Phân tích thống kê dữ liệu nhiều biến Seminar

Lưu Nam Đạt
22127062
lndat22@clc.fitus.edu.vn

Nguyễn Bá Công
22127046
nbcong22@clc.fitus.edu.vn

Nguyễn Huỳnh Hải Đăng
22127052
nhhdang22@clc.fitus.edu.vn

Đặng Trần Anh Khoa
22127024
dtakhoa22@clc.fitus.edu.vn

Lời Giới Thiệu

None

Mục lục

1 phân tích tương quan chính tắc

Phương pháp:

Tương quan giữa các biến chính tắc và các biến gốc:

- Các biến chính tắc không có tính chất tự nhiên nên khó phân ích

ý nghĩa hình học:

- Xét đại lượng mang giá trị thực

Cho A = [3 3 3 \ 3 3 3 \ 3 3 3]

$\lambda_1\lambda_2\lambda_3$

A^(1/2) = sqrt(lambda1) e1 e1^T + sqrt(lambda2) e2 e2^T + sqrt(lambda2) e2 e2^T

2 PCA - PHÂN TÍCH THÀNH PHẦN CHÍNH

pca giảm chiều dữ liệu

tối ưu hiệu suất mô hình

–> phải nêu ra thách thức khi số chiều gia tăng, yêu cầu cấu hình lớn ra sao. nhóm chưa nêu rõ vì sao cần giảm nhiễu

ý nghĩa khoa học:

- trích xuất thông tin quan trọng:

thiếu: trong p biến, phải tìm được 1 không gian mới, không gian p chiều, về khoa học là tìm ra được một không gian mới có số chiều thấp hơn dữ liệu đầu vào

dữ liệu nào không có tương làm như vậy, giúp dữ liệu có tính khả tắc

xấp xỉ sao cho sai số là thấp nhất có thể

- tối ưu hoá dữ liệu (how?)

ý nghĩa thực tiễn:
- xử lý ảnh và thị giác máy tính:

thiếu: tìm không gian con để chiếu dữ liệu mặt xuống không gian con mới từ 1tr xuống còn 100 chiều, sau đó được lấy đi nhận dạng.

(nói chung chung)

- y học: giúp giảm chiều
- NLP: (dữ liệu
- Tài

2.1 PHÁT BIỂU BÀI TOÁN

(tìm tập hợp mới các biến số)

(chưa biết thành phần chính là gì)

("giữ lại phần lớn phương sai còn lại" ???)

Tìm ra một không gian con `m` chiều thấp hơn số chiều của không gian của tập mẫu ban đầu
- `m` biến không tương quan với nhau
- phương sai của `m` biến lớn nhất có thể
- sai số xấp xỉ trong không gian con `m` chiều là bé nhất

Input:
- tập dữ liệu dưới dạng ma trận X
- n là số mẫu
- p la số biến

Output: Ma trận mới Z chứa các thành phần chính (???) giữ lại phần lớn phương sai (???) của dữ liệu gốc.

2.2 Các công đoạn chính của hệ thống

PHƯƠNG PHÁP: ???

Tập mẫu, mỗi điểm đang được tính từ gốc toạ độ.

Ví dụ: Toạ độ

Chọn 1 điểm mới là sample mean, tức trung bình mẫu. Từ điểm này tạo thành trục mới e1, e2

$X = X - \mu$

Subtract mean: toạ độ của tất cả các điểm còn lại sẽ được tính theo hệ toạ độ mới.

tính ma trận hiệp phương sai

- Tính trị riêng và vector riêng

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

Chọn K vector riêng có giá trị riêng lớn nhất để xây dựng ma trận U

mỗi trục toạ độ của không gian con là 1 vector riêng.

chiếu sample lên không gian mới.

$Z = \hat{X}U_k$

2.3 Phương pháp

$X = U_K Z + \bar{U}_K Y$

(mạch logic...?)

eigenface pca

kernel PCA: PCA kernel phi tuyến

Robust PCA tách dữ liệu thành phần chính và nhiễu: N mẫu + p biến –> ma trận có k vector đặc trưng, có giảm số biến Incremental PCA

==> Cần giải thích ra thành phẩm cụ thể của từng công trình, không hiểu rõ phương pháp và sự liên quan của nó với PCA gốc

3 Phân lớp

Gắn nhãn cho

Định nghĩa: nói quá nhanh, ko hiểu nói gì

Công đoạn 1:

Công đoạn 2: Tập dữ liệu đã tiền xử lý

Input:
- tập dữ liệu đã tiền xử lý
- nhãn (labels)

Output:
- Không gian đặc trưng tối ưu giúp phân tách rõ ràng giữa các lớp

==> Tập mẫu trong không gian mới có tính khả tắc

Công đoạn 3: Tìm luật phân lớp

CĐ 3 mới là tìm luật phân lớp.

Ý nghĩa khoa học:

Ví dụ cũng chưa cụ thể.

Phân lớp, trong bài toán thực tế vì có liên quan đến tiên đoán (prediction).

Ví dụ: Ảnh u xương –> u đại bào, u tiểu bào, lành tính hay ác tính → dẫn đến bài toán phân lớp

3.1 Công trình nghiên cứu liên quan

LDA: Linear Discriminant Analysis

k-NN

SVM

Decision Tree

Tham khảo