Xử lí đa chiều Thực hành tuần 1

Trình bày: Đặng Bách Phố, 20110079.

Bài tập 1: Xét một tập dữ liệu gồm các quan sát x_n trong đó $n = 1, \dots, N$ và x_n là một biến Euclide có số chiều D. Mục tiêu là chiếu dữ liệu lên một không gian có số chiều $M \leq D$ bằng cực đại hóa phương sai của dữ liệu được chiếu.

Đưa ra một chứng minh chi tiết của PCA bằng phương pháp cực đại hóa phương sai với số chiều $M = 2$. Trả lời các câu hỏi sau để làm rõ chứng minh:

a) Chỉ ra bài toán tối ưu cụ thể, ghi rõ công thức của ma trận Hiệp phương sai.
b) Sử dụng hàm lagrange để giải bài toán tối ưu đã nêu – ghi rõ biểu diễn của v_1, v_2 (hai thành phần chính).

Giải

a) Việc tối ưu hóa cho PCA với M = 2 là tìm không gian con hai chiều tối đa hóa phương sai của dữ liệu. Cho X là ma trận dữ liệu D chiều. Ma trận hiệp phương sai của X có dạng:

$$S = \frac{1}{N}\sum (x_i - \mu)(x_i - \mu)^T$$

trong đó: x_i là quan sát thứ i và N là số quan sát.

Phương sai của y được cho bởi:

$$Var(y) = \frac{1}{N}\sum \left(u^T(x_i - \mu)\right)^2$$

Mục tiêu là tìm ra hai vectơ đơn vị u_1 và u_2 tối đa hóa tổng phương sai của các phép chiếu của chúng lên không gian con mà chúng trải dài. Điều này có thể được thể hiện như vấn đề tối ưu hóa sau:

tối đa u_1, u_2:

$$u_1^T S u_1 + u_2^T S u_2$$

với, $u_1^T u_1 = 1$ và $u_2^T u_2 = 1$ $u_1^T u_2 = 0$

b) Chúng ta có thể sử dụng phương pháp nhân Lagrange để giải quyết vấn đề tối ưu hóa.Với các hệ số nhân Lagrange λ_1 và λ_2:

$$L(u_1, u_2, \lambda_1, \lambda_2) = u_1^T S u_1 + u_2^T S u_2 - \lambda_1(u_1^T u_1 - 1) - \lambda_2(u_2^T u_2 - 1) - \lambda_3 u_1^T u_2$$

$$\frac{\partial L}{\partial u_1} = 2Su_1 - 2\lambda_1 u_1 - \lambda_3 u_2 = 0$$

$$\frac{\partial L}{\partial u_2} = 2Su_2 - 2\lambda_2 u_2 - \lambda_3 u_1 = 0$$

$$\frac{\partial L}{\partial \lambda_1} = u_1^T u_1 - 1 = 0$$

$$\frac{\partial L}{\partial \lambda_2} = u_2^T u_2 - 1 = 0$$

$$\frac{\partial L}{\partial \lambda_3} = u_1^T u_2 = 0$$

Nhân phương trình đầu tiên với u_2^T và phương trình thứ hai với u_1^T, lấy chúng trừ nhau, ta có:

$$2Su_1 u_2^T - \lambda_3 3(u_2 u_1^T + u_1 u_2^T) = 0$$

Vì u_1 và u_2 là trực giao, $u_1 u_2^T + u_2 u_1^T = 2I$, trong đó I là ma trận đơn vị. Do đó, chúng ta có:

$$Su_1 = \lambda_1 u_1 Su_2 = \lambda_2 u_2 Su_1 u_2^T = 0$$

Để tìm u_1 và u_2, chúng ta có thể giải bài toán eigenvalue của ma trận hiệp phương sai S:

$$Su \;=\; \lambda u$$

$$(S \;-\; \lambda I)u \;=\; 0$$

$$det(S \;-\; \lambda I) \;=\; 0$$

Công thức cho u_1 và u_2 có thể được viết là:

$u_1 = Argmax\big||Xw|\big|^2$, với$\big||w|\big| = 1. u_2 = Argmax\big||Xw|\big|^2$, với $||\, w||\; =\; 1$ và $w \perp u_1$

Với y_i là phép chiếu hai chiều của x_i lên không gian con kéo dài bởi u_1 và u_2.

$$y_i = [u_1^T x_i, u_2^T x_i]$$

Trong đó y_i là phép chiếu hai chiều của x_i lên không gian con kéo dài bởi u_1 và u_2. Phương sai của dữ liệu dự kiến:

$$Var(y) = \frac{1}{N}\sum_{i=1}^{N}\big(y_{i_1} - \mu_1\big)^2 + \big(y_{i_2} - \mu_2\big)^2$$