Khoa Khoa Học & Kỹ Thuật Máy Tính
Trường Đại Học Bách Khoa Tp. Hồ Chí Minh



# Chương 3: Hồi qui dữ liệu

KHAI PHÁ DỮ LIỆU

(DATA MINING)

# Nội dung



- 3.1. Tổng quan về hồi qui
- 3.2. Hồi qui tuyến tính
- 3.3. Hồi qui phi tuyến
- 3.4. Ứng dụng
- 3.5. Các vấn đề với hồi qui
- 3.6. Tóm tắt

# Tài liệu tham khảo



**[1] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers, 2006.**

- **6.11 Prediction (pp. 354 -> pp. 359)**
- **6.12 Accuracy and Error Measures (pp. 359 -> pp.363)**
- **6.13 Evaluating the Accuracy of Classifier, Predictor (pp 363 -> 366)**

# 3.0. Tình huống 1



Ngày mai giá cổ phiếu STB sẽ là bao nhiêu???

Microsoft Excel - stb.csv

File Edit View Insert Format Tools Data Window Help

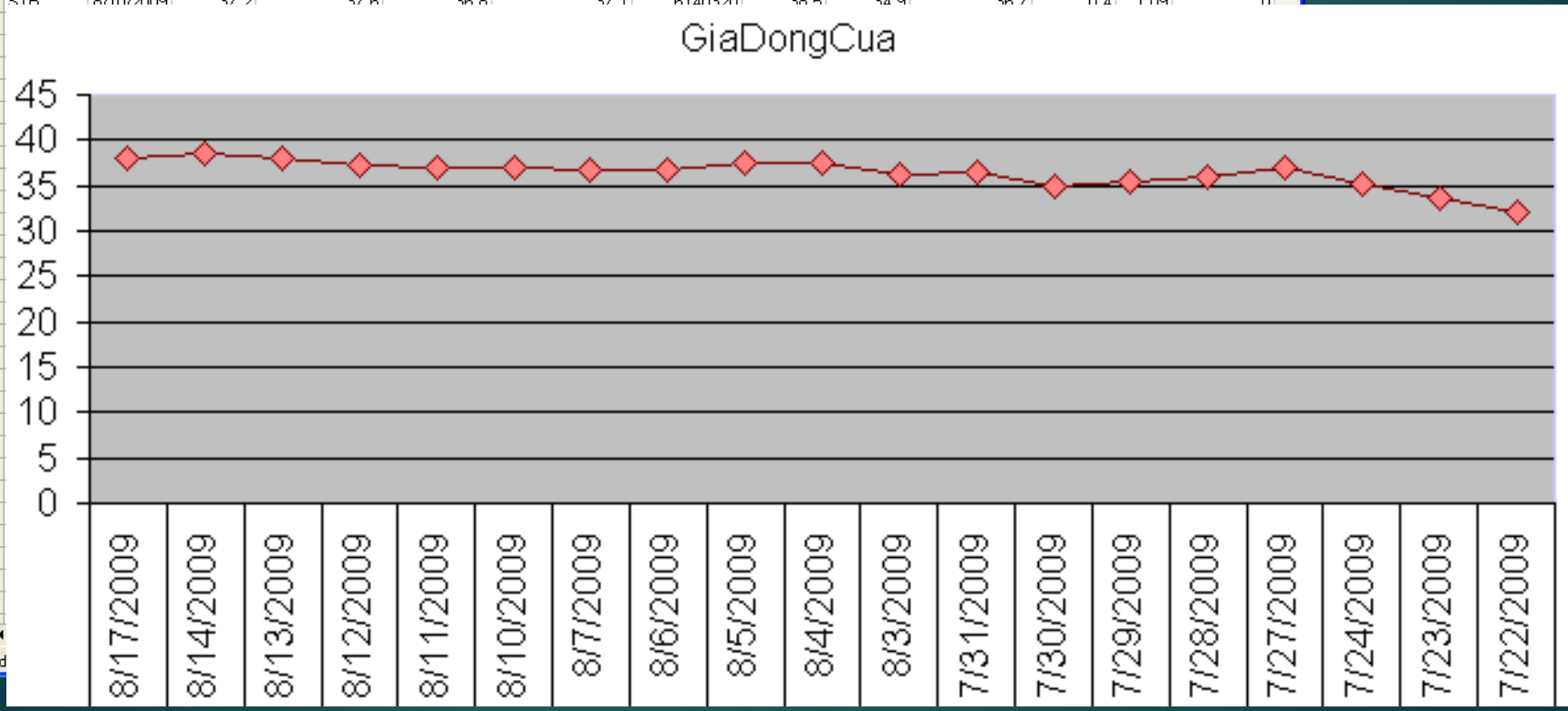
Type a question for help

Arial 10

Reply with Changes... End Review...

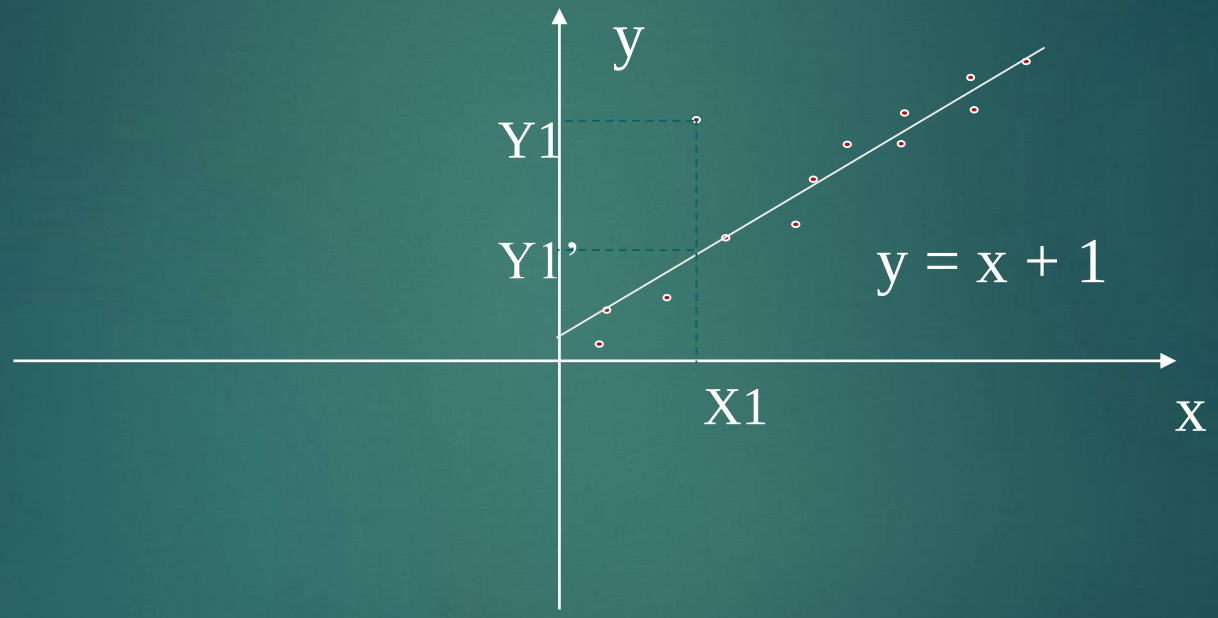
A1 MaCK

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | MaCK | Ngay | GiaMoCua | GiaCaoNhat | GiaThapNhat | GiaDongCua | KhoiLuongGD | GiaTran | GiaSan | GiaThamChieu | TangGiam | % | GDThoaThua |
| 2 | STB | 8/17/2009 | 38.5 | 38.8 | 38.1 | 38.1 | 5986700 | 40.4 | 36.6 | 38.5 | -0.4 | -1.04 | 24343 |
| 3 | STB | 8/14/2009 | 38 | 38.7 | 38 | 38.5 | 6886430 | 39.9 | 36.1 | 38 | 0.5 | 1.32 | 340000 |
| 4 | STB | 8/13/2009 | 38 | 38.5 | 37.6 | 38 | 8716920 | 39 | 35.4 | 37.2 | 0.8 | 2.15 | 188000 |
| 5 | STB | 8/12/2009 | 37.3 | 37.4 | 37 | 37.2 | 5361890 | 38.7 | 35.1 | 36.9 | 0.3 | 0.81 | 200000 |
| 6 | STB | 8/11/2009 | 37.1 | 37.3 | 36.9 | 36.9 | 3675610 | 38.9 | 35.3 | 37.1 | -0.2 | -0.54 | 0 |
| 7 | STB | 8/10/2009 | 37.2 | 37.6 | 36.8 | 37.1 | 6140320 | 38.5 | 34.9 | 36.7 | 0.4 | 1.09 | 0 |



# 3.0. Tình huống 2






Mô hình phân bố dữ liệu của *y* theo *x*???

# 3.1. Tổng quan về hồi qui



- Định nghĩa - Hồi qui (regression)
  - J. Han et al (2001, 2006): Hồi qui là kỹ thuật thống kê cho phép dự đoán các trị (số) liên tục.
  - Wiki (2009): Hồi qui (Phân tích hồi qui – regression analysis) là kỹ thuật thống kê cho phép ước lượng các mối liên kết giữa các biến
  - R. D. Snee (1977): Hồi qui (Phân tích hồi qui) là kỹ thuật thống kê trong lĩnh vực phân tích dữ liệu và xây dựng các mô hình từ thực nghiệm, cho phép mô hình hồi qui vừa được khám phá được dùng cho mục đích dự báo (prediction), điều khiển (control), hay học (learn) cơ chế đã tạo ra dữ liệu.

# 3.1. Tổng quan về hồi qui



- Mô hình hồi qui (regression model): mô hình mô tả mối liên kết (relationship) giữa một tập các biến dự báo (predictor variables/independent variables) và một hay nhiều đáp ứng (responses/dependent variables).

- Phân loại
  - Hồi qui tuyến tính (linear) và phi tuyến (nonlinear)
  - Hồi qui đơn biến (single) và đa biến (multiple)

# 3.1. Tổng quan về hồi qui



- Phương trình hồi qui: $\mathbf{Y} = f(\mathbf{X}, \boldsymbol{\beta})$
  - **X**: các biến dự báo (predictor/independent variables)
  - **Y**: các đáp ứng (responses/dependent variables)
  - **β**: các hệ số hồi qui (regression coefficients)

X dùng để giải thích sự biến đổi của các đáp ứng Y.

Y dùng đề mô tả các hiện tượng (phenomenon) được quan tâm/giải thích.

Quan hệ giữa Y và X được diễn tả bởi sự phụ thuộc hàm của Y đối với X.

β mô tả sự ảnh hưởng của X đối với Y.

# 3.1. Tổng quan về hồi qui



- Phân loại
  - Hồi qui tuyến tính (linear) và phi tuyến (nonlinear)
    - Linear in parameters: kết hợp tuyến tính các thông số tạo nên Y
    - Nonlinear in parameters: kết hợp phi tuyến các thông số tạo nên Y
  - Hồi qui đơn biến (single) và đa biến (multiple)
    - Single: $X = (X_1)$
    - Multiple: $X = (X_1, X_2, \ldots, X_k)$

# 3.2. Hồi qui tuyến tính



- Hồi qui tuyến tính đơn biến
- Hồi qui tuyến tính đa biến

# 3.2.1. Hồi qui tuyến tính đơn biến

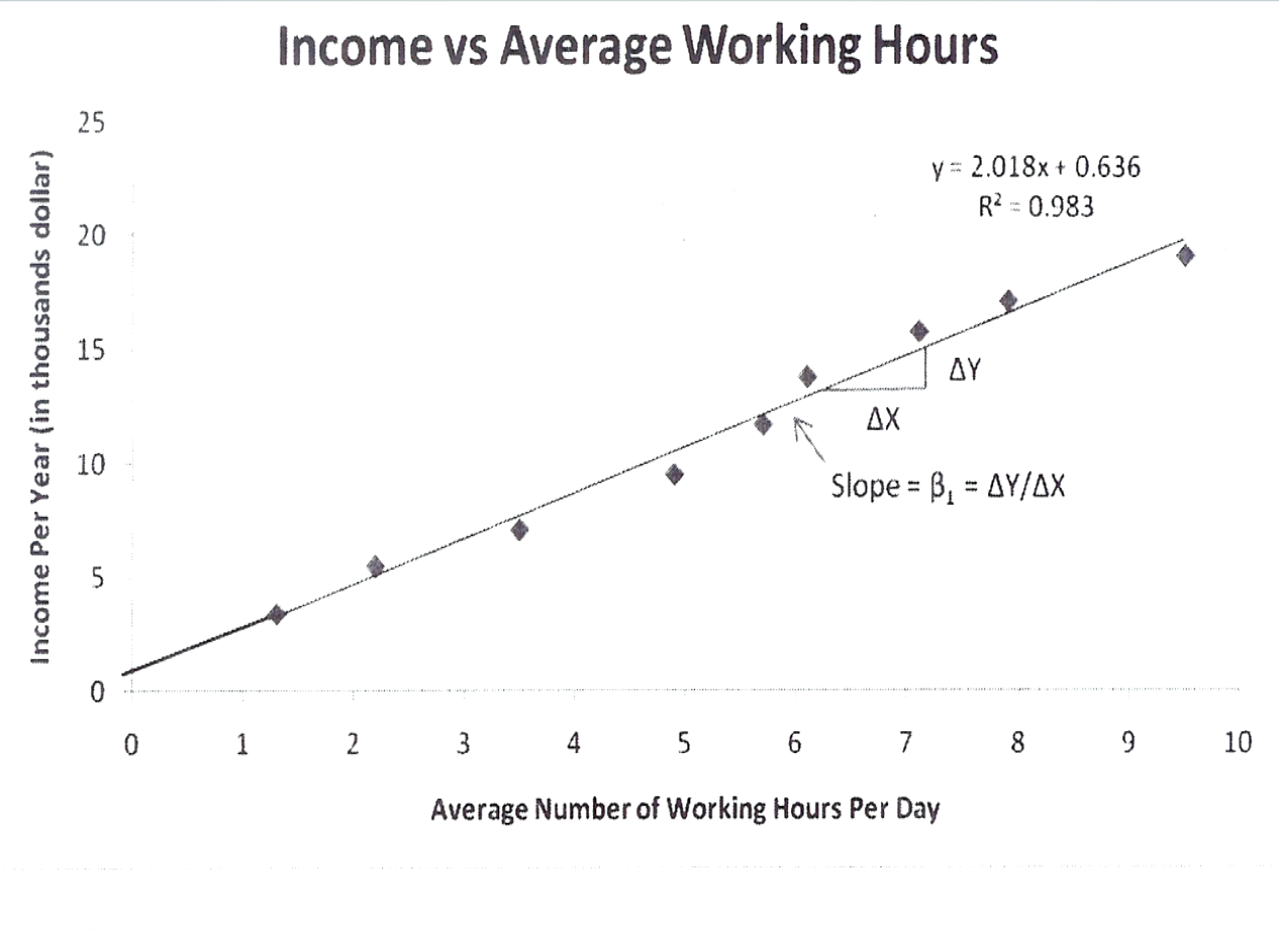


Cho N đối tượng đã được quan sát, mô hình hồi qui tuyến tính đơn biến được cho dưới dạng sau:

$$y = w_0 + w_1 x.$$

# 3.2.1. Hồi qui tuyến tính đơn biến






- $Y = \beta 0 + \mathbf{\beta 1} * X1 \rightarrow Y = 0.636 + 2.018 * X$
- Dấu của β1 cho biết sự ảnh hưởng của X đối với Y.

# 3.2.1. Hồi qui tuyến tính đơn biến



- Ước lượng bộ thông số để đạt được mô hình hồi qui tuyến tính đơn biến

$$y = w_0 + w_1 x.$$

$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2}$$

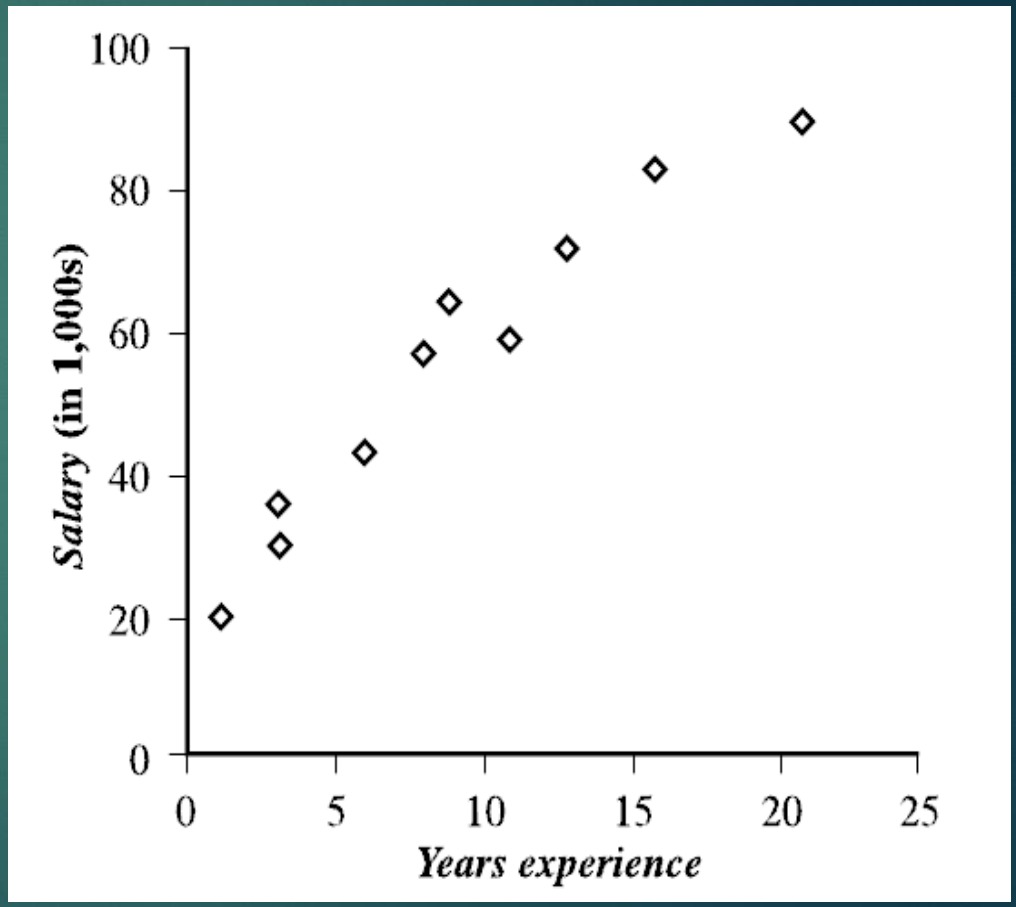
$$w_0 = \bar{y} - w_1 \bar{x}$$

# 3.2.1. Hồi qui tuyến tính đơn biến (example)



Salary data.

| *x* years experience | *y* salary (in $1000s) |
|---|---|
| 3 | 30 |
| 8 | 57 |
| 9 | 64 |
| 13 | 72 |
| 3 | 36 |
| 6 | 43 |
| 11 | 59 |
| 21 | 90 |
| 1 | 20 |
| 16 | 83 |



# 3.2.2. Hồi qui tuyến tính đa biến



- Hồi qui tuyến tính đa biến: phân tích mối quan hệ giữa biến phụ thuộc (response/dependent variable) và hai hay nhiều biến độc lập (independent variables)

$$y_i = b_0 + b_1x_{i1} + b_2x_{i2} + \ldots + b_kx_{ik}$$

i = 1..n với n là số đối tượng đã quan sát
k = số biến độc lập (số thuộc tính/tiêu chí/yếu tố…)
Y = biến phụ thuộc
X = biến độc lập
$b_{0..k}$ = trị của các hệ số hồi qui

# 3.2.2. Hồi qui tuyến tính đa biến

Trị ước lượng của Y

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

Trị ước lượng của bộ thông số b

$$\mathbf{b} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{Y}$$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}$$

# 3.2.2. Hồi qui tuyến tính đa biến



- Example: a sales manager of Tackey Toys, needs to **predict sales** of Tackey products in selected market area. He believes that **advertising expenditures** and **the population in each market area** can be used to predict sales. He gathered sample of toy sales, advertising expenditures and the population as below. **Find the linear multiple regression equation which the best fit to the data.**

# 3.2.2. Hồi qui tuyến tính đa biến



| Market Area | Advertising Expenditures (Thousands of Dollars) $x_1$ | Population (Thousands) $x_2$ | Toy sales (Thousands of Dollars) $y$ |
|---|---|---|---|
| A | 1.0 | 200 | 100 |
| B | 5.0 | 700 | 300 |
| C | 8.0 | 800 | 400 |
| D | 6.0 | 400 | 200 |
| E | 3.0 | 100 | 100 |
| F | 10.0 | 600 | 400 |

# 3.2.2. Hồi qui tuyến tính đa biến

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.98674032 |
| R Square | 0.97365646 |
| Adjusted R Square | 0.9560941 |
| Standard Error | 28.8827296 |
| Observations | 6 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 2 | 92497.3638 | 46248.68 | 55.43996 | 0.004275739 |
| Residual | 3 | 2502.636204 | 834.2121 | | |
| Total | 5 | 95000 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | 6.39718805 | 25.98627725 | 0.246176 | 0.821429 | -76.30282156 | 89.097198 | -76.302822 | 89.0971977 |
| X Variable 1 | 20.4920914 | 5.882177186 | 3.48376 | 0.039947 | 1.772360776 | 39.211822 | 1.77236078 | 39.211822 |
| X Variable 2 | 0.28049209 | 0.068601659 | 4.088707 | 0.026442 | 0.062170791 | 0.4988134 | 0.06217079 | 0.49881339 |

$$\hat{y} = 6.3972 + 20.4921x_1 + 0.2805x_2$$

# 3.3. Hồi qui phi tuyến



- **Y** = f(**X**, **β**)
  - Y là hàm phi tuyến cho việc kết hợp thông số β.
  - Ví dụ: hàm mũ, hàm logarit, hàm Gauss, …

$$f(x, \beta) = \frac{\beta_1 x}{\beta_2 + x}$$

  - Biến đổi sang hàm tuyến tính

$$y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

$$x_1 = x \qquad x_2 = x^2 \qquad x_3 = x^3$$

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$$

# 3.4. Ứng dụng

- Quá trình khai phá dữ liệu
  - Giai đoạn tiền xử lý dữ liệu
  - Giai đoạn khai phá dữ liệu
    - Khai phá dữ liệu có tính mô tả
    - Khai phá dữ liệu có tính dự báo
- Các lĩnh vực ứng dụng: sinh học (biology), nông nghiệp (agriculture), xã hội (social issues), kinh tế (economy), kinh doanh (business), …

# 3.5. Các vấn đề với hồi qui

- Các giả định (assumptions) đi kèm với bài toán hồi qui.
- Lượng dữ liệu được xử lý.
- Đánh giá mô hình hồi qui.
- Các kỹ thuật tiên tiến cho hồi qui:
  - Artificial Neural Network (ANN)
  - Support Vector Machine (SVM)

# 3.6. Tóm tắt

- Hồi qui
  - Kỹ thuật thống kê, được áp dụng cho các thuộc tính liên tục (continuous attributes/features)
  - Có lịch sử phát triển lâu đời
  - Đơn giản nhưng rất hữu dụng, được ứng dụng rộng rãi
  - Cho thấy sự đóng góp đáng kể của lĩnh vực thống kê trong lĩnh vực khai phá dữ liệu
- Các dạng mô hình hồi qui: tuyến tính/phi tuyến, đơn biến/đa biến, đối xứng/bất đối xứng

# Hỏi & Đáp …