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

1.1 Khái niệm

1.2 Phương pháp

1.3 Nhận xét

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

1.4 Thực nghiệm

- Tương quan truyền thống là gì?
- Cần thực hiện so sánh phương pháp truyền thống và phương pháp mới, song song
- Cần có điều kiện bài toán và thông tin kiểm nghiệm rõ ràng hơn

Tương quan chính tắc Mẫu

Kiểm định mô hình

2 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 (2)

Đây là đồ án của nhóm Vstatic.

3.1 Ý nghĩa ứng dụng

- Phân loại đối tượng
- Phân loại hình ảnh
- Phân loại khách hàng
- Nhận diện cảm xúc

3.2 Mục tiêu

Xây dựng một hàm phân biệt để gán một quan sát x có d đặc trưng, vào 1 trong k lớp, sao cho tối ưu được độ phân biệt giữa các lớp.

Xét trường hợp 2 lớp và đa lớp

3.3 Công trình liên quan

3.3.1 Phân biệt Triển vọng cực đại - Maximum Likelihood Discriminant:

Gán x vào hàm có mật độ xác suất lớn nhất, tức là lớp có khả năng sinh ra x cao nhất theo mô hình giả định:

$$C* = \arg\max(C_k \in C)p(x \mid C_k) \tag{1}$$

3.3.2 Phân loại bằng định lý Bayes

Áp dụng định lý Bayes để tính xác suất hậu nghiệm, và chọn xác suất hậu nghiệm lớn nhất. Xác suất tiên nghiệm là xác suất ta có được trước khi thực hiện quan sát x và thực hiện phân lớp x vào tập cụ thể w. (công thức xác suất hậu nghiệm)

3.3.3 Phân biệt tuyến tính LDA

Phân tích Thành phần Chính (Principal Component Analysis - PCA), vốn chỉ tập trung vào việc giữ lại phương sai tổng thể của dữ liệu mà không xét đến thông tin về nhãn lớp. Tuy nhiên, việc giữa lại thông tin nhiều nhất không đồng nghĩa với việc giúp phân loại tốt nhất. Để giải quyết bài toán này, LDA giúp tăng cường khả năng phân biệt giữa các nhóm dữ liệu, từ đó cải thiện hiệu suất của các thuật toán phân loại.

3.4 Thực nghiệm

3.4.1 Bài toán 2 lớp

Xác định xem một tờ tiền là giả hay thật. Đầu vào: hình ảnh từ máy quét tiền Đầu ra: Giá trị boolean: Đây là tiền thật hay tiền giả?

Linear Discriminant Analysis Phương pháp:

1. Tính toán trung tâm lớp
2. Tính ma trận hiệp phương sai trong lớp (within-class covariance)
3. Tính ma trận hiệp phương sai giữa các lớp (between-class covariance)
4. Tính vector chiếu tối ưu (tìm vector riêng tương ứng với trị riêng lớn nhất)
5. Chiếu dữ liệu sang không gian mới (bằng vector w)
6. Phân loại bằng Nearest Centroid Classifier: Tức là gán nhãn

⇒ Đạt được độ chính xác 96.73% trên tập dữ liệu kiểm tra

3.4.2 Bài toán đa lớp

Tên bài toán: Khảo sát dữ liệu cho NASA

Thách thức:

- Chiều dữ liệu rất cao: rất nhiều đặc trưng gắn với mỗi ngôi sao
- Chồng chéo giữa các lớp
- Khối lượng dữ liệu khổng lồ, không thể phân loại bằng tay

Bài toán phân lớp ngôi sao:

- Dựa vào đặc trưng vật lý: nhiệt độ, độ sáng, bán kính, độ lớn tuyệt đối, màu sắc, phân loại quang phổ
- Mục tiêu phân loại: Giá trị phân loại ứng với các loại ngôi sao: Red Dwarf, Brown Dwarf, White, Main Sequence, Super Giants, Hyper Giants
- Đầu vào: Các tham số về ngôi sao
- Đầu ra: Giá trị phân loại ngôi sao

Giải bài toán:

- Within-class Variance: thể hiện độ phân tán của các điểm dữ liệu trong 1 class. Kết quả là các vector riêng cùng phương trình giá trị riêng.

- Lựa chọn và sắp xếp các vector riêng: Nhằm giữ lại nhiều thông tin phân biệt nhất giữa các lớp
- Chiếu dữ liệu vào không gian mới.
- Phân loại bằng Nearest Centroid Classifier

Nhận xét dữ liệu trước khi thực hiện bài toán:
- Có sự chồng chéo đáng kể
- Đặc trưng 0 không phải là một đặc trưng tốt để phân biệt rõ ràng tất cả các lớp

Sau khi thực hiện:
- Phân tách đáng kể
- Vẫn còn tồn tại sự chồng chéo

3.5 Nhận xét

4 CLUSTERING

4.1 Giới thiệu

Clustering là kỹ thuật quan trọng trong Phân tích thống kê đa biến, giúp nhóm các quan sát dựa trên sự tương đồng mà không cần gán nhãn trước.

Mục tiêu:
- Tìm cấu trúc ẩn trong dữ liệu đa biến
- Hỗ trợ khám phá mô hình
- Phân tích dữ liệu thăm dò
- Giảm chiều dữ liệu

Thách thức:
- Xác định số cụm tối ưu
- Xử lý dữ liệu nhiều chiều

4.2 Ý nghĩa khoa học

- Hiểu bản chất phức tạp của các mối quan hệ đa biến
- Đánh giá mức độ đa chiều của dữ liệu
- Xác định ngoại lai
- Đề xuất các giả thuyết thú vị về mối quan hệ giữa các đối tượng

4.3 Ý nghĩa ứng dụng

- Marketing: Chọn các thị trường thử nghiệm; Phân loại và cơ cấu công ty theo tổ chức
- Tâm lý học: Tìm ra các loại tính cách trên cơ sở các bảng câu hỏi
- Khảo cổ học: Phân loại các đồ vật nghệ thuật trong các thời kỳ khác nhau

Các ngành khoa học khác: y học, sinh học, xã hội học, ngôn ngữ học.

Trong mỗi trường hợp, một mẫu các đối tượng không đồng nhất được phân tích với mục đích xác định các nhóm con đồng nhất

4.4 Phát biểu bài toán

Đầu vào:
- Một tập hợp gồm N điểm dữ liệu không gán nhãn

$$X = \{x_i \mid x_i \in \mathbb{R}^p, 0 \leq i \leq N\} \tag{2}$$

- Hàm đo độ tương đồng giữa 2 điểm dữ liệu

$$s(x_i, x_j) \tag{3}$$

- Hàm đo khoảng cách giữa 2 điểm dữ liệu

$$d(x_i, x_j) \tag{4}$$

Đầu ra: Một tập hợp gồm M nhãn tương ứng với N điểm dữ liệu

$$Y = \{y_i \mid y_i \in \{0, 1, ..., M-1\}, 0 \leq i \leq N\} \tag{5}$$

4.5 Phương pháp

Framework:

1. Nhập dữ liệu đầu vào (Input)
2. Tiền xử lý dữ liệu: Normalization, Standardization, Đơn giản hoá dữ liệu
3. Tính toán độ tương đồng (Similarity Measure): l1-norm (Manhattan), l2-norm (Euclidean), Jaccard, X^2 matrice
4. Gom nhóm (Clustering)
5. Trực quan hoá dữ liệu: Dendrogram, Chernoff, Star
6. Trả kết quả: `.csv` file cùng đồ thị biểu diễn

4.5.1 Similarity Measure

Đo độ tương đồng dựa vào khoảng cách (distance) hoặc mức độ tương quan Coefficient giữa các item/variable.

Một số yếu tố cần xem xét khi lựa chọn một loại thang đo độ tương đồng:

- Biến rời rạc, liên tục, hay nhị phân?
- Thang đo của feature
- Kiến thức chuyên môn của lĩnh vực đang thực hiện
- Có nhiều cách để đo lường sự tương đồng giữa các cặp đối tượng. Hầu hết các nhà nghiên cứu sử dụng tiêu chí khoảng cách (distance measures) hoặc hệ số tương đồng (correlation coefficient)
- Đối với kiểu biến nhị phân, có thể tính distance measure theo l1-norm, l2-norm. Tính hệ số tương quan coefficient theo Simple Matching, Jaccard, Russell & Rao, Dice,…
- Đối với biến liên tục: Tính khoảng cách bằng lr-norm (Minkowski distance), Mahalanobis distance, Cosine distance. Tính hệ số tương quan theo Pearson.

4.5.2 Thuật toán Gom nhóm

1. **Hierarchical**

1.1. Agglomerative - Gom cụm: Bắt đầu ở phân vùng nhỏ nhất, mỗi mẫu là 1 cụm, sau đó đi gom dần dữ liệu lại thành cụm.

- Single Linkage Distance: Khoảng cách giữa hai cụm được tính là khoảng cách ngắn nhất giữa hai điểm ở hai cụm.
- Complete Linkage Distance: Khoảng cách giữa hai cụm được tính là khoảng cách xa nhất giữa hai điểm ở hai cụm.
- Average Linkage Distance: Khoảng cách giữa hai cụm được tính là trung bình các khoảng cách giữa tất cả các điểm ở hai cụm.
- Ward method: Tìm cách giảm thiểu tổng bình phương sai số (Total Within-Cluster Variance) khi gom cụm.

Thuật toán Gom cụm phân cấp:

1. Xây dựng phân cụm nhỏ nhất, mỗi đối tượng là một cụm riêng lẻ
2. Tính toán ma trận khoảng cách D

3. Tìm 2 cụm có khoảng cách nhỏ nhất
4. Gom 2 cụm đó thành 1 cụm mới
5. Cập nhật lại ma trận khoảng cách D sau khi hợp nhất cụm

Lặp lại các bước 3 - 5 cho đến khi tất cả các cụm được hợp nhất thành 1 cụm X duy nhất.

Kết luận:

- Không xem xét về sai số và hàm lỗi nên phương pháp này nhạy cảm với các giá trị ngoại lai và nhiễu
- Khi đối tượng đã được phân cụm thì không thể sửa lại, vì vậy cần cẩn trọng sai sót từ giai đoạn đầu.
- Những giá trị bằng nhau trong ma trận khoảng cách hoặc độ tương đồng có thể tạo ra nhiều lời giải khác nhau cho bài toán gom cụm phân cấp.

1.2. Splitting - Phân cụm: Cho tất cả mẫu thành 1 nhóm, sau đó tách ra dần dần thành các cụm nhỏ hơn.

2. **Partitioning** (Non-hierarchical): Khởi đầu với k cụm bằng cách gom ngẫu nhiên dữ liệu ở gần tâm cụm lại, chấp nhận rủi ro gán nhầm; sau đó hoán đổi các phần tử giữa các cụm theo 1 tiêu chí nhất định (ví dụ: Khoảng cách so với tâm cụm mới).

Kỹ thuật này được thiết kế để nhóm các đối tượng thay vì biến số vào 1 tập hợp gồm K cụm. Số cụm K có thể được xác định trước hoặc được xác định trong quá trình phân cụm.

Do không cần xây dựng ma trận khoảng cách (độ tương đồng), và dữ liệu cơ bản không cần phải lưu trữ trong quá trình thực thi, các phương pháp phân cụm phi phân cấp có thể áp dụng trên các tập dữ liệu lớn hơn nhiều so với các phương pháp phân cụm phân cấp.

Các phương pháp phân cụm phi phân cấp thường bắt đầu bằng:

- Phân nhóm ban đầu các đối tượng vào các cụm
- Chọn một tập hợp các điểm trung tâm ban đầu, chúng sẽ hình thành hạt nhân của các cụm.

Thuật toán K-Means Clustering:

1. Phân chia dữ liệu thành K cụm ban đầu
2. Duyệt qua danh sách các đối tượng, gán mỗi đối tượng vào cụm có trung tâm gần nhất.
   - Sử dụng khoảng cách Euclidean, với dữ liệu chuẩn hoá hoặc không chuẩn hoá.
   - Cập nhật lại trung tâm cụm khi 1 điểm dữ liệu mới được thêm vào và khi 1 điểm rời khỏi cụm trước đó.
3. Lặp lại bước 2 cho đến khi không còn sự thay đổi nào trong việc gán cụm.

Sự khác biệt giữa Hierarchical và Partitioning:

- Hierarchical: Nếu dữ liệu đã gán cụm rồi thì không thể thay đổi được nữa.
- Partitioning: Có thể thay đổi nhãn của cụm dữ liệu, tuỳ vào quá trình chạy

4.6 Nhận xét

- Chọn chiến lược linkage, cần thêm hình ảnh minh hoạ, cái nào mạnh, cái nào yếu
- Trình bày chưa rõ ràng về mục đích thực hiện phân cụm
- Chưa biết được khi nào kết thúc gom cụm

5 PHÂN TÍCH PHÂN BIỆT (Linear Discriminant Analysis)

5.1 Giới thiệu

LDA được giới thiệu lần đầu bởi R.A.Fisher vào năm 1936 và được mở rộng bởi C.R.Rao cho bài toán phân lớp đa nhóm vào những năm 1940. LDA được ứng dụng rộng rãi trong các bài toán phân loại dữ liệu đa biến với giả định toán học thuần tuý.

Ý nghĩa khoa học:

Ý nghĩa ứng dụng:

- Y tế và chăm sóc sức khoẻ: Chẩn đoán bệnh, phân tích hình ảnh y tế
- Tài chính - Ngân hàng: Đánh giá rủi ro, phát hiện gian lận
- Marketing và bán hàng: Phân khúc khách hàng, dự đoán xu hướng
- An ninh mạng: Phát hiện xâm nhập
- Xử lý ngôn ngữ: Phân loại văn bản
- Khoa học xã hội và Nghiên cứu thị trường: Phân tích khảo sát, dự đoán hành vi

5.2 Phát biểu bài toán

Đầu vào:

- Tập dữ liệu đã tiền xử lý
- Nhãn lớp (Labels)

Đầu ra:

- Các quy tắc hoặc hàm phân biệt cho

5.3 Phương pháp

Mục tiêu của LDA là tìm ra một tổ hợp tuyến tính của các biến ban đầu sao cho khi chiếu dữ liệu sang 1 không gian có chiều thấp hơn, các lớp được tách biệt rõ ràng hơn. Điều này được thể hiện thông qua việc tối đa hoá tỷ số giữa độ phân tán giữa các lớp và độ phân tán trong lớp.

Hàm mục tiêu được diễn đạt như sau:

Giả định 1: Giả sử tồn tại K lớp, và dữ liệu của mỗi lớp C_j được mô hình hoá bởi phân phối chuẩn đa biến, nghĩa là:

$$x \mid C_j \sim N(\mu_j, \Sigma_j), j = 1, ..., K \tag{6}$$

hay hàm mật độ của lớp C_j được cho bởi:

$$f_{j(X)} = \frac{1}{(2\pi)^{\frac{P}{2}} \ |\Sigma_j|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\right\} \tag{7}$$

trong đó:

- x là vector quan sát có p thành phần
- μ_j là vector trung bình của lớp C_j
- Σ_j là ma trận hiệp phương sai của lớp C_j.

Các đại lượng đo độ phân tán: Giả sử dữ liệu được chia thành 2 lớp C_1 và C_2 với số mẫu N_1, N_2 tương ứng, mỗi quan sát $x \in \mathbb{R}^p$. Khi chiếu dữ liệu lên một hướng w, trung bình sau chiếu của mỗi lớp được tính như sau:

$$\begin{aligned} m_1 &= \frac{1}{N_1} \sum_{x \in C_1} w^T x \\ m_2 &= \frac{1}{N_2} \sum_{x \in C_2} w^T x \end{aligned} \tag{8}$$

Tương tự, độ phân tán (sai số) của mỗi lớp sau chiếu được định nghĩa là:

$$\begin{aligned} s_1^2 &= \sum_{x \in C_1} (w^T x - m_1)^2 \\ s_2^2 &= \sum_{x \in C_2} (w^T x - m_2)^2 \end{aligned} \tag{9}$$

Between-class covariance matrix: Khoảng cách giữa 2 điểm sau chiếu được đo bằng hiệu của trung bình các điểm sau chiếu, cụ thể:

$$(m_1 - m_2)^2 = w^T(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w \tag{10}$$

Ta định nghĩa between-class covariance matrix:

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \tag{11}$$

Như vậy ta có:

$$(m_1 - m_2)^2 = w^T S_B w \tag{12}$$

Hàm mục tiêu của LDA: *Mục tiêu:* Mục tiêu của LDA là tìm một hướng w sao cho tỷ số giữa độ phân tán giữa lớp và độ phân tán trong lớp được tối đa hoá. Ban đầu, ta xây dựng hàm mục tiêu dựa trên đại lượng đo sự phân tán:

$$J(w) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2} \tag{13}$$

trong đó:

- m_1, m_2 là trung bình các điểm sau chiếu của từng lớp
- s_1^2, s_2^2 là tổng bình phương sai lệch của các điểm so với trung bình trong từng lớp

Hàm mục tiêu biểu diễn qua S_B và S_W:

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \tag{14}$$

Các bước thực hiện LDA cho 2 lớp:

- Giải bài toán eigenvalue:

$$S_W^-1 S_B w = \lambda w \tag{15}$$

- Chọn nghiệm tối ưu: Sắp xếp các trị riêng (eigenvalue) theo thứ tự giảm dần rồi chọn vector riêng (eigenvector) ứng với trị riêng lớn nhất làm nghiệm tối ưu w^*
- Chiếu dữ liệu: Với mỗi x, tính giá trị chiếu:

$$y = (w^*)^T x \tag{16}$$

- Phân lớp: Xác định ngưỡng (thường là trung điểm của $(w^*)^T\mu_1$ và $(w^*)^T\mu_2$) để gán nhãn cho các điểm dữ liệu

5.4 Nhận xét

- Không có khái niệm phương sai giữa các lớp
- Nhiều chi tiết toán không rõ?

6 SUY DẪN KẾT QUẢ LIÊN QUAN ĐẾN QUẦN THỂ DỰA TRÊN THÔNG TIN MẪU

6.1 Giới thiệu

Thống kê suy luận là một phương pháp thống kê được sử dụng để đưa ra dự đoán hoặc suy luận về một quần thể dựa trên mẫu dữ liệu.

Vai trò: Diễn giải dữ liệu, tiến đến đưa ra kết luận thực tiễn, hữu ích, tiếp tục tiến đến sử dụng cho các quyết định trong tương lai.

Ý nghĩa khoa học:

- Khả năng tổng quát hoá (generalization) và dự đoán (prediction)
- Tăng hiệu quả và tính khả thi trong việc thu thập dữ liệu trong nghiên cứu

Ý nghĩa ứng dụng:

- Kiểm soát chất lượng và quản lý rủi ro: Kiểm tra một nhóm mẫu từ mỗi lô sản xuất, nhằm giám sát chất lượng sản phẩm
- Phục vụ y khoa và dịch vụ y tế công cộng: Thu thập dữ liệu mẫu để theo dõi mức độ lây lan của bệnh trong một vùng, giúp lên kế hoạch cho các biện pháp can thiệp và phân bố nguồn lực hợp lý.
- Nghiên cứu về môi trường và xã hội: Thu thập dữ liệu về chất lượng không khí từ một vài trạm giám sát để đánh giá mức độ ô nhiễm trong 1 thành phố, qua đó đề xuất các chính sách cải thiện chất lượng không khí.

6.2 Phát biểu bài toán

Đầu vào:

- Một tập mẫu ngẫu nhiên $(X_1, X_2, X_3, ..., X_n)$ được lấy từ một quần thể.
- Giả thuyết không (H_0): Vector trung bình thực sử của quần thể (μ) bằng với vector trung bình kiểm định của quần thể (μ_0).

$$H_0 : \mu = \mu_0 \tag{17}$$

- Giả thuyết thay thế (H_1): Vector trung bình thực sự của quần thể (μ) khác với vector trung bình của quần thể mà ta kiểm định (μ_0)

$$H_1 : \mu \neq \mu_0 \tag{18}$$

- Mức ý nghĩa α: Xác suất cho phép mắc lỗi loại I (tức lỗi bác bỏ H_0 mặc dù H_0 đúng).

Đầu ra: Bác bỏ hoặc không bác bỏ giả thuyết H_0, dựa vào kiểm định thống kê.

6.3 Phương pháp

6.3.1 Một số phân phối cần thiết

Phân phối t (hay còn gọi là phân phối Student's t) là một loại phân phối xác suất được sử dụng khi:

- Kích thước mẫu (n) nhỏ.
- Phương sai tổng thể (σ^2) chưa biết.

Nó thường được dùng để kiểm định giả thuyết hoặc xây dựng khoảng tin cậy cho giá trị trung bình của một quần thể.

Đặc điểm chính:

- Đối xứng quanh 0.
- Có hình chuông.
- Phụ thuộc vào bậc tự do $n - 1$ (degrees of freedom - df), với n là kích thước mẫu.
- Khi bậc tự do tăng lên (kích thước mẫu lớn), phân phối t tiến gần đến phân phối chuẩn.
- Mức ý nghĩa α là một ngưỡng mà ta đặt ra để quyết định xem có bác bỏ giả thuyết ban đầu H_0 hay không.
- Sau khi xác định bậc tự do và mức ý nghĩa, ta có thể sử dụng bảng phân phối T để tra giá trị tương ứng.

Phân phối Fisher là một loại phân phối xác suất được đặt tên theo nhà thống kê Ronald Fisher. Sử dụng phân phối F khi:

- So sánh hai phương sai.
- Kiểm định giả thuyết liên quan đến nhiều biến.

Để tra bảng phân phối Fisher, chúng ta cần biết hai giá trị:

- Bậc tự do của tử số (numerator degrees of freedom) được tính bằng công thức: $p-1$, với p là số nhóm dữ liệu mà chúng ta đang so sánh.
- Bậc tự do của mẫu số (denominator degrees of freedom) được tính bằng công thức: $n-p$, với n là tổng số quan sát và p là số nhóm dữ liệu.

6.3.2 Mẫu một biến

Ta có các mẫu ngẫu nhiên từ một quần thể chuẩn. Để kiểm chứng xem giả thuyết có hợp lý hay không, ta sử dụng thống kê kiểm định t.

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \tag{19}$$

Trong đó:
- μ_0 là trung bình của quần thể theo giả thuyết.
- $\bar{X} = \frac{1}{N}\sum_{j=1}^{n} X_j$ là trung bình mẫu.
- $s^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2$ là phương sai của mẫu.
- n là số phần tử có trong mẫu.

Kết quả này tuân theo phân phối t với $n-1$ bậc tự do.

Ta bác bỏ H_0 khi $|t|$ lớn hay bình phương của nó lớn, tức là bác bỏ H_0 để chấp nhận H_1 ở mức ý nghĩa α nếu:

$$n(\bar{x} - \mu_0)(s^2)^{-1}(\bar{x} - \mu_0) > t^2_{n-1}\left(\frac{\alpha}{2}\right) \tag{20}$$

Ta không bác bỏ H_0 khi:

$$\left|\frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}\right| < t_{n-1}\left(\frac{\alpha}{2}\right) \tag{21}$$

Tổng kết: Với giá trị kiểm định thống kê t có phân phối t với $n-1$ bậc tự do:
- Tính $t_{n-1}\left(\frac{\alpha}{2}\right)$ bằng cách tra bảng phân phối t với $n-1$ bậc tự do.
- Nếu $|t| > t_{n-1}\left(\frac{\alpha}{2}\right)$ thì ta bác bỏ H_0.
- Nếu $|t| \leq t_{n-1}\left(\frac{\alpha}{2}\right)$ thì ta không bác bỏ H_0.

6.3.3 Mẫu nhiều biến

Với trường hợp mẫu có nhiều biến, giả sử số biến là p.

$$T^2 = n\left(\bar{X} - \mu_0\right)' S^{-1}\left(\bar{X} - \mu_0\right) \tag{22}$$

Dấu chấm phẩy là thay cho chữ T, tức chuyển vị ma trận.

Trong đó:
- μ_0 là trung bình của quần thể theo giả thuyết. $(p \times 1)$
- $\bar{X} = \frac{1}{N}\sum_{j=1}^{n} X_j$ là trung bình nhóm mẫu được chọn. $(p \times 1)$
- $s^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \bar{X}\right)\left(X_j - \bar{X}\right)'$ là ước lượng hiệp phương sai của tập mẫu được chọn $(p \times p)$. $\left(X_j - \bar{X}\right)$ có $p \times 1$ chiều.
- n là số phần tử có trong mẫu.

6.3.4 Các bước giải bài toán kiểm định

6.4 Nhận xét

- Ví dụ xác thực, code đạt chuẩn
- Thiếu CDF

- Cần giải thích phân phối student

7 TÁI LẤY MẪU (RESAMPLING)

Suy luận thống kê

- Lấy ví dụ về tái lấy mẫu
- Chưa rõ về việc tái lấy mẫu: mẫu xuất hiện mới sẽ từ đâu ra? mẫu có rồi thì sao?
- Ví dụ tái lấy mẫu nhằm mục đích gì?
- Phát biểu đồ án không rõ ngay từ đầu?

trung bình của nhóm có dùng thuốc mới và nhóm không dùng thuốc mới

8 CÁC KHÁI NIỆM CƠ BẢN VỀ PHÂN TÍCH THỐNG KÊ DỮ LIỆU NHIỀU BIẾN

8.1 Phát biểu bài toán

Tên bài toán: Ứng dụng kiểm định vào View Synthesis

Đầu vào: Tập dữ liệu ảnh chụp từ nhiều vị trí / góc nhìn

Đầu ra: Khẳng định tập dữ liệu ảnh có đủ "dày" không

8.2 Phương pháp 1: Hình học mẫu

Cách biểu diễn hình học của dữ liệu nhiều biến trong không gian Euclid, giúp trực quan hóa mối quan hệ giữa các biến và các quan sát.

Vai trò:

- Trực quan hóa dữ liệu phức tạp
- Đo lường mối quan hệ giữa các biến và quan sát
- Hỗ trợ phương pháp phân tích đa biến .

8.3 Phương pháp 2: Phân phối chuẩn đa biến

Phân phối chuẩn đa biến với trung bình μ và ma trận hiệp phương sai có hàm mật độ xác suất là:

$$f(x) = \frac{1}{\sqrt{|2\pi\ |\Sigma\|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{23}$$

Trong đó:

- x là vector điểm dữ liệu cần tính mật độ xác suất
- μ là vector trung bình
- Σ là ma trận hiệp phương sai
- $(x-\mu)^T\Sigma^{-1}(x-\mu)$ là khoảng cách Mahalanobis giữa x và μ.

Công thức này mở rộng hàm mật độ của phân phối chuẩn một chiều sang nhiều chiều, với khoảng cách Mahalanobis thay thế cho khoảng cách Euclid để phản ánh mối quan hệ giữa các biến.

Phép biến đổi Mahalanobis: Chuẩn hoá phân phối chuẩn đa biến thành phân phối chuẩn độc lập

Tính chất:

- Bảo toàn phân phối chuẩn dưới phép biến đổi tuyến tính
- Phân phối chuẩn nhiều chiều có điều kiện

8.4 Phương pháp 3: Hàm hợp lý và Thống kê kiểm định

8.4.1 Hàm hợp lý

Xét một mẫu độc lập, phân phối giống nhau $\{x_i\}_{i=1}^n$. Mỗi giá trị x_i được giả định tuân theo một phân phối xác suất với hàm mật độ xác suất (PDF) $f(x;0)$ với θ là tham số chưa biết cần ước lượng.

Hàm hợp lý được định nghĩa là xác suất (hoặc mật độ) của toàn bộ mẫu dữ liệu đã quan sát, được xem như là một hàm của θ:

$$L(X;\theta) = \Pi_{i=1}^n f(x_i;\theta) \tag{24}$$

Hàm này phản ánh mức độ "hợp lý" của tham số θ trong việc tạo ra dữ liệu đã quan sát.

Ước lượng hợp lý cực đại: Là tìm giá trị tham số θ thoả mãn $L(X;\theta)$ đạt giá trị lớn nhất.

$$\hat{\theta} = \arg\max_{\theta} L(X;\theta) \tag{25}$$

Để đơn giản hóa việc tính toán, đặc biệt khi làm việc với tích của nhiều hàm, ta thường sử dụng hàm hợp lý logarit:

$$l(X;\theta) = \log L(X;\theta) = \sum_{i=1}^{N} \log f(x_i;\theta) \tag{26}$$

8.4.2 Kiểm định giả thuyết

Trong bài toán kiểm định giả thuyết, ta có:

- Giả thuyết không (H_0) là giả thuyết cần kiểm định. Tập hợp các giá trị thuộc H_0 được kí hiệu là Ω_0.
- Giả thuyết đối (H_1) là giả thuyết đối lập với giả thuyết không. Tập hợp các giá trị thuộc H_1 được kí hiệu là Ω_1.

Hai loại sai lầm trong kiểm định giả thuyết:

- Sai lầm loại I xảy ra khi bác bỏ H_0, mặc dù nó đúng.
- Sai lầm loại II xảy ra khi bác bỏ H_1, mặc dù nó sai.

Xác suất xảy ra sai lầm loại I được gọi là mức ý nghĩa, ký hiệu là α.

Miền bác bỏ: Tập hợp các giá trị của thống kê kiểm định mà khi rơi vào đó, ta bác bỏ giả thuyết không.

Kiểm định tỷ số hợp lý: Kiểm định một tham số θ trong mô hình phân phối đa biến chuẩn $N_{p(\theta,I)}$:

- H_0: $\theta = \theta_0$ (giá trị cụ thể đã biết)
- H_1: $\theta \neq \theta_0$

Để đánh giá mức độ phù hợp của H_0, ta so sánh giá trị hợp lý cực đại khi H_0 đúng và khi H_1 đúng. Với mẫu ngẫu nhiên X, ta có tỷ số hợp lý:

$$\lambda(X) = \frac{L_0^\star}{L_1^\star} \tag{27}$$

Trong đó:

- $L_0^\star$ là giá trị cực đại của hàm hợp lý khi ủng hộ H_0.
- $L_1^\star$ là giá trị cực đại của hàm hợp lý khi ủng hộ H_1.

Kiểm định tỷ số hợp lý với mức ý nghĩa α để kiểm định H_0 so với H_1 có miền bác bỏ R:

$$R = \{X : \lambda(X) < c\} \tag{28}$$

với c là một ngưỡng xác định, sao cho xác suất bác bỏ H_0 không vượt quá mức ý nghĩa.

8.5 Nhận xét

- Nhiều chi tiết chưa rõ ràng
- Giải thích tại sao là "hàm hợp lý"? Từ "hợp lý" từ đâu ra? Likelihood? Phải dịch là hàm triển vọng. Ước lượng triển vọng cực đại
- Tại sao phải ước lượng triển vọng cực đại?

Giải đáp: Ta cần ước lượng các tham số chưa biết sao cho hàm hợp lý - tức là tích của các hàm mật độ xác suất (hoặc hàm khối xác suất) của n mẫu quan sát - đạt giá trị lớn nhất. Khi đó, mô hình với các tham số này giải thích dữ liệu đã quan sát là hợp lý nhất.

- Trong tính toán hàm vật lý, vì hai hàm đồng biến, không thể lấy log vì "để bỏ đi e mũ".

Kiểm định giả thuyết:

- Kiểm định tỷ số hợp lý: l_0 ở đâu ra? Trang sau có liên quan gì đến trang trước
- Kiểm định giả thuyết khi biết ma trận hiệp phương sai: ta cần biết gì? μ

Tham khảo