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 trong đó và là một biến Euclide có số chiều D. Mục tiêu là chiếu dữ liệu lên một không gian có số chiều 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 . Trả lời các câu hỏi sau để làm rõ chứng minh:

1. Chỉ ra bài toán tối ưu cụ thể, ghi rõ công thức của ma trận Hiệp phương sai.
2. Sử dụng hàm lagrange để giải bài toán tối ưu đã nêu – ghi rõ biểu diễn của (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:

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

Phương sai của được cho bởi:

Mục tiêu là tìm ra hai vectơ đơn vị và 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

với, và

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 và :

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

Vì và là trực giao, , trong đó I là ma trận đơn vị. Do đó, chúng ta có:

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

Công thức cho và có thể được viết là:

với, với và

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

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