằm trong khoảng: $-1 \leq \rho \leq 1$
- $\rho = 1$: Tương quan dương hoàn hảo
- $\rho = -1$: Tương quan âm hoàn hảo
- $\rho = 0$: Không có tương quan tuyến tính
- **Không nhạy cảm với đơn vị đo**

Để hiểu vì sao cần Correlation Coefficient, ta sẽ đi từ Covariance đến Correlation Coefficient
Hiệp phương sai (Covariance) là chỉ số mô tả hướng thay đổi giữa hai biến:

- Nếu Cov > 0: hai biến cùng chiều (khi một tăng, cái kia cũng tăng).
- Nếu Cov < 0: hai biến ngược chiều (một tăng, cái kia giảm).

Tuy nhiên, giá trị Covariance rất khó diễn giải vì **phụ thuộc vào đơn vị đo lường của từng biến**:

- Ví dụ: Nếu X là chiều cao (đơn vị *cm*), và Y là cân nặng (đơn vị *kg*), thì giá trị Cov sẽ nằm trong đơn vị phức tạp: *cm·kg*.
- Nếu ta đổi đơn vị chiều cao từ *cm* sang *m*, giá trị Covariance cũng thay đổi theo (do nhân với hệ số 0.01).
- Điều này làm cho việc so sánh giữa các tập dữ liệu khác nhau hoặc giữa các bài toán khác nhau trở nên không đáng tin cậy.

Hệ số tương quan Pearson (Correlation Coefficient) là một cách *chuẩn hóa* Covariance, bằng cách chia cho độ lệch chuẩn của từng biến:

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

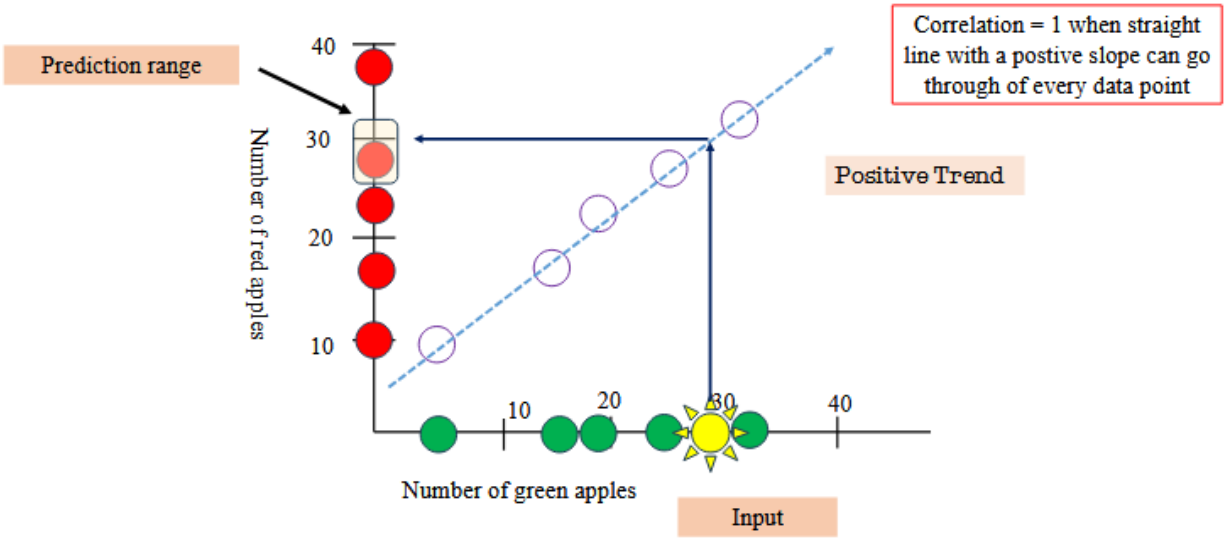
Ưu điểm:

- Không phụ thuộc vào đơn vị đo: Dù bạn đo chiều cao bằng cm hay m, kết quả ρ vẫn giữ nguyên.
- Dễ hiểu: ρ nằm trong đoạn $[-1, 1]$.
  - $\rho \approx 1$: mối liên hệ tuyến tính dương rất mạnh.
  - $\rho \approx 0$: không có mối liên hệ tuyến tính.
  - $\rho \approx -1$: mối liên hệ tuyến tính âm rất mạnh.

Do đó, hệ số tương quan được sử dụng rất phổ biến trong các kỹ thuật học máy như:

- Phân tích thành phần chính (PCA)

- Hồi quy tuyến tính (Linear Regression)
- Phát hiện đặc trưng tương quan trong dữ liệu lớn



Hình trên minh họa mối liên hệ tuyến tính hoàn hảo:

- Mỗi điểm dữ liệu đều nằm trên một đường thẳng dốc lên.
- Khi nhập số lượng táo xanh (Input = 30), có thể dự đoán số lượng táo đỏ nằm trong **Prediction Range** xung quanh xu hướng chính.
- Đây là một mối quan hệ có hệ số tương quan $r = 1$.

Ghi nhớ

Correlation Coefficient không chỉ mô tả mối liên hệ tuyến tính mạnh hay yếu, mà còn giúp đưa ra dự đoán có cơ sở (như trong học máy).

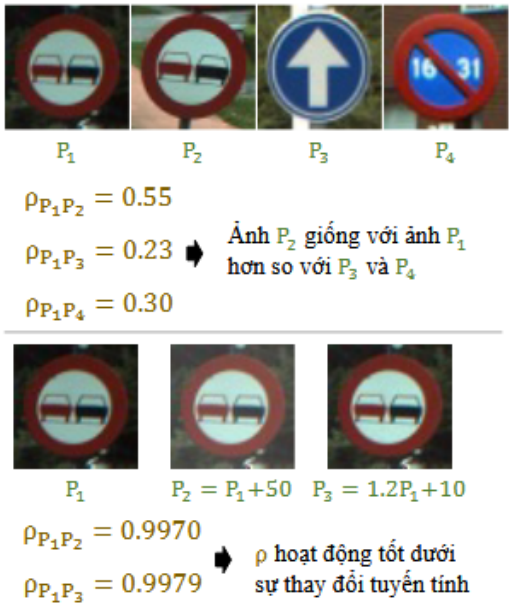
Tóm tắt phần 2:

- Covariance cho biết hai biến có cùng chiều thay đổi hay không.
- Correlation mô tả xu hướng mối liên hệ (dương, âm, không liên hệ).
- Hệ số tương quan Pearson giúp đo lường mức độ và hướng tương quan tuyến tính một cách chuẩn hóa.

Phần 3: Ứng dụng của Thống kê mô tả trong Học Máy

3.1 So sánh mức độ tương đồng giữa ảnh

Hệ số tương quan có thể dùng để đo độ giống nhau giữa các ảnh dưới dạng vector.




```python
# aivietnam.ai

import numpy as np
from PIL import Image

# load ảnh và chuyển về kiểu list
image1 = Image.open('images/img1.png')
image2 = Image.open('images/img2.png')
image3 = Image.open('images/img3.png')
image4 = Image.open('images/img4.png')

image1_list = np.asarray(image1).flatten().tolist()
image2_list = np.asarray(image2).flatten().tolist()
image3_list = np.asarray(image3).flatten().tolist()
image4_list = np.asarray(image4).flatten().tolist()


# tính correlation coefficient
corr_1_2 = find_corr_x_y(image1_list, image2_list)
corr_1_3 = find_corr_x_y(image1_list, image3_list)
corr_1_4 = find_corr_x_y(image1_list, image4_list)

print('corr_1_2:', corr_1_2)
print('corr_1_3:', corr_1_3)
print('corr_1_4:', corr_1_4)
```

Ví dụ: Giả sử ta có các ảnh biển báo P_1, P_2, P_3, P_4.

- $r_{P_1,P_2} = 0.55$: Ảnh P_2 có vẻ giống P_1
- $r_{P_1,P_3} = 0.23$ và $r_{P_1,P_4} = 0.30$: ít giống hơn
- Ta có thể kết luận: P_2 gần giống P_1 nhất

Hệ số tương quan vẫn hoạt động tốt dưới các phép biến đổi tuyến tính:

- $P_2 = P_1 + 50$
- $P_3 = 1.2P_1 + 10$
- $\Rightarrow r \approx 0.9970$: vẫn giữ mức độ tương đồng rất cao

Ví dụ code minh họa (Python):

```python
import numpy as np
from PIL import Image

# Load ảnh và ểchuyn sang vector
image1 = Image.open('images/img1.png')
image2 = Image.open('images/img2.png')
image1_list = np.asarray(image1).flatten().tolist()
image2_list = np.asarray(image2).flatten().tolist()

# Tính ệh ố ươtng quan
def find_corr_x_y(x, y):
```

```
    return np.corrcoef(x, y)[0, 1]

corr = find_corr_x_y(image1_list, image2_list)
print("Correlation:", corr)
```

Listing 1: So sánh độ tương đồng ảnh bằng hệ số tương quan

3.2 Template Matching – Tìm mẫu trong ảnh lớn

Một ứng dụng quan trọng khác của hệ số tương quan là tìm kiếm mẫu nhỏ trong ảnh lớn — còn gọi là **template matching**.

Trong ví dụ trên:

- Ảnh nhỏ bên trái là **template** cần tìm.
- Ảnh bên phải là **output**, nơi vị trí có tương quan cao nhất (ô đỏ) là vị trí khớp với template.
- Hệ số tương quan được tính tại mỗi điểm trượt của template trên ảnh.

OpenCV hỗ trợ sẵn hàm 'matchTemplate':

```
import cv2

output = cv2.matchTemplate(image, template, cv2.TM_CCOEFF_NORMED)
min_val, max_val, min_loc, max_loc = cv2.minMaxLoc(output)
```

Listing 2: Template Matching với hệ số tương quan chuẩn hóa

Kết luận

Hệ số tương quan không chỉ giúp phân tích mối quan hệ giữa các biến, mà còn có ứng dụng mạnh trong thị giác máy tính như:

- So sánh độ giống nhau giữa hình ảnh
- Phát hiện mẫu trong ảnh (template matching)