1 Bài tập Lớp nhà

1.1 Tuần 1: 08/01/2024

1.1.1 Giới thiệu về MSA

1. Xác suất và thống kê khác nhau như thế nào?

   Trả lời:

   - **Xác suất** là về "điều gì có thể xảy ra".
   - **Thống kê** là về "điều gì đã xảy ra và ý nghĩa của nó".
   - Mục tiêu Xác suất là từ thông tin từ quần thể (population) có thể suy diễn từ mẫu (sample).
   - Thống kê là từ mẫu suy diễn ngược lại đặc tính, bản chất của mẫu (sample).

2. Cho ví dụ người ta cần xử lý như thế nào trong các lĩnh vực của Intended Audience liên quan tới MSA (Computer Science, Biology, Physics, Social Sciences,...)?

   Trả lời

   Computer Science (Khoa học Máy tính):

   Ví dụ: Phân tích hiệu suất hệ thống phần mềm.

   Xử lý:

   - Thu thập dữ liệu từ nhiều nguồn (CPU usage, memory, network latency).
   - Áp dụng MSA (như PCA - Principal Component Analysis) để giảm chiều dữ liệu và xác định các yếu tố chính ảnh hưởng đến hiệu suất.
   - Sử dụng phân tích cụm (Cluster Analysis) để nhóm các cấu hình hệ thống tương tự nhau.

   Mục tiêu: Tối ưu hóa hiệu suất hệ thống và xác định các vấn đề tiềm ẩn.

1.1.2 Các khái niệm cơ bản về MSA

1. Ý nghĩa của Sample Mean là gì? Nó dùng để tính gì?

   Trả lời: Sample Mean dùng để tính giá trị trung bình của một mẫu dữ liệu, đại diện cho xu hướng trung tâm của tập mẫu, từ đó suy ra đặc điểm chung của tổng thể.

2. Cho ví dụ cụ thể về sample mean? Vì sao chỉ dùng sample mean lại phản ánh không đúng bản chất của tập dữ liệu?

   Trả lời:

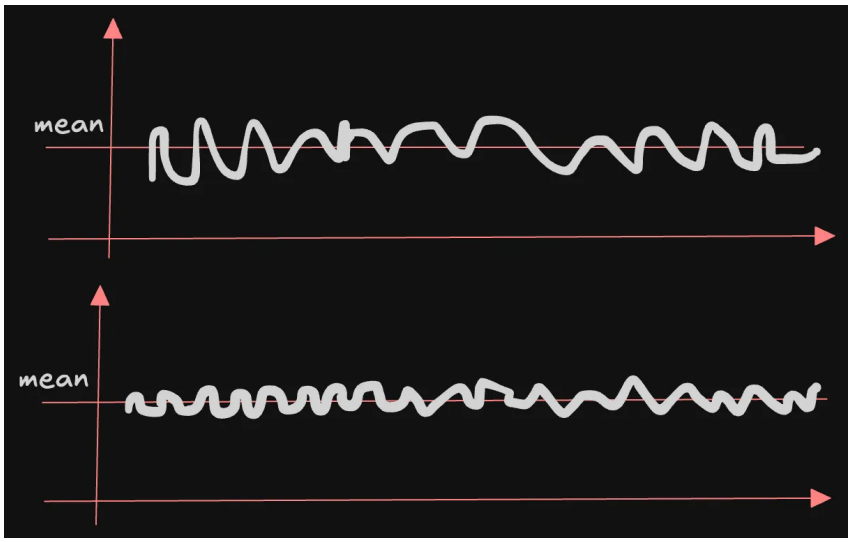

Figure 1: Ảnh minh hoạ về 2 mẫu là 2 tần số khác nhau nhưng cùng giá trị sample mean

Ảnh bên trên mô phỏng 2 tính hiệu có biên độ dao động khác nhau, nhưng lại cùng 1 giá trị trung bình (`mean`). Từ ví dụ trên, sample Mean không phản ánh đầy đủ tập dữ liệu vì nó không thể hiện sự phân bố hoặc độ biến thiên (variability) của dữ liệu.

3. Vì sao công thức của sample variance lại cần bình phương?

   Trả lời:

   - Tránh triệt tiêu độ lệch
     - Nếu không bình phương, độ lệch $(x_i - \bar{x})$ có thể âm hoặc dương.
     - Khi cộng tất cả các độ lệch này, kết quả luôn bằng 0 vì các giá trị âm và dương sẽ triệt tiêu lẫn nhau.
     - Việc bình phương loại bỏ dấu âm, đảm bảo rằng phương sai chỉ thể hiện "độ lớn" của sự chênh lệch mà không bị ảnh hưởng bởi hướng.
   - Muốn sai số bé có thể làm nhỏ xuống, và sai số lớn được đẩy lớn lên. Điều này đẩy sai số lên lớn hơn để có thể thể hiện rõ độ chênh lệch của dữ liệu với giá trị trung bình.

4. Cho ví dụ trong thị giác máy tính (xử lý ảnh, đồ hoạ máy tính, thị giác máy tính) có dùng sample mean, sample variance?

5. Chứng minh biểu thức $r_{ik} \leq 1$. (Gợi ý: $r_{ik} = cos(v_i, v_k)$)

$$r_{ik} = \frac{S_{ik}}{\sqrt{S_{ii}}\sqrt{S_{kk}}} = \frac{\sum_{j=1}^{n}(x_{ji} - \overline{x_i})(x_{jk} - \overline{x_k})}{\sqrt{\sum_{j=1}^{n}(x_{ji} - \overline{x_i})^2}\sqrt{\sum_{j=1}^{n}(x_{jk} - \overline{x_k})^2}},$$

$$\mathrm{i,k} = 1,2,...,\mathrm{p}$$

$$r_{ik} = r_{ki} \quad \forall i,k$$

6. Vì sao phải tính thêm phần mẫu số trong công thức Sample Correlation Coeffiicient?

   Trả lời:

   Phần mẫu số $s_1 s_2$ là cần thiết vì nó chuẩn hóa giá trị của **sample covariance** (s_{xy}), giúp điều chỉnh cho sự khác biệt về độ biến thiên (variance) của từng biến. Điều này giúp xác định mối tương quan một cách chính xác và không bị ảnh hưởng bởi đơn vị đo lường hoặc mức độ biến thiên của từng biến.

   Công thức trên và *cos* của góc giữa 2 vector i, k. Khi đó, xét được sự tương quan giữa 2 biến mạnh hay yếu. Cho biết 2 biến và i và k khớp nhau như thế nào.

7. Ví dụ:

$$X = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$

   Tính sample mean, sample variance + sample standard deviation, sample covariance, sample correlation coefficient.

8. **Ưu và nhược điểm của biểu đồ phân tán (scatter plot)? So sánh hai loại biểu đồ: scatter plot và dot diagram?** hint: áp dụng vào xử lý ảnh, khi có histogram thì 2 biểu đồ biểu diễn như thế nào?

   Trả lời:

9. Hãy cho ví dụ và minh họa bằng biểu đồ Star Plot và Chernoff Faces.

   Trả lời: