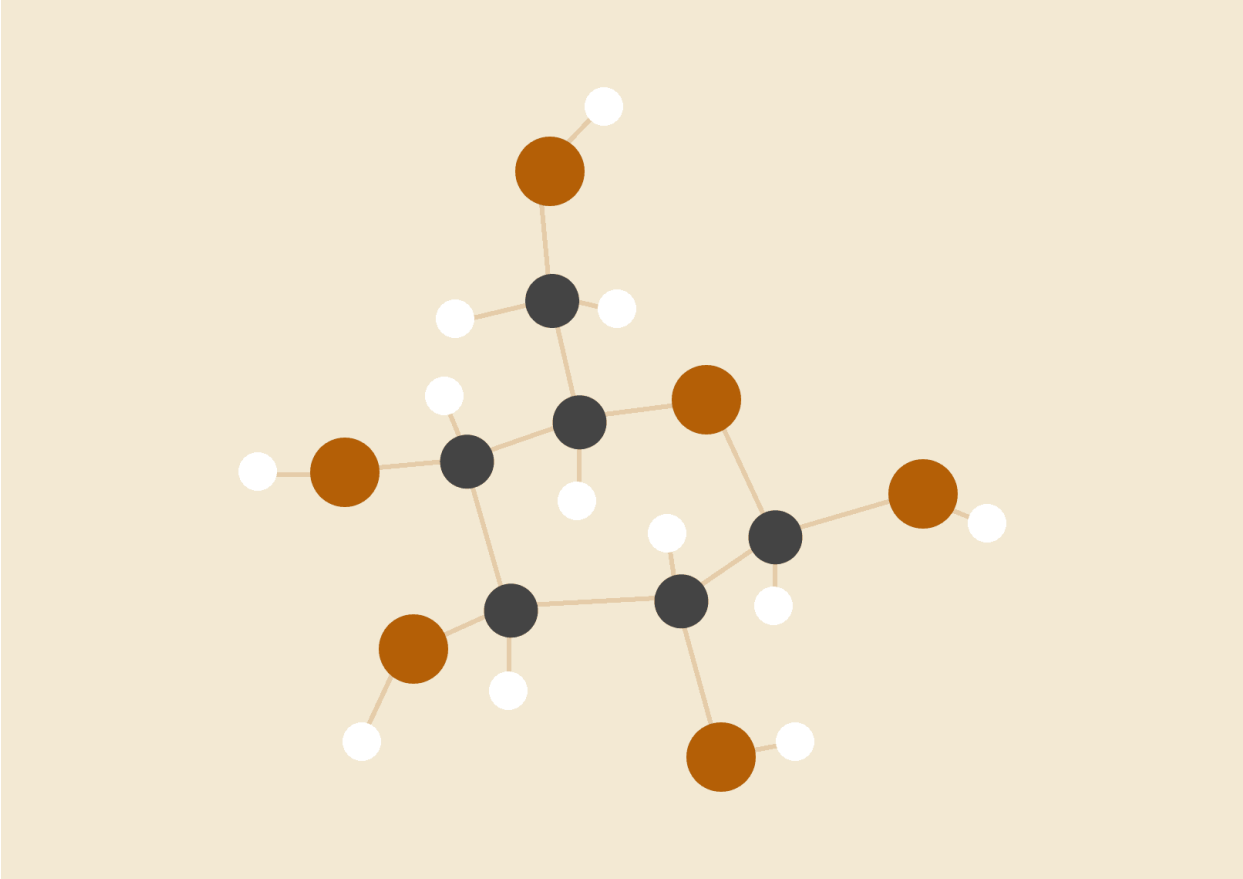
Nhóm Nhà



Team 404 - NoName

19127614 - Nguyễn Anh Tuấn
19127517- Hồ Thiên Phước
19127165 - Võ Gia Huy

09/01/2022
K19-HCMUS-CLC-PPTKDLNB

MỤC LỤC

TUẦN 1

Bài 1

1) Phân biệt xác suất và thống kê

Xác suất là thước đo khả năng xảy ra một sự kiện. Vì xác suất là một thước đo được lượng hóa nên nó phải được phát triển với nền tảng toán học. Cụ thể, cấu trúc toán học về xác suất này được gọi là lý thuyết xác suất. Thống kê là bộ môn thu thập, tổ chức, phân tích, giải thích và trình bày dữ liệu

Xác suất: thông tin,quy luật của quần thể suy ra các thông tin cơ bản từng trường hợp.
Populations -> Sample
Thống kê: Từ thông tin cơ bản từng trường hợp suy ra các quy luật cho quần thể.
Sample -> Populations

2) Phân biệt thống kê với học máy

Khác biệt chính giữa học máy và thống kê là mục đích của chúng:

- Các mô hình học máy được thiết kế để đưa ra những dự đoán chính xác nhất có thể.
- Các mô hình thống kê được thiết kế để suy luận về mối quan hệ giữa các biến.

3) Liệt kê yếu tố thống kê trong dữ liệu thị giác

-Hàm phân bố xác suất trong histogram equalization

4) Cho ví dụ về thống kê

Dịch bệnh Covid - 19:

- Thống kê tình hình số ca nhiễm Covid - 19 theo vùng:
    - Số người mất, khỏi bệnh, tái bệnh, … => Phân vùng số người mắc Covid (vùng xanh, đỏ, vàng,...)
- Thống kê về Vaccine:
    - Số người tiêm (từng loại Vaccine),số tuổi, giới tính , triệu chứng từng người,... => Triệu chứng sau tiêm tùy thuộc tính chất (từng loại Vaccine) => Phân bổ Vaccines

5) Ví dụ ứng dụng thống kê: Sample -> Population

Ví dụ về nhà ở sinh viên

Một cuộc khảo sát ước tính tỷ lệ tất cả sinh viên đại học sống ở nhà trong học kỳ hiện tại. Trong số 3.838 sinh viên đại học theo học tại trường, một mẫu ngẫu nhiên gồm 100 người đã được khảo sát.
Quần thể: Tất cả 3.838 sinh viên đại học
Mẫu: 100 sinh viên đại học được khảo sát
=>Chúng ta có thể sử dụng dữ liệu thu thập được từ mẫu 100 sinh viên để suy luận về quần thể 3,838 sinh viên.

Bài 2

1. **Sample mean có ý nghĩa thống kê không, đã đủ để thể hiện bản chất dữ liệu chưa**

Ý nghĩa: Chưa đủ để thể hiện được bản chất của dữ liệu. Thứ nhất vì là trung bình mẫu chứ không phải trung bình từng cá thể trong quần thể. Thứ hai vì cách lấy, chọn data tác động rất nhiều lên Sample mean, Do đó nếu chỉ dựa trên mean, bản chất dữ liệu sẽ bị tác động dựa trên cách chọn data, khiến cho bản chất không còn chính xác và dễ bị thay đổi.

2. **Tại sao dùng phép nhân để biết quan hệ giữa 2 biến (Tại sao xài phép nhân trong Sample covariance)**

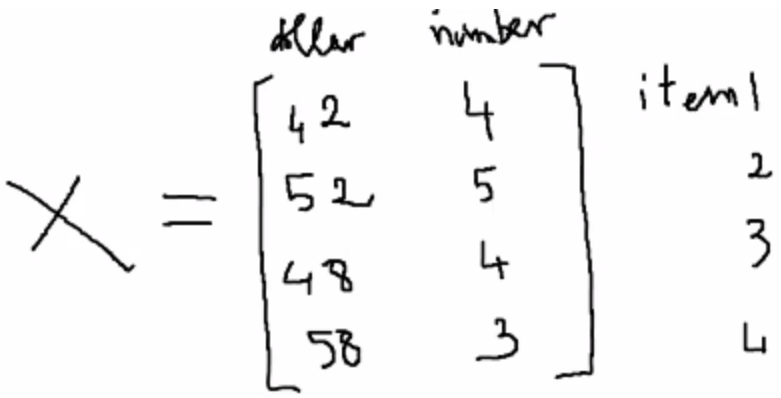
3. **Chứng minh Rik nằm trong khoảng nào?(Sample correlation coefficient)**

Áp dụng bất đẳng thức bunhiacopxki ta có

$$\sum_{j=1}^{n}(x_{ji}-\overline{x_i})(x_{jk}-\overline{x_k}) \quad <= \quad \sqrt{\sum_{j=1}^{n}(x_{ji}-\overline{x_i})^2}\sqrt{\sum_{j=1}^{n}(x_{jk}-\overline{x_k})^2}$$

Nên ta có |Rik|<=1 → -1<=Rik<=1

4. **Bài tập thống kê:**
   V1 (dollar sales) - Giá tiền
   V2 (numbers of books) - Số sách



a)Tính Sample mean

$$\overline{x_k} = \frac{1}{n}\sum_{j=1}^{n} x_{jk} \quad k = 1, 2, \ldots p$$

Với n = 4, p = 2

=> $\overline{x_1} = \frac{1}{4}\sum_{j=1}^{4} x_{j1} = \frac{1}{4}(42 + 52 + 48 + 58) = 50$

=> Số tiền bán trung bình (doanh thu trung bình)

=>$\overline{x_2} = \frac{1}{4}\sum_{j=1}^{4} x_{j2} = \frac{1}{4}(4 + 5 + 4 + 3) = 4$

=> Số lượng sách bán trung bình

$$\overline{x} = \begin{bmatrix} 50 \\ 4 \end{bmatrix}$$

b)Tính Sample Variance, Standard Variance

$$S_{kk} = \frac{1}{n}\sum_{j=1}^{n} (x_{jk} - \overline{x_k})^2 \quad k = 1, 2, \ldots p$$

Với n = 4, p = 2

$$S_{11} = \frac{1}{4}\sum_{j=1}^{4}(x_{j1} - \overline{x_1})^2$$

$$= \frac{1}{4}[(42 - 50)^2 + (52 - 50)^2 + (48 - 50)^2 + (58 - 50)^2$$

$$= 34 \Rightarrow S_1 = \sqrt{34}$$

=> Sự chênh lệch số lượng giữa các loại sách nhiều.

$$S_{22} = \frac{1}{4}\sum_{j=1}^{4}(x_{j2} - \overline{x_2})^2$$

$$= \frac{1}{4}[(4 - 4)^2 + (5 - 4)^2 + (4 - 4)^2 + (3 - 4)^2]$$

$$= \frac{1}{2} \Rightarrow S_2 = \sqrt{\frac{1}{2}}$$

=> Sự chênh lệch số lượng giữa giá các loại sách ít.

c)Tính Sample Covariance

$$S_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \overline{x_i})(x_{jk} - \overline{x_k}) \quad i, k = 1, 2$$

$$S_{12} = \frac{1}{4}\sum_{j=1}^{4}(x_{j1} - \overline{x_1})(x_{j2} - \overline{x_2}) = \frac{-3}{2}$$

=> Số tiền và số sách có quan hệ ngược nhau

$$S_{21} = S_{12} = \frac{-3}{2}$$

d)Tính Correlation Coefficient

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

$$r_{ik} = r_{ki}$$

$$r_{11} = \frac{S_{11}}{\sqrt{S_{11}}\sqrt{S_{11}}} = 1$$

$$r_{12} = \frac{S_{12}}{\sqrt{S_{11}}\sqrt{S_{22}}} = \frac{\frac{-3}{2}}{\sqrt{34}\sqrt{\frac{1}{2}}} = \frac{-3\sqrt{17}}{34}$$

$$r_{21} = r_{12} = \frac{-3\sqrt{17}}{34}$$

=> Cường độ tương quan giữa hai biến (giá tiền - sách)

$$r_{22} = \frac{S_{22}}{\sqrt{S_{22}}\sqrt{S_{22}}} = 1$$

5. Diễn giải ý nghĩa của các biến(Sample) ở bài trên

-Sample mean: Trung bình mỗi loại sách bán được

Trung bình giá tiền mỗi loại sách

-Sample Variance, Standard Variance:

Độ lệch phương sai

Lấy số sách trung bình trừ độ lệch phương sai sách = số sách nên nhập mỗi ngày

Lấy số tiền trung bình trừ độ lệch phương sai tiền = số tiền nên sài từ số tiền bán sách

-Sample Covariance:

Bán ít sách nhưng doanh thu nhiều và ngược lại

-Correlation Coefficient : chỉ rõ hơn độ tương quan giữa số tiền thu được và số sách bán được.

TUẦN 2

1) Nhìn vào đồ thị họ hiểu về 16 công ty như thế nào

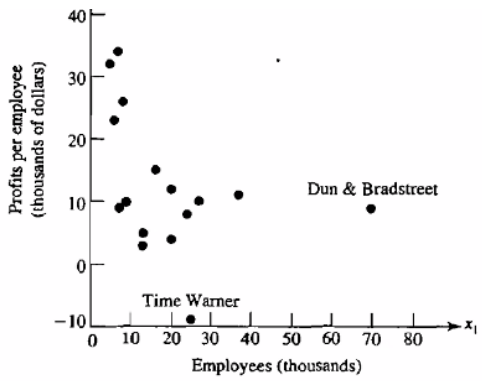


Figure 1.3 Profits per employee and number of employees for 16 publishing firms.

The sample correlation coefficient computed from the values of x_1 and x_2 is

$$r_{12} = \begin{cases} -.39 & \text{for all 16 firms} \\ -.56 & \text{for all firms but Dun \& Bradstreet} \\ -.39 & \text{for all firms but Time Warner} \\ -.50 & \text{for all firms but Dun \& Bradstreet and Time Warner} \end{cases}$$

Số lượng công nhân không tỉ lệ thuận với năng suất lao động trên đầu người (scc không đổi so với mức trung bình -0.39 khi thiếu công ty Time Warner).

2) Đọc ví dụ 1.9 rồi nêu ra kết luận:

Xem biểu đồ 3D Scatter plot xoay theo các chiều hướng khác nhau giúp dễ dàng nhìn ra những trường hợp ngoại lệ.
Tại vị giao điểm của ba trục sẽ có chỉ số ở mức thấp nhất.

3) Vẽ trục hình sao điểm 8 môn học :

SE	AI	OOP	MSA	OS	DHMT	XLAV	VLDC2
5	7	6	8	6	7	8	10

4) Viết python để vẽ đồ thị hình sao:

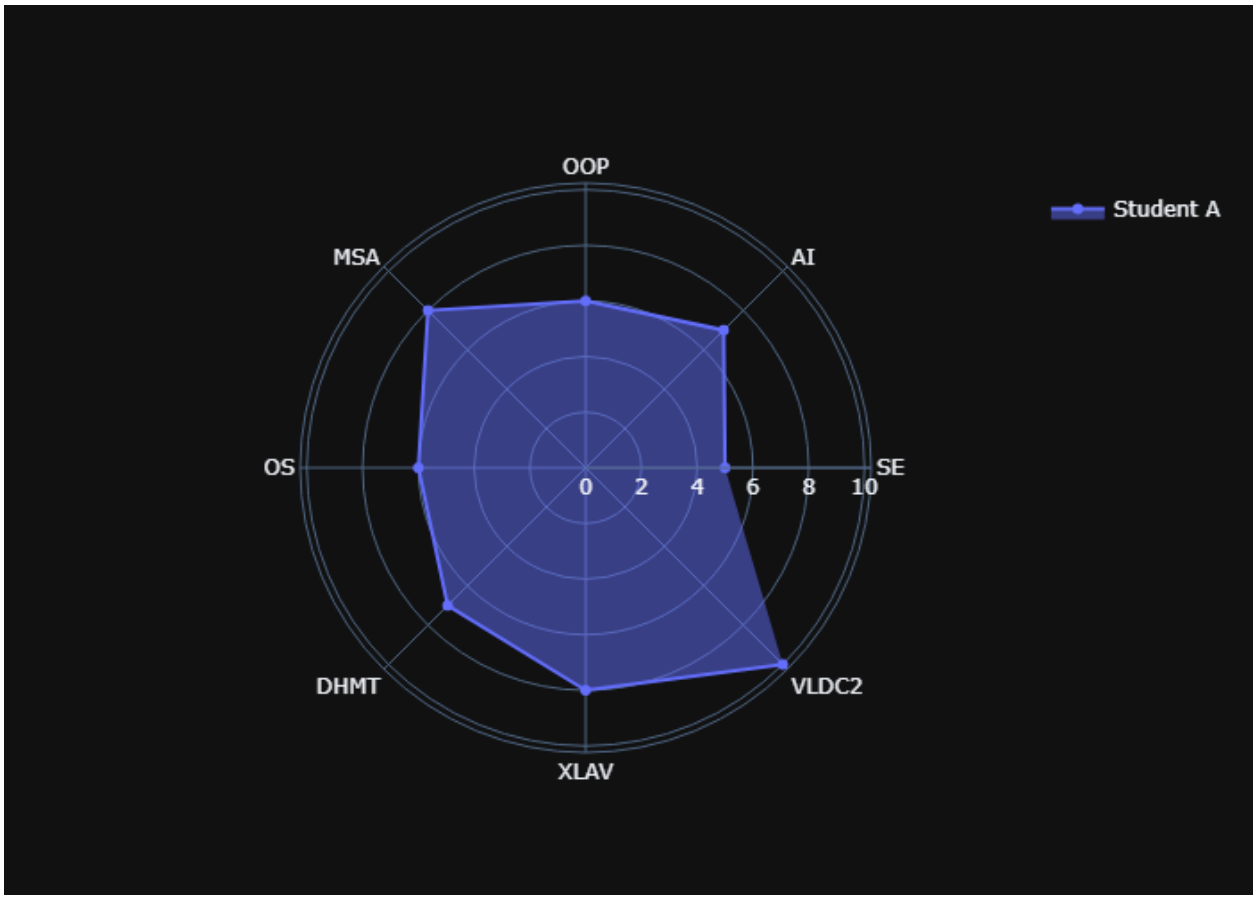
```
import pandas as pd
import numpy as np
import plotly.express as px
import plotly.graph_objects as go

# read dataset from csv and perform preprocessing
data = pd.read_csv('data.csv')
categories = ['SE','AI','OOP','MSA','OS','DHMT','XLAV','VLDC2']
# plot unfilled scatter plot
fig = go.Figure()
```

```
fig.add_trace(go.Scatterpolar(
  r=pd.Series(data.loc[0, categories].values),
    theta=categories,
    fill='toself',
    name='Student A'
))
fig.update_layout(
  polar=dict(
    radialaxis=dict(
      visible=True,
    ),
  ),
  template='plotly_dark',
  showlegend=True,

)
fig.show()
fig.write_image("raderChart.png")
```

Kết quả:



5) Vẽ biểu tượng khuôn mặt từ năm 1 -> năm 4:

6) Ví dụ về tầm quan trọng của statistical distance:

Nếu tính điểm trung bình học tập mà không sử dụng số tín chỉ để tính thì điểm trung bình đó không thể hiện được tầm quan trọng của những môn quan trọng hơn(có số tín chỉ nhiều hơn)

=> Để đánh giá dữ liệu đúng đắn thì phải xét với khoảng biến thiên dữ liệu (phương sai σ^2).

7) Viết biểu thức sau dưới dạng ma trận

$$d(P,Q) = \sqrt{\begin{aligned}&[a_{11}(x_1 - y_1)^2 + a_{22}(x_2 - y_2)^2 + \cdots + a_{pp}(x_p - y_p)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2)\\&+ 2a_{13}(x_1 - y_1)(x_3 - y_3) + \cdots + 2a_{p-1,p}(x_{p-1} - y_{p-1})(x_p - y_p)]\end{aligned}} \quad (1\text{-}23)$$

$$d(P,Q) = \begin{bmatrix} x_1 - y_1 & x_2 - y_2 & \cdots & x_p - y_p \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{j1} & s_{j2} & \cdots & a_{jp} \end{bmatrix} \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \\ \cdot \\ \cdot \\ x_p - y_p \end{bmatrix}$$

8) Giải thích vì sao OP, OQ, OR là như nhau nếu nhìn dưới con mắt thống kê thay vì khoảng cách Euclide.

Statistical distance

9) Tính mean vector/ covariance matrix 2.13

x_1 \ x_2	0	1	$p_1(x_1)$
−1	.24	.06	.3
0	.16	.14	.3
1	.40	.00	.4
$p_2(x_2)$	.8	.2	1

E(X1)=-1*0.3+0*0.3+1*0.4=0.1
E(X2)=0*0.8+1*0.2=0.2

Mean vector
[0.1
0.2]

E(X1-E(X1))^2= (-1-0.1)*(-1-0.1)*0.3+ (0-0.1)*(0-0.1)*0.3 +(1-0.1)*(1-0.1)*0.4=0.69
E(X2-E(X2))^2=(0-0.2)*(0-0.2)*0.8 + (1-0.2)*(1-0.2)*0.2 = 0.16
E(X1-E(X1))(X2-E(X2)) = (-1-0.1)*(0-0.2)*0.24 + (-1-0.1)*(1-0.2)*0.06 +
…+(1-0.1)*(1-0.2)*0.0=-0.08
E(X2-E(X2))(X1-E(X1)) = E(X1-E(X1))(X2-E(X2)) = -0.08

Covariance matrix

[0.69 -0.08
-0.08 0.16]

TUẦN 3

1) Người ta dùng Maximum Likelihood Estimation ở đâu, khi nào

Sử dụng MLE khi:

- Nếu mô hình được giả định chính xác, Maximum Likelihood Estimation là công cụ ước tính hiệu quả nhất.

- Nó cung cấp một cách tiếp cận nhất quán nhưng linh hoạt, phù hợp với nhiều loại ứng dụng, bao gồm cả những trường hợp các giả định của các mô hình khác bị vi phạm.
- Nó dẫn đến các ước tính không chệch trong các mẫu lớn hơn.

2) Giải thích Luật số lớn (Law of large numbers)

Luật số lớn chỉ ra rằng, khi ta chọn ngẫu nhiên các giá trị (mẫu thử) trong một dãy các giá trị (quần thể), kích thước dãy mẫu thử càng lớn thì các đặc trưng thống kê (trung bình, phương sai,...) của mẫu thử càng "gần" với các đặc trưng thống kê của quần thể

3) Cho ví dụ về luật số lớn

Tung một con xúc xắc, Trong đó, kết quả của xác suất xuất hiện các mặt có 1, 2, 3, 4, 5 và 6 chấm là như nhau. Giá trị kỳ vọng của các kết quả là:

$$(1 + 2 + 3 + 4 + 5 + 6) / 6 = 3,5$$

4) Có mấy loại luật số lớn

2 Loại:

- Quy luật số lớn dạng yếu

Xét n biến ngẫu nhiên X1, X2, ..., Xn là các biến ngẫu nhiên độc lập với nhau có cùng phân phối xác suất với kỳ vọng E(X) luật số lớn yếu phát biểu rằng, với mọi số thực dương, xác suất để khoảng cách giữa trung bình tích lũy Yn và kì vọng E(X) lớn hơn là tiến về 0 khi n tiến về vô cực.

- Quy luật số lớn dạng mạnh

Xét n biến ngẫu nhiên độc lập cùng phân phối xác suất, khả tích (nghĩa là $E(|X|) < \infty$), luật số lớn dạng mạnh phát biểu rằng trung bình tích lũy Yn hội tụ gần như chắc chắn về E(X).

5) Giải thích Định lý giới hạn trung tâm (The Central Limit Theorem)

Trong xác suất, định lý giới hạn trung tâm là định lý nổi tiếng và có vai trò quan trọng. Nó là kết quả về sự hội tụ yếu của một dãy các biến ngẫu nhiên. Với định lý này, ta có kết quả là tổng của các biến ngẫu nhiên độc lập và phân phối đồng nhất theo cùng một phân phối xác suất, sẽ hội tụ về một biến ngẫu nhiên nào đó.
Trong trường hợp đơn giản nhất, được dùng dưới đây trong phần chứng minh của định lý, các biến ngẫu nhiên là độc lập, có cùng kỳ vọng và phương sai. Một cách tổng quát, tổng của các biến ngẫu nhiên sẽ tăng vô định khi số biến ngẫu nhiên tăng. Do đó để có một kết quả hữu hạn, ta hạn chế sự tăng của tổng bằng cách lấy tổng trừ đi giá trị trung bình và rút gọn bằng cách chia cho căn bậc hai của phương sai. Với một số các điều kiện nữa thì phân phối xác suất của biến ngẫu nhiên giản lược sẽ hội tụ về một phân phối chuẩn.

6) So sánh Định lý giới hạn trung tâm và Luật số lớn

Định lý giới hạn trung tâm	Luật số lớn
Giống với tính trung bình z=($\bar{x}$ - kỳ vọng)/căn(phương sai/n). Định lý giới hạn trung tâm cho chúng ta hình dạng gần đúng của phân phối. Độ tuyến tính của kỳ vọng cho chúng ta Giá trị trung bình kỳ vọng / Phương sai của phân phối lấy mẫu.	Luật số lớn chỉ nói về giá trị gần đúng của trung bình mẫu, tất nhiên giá trị nào càng ngày càng gần với trung bình tổng thể khi kích thước trung bình mẫu trở nên lớn.

7) χ_p^2 trong Định lý giới hạn trung tâm là gì? Viết công thức.

Là phân phối chi bình phương đơn giản nhất là bình phương của phân phối chuẩn (định lý giới hạn trung tâm). Công thức:

$$f(u) = \frac{1}{\left(\frac{k}{2}-1\right)!}\frac{1}{2^{k/2}}u^{(k/2)-1}e^{-(u/2)}, \quad u > 0$$

TUẦN 4

1) **Điều chỉnh 2 thông số kia thế nào để bán hàng nhiều hơn? (hint: vẽ biểu đồ khi gia tăng giá sản phẩm và chi phí quảng cáo trong khoảng nhất định) [Mô hình hồi quy tuyến tính nhiều biến]**

TUẦN 5 SEMINAR ĐỢT 1 (22/2/2022)

HỒI QUY TUYẾN TÍNH:

1) **Nếu ta muốn tăng số sản phẩm bán ra hằng tuần cao hơn max hiện tại là 689, thì ta cần phải làm sao? Hiệu chỉnh x, y thế nào?**

..

Hồi quy tuyến tính thích hợp với bài toán có số phương trình nhiều hơn số ẩn số.

NHẬN XÉT: Nhóm chưa xác định được mục tiêu của bài toán là tiên đoán.

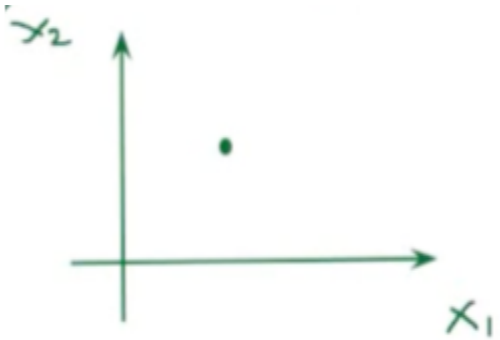
PHÂN TÍCH THÀNH PHẦN CHÍNH PCA

Tính Variance là khai thác, làm cho khoảng biến thiên lớn nhất có thể. Variance thể hiện độ lệch so với giá trị trung bình.
Khoảng biến thiên thể hiện tính khả tách, khoảng biến thiên càng lớn thì tính khả tách càng cao.
Covariance nói lên sự tương quan của 2 biến.
Ràng buộc:

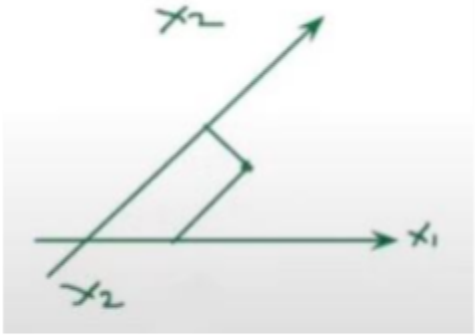
1)e_i (vector cơ sở, vector của hệ tọa độ mới sau khi dùng phép biến đổi tuyến tính chiếu xuống) phải là vector đơn vị, nếu không là vector đơn vị thì ta tiến hành chuẩn hóa theo chiều dài. Sau đây là ràng buộc cho cực trị:

$$e_i{}'e_i = 1 \; hoặc \; e_i{}'e_j = 0, \; \forall i \neq j$$

2)



Góc vuông (trực giao đôi một), thể hiện độc lập tuyến tính



Góc méo, thể hiện sự phụ thuộc tuyến tính

$$Cov(Y_i, Y_j) = 0, \; \forall i \neq j$$

NHẬN XÉT: Nhóm cần xác định rõ ràng buộc ngay tại Phát biểu bài toán:

1)Không gian mới có số chiều nhỏ hơn.

2)Các thành phần giữ lại sau khi giảm chiều phải có tính không tương quan (độc lập tuyến tính). Các vector tương ứng với các trục trong không gian mới đòi hỏi trực giao đôi một. Vì thế chúng ta tính Covariance.

3)Khoảng biến thiên là lớn nhất có thể.

Nhóm chưa tìm được các hệ vector riêng. $e_i{}' \sum e_i = e_i{}' \lambda_i e_i$ (vì $\sum e_i = \lambda_i e_i$) = λ_i (vì e_i là vector đơn vị)
Nhóm chưa hiểu giảm chiều là gì. Giảm chiều là sau khi tìm được các hệ vector riêng và giá trị riêng sắp xếp giảm dần, ta giữ lại k giá trị riêng ứng với k vector riêng lớn nhất, còn lại sẽ bị bỏ đi.

2) Tính

$$X'=[X_1, X_2, ..., X_p]$$

Ma trận hiệp phương sai Cov là ∑

$$Y_1 = a_1'X = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$$
$$Y_2 = a_2'X = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p$$
$$\vdots$$
$$Y_p = a_p'X = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p$$
$$Var(Y_i) = a_i' \sum a_i, \ i = 1,2,...,p$$
$$Cov(Y_i, Y_j) = a_i' \sum a_j, i, j = 1,2,...,p$$

Độ lệch của từng vector (điểm đặc trưng) so với vector trung bình centroid thể hiện ở:

$Var(Y_i) = a_i' \sum a_i = a_i' \lambda_i a_i$ (vì $\sum a_i = \lambda_i a_i$) = λ_i (vì a_i là vector đơn vị)

Khoảng biến thiên cho biết tính khả tách thể hiện ở:

$Cov(Y_i, Y_j) = a_i' \sum a_j = a_i' \lambda_j a_j = 0$ (do điều kiện các vector phải trực giao đôi một)

TUẦN 6

Phân tích tương quan chính tắc (CCA)

1) Chứng minh 4 công thức bên dưới: và

- Trong đó a,b là các hệ số tương ứng, Ta có tiếp
  - $\mathrm{Var}(U) = a'\mathrm{Cov}(X^{(1)})a = a'\sum_{11} a$
  - $\mathrm{Var}(V) = b'\mathrm{Cov}(X^{(2)})b = b'\sum_{22} b$
  - $\mathrm{Cov}(U,V) = a'\mathrm{Cov}(X^{(1)}, X^{(2)})b = a'\sum_{12} b$
- Để tìm mối tương quan giữa U và V,ta tính hệ số tương quan
  - $\mathrm{Corr}(U,V) = \dfrac{\mathrm{Cov}(U,V)}{\sqrt{\mathrm{Var}(U)}\sqrt{\mathrm{Var}(V)}} = \dfrac{a'\sum_{12} b}{\sqrt{a'\sum_{11} a}\sqrt{b'\sum_{22} b}}$

2) Phân biệt Sample Mean và Mean Vector

	sample mean	kì vọng E, sigma
Dữ liệu	mẫu	quần thể
Lượng dữ liệu	Phải lấy đủ mẫu mới có thể tính được vì không có hàm phân bố mẫu.	Không cần phải đi lấy đủ mẫu để tính vì đã có hàm phân bố xác suất/ phân bố mẫu (kì vọng), hàm này cho chúng ta biết mẫu được phân bố như thế nào

3) Trình bày ý nghĩa hình học của Corr(U,V)

Góc giữa 2 vector u,v bé nhất => cos(u,v) lớn nhất.

4) Lập bảng phân biệt PCA và CCA và linear regression

	PCA	**CCA**	**Linear Regression**
Giảm số chiều	Có	Có	
Tương quan các biến	TÌm được các thành phần không tương quan	Tương quan giữa tập biến sao cho lớn nhất	
ưu			
khuyết			

Phân tích dữ kiện (FA)

5) So sánh 4 phuong pháp pca, fa , linear regression, cca

	pca	fa	hồi quy tt	cca
	xấp xỉ tuyến tính vector gốc thành tổ hợp vector riêng ứng với trị riêng max			
sự phụ thuộc tuyến tính	có	giữa x và các vector có sự phụ thuộc.	Có ở quan hệ giữa các biến độc lập và các	có

			biến phụ thuộc	
phạm vi ứng dụng	Bài toán có dữ liệu lớn (số chiều lớn) hoặc dư thừa dữ liệu ảnh (pixel); 2 lớp và có hơn 30 thuộc tính => rút trích thành phần quan trọng để biểu diễn dữ liệu lại. Tiêu chuẩn: tỷ số % giá trị riêng L max chia với tổng các L (phương sai max tập trung tại những giá trị riêng nào thì giữ lại vector riêng/không gian con đó)			

6) **Trong 4 phương pháp chúng ta đã sử dụng thì vai trò của sample và population ở đâu ?**

	LR	**PCA**	**CCA**	**FA**
dữ liệu	mẫu ở cực tiểu sai số	quần thể ở cực đại sigma variance, sigma covariance	quần thể ở cực đại sigma correlation	quần thể ở cực tiểu sai số

(**sigma**: hàm phân bố xác suất**)**

7) **Trong 4 phương pháp có đụng tới vector riêng, giá trị riêng không? Ở đâu?**

Hồi quy tuyến tính không có

TUẦN 7

1) **Trong bài toán phân lớp, quá trình cực tiểu sai số ECM và chuyển ECM từ 2 miền thành 1 miền, Tại sao cần tích phân trong ECM <= 0?**

$$ECM = \int_{R_1} [c(1 \mid 2)p_2 f_2(\mathbf{x}) - c(2 \mid 1)p_1 f_1(\mathbf{x})]d\mathbf{x} + c(2 \mid 1)p_1$$

Vì 'số cá thể lớp 1 bị nhầm vào lớp 2' >= 0
nhân với 'số cá thể lớp 1'>0
=> c(2|1) p_1>=0 : không âm

Để cực tiểu sai số ECM thì ECM = 0, suy ra 'tích phân miền R_1' = 0 - không âm = không dương
suy ra 'tích phân miền R_1' <=0

2) **Làm thế nào đề ra được tiêu chí phân lớp như hình dưới? Tiêu chí này gán nhãn là lớp R_2 được hay không?**

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{c(1 \mid 2)}{c(2 \mid 1)} \frac{p_2}{p_1}$$

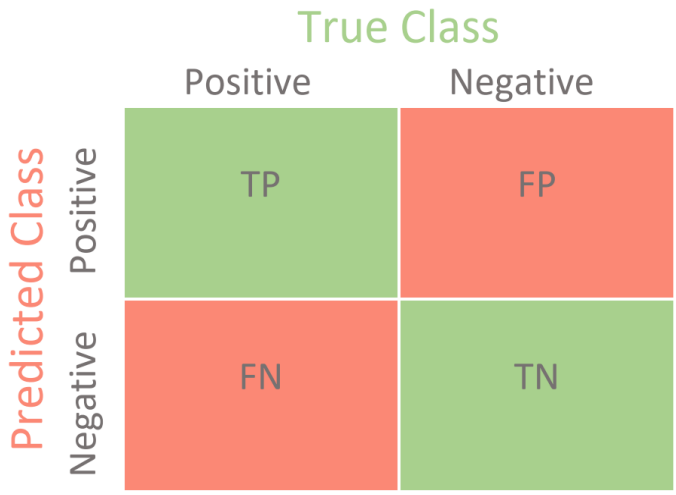
Lớp gồm 'gán nhãn đúng' và 'bị nhầm thành'.
pt ⇔ f1 * c(2|1) * p1 >= f2 * c(1|2) * p2
Vế trái: 'gán nhãn đúng 1' và 'lớp 1' và '1 bị nhầm thành 2'.
Vế phải: 'lớp 2' và 'gán nhãn đúng 2' và "2 bị nhầm thành 1'.
VT >= VP đồng nghĩa tỉ lệ nghiêng về lớp 1 nhiều hơn và ngược lại.
Cho nên tiêu chí trên KHÔNG ĐƯỢC gán nhãn là lớp R_2.

3) **Vì sao ở bài toán phân lớp 2 quần thể đa biến (con cá có nhiều thuộc tính), thì tiêu chí phân lớp chuyển từ $x - \mu_1$, $x - \mu_2$ thành $(\mu_1 - \mu_2)$, $(\mu_1 - \mu_2)^T$? Nó có lợi gì cho mình sau này?**

Bởi vì mục tiêu đạt được hiệu quả trong việc phân lớp, người ta mong muốn 2 điều:
Một, mức độ tập trung của các cá thể trong 1 lớp là lớn (tức phương sai đủ nhỏ).
Hai, 2 lớp phải khác biệt rõ ràng để tránh nhầm lớp (khoảng cách 2 kì vọng lớn).
Việc chuyển $x - \mu_1$, $x - \mu_2$ thành $(\mu_1 - \mu_2)$, $(\mu_1 - \mu_2)^T$ và làm cực đại $(\mu_1 - \mu_2)$, $(\mu_1 - \mu_2)^T$ sẽ thoả được mong muốn thứ hai.

4) **Độ đo sai số khi phân lớp từ công thức lỗi sai TNR, apparent, position recall, accuracy, sensitivity?**tham khảo

Ta có Confusion Matrix:



True Negative Rate (TNR):

TNR = TN/(TN+FP)

Precision, Positive Predictive Value (PPV):

PPV = TP/(TP+FP)

Recall, Sensitivity, Hit Rate, True Positive Rate (TPR)

TPR = TP / (TP+FN)

F Measure

F = (PPV*TPR)/(PPV+TPR)

Hoặc

F = 2TP/(2TP+FP+FN)

Accuracy (ACC)

ACC = (TP+TN)/(TP+FP+FN+TN)

5) Gom nhóm K-Mean, 'không có điểm nào thay đổi' liên quan gì đến cực tiểu lỗi?

Trong thuật toán K-means clustering là một thuật toán của học không giám sát do đó chúng ta không biết nhãn (label) của từng điểm dữ liệu.

Mục đích là làm thế nào để phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau.

Bài toán (2) là một bài toán khó tìm điểm tối ưu vì nó có thêm các điều kiện ràng buộc. Bài toán này thuộc loại mixed-integer programming (điều kiện biến là số nguyên) - là loại rất khó tìm nghiệm tối ưu toàn cục (global optimal point, tức nghiệm làm cho hàm mất mát đạt giá trị nhỏ nhất có thể). Tuy nhiên, trong một số trường hợp chúng ta vẫn có thể tìm được phương pháp để tìm được nghiệm gần đúng hoặc điểm cực tiểu. (Nếu chúng ta vẫn nhớ chương trình toán ôn thi đại học thì điểm cực tiểu chưa chắc đã phải là điểm làm cho hàm số đạt giá trị nhỏ nhất).

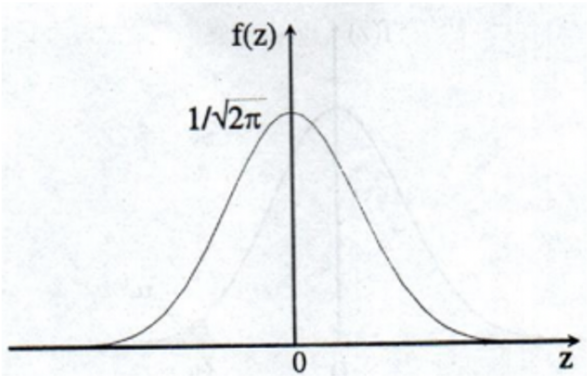
6) Tìm hiểu về các loại phân phối (phân phối student - Chi bình phương - Fisher)

Phân phối Chuẩn

Định nghĩa 1: Biến ngẫu nhiên Z được gọi là có *phân phối chuẩn tắc*, ký hiệu $Z \sim N(0;1)$ nếu nó có hàm mật độ xác suất:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty \text{ (z là số thực)}$$

Đồ thị của hàm $f(z)$ cũng có dạng hình chuông, đối xứng qua trục tung (Hình 1).



Hình 1. Đồ thị phân phối chuẩn tắc

Định nghĩa 2: Biến ngẫu nhiên X được gọi là có *phân phối chuẩn*, ký hiệu $X \sim N(\mu; \sigma^2)$ nếu nó có hàm mật độ xác suất (Normal density probability function):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty \text{ (x là số thực)}$$

Trong đó:

$\pi = 3{,}14159\ldots$

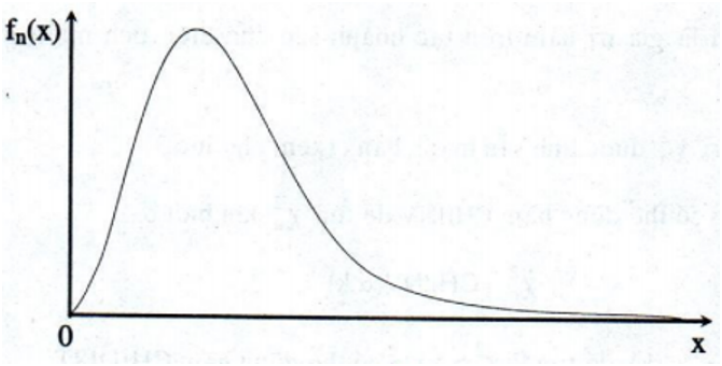
$e = 2{,}71828\ldots$ (*cơ số Logarit Neper*)

μ: *trị số trung bình*

σ : độ *lệch chuẩn*

Phân phối Chi - bình phương

Hàm mật độ: $f(x) = \frac{x^{\frac{k}{2}-1} e^{\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)}$ với $x > 0$



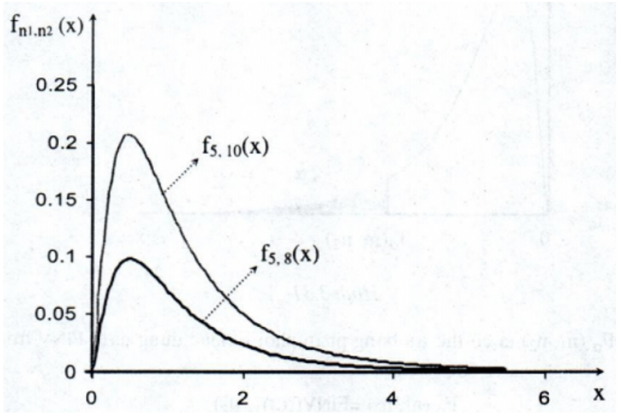
Phân phối Fisher

$$f_{n,m}(x) = \begin{cases} 0 & x \leq 0 \\ C\dfrac{x^{\frac{(n-2)}{2}}}{(m+nx)^{\frac{(n+m)}{2}}} & x > 0 \end{cases}$$

Trong đó:

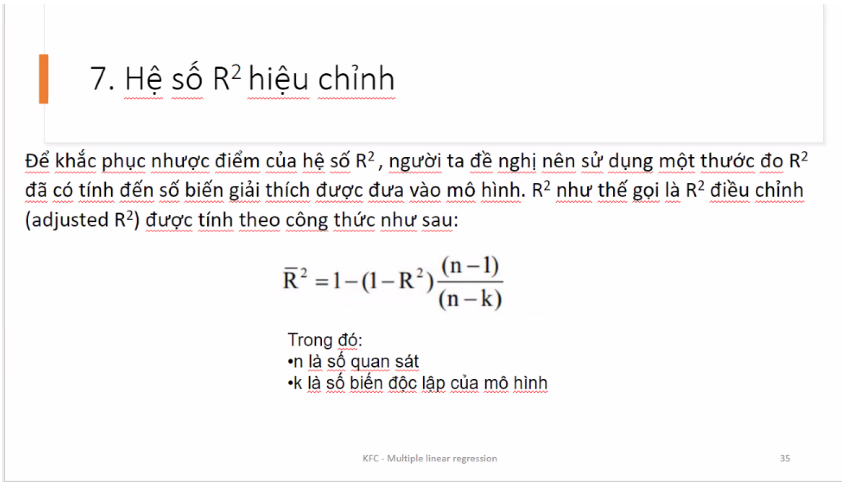
$$C = \frac{\Gamma\left(\frac{n+m}{2}\right) n^{\frac{n}{2}} m^{\frac{m}{2}}}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)}$$

Đồ thị của hàm số $f_{n,m}(x)$



TUẦN 8

1. **Giải thích hệ số điều chỉnh R^2**



7. Hệ số R^2 hiệu chỉnh

Để khắc phục nhược điểm của hệ số R^2, người ta đề nghị nên sử dụng một thước đo R^2 đã có tính đến số biến giải thích được đưa vào mô hình. R^2 như thế gọi là R^2 điều chỉnh (adjusted R^2) được tính theo công thức như sau:

$$\bar{R}^2 = 1-(1-R^2)\frac{(n-1)}{(n-k)}$$

Trong đó:
•n là số quan sát
•k là số biến độc lập của mô hình

KFC - Multiple linear regression 35

TUẦN 9

1) **Phân lớp LDA/Fisher: Tư duy đưa mong muốn vào phân số và cực đại khoảng cách 2 lớp?**

Mong muốn của Fisher là đồng thời 'cực đại bình phương khoảng cách của 2 lớp và cực tiểu tổng 2 phương sai của 2 lớp' là 2 mong muốn tỉ lệ với nhau. Để fisher tối ưu thì fisher phải đạt cực đại. Vậy khi chuyển 2 mong muốn này vào phương trình toán học, ta sẽ để mong muốn đạt cực đại vào tử số của phân số và mong muốn cực tiểu sẽ để ở mẫu số của phân số, và khi ta cực đại phân số này, fisher sẽ tối ưu, mẫu số sẽ tự khắc cực tiểu và tử số sẽ tự khắc cực đại.

2) **Gán nhãn nằm ở biểu thức nào trong LDA/Fisher?**

(m1-m2)(m1-m2)^T

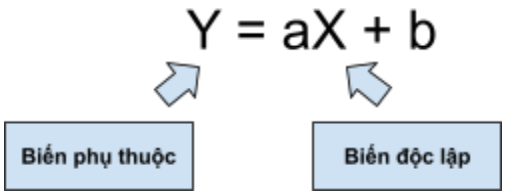
TỔNG HỢP ĐỒ ÁN

1. Linear Regression (Hồi quy tuyến tính)

Ý NGHĨA:

Khoa học:

Phương pháp phân tích hồi quy tuyến tính là phương pháp thống kê để dự đoán các giá trị của một hoặc nhiều biến phụ thuộc - biến kết quả (Dependent variable) từ tập hợp các giá trị của biến độc lập - biến dự báo (Independent variable)



Ứng dụng:

- Kinh doanh, kinh tế - Dự đoán tăng giảm thị trường qua các thông số
- Y học - Dự đoán bệnh qua các thông của bệnh nhân
- Trồng trọt - Dự đoán năng suất cây trồng thông qua các yếu tố thời tiết, đất đai, …

=> Nhìn chung, các ứng dụng của phân tích hồi quy đa phần là dự đoán một kết quả khó biết trước dựa trên các biến độc lập sẵn có.

PHÁT BIỂU BÀI TOÁN

Cho z_{i1}, z_{i2}, z_{i3}, … z_{ir} là các biến dự báo liên quan tới biến kết quả Y_i, với i=1…n

Mô hình hồi quy tuyến tính với n biến kết quả sẽ có dạng sau:

$$Y_1 = \beta_0 + \beta_1 z_{11} + \ldots + \beta_r z_{1r} + \varepsilon_1$$
$$Y_2 = \beta_0 + \beta_1 z_{21} + \ldots + \beta_r z_{2r} + \varepsilon_2$$
$$\cdot$$
$$\cdot$$
$$Y_n = \beta_0 + \beta_1 z_{n1} + \ldots + \beta_r z_{nr} + \varepsilon_n$$

Với ε_i giả định là:

- $E(\varepsilon j) = 0$
- $Var(\varepsilon j) = \sigma^2$ (hằng số)
- $Cov(e_j, e_k) = 0, j \neq k$

Mô hình hồi quy tuyến tính với n biến kết quả dưới dạng ma trận:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & z_{12} \cdots & z_{1r} \\ 1 & z_{21} & z_{22} \cdots & z_{2r} \\ \vdots & \vdots & & \vdots \\ 1 & z_{n1} & z_{n2} \cdots & z_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y = Z.\beta + \varepsilon$$

Khi đó giả định trở thành

- $E(\varepsilon) = \mathbf{0}$
- $Cov(\varepsilon) = E(\varepsilon\varepsilon') = \sigma \mathrm{I}$

PHƯƠNG PHÁP

ƯỚC LƯỢNG THAM SỐ MÔ HÌNH

Giả sử $\boldsymbol{b}$ là giá trị thử của $\boldsymbol{\beta}$.
Xét hiệu của y_j và $b_0 + b_1 z_{j1} + \cdots + b_r z_{jr}$ là sai lệch giữa giá trị thực và giá trị tiên đoán của mô hình
Nếu hiệu này bằng 0, điểm dữ liệu sẽ nằm trên "đường thẳng" của mô hình hồi quy tuyến tính.
Thông thường hiệu này sẽ không bằng 0, vì các điểm dữ liệu thường dao động xung quanh "đường thẳng" do nhiễu (thể hiện qua tham số ε của mô hình). Do đó ta cần các phương pháp để cực tiểu hóa sai lệch.

Least square estimation

Là một trong những phương pháp dùng để cực tiểu hóa sai lệch của mô hình hồi quy tuyến tình trên.

2. Principal Component Analysis (Phân tích thành phần chính)

Ý NGHĨA:

Khoa học:

Một trong những vấn đề thường gặp nhất trong các bài toán phân tích dữ liệu là dữ liệu có rất nhiều chiều và thưa thớt. Khi dữ liệu có quá nhiều chiều như vậy, việc xây dựng một mô hình để đặc tả hết các tính chất của dữ liệu là một điều không thể. Do đó người ta quyết định sẽ tìm và sử dụng các phương pháp để giảm chiều dữ liệu mà không quá ảnh hưởng đến các bài toán mà dữ liệu tham gia => chỉ giữ lại các đặc trưng, chiều quan trọng.

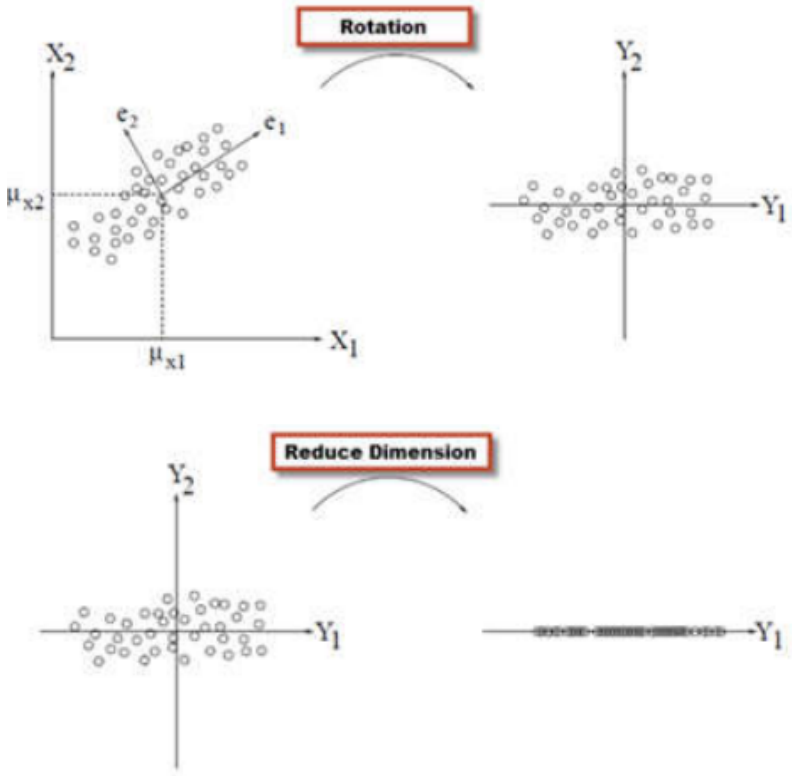
Do đó ta, có được phương pháp phân tích thành phần chính (PCA), một kỹ thuật giảm chiều phổ biến. PCA là một thuật toán thống kê sử dụng phép biến đổi trực giao để biến đổi một tập hợp dữ liệu từ một không gian nhiều chiều sang một không gian mới ít chiều hơn nhằm tối ưu hóa việc thể hiện sự biến thiên của dữ liệu.

Ứng dụng:

Các ứng dụng thực tế của PCA đa phần liên quan đến việc giảm số chiều của dữ liệu, giữ lại các đặc trưng và tính chất quan trọng của các tập dữ liệu. Qua đó giúp việc trực quan hóa dữ liệu dễ dàng hơn, các bài toán phân tích cũng như dự đoán từ các tập dữ liệu nhiều chiều sẽ hoạt động nhanh và trơn tru hơn.

PHÁT BIỂU BÀI TOÁN:

Mục tiêu chính của phương pháp PCA là tìm được một không gian mới với số chiều nhỏ hơn không gian cũ để biểu diễn dữ liệu. Các trục tọa độ được xây dựng sao cho độ biến thiên dữ liệu trên đó là lớn nhất có thể (**maximize the variability**)



Mô hình PCA cuối cùng có được:

$$U = [u_1 | u_2 | .... | u_k]$$
$$F = XU$$

Với:

- U là ma trận có vector PCA
- F là dữ liệu trong không gian mới.

NGUYÊN LÝ:

Gọi α là một trục trong không gian mới cần tìm. Khi đó tọa độ của x_i trên trục chính α là tích vô hướng $\varphi_i = \alpha^T x_i$

Mục tiêu của PCA là tìm α sao cho nó biểu diễn tốt nhất, nghĩa là sao cho φ_i lớn nhất và điều này phải đúng với tất cả n điểm của X.

=> Mục tiêu của PCA là tìm α sao cho tất cả các $\varphi_i = \alpha^T x_i$ với i = 1...n là phải cực đại.

=> Mục tiêu của PCA là cực đại tổng $\sum_{i=1}^{n} \varphi_i^2$. Mà ta lại có:

$$\sum_{i=1}^{n} \varphi_i^2 = \sum_{i=1}^{n} \varphi_i \varphi_i = \sum_{i=1}^{n} (\alpha^T x_i)(\alpha^T x_i) = \sum_{i=1}^{n} (\alpha^T x_i)(x_i^T \alpha) = \sum_{i=1}^{n} \alpha^T [x_i x_i^T] \alpha = \alpha^T X X^T \alpha$$

Vậy ta cần phải thực hiện:

$$\max_{\alpha \in \mathbb{R}^m, \alpha^T \alpha = 1} \alpha^T X X^T \alpha$$

PHƯƠNG PHÁP:

Cho biến ngẫu nhiên X có k chiều, $X' = (x_1, x_2, \dots, x_k)$ mô tả dữ liệu ban đầu và ta có ma trận hiệp phương sai là $\sum = XX^T$.

Phương pháp toán học:

- Ta sẽ tìm các biến ngẫu nhiên mới Y có dạng là tổ hợp tuyến tính của các thành phần của X sao cho y có phương sai càng lớn càng tốt.

$$Y = \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_k x_k = \alpha . X'$$

- Như vậy, vấn đề là tìm được vector α phù hợp. Như đã chứng minh ở phần nguyên lý, α phù hợp sẽ thỏa.

$$\max_{\alpha \in \mathbb{R}^m, \alpha^T \alpha = 1} \alpha^T X X^T \alpha$$

- Bằng phương pháp nhân tử Largrange ta chuyển bài toán này thành:

$$\max_{\alpha \in \mathbb{R}^m, \alpha^T \alpha = 1} \alpha^T X X^T \alpha - \lambda\left(1 - \alpha^T \alpha\right) \qquad (\text{vì } \alpha^T \alpha = 1)$$

- Sử dụng phương pháp Largrange để tìm cực trị của hàm số:

$$L(\alpha, \alpha^T \lambda) = \alpha^T X X^T \alpha - \lambda\left(1 - \alpha^T \alpha\right)$$

- Ta được α là nghiệm của $\Sigma\alpha = \lambda\alpha$
- Như vậy, theo định nghĩa vector riêng và trị riêng, α chính là vector riêng, còn λ chính là trị riêng tương ứng của ma trận Σ

3. Factor Analysis

Ý NGHĨA:

Phân tích dữ kiện(Factor Analysis)là một phương pháp thống kê dùng để mô tả sự biến thiên của những biến có tương quan được quan sát bằng một số nhỏ hơn các biến không quan sát được gọi là nhân tố.

Phân tích dữ kiện được sử dụng giống như một phương pháp giảm chiều dữ liệu hoặc như là một phương pháp phân tích cấu trúc bên dưới dữ liệu. Phân tích dữ kiện được áp dụng vào các lĩnh vực trong khoa học hành vi, khoa học xã hội, tiếp thị, quản lý sản phẩm, vận trù học và các ngành khoa học dữ liệu.

Ngoài việc giảm chiều dữ liệu, phân tích dữ kiện được áp dụng nhiều trong tâm lý lượng, kinh tế học, khoa học hành vi...

PHÁT BIỂU BÀI TOÁN:

NGUYÊN LÝ:

PHƯƠNG PHÁP:

CODING:

4. Canonical Correlation Analysis

Ý NGHĨA:

Trong quá trình phân tích dữ liệu thống kê nhiều biến, ta luôn có nhu cầu phân tích mối tương quan giữa các biến để có thể hiểu sâu dữ liệu cần phân tích hay tìm những thứ giá trị trong tập dữ liệu đó

Giả sử một nhà nghiên cứu đã thu thập dữ liệu về ba biến tâm lý, bốn biến học tập (điểm kiểm tra chuẩn hóa) và giới tính của 600 sinh viên năm nhất đại học. Cô ấy quan tâm đến việc tập hợp các biến số tâm lý liên quan như thế nào đến các biến số học tập và giới tính. Đặc biệt, nhà nghiên cứu quan tâm đến việc có bao nhiêu chiều (biến chính tắc) là cần thiết để hiểu được mối liên hệ giữa hai tập hợp biến.

Để thực hiện phân tích mối quan hệ giữa 2 tập biến, ta cần một công cụ phù hợp. Và công cụ phù hợp nhất với bài toán này là phân tích tương quan chính tắc

PHÁT BIỂU BÀI TOÁN:

Input

- Cho 2 tập vector ngẫu nhiên

$$(p\ x\ n)X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \text{ và } (q\ x\ n)Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{pmatrix}$$

- Trong đó, mỗi X,Y đại diện cho 2 tập biến và giả sử p < q

Và $X_i = [x_{i1}, x_{i2}, \ldots x_{in}]$ Là các giá trị của biến X_i

$Y_i = [y_{i1}, y_{i2}, \ldots y_{in}]$

Output

Mối tương quan chính tắc giữa 2 tập vector ngẫu nhiên

Hai biến chính tắc

U=a'X

V=b'Y

NGUYÊN LÝ:

- Tổ hợp tuyến tính cung cấp các độ đo tóm tắt giữa tập biến, do đó ta đặt

$$U = a'X^{(1)}$$
$$V = b'X^{(2)}$$

- Trong đó a,b là các hệ số tương ứng, Ta có tiếp

$$Var(U) = a'Cov(X^{(1)})a = a'\sum_{11}a$$
$$Var(V) = b'Cov(X^{(2)})b = b'\sum_{22}b$$
$$Cov(U,V) = a'Cov(X^{(1)},X^{(2)})b = a'\sum_{12}b$$

- Để tìm mối tương quan giữa U và V,ta tính hệ số tương quan

$$Corr(U,V) = \frac{Cov(U,V)}{\sqrt{Var(U)}\sqrt{Var(V)}} = \frac{a'\sum_{12}b}{\sqrt{a'\sum_{11}a}\sqrt{b'\sum_{22}b}}$$

Sao cho đạt giá trị lớn nhất

5. Classification

Ý NGHĨA:

KHOA HỌC:

Định nghĩa:

Classification là một kĩ thuật nhằm phân tích các dữ liệu (observation), sau đó phân bố các đối tượng mới (observation) cho các class khác nhau đã được xác định trước đó.

Nhiệm vụ của bài toán phân lớp gồm 2 phần chính:

- Sắp xếp các đối tượng (observation) vào 2 hay nhiều lớp đã được dán nhãn.
- Tìm ra một phương pháp tối ưu để dán nhãn đối tượng cho các lớp đã được dán nhãn trước. (Trọng tâm)

ỨNG DỤNG THỰC TẾ:

- Nhận dạng chữ số viết tay
- Nhận dạng khuôn mặt
- Khai thác văn bản (Text mining)
- Truy vấn ảnh
- Xếp loại học lực dựa trên ĐTB
- Tìm và đánh giá tỷ lệ để dự đoán xác suất mắc bệnh tiểu đường của một người ở một độ tuổi nhất định dựa trên tuổi tác, tình trạng sức khỏe, truyền thống gia đình, thói quen ăn

uống, thời kỳ mang thai, huyết áp, độ dày của da, insulin, chỉ số khối cơ thể,...

PHÁT BIỂU BÀI TOÁN:

PHÁT BIỂU:

Ta có n lớp (population) R như sau:

- Lớp R_1 có label π_1 , hàm mật độ xác suất $f_1(x)$, xác suất tiên nghiệm p_1
- Lớp R_2 có label π_2, hàm mật độ xác suất $f_2(x)$, xác suất tiên nghiệm p_2

  …
- Lớp R_n có label π_n, hàm mật độ xác suất $f_n(x)$, xác suất tiên nghiệm p_n

Bài toán:
Ta có tập vector X = [X_1, X_2, X_3… X_p] và muốn phân bố tập vector này vào n lớp R.
Để làm việc trên, ta phải tìm ra một Mô Hình Phân Lớp để biết nên phân bố được các vector trong tập X này vào lớp nào trong n lớp R.

INPUT:

Vector X^T = [X_1, X_2, X_3… X_p]

OUTPUT:

X được dán nhãn π_1 hoặc π_2 hoặc …. π_n

NGUYÊN LÝ:

LDA:

Cơ chế hoạt động:

- B1: Tính ma trận within s_w
- B2: Tính ma trận between s_B
- B3 Tìm vector phép chiếu tốt nhất (tính trị riêng, tính vector riêng, tính vector cơ sở theo vector riêng) $s_w^{-1}.s_B.w = \lambda w$
- B4: Giảm số chiều $y = w^T x$

Tiêu chí phân lớp cho cá thể mới:

Bước 1- Chiếu điểm mới với w^T vừa tìm được

$$new = w^T.[x_{new}, y_{new}]$$

Bước 2- Tính d_1, d_2 - độ lệch của điểm mới với từng centroid của mỗi lớp:

$$d_1 = |new - m_1|$$

$$d_2 = |new - m_2|$$

Bước 3- Gần với centroid của lớp nào thì thuộc lớp đó:

$$d_1 < d_2 : R_1$$

$$d_2 < d_1 : R_2$$

6. Clustering

Ý NGHĨA:

Trong khoa học, phương pháp gom nhóm được sử dụng để phân tích cấu trúc dữ liệu để hỗ trợ người dùng trong việc phân tích và khám phá dữ liệu như:

- Gom gen thành các nhóm có các chức năng giống nhau.
- Phân tích các nhóm thành các danh mục con để tạo ra cấu trúc phân cấp để hỗ trợ người dùng trong việc khám phá dữ liệu.
- Phát hiện các mối quan hệ giữa các loại bệnh với nhau.

Trong ứng dụng, các nhóm được hình thành từ phương pháp gom nhóm có thể được sử dụng để làm bước đầu cho các mục đích khác như:

- Tóm tắt dữ liệu
- Nén dữ liệu
- Phân đoạn ảnh
- Tìm láng giềng gần nhất
- Tiếp thị
- Các thủ tục chống gian lận
- Phân tích tác động

PHÁT BIỂU BÀI TOÁN:

Input:

- Tập dữ liệu (quan sát) X = {X1, X2, ..., Xn}
- Xi = (Xi1, Xi2, ..., Xim) mô tả tả thuộc tính của phần tử thứ i

Output:

Phân các dữ liệu thuộc tập X thành các nhóm sao cho:

- Các phần tử trong cùng một nhóm có tính chất giống nhau (gần nhau)
- Các phần tử trong các nhóm khác nhau có tính chất khác nhau (xa nhau)

NGUYÊN LÝ:

PHƯƠNG PHÁP:

CODING:

7. Inference from sample

Ý NGHĨA:

Trong thống kê, một tổng thể (population) được biểu diễn bằng các đặc điểm số học, và được đại diện bởi các tham số (parameter) của tổng thể. Vì thế, việc tìm hiểu về các tham số cực kì quan trọng để hiểu đặc điểm của tổng thể. Thông thường, ta khó xác định chính xác tham số của 1 tổng thể vì giới hạn thời gian, chi phí để thu thập và phân tích toàn bộ tổng thể, hay số lượng phần tử của tổng thể quá lớn, nên người ta sẽ tiến hành chọn mẫu (Sampling), tính toán các tham số của mẫu (sampling mean, variance, proportion), và dựa vào đó, để đưa ra các đặc điểm của các tham số tổng thể. Suy luận thống kê (Statistic inference) được sử dụng để đưa ra các quyết định về tổng thể dựa trên chọn mẫu.

Suy luận thống kê thường được chia làm 2 loại:

- Ước lượng tham số (Parameter estimation):
- Kiểm định giả thuyết (Hypothesis Testing):

PHÁT BIỂU BÀI TOÁN:

Bài toán kiểm định giả thuyết

Thông qua bài toán này ta có thể kiểm tra trung bình mẫu của quần thể (population) được dự đoán trước đó có phù hợp với quần thể không? Trung bình mẫu này ta có thể ước lượng bằng cách tính trung bình các mẫu đã thu thập được.

Input:

- Tập các mẫu (sample) X_i $(i = 1,2, \dots, p)$ trong mỗi tập mẫu X_i bao gồm các mẫu con x_j $(j = 1,2, \dots, n)$

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{bmatrix}$$

- Độ tin cậy % (mức ý nghĩa α) cần xét
- Giá trị μ_0 (Giả thuyết cho trước): $\underset{(p \times 1)}{\mu_0} = \begin{bmatrix} \mu_{10} \\ \vdots \\ \mu_{p0} \end{bmatrix}$

Output:

Bác bỏ giả thuyết H_0: μ_0? (Bác bỏ kỳ vọng ban đầu)

Bài toán Ước lượng trung bình

Trong thực tế, đôi khi ta sẽ không biết được giá trị trung bình mẫu của quần thể(population) vậy nên ta chỉ có thể ước lượng được giá trị nằm trong khoảng bao nhiêu.

Input:

- Tập các mẫu (sample) X_i $(i = 1,2,\dots,p)$ trong mỗi tập mẫu X_i bao gồm các mẫu con x_j $(j = 1,2,\dots,n)$

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{bmatrix}$$

- Độ tin cậy % (mức ý nghĩa α)cần xét
- Giá trị μ_0 (Giả thuyết cho trước): $\underset{(p \times 1)}{\mu_0} = \begin{bmatrix} \mu_{10} \\ \vdots \\ \mu_{p0} \end{bmatrix}$ (Nếu có)

Output:

Khoảng của trung bình $\underset{(p \times 1)}{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}$

PHƯƠNG PHÁP:

Bài toán kiểm định giả thuyết

1. Tính trung bình mẫu
2. Tính ma trận nghịch đảo của phương sai mẫu:
3. Tính T^2 (tỷ lệ với khoảng cách giữa giá trị trung bình của mẫu và μ):

Kết luận: Nếu $T^2 > \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$ thì bác bỏ H_0 (Với $F_{p,n-p}$ tra trong bảng phân phối Fisher)

Vì phân phối của T^2 tương đương ngụ ý rằng $\frac{n-p}{(n-1)p} T^2 \sim F_{p,n-p}$.

Bài toán Ước lượng tham số

Tính trung bình mẫu

Tính phương sai mẫu

Tính ma trận nghịch đảo của phương sai mẫu:

Tính T^2 (Nếu có giá trị μ_0 được cho trước):

Kết luận: Nếu $T^2 > \chi_p^2(\alpha)$ thì bác bỏ H_0 (Với $\chi_p^2(\alpha)$ tra trong bảng phân phối Chi – bình phương)