Statistics: Correlation Coefficient

Hoàng-Nguyên Vũ

1. Lý thuyết: Correlation Coefficient

1.1 Định nghĩa

Hệ số tương quan Pearson đo lường mức độ liên hệ tuyến tính giữa hai biến ngẫu nhiên X và Y.

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

Trong đó:

- $\text{Cov}(X, Y)$: hiệp phương sai giữa X và Y
- σ_X, σ_Y: độ lệch chuẩn của X và Y

1.2 Công thức tính từ dữ liệu

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

1.3 Ý nghĩa của hệ số r

- $r = 1$: Tương quan tuyến tính dương hoàn hảo
- $r = -1$: Tương quan tuyến tính âm hoàn hảo
- $r = 0$: Không có tương quan tuyến tính

1.4 Ứng dụng trong AI

- Phân tích tương quan giữa đặc trưng và nhãn (feature selection)
- So sánh embedding vectors
- Tìm kiếm văn bản sử dụng TF-IDF và correlation

2. Bài tập Correlation Coefficient

Bài 1. Tương quan tuyến tính hoàn hảo Cho $x = [1, 2, 3, 4, 5]$, $y = [2, 4, 6, 8, 10]$. Tính hệ số tương quan Pearson giữa x và y.

Bài 2. Tương quan âm hoàn hảo Cho $x = [1, 2, 3, 4, 5]$, $y = [10, 8, 6, 4, 2]$. Tính hệ số tương quan và giải thích kết quả.

Bài 3. Không tương quan tuyến tính Tạo $x \in [0, 10]$, $y = \sin(x)$ với 100 điểm. Tính Pearson correlation giữa x và y.

Bài 4. Tương quan giữa đặc trưng và nhãn Cho feature $= [1.1, 1.9, 3.2, 4.5, 5.1]$, label $= [1.0, 2.0, 3.0, 4.1, 5.3]$. Tính tương quan giữa feature và label.

Bài 5. Tương quan trong bảng dữ liệu Cho height $= [150, 160, 170, 180, 190]$, weight $= [50, 60, 70, 80, 90]$. Tính tương quan giữa chiều cao và cân nặng.

Bài 6. Embedding similarity Cho $\text{embed}_A = [0.3, 0.5, 0.7, 0.8]$, $\text{embed}_B = [0.9, 1.4, 2.1, 2.4]$. Tính tương quan giữa hai vectors.

Bài 7. Tương quan ngẫu nhiên Sinh hai vector ngẫu nhiên $x, y \in [0, 1]$ gồm 100 giá trị. Tính Pearson correlation và nhận xét.

Bài 8. Tương quan với nhiễu (noise) Cho $x = [0, 1, 2, ..., 99]$, $y_{\text{clean}} = x$, $y_{\text{noisy}} = x + \text{noise}$. So sánh tương quan giữa x và hai biến y.

Bài 9. Dữ liệu thời gian: nhiệt độ và doanh số Cho temperature $= [22, 24, 23, 25, 26]$, sales $= [100, 110, 105, 115, 120]$. Tính tương quan giữa nhiệt độ và doanh số.

Bài 10. Retrieval văn bản với TF-IDF + Pearson (không dùng thư viện) Cho các văn bản:

```
doc1 = "deep learning for natural language processing"
doc2 = "transformer models improve language understanding"
doc3 = "convolutional neural networks for image classification"
query = "language models for text understanding"
```

Tính TF-IDF, sau đó tính hệ số tương quan Pearson giữa truy vấn và từng văn bản. Xếp hạng các văn bản theo mức độ liên quan.