Hồi quy tuyến tính
(Linear Regression- Supervised learning)

Trường Đại học Công nghệ Thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh
Tài liệu nội bộ

Tháng 2 năm 2020

Tổng quan

1. Giới thiệu về hàm hồi quy
2. Phương pháp bình phương bé nhất-Least square method (LSM)
3. Thực hành với python

Nội dung trình bày

1. Giới thiệu về hàm hồi quy

Hồi quy là gì?

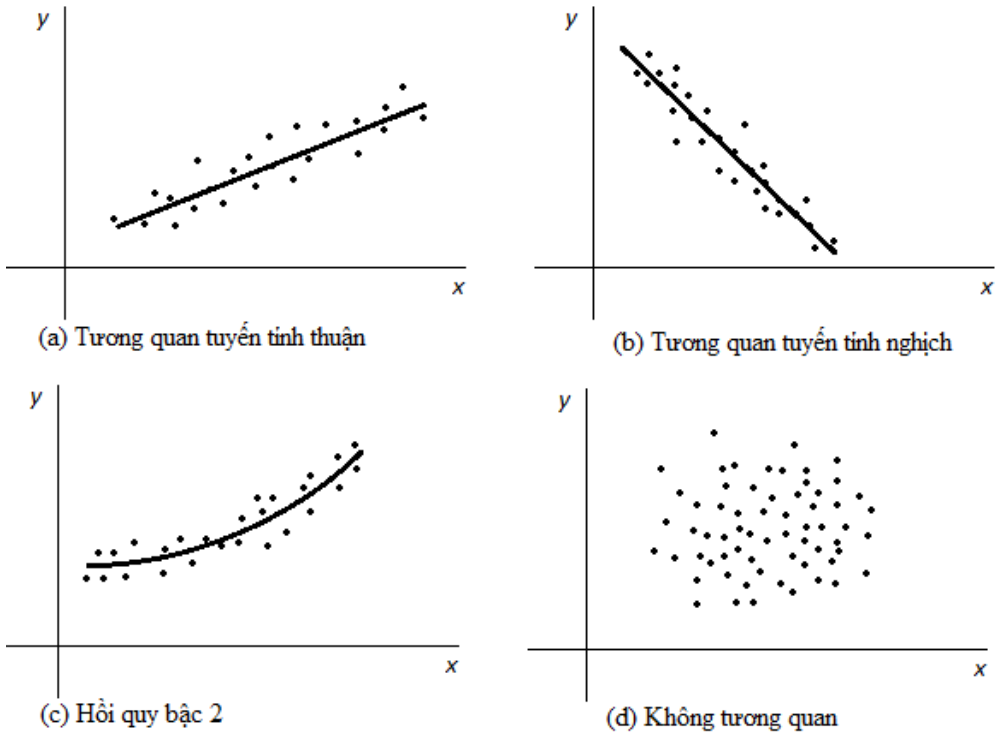
Khái niệm hồi qui dùng để mô tả quan hệ thống kê giữa các biến.
Để "đọc" được mối liên hệ giữa X và Y và dự đoán được Y khi biết giá trị của X người ta theo các bước sau:

- Biểu diễn mỗi quan sát $(x_i; y_i)$ bởi một điểm trên mặt phẳng toạ độ, ta còn gọi nó là *đồ thị phân tán*
- "Vẽ" một đường cong để mô tả mối quan hệ giữa hai đại lượng và dùng nó để dự đoán xu hướng của Y cũng như giá trị của nó khi biết giá trị của X. Đường cong như vậy được gọi là đường hồi quy (hay đường cong xấp xỉ). Thuật toán hồi quy tuyến tính thuộc vào nhóm học có giám sát (supervised learning)

Cách chọn hàm hồi quy

Chọn hàm hồi quy tùy thuộc hình dáng đám mây điểm. Hàm hồi quy có thể tuyến tính $y = a + bx$, hay bậc hai $y = a + bx + cx^2$,.... Các trường hợp kể trên thuộc họ hồi qui có tham số.

Phương pháp chọn một đường cong như vậy gọi là phương pháp hồi qui, còn phương trình của đường cong được gọi là phương trình hồi qui.



Hình 1: Đám mây điểm (trên mặt phẳng) và đường hồi quy

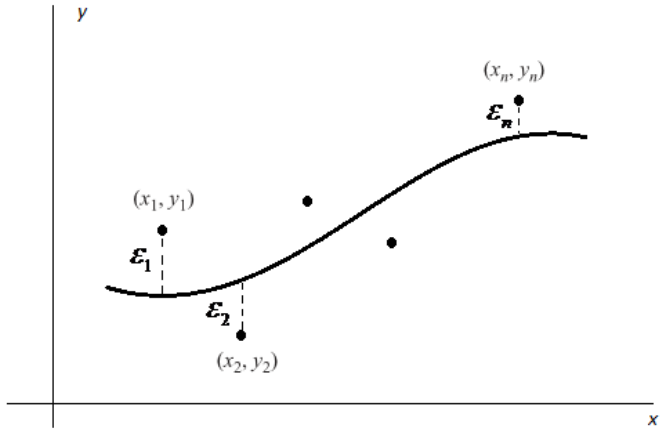
Tại sao chọn mô hình tuyến tính

Bằng phép biến đổi, nhiều đường cong có thể được "tuyến tính hóa"

- $y = a + bx + cx^2 = A + B(C + Dx)^2$. Đặt $X = (A + Bx)^2$ ta được $y = A + BX$
- Hoặc $y = a + bx + cx^2 = A + B(C + Dx)^2$, đặt $x2 = x^2$ ta sẽ được "đường tuyến tính" $y = a + bx + cx2$ (chính là phương trình mặt phẳng).
- $y = a^{b+cx}$. Lấy log hai vế ta được $\ln y = (b + cx) \ln a = b \ln a + (c \ln a)x$. Đặt $Y = \ln y$, $A = b \ln a$, $B = c \ln a$ ta được $Y = A + Bx$
-

Nội dung trình bày

❷ Phương pháp bình phương bé nhất-Least square method (LSM)



Hình 2: Sai số của dữ liệu và đường hồi quy

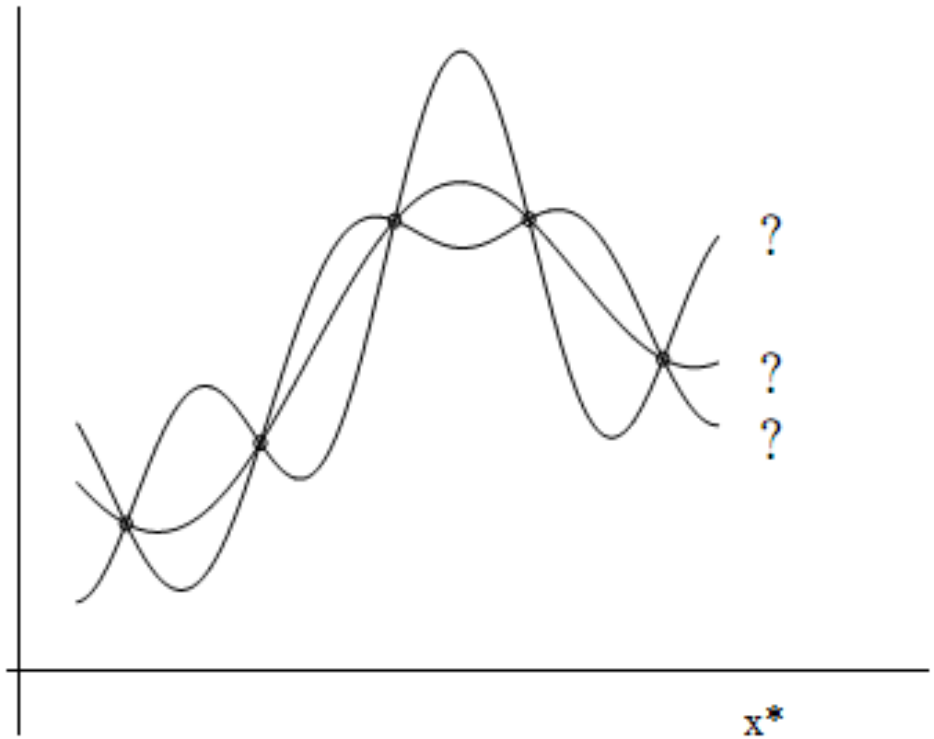
- Xác định trước dạng của hàm hồi quy $y = E(Y/X = x) = f(x)$.
- Đặt hàm mất mát (lost function) (còn gọi là tổng bình phương sai số)

$$L = \sum \varepsilon_i^2 = \sum_i \left[y^{(i)} - f\left(x^{(i)}\right) \right]^2 \tag{1}$$

Trong đó ε_i là độ lệch giữa giá trị quan sát thực tế và giá trị dự đoán bởi hàm hồi qui.

- Ta chọn các tham số của mô hình sao cho L là bé nhất.

Mô hình quá khớp với dữ liệu (Overfitting a model): Nếu ta muốn chọn hàm hồi quy sao cho hàm mất mát bằng 0, tức là phương trình hồi quy đi qua tất cả các điểm số liệu, thì có nhiều cách để chọn hàm hồi quy thỏa mãn điều này. Chính vì vậy mà xu hướng của dữ liệu và việc dự đoán số liệu mới trở nên khó đoán hơn.



Hình 3: Mô hình quá khớp với dữ liệu

Hồi quy tuyến tính

Giả sử biến đáp ứng y phụ thuộc vào $n-1$ biến độc lập (hay *biến đầu vào*) $x_i, i = 1, ..., n-1$. Khi đó phương trình hồi quy của y theo x_i được gọi là PTHQ đa biến.
Mô hình tuyến tính có dạng

$$\hat{y}(\mathbf{x}, \boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \ + \theta_{n-1} x_{n-1} = \boldsymbol{\theta}^T \begin{bmatrix} 1 \\ \mathbf{x}^- \end{bmatrix} = \boldsymbol{\theta}^T \mathbf{x} \tag{2}$$

trong đó $\mathbf{x}^- \in \mathbb{R}^{n-1}$ là véc tơ biến đầu vào và $\boldsymbol{\theta} \in \mathbb{R}^n$ là véc tơ tham số của mô hình. Giá trị của tham số sẽ được ước lượng bằng cách sử dụng m cặp giá trị $(\mathbf{x}^{-(i)}, y^{(i)})$ của dữ liệu đã quan sát (hay còn gọi là *tập huấn luyện*).
Các tên gọi khác trong ứng dụng:

- $n-1$ là số thuộc tính (feature)
- x_i là giá trị thuộc tính thứ i
- θ_i là trọng số của thuộc tính thứ i, $i \geq 1$

Hồi quy tuyến tính

Giả sử ta có m cặp dữ liệu huấn luyện $(\mathbf{x}^{(i)}, y^{(i)}), i = \overline{1, m}$ được biểu diễn tương ứng bằng các véc tơ $\mathbf{X} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, , \mathbf{x}^{(m)}]^\intercal, \mathbf{y} = [y^{(1)}, y^{(2)}, ..., y^{(m)}]^\intercal$ - được gọi là ma trận mẫu (design matrix), và $\hat{\mathbf{y}} \in \mathbb{R}^m$ là kết quả dự đoán tương ứng.

$$\mathbf{X} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, , \mathbf{x}^{(m)}]^\intercal = \begin{bmatrix} 1 & x_1^{(1)} & \dots & x_{n-1}^{(1)} \\ 1 & x_1^{(2)} & \dots & x_{n-1}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & \dots & x_{n-1}^{(m)} \end{bmatrix} \tag{3}$$

Để ý rằng ở ma trận $\mathbf{X}$, ta sắp mỗi dữ liệu huấn luyện theo hàng (m hàng) và các thuộc tính của chúng theo cột (n cột).

Để xác định tham số $\boldsymbol{\theta}$ theo phương pháp bình phương bé nhất, ta cần cực tiểu hàm mất mát $J(\boldsymbol{\theta})$ - (lost function)- như sau:

$$J(\boldsymbol{\theta}) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2 = \frac{1}{2m} \sum_{i=1}^{m} \left(\boldsymbol{\theta}^\intercal \mathbf{x}^{(i)} - y^{(i)} \right)^2 \tag{4}$$

Biểu thức có chia cho

Tổng quát hóa mô hình tuyến tính đa biến

Mô hình hồi quy tuyến tính đa biến tổng quát có dạng:

$$\hat{y}(\mathbf{x}, \boldsymbol{\theta}) = \theta_0 + \theta_1 \varphi_1(\mathbf{x}) + \ + \theta_{n-1} \varphi_{n-1}(\mathbf{x}) \tag{5}$$

trong đó các hàm theo biến $\mathbf{x}$ là $\varphi_i(\mathbf{x})$ được gọi là các hàm cơ bản (basic function). Thường người ta sẽ đặt $\varphi_0(\mathbf{x}) = 1$ và viết lại công thức trên như sau:

$$\hat{y}(\mathbf{x}, \boldsymbol{\theta}) = \sum_{i=0}^{n-1} \theta_i \varphi_i(\mathbf{x}) = \boldsymbol{\theta}^\intercal \varphi(\mathbf{x}) \tag{6}$$

Trong các công thức trên véc tơ được đề cập được hiểu là véc tơ cột.
Mô hình hồi quy tuyến tính tương ứng với chọn hàm $\varphi(\mathbf{x}) = \mathbf{x}$

Hồi quy tuyến tính - Công thức nghiệm

Vậy tham số $\boldsymbol{\theta}$ được ước lượng bằng giá trị làm cực tiểu hàm mất mát:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{7}$$

Giải phương trình đạo hàm $J'(\boldsymbol{\theta}) = 0$ ta được giá trị ước lượng của $\boldsymbol{\theta}$ là:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{8}$$

Nhận xét Công thức trên??

Ví dụ (mô phỏng quá trình tính toán ma trận)

Lập phương trình hồi quy tuyến tính của Y theo X từ tập huấn luyện sau:
(x_i, y_i)= (147,49), (150,53), (153,51), (160,54).

Giải: Ma trận mẫu và ma trận đáp ứng

$$\mathbf{X} = \begin{bmatrix} 1 & 147 \\ 1 & 150 \\ 1 & 153 \\ 1 & 160 \end{bmatrix}; \mathbf{y} = \begin{bmatrix} 49 \\ 53 \\ 51 \\ 54 \end{bmatrix}$$

Các ma trận trung gian:

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 4 & 610 \\ 610 & 93118 \end{bmatrix}; (\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{186}\begin{bmatrix} 46559 & -305 \\ -305 & 2 \end{bmatrix}$$

Thay các ma trận trên vào công thức nghiệm ta được

$$\hat{\theta} \approx \begin{bmatrix} 5.02 \\ 0.31 \end{bmatrix}$$ Vậy phương trình hồi quy cần tìm là: $y = 5.02 + 0.31x$

Điều kiện sử dụng phương pháp LSM

Xét mô hình tuyến tính

$$y(\mathbf{x}, \boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \ + \theta_{n-1} x_{n-1} + u = \boldsymbol{\theta}^T \mathbf{x} + u \tag{9}$$

- $u \sim N(0, \sigma^2)$
- $E(u_i) = 0, cov(u_i, u_j) = 0, \forall i \neq j, var(u_i) = \sigma^2 \ \forall i$
- $cov(u_i, x_j) = 0$

Đánh giá sự phụ thuộc tuyến tính- hệ số xác định R^2

- $TSS(\text{total sum squares}) = \sum(y_i - \bar{y})^2$ (SS_{TOT})
  Tổng bình phương tất cả sai lệch giữa y_i và giá trị trung bình.
- $ESS(\text{explained sum of squares}) = \sum(\hat{y}_i - \bar{y})^2$ (SS_{REG})
  Tổng bình phương các sai lệch giữa giá trị dự đoán của biến phụ thuộc y và giá trị trung bình => đo độ chính xác của hàm hồi qui.
- $RSS(\text{residual sum of squares}) = \sum(e_i^2)$ (SS_{ERR})
  Tổng bình phương sai số.

Từ quan hệ

$$TSS = ESS + RSS \tag{10}$$

Chia hai vế cho TSS ta được hệ số xác định (hay giá trị thống kê "good of fit")

$$R^2 = 1 - \frac{RSS}{TSS} \tag{11}$$

R^2 càng cao (càng gần 1) thì mô hình càng giải thích được biến động của biến phụ thuộc (VD: $R^2 = 0.85$ cho thấy biến độc lập giải thích được 85% sự thay đổi của biến phụ thuộc, còn 15% còn lại là do các yếu tố ngẫu nhiên gây ra.

Đánh giá sự phụ thuộc tuyến tính-Hệ số xác định hiệu chỉnh (Adjusted R-squared)

Khi sử dụng nhiều biến độc lập trong mô hình hồi quy thì số bậc tự do sẽ giảm đi (do df=m-n, m là cỡ mẫu, n là số hệ số của mô hình). Để khắc phục điều này điều chỉnh hệ số xác định R^2 bằng cách đưa thêm bậc tự do của các tổng bình phương vào công thức hệ số xác định:

$$\bar{R}^2 = 1 - \frac{RSS/(m-n)}{TSS/(m-1)} = R^2 + (1-R^2)\frac{1-n}{m-n} \tag{12}$$

Thông thường, biến độc lập đưa thêm vào mô hình là xác đáng nếu nó làm tăng giá trị $\bar{R}^2$ và hệ số hồi quy của biến này khác 0 có ý nghĩa thống kê.
Ghi chú: $\bar{R}^2$ có thể âm, khi đó ta gán giá trị 0.

Bảng phân tích phương sai một biến (one-way ANOVA hay univariate ANOVA

Source	Sum of squares	Degrees of freedom	Mean squares	F
Model	$SS_{\text{REG}} = \sum(\hat{y}_i - \bar{y})^2$	1	$MS_{\text{REG}} = SS_{\text{REG}}$	$\frac{MS_{\text{REG}}}{MS_{\text{ERR}}}$
Error	$SS_{\text{ERR}} = \sum(y_i - \hat{y}_i)^2$	$\mathrm{m}-2$	$MS_{\text{ERR}} = \frac{SS_{\text{ERR}}}{\mathrm{m}-2}$	
Total	$SS_{\text{TOT}} = \sum(y_i - \bar{y})^2$	$\mathrm{m}-1$		

Bảng phân tích phương sai nhiều biến (Multivariate ANOVA

Source	Sum of squares	Degrees of freedom	Mean squares	F
Model	$SS_{\text{REG}} = (\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})^T(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})$	n-1	$MS_{\text{REG}} = \frac{SS_{\text{REG}}}{\text{n-1}}$	$\frac{MS_{\text{REG}}}{MS_{\text{ERR}}}$
Error	$SS_{\text{ERR}} = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T(\boldsymbol{y} - \hat{\boldsymbol{y}})$	m-n	$MS_{\text{ERR}} = \frac{SS_{\text{ERR}}}{\text{m-n}}$	
Total	$SS_{\text{TOT}} = (\boldsymbol{y} - \bar{\boldsymbol{y}})^T(\boldsymbol{y} - \bar{\boldsymbol{y}})$	$\text{m} - 1$		

Hình 5

Kiểm định mức ý nghĩa toàn bộ mô hình- Xem có hệ số nào khác 0 không, mức α

- Đặt giả thuyết H_0: "tất cả các hệ số θ_i đều bằng 0". H_1: "Có một hệ số nào đó khác 0".
- Tính kiểm định F-Stat: Tuân theo phân phối Fisher-Snedecor (n-1,m-n)
- Nếu $P-$ value for F $<\alpha$ thì bác giả thuyết H_0

Các giá trị được tính sẵn trong python.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.733
Model:	OLS	Adj. R-squared:	0.730
Method:	Least Squares	F-statistic:	268.6
Date:	Tue, 17 Mar 2020	Prob (F-statistic):	7.88e-30
Time:	16:50:19	Log-Likelihood:	-147.81
No. Observations:	100	AIC:	299.6
Df Residuals:	98	BIC:	304.8
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.1648	0.203	20.503	0.000	3.762	4.568
X	2.9328	0.179	16.389	0.000	2.578	3.288

Omnibus:	1.707	Durbin-Watson:	1.886
Prob(Omnibus):	0.426	Jarque-Bera (JB):	1.200
Skew:	-0.241	Prob(JB):	0.549
Kurtosis:	3.237	Cond. No.	3.54

Kiểm định giả thuyết hệ số của mô hình

Để biết hệ số θ_i của mô hình tuyến tính thực sự có ý nghĩa không, ta cần kiểm định giả thuyết $H_0 : \theta_i = 0$, với đối thuyết $H_1 : \theta_i <> 0$ ở mức ý nghĩa α cho trước (thường được chọn là 0.05).

- Tính

$$t = |\frac{\theta_i}{S(\theta_i)}|, \; S(\theta_i) = \frac{\sigma}{\sqrt{\sum x_i^2}} \tag{13}$$

- Tra bảng t-Student giá trị $t(m-n, \alpha)$, n là số tham số cần xác định trong mô hình, m là kích thước tập huấn luyện.
- Nếu $t \geq t(m-n, \alpha)$ thì bác giả thuyết H_0, nghĩa là biến x_i có ý nghĩa.

Trong python hay sử dụng giá trị $P - value$:

$$P - value = P(T \geq t) \tag{14}$$

Nếu $P - value \leq \alpha$ thì bác giả thuyết H_0, nghĩa là biến x_i có ý nghĩa.

Nội dung trình bày

3. Thực hành với python

BT1. Tìm hiểu các lệnh thống kê trong chương trình python, giải thích các thông số trong bảng phân tích hồi quy:

```
4 import statsmodels.formula.api as smf
5 from pandas import DataFrame
6 import numpy as np
7 #Đọc dữ liệu
8 xl = pd.ExcelFile('demo_data.xls')
9 # get the first sheet as an object
10 df = pd.read_excel(xl, 0, header=0)
11 #print(df.head())
12 #print(df)
13
14 results = smf.ols('y ~ X', data=df).fit()
15 print(results.summary())
16
```

Hình 7: Phân tích hồi quy bằng python

BT2. Thực hành trên python tính phương trình hồi quy đơn (tr 117).
BT3. Sử dụng CT python tr 117 tính lại VD mô phỏng ở slide 11.
BT4. Thực hành với dữ liệu cho trong Sheet1 file Excel demo-data.xls.
- Vẽ biểu đồ phân tán (biểu đồ Scatter) và nhận định về quan hệ giữa x và y.
- Thay đổi một giá trị của y sao cho thật khác biệt. Chạy CT và quan sát.

BT5. (Phân tích hồi quy đa biến) Cho dữ liệu về giá cổ phiếu demo-data-mul.xls.
- Vẽ biểu đồ Scatter giữa giá cổ phiếu và các thuộc tính Interest Rate và Unemployment Rate
- Tìm phương trình hồi quy của giá cổ phiếu theo hai thuộc tính (biến) Interest Rate và Unemployment Rate.

BT6. Thực hành với dữ liệu cho trong Sheet2 file Excel demo-data.xls.
- Thực hiện các yêu cầu tương tự BT5.
- Bỏ bớt biến và quan sát sự thay đổi các thông số trong bảng phân tích hồi quy
- Xác định phương trình hồi quy có chứa $X2^2$ và nhận xét các thông số (Tạo thêm 1 cột $X2^2$)