LINEAR REGRESSION

Classification and Prediction

Lê Hồng Phương

<*phuonglh@gmail.com*>
Vietnam National University of Hanoi
Hanoi University of Science

January 2015

Nội dung

Nội dung

Nội dung

Nội dung

Nội dung

Phân tích hồi quy

- **Phân tích hồi quy** nghiên cứu sự phụ thuộc của một *biến trả lời* (response variable) vào một hoặc nhiều *biến dự báo* (predictors).
  - y: biến trả lời
  - $(x_1, x_2, \ldots, x_D)$: các biến dự báo
- Đây là kĩ thuật rất cơ bản của ngành thống kê toán học. Để hiểu được nhiều mô hình phân loại, dự báo hiện đại, ta cần hiểu rõ các phương pháp phân tích hồi quy cổ điển.
- Có 2 dạng phân tích hồi quy:
  - hồi quy tuyến tính
  - hồi quy phi tuyến
- Phân tính hồi quy đơn: $y \in \mathbb{R}$; phân tích hồi quy bội $y \in \mathbb{R}^n, n > 1$.

Phân tích hồi quy

- Biến trả lời y luôn là biến liên tục.
- Các biến dự báo có thể liên tục hoặc rời rạc.
- Hồi quy đơn biến: $D = 1$, hồi quy nhiều biến $D > 1$.

Hồi quy tuyến tính

Hàm dự báo $h_\theta(\mathbf{x})$ được xấp xỉ bởi một hàm tuyến tính của $\mathbf{x}$:

$$h_\theta(\mathbf{x}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_D x_D.$$

Nếu bổ sung thêm đặc trưng cố định $x_0 \equiv 1$ thì ta có thể biểu diễn h dưới dạng

$$h_\theta(\mathbf{x}) = \sum_{j=0}^{D} \theta_j x_j = \theta^T \mathbf{x}\,.$$

Hồi quy tuyến tính

Tập dữ liệu huấn luyện:

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$$

Để ước lượng tham số $\theta \in \mathbb{R}^{D+1}$, ta cực tiểu hóa sai số của mô hình trên tập dữ liệu huấn luyện:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{N} [h_\theta(\mathbf{x}_i) - y_i]^2. \tag{1}$$

Phương pháp này được gọi là **bình phương tối thiểu** (OLS – *Ordinary Least Squares*)

$$J(\theta) \to \min .$$

Hồi quy tuyến tính

Để dự đoán giá trị y cho mỗi đối tượng $\mathbf{x}$, ta dùng hàm $h_\theta(\mathbf{x})$ với sai khác là một *nhiễu ngẫu nhiên* ϵ:

$$y = h_\theta(\mathbf{x}) + \epsilon.$$

Giả sử ϵ tuân theo phân phối chuẩn một chiều

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

Hồi quy tuyến tính

Hàm mật độ của ϵ:

$$P(\epsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right).$$

Từ đó ta có

$$P(y|\,\mathbf{x};\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - h_\theta(\mathbf{x}))^2}{2\sigma^2}\right).$$

Hồi quy tuyến tính

- Log-hợp lí của dữ liệu là:

$$L(\theta) = \sum_{i=1}^{N} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{(y_i - h_\theta(\mathbf{x}_i))^2}{2\sigma^2} \right) \right]$$

$$= N \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - h_\theta(\mathbf{x}_i))^2.$$

- Phương pháp hợp lí cực đại cực tiểu hàm mục tiêu

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{N} [h_\theta(\mathbf{x}_i) - y_i]^2.$$

- Đây chính là hàm sai số trong phương pháp bình phương tối thiểu.

Hồi quy tuyến tính

- Log-hợp lí của dữ liệu là:

$$L(\theta) = \sum_{i=1}^{N} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{(y_i - h_\theta(\mathbf{x}_i))^2}{2\sigma^2} \right) \right]$$

$$= N \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - h_\theta(\mathbf{x}_i))^2.$$

- Phương pháp hợp lí cực đại cực tiểu hàm mục tiêu

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{N} [h_\theta(\mathbf{x}_i) - y_i]^2.$$

- Đây chính là hàm sai số trong phương pháp bình phương tối thiểu.

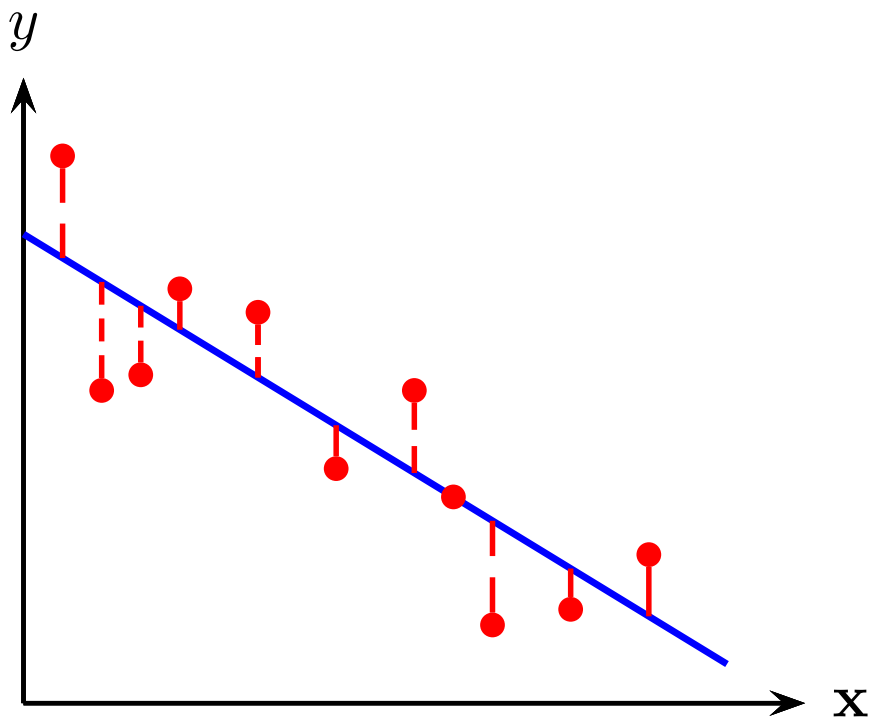
Hồi quy tuyến tính

Như vậy:

- Nếu giả định nhiễu ngẫu nhiên tuân theo phân phối chuẩn thì phương pháp ước lượng hợp lí cực đại dẫn tới phương pháp hồi quy bình phương tối thiểu.
- Ta thấy θ không phụ thuộc vào phương sai σ^2, ngay cả nếu không biết σ^2 thì ta vẫn ước lượng được θ.

Hồi quy tuyến tính

Các điểm dữ liệu $(\mathbf{x}, y)$ được dự báo bởi một siêu phẳng trong không gian nhiều chiều $h_\theta(\mathbf{x}) = \theta^T \mathbf{x}$.



Nội dung

Phương trình chuẩn

Ta có thể tìm được nghiệm đúng dạng giải tích của θ bằng phương trình chuẩn.

- Ta có **ma trận thiết kế** X cỡ $N \times (D+1)$

$$\mathbf{X} = \begin{pmatrix} (\mathbf{x}_1)^T \\ (\mathbf{x}_2)^T \\ \vdots \\ (\mathbf{x}_N)^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1D} \\ 1 & x_{21} & \cdots & x_{2D} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & \cdots & x_{ND} \end{pmatrix}.$$

- Đặt $\mathbf{y}$ là véc-tơ cột chứa tất cả các giá trị của biến dự báo:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}.$$

Phương trình chuẩn

Viết lại hàm mục tiêu dưới dạng ma trận:

$$J(\theta) = \frac{1}{2}(\mathbf{X}\theta - \mathbf{y})^T(\mathbf{X}\theta - \mathbf{y}).$$

Đạo hàm của $J(\theta)$ ứng với θ là

$$\nabla J(\theta) = \mathbf{X}^T\mathbf{X}\theta - \mathbf{X}^T\mathbf{y}.$$

Để cực tiểu hoá J, ta tìm θ thoả mãn $\nabla J(\theta) = 0$, từ đó có nghiệm đúng cho θ là

$$\boxed{\widehat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.}$$

Phương trình trên được gọi là **phương trình chuẩn**.

Phương trình chuẩn

- Nếu tồn tại nghịch đảo của ma trận $\mathbf{X}^T\mathbf{X}$ thì ta mới tìm được nghiệm duy nhất θ theo phương trình chuẩn.
- Nếu nghịch đảo không tồn tại, tức là $\mathbf{X}^T\mathbf{X}$ không đủ hạng thì ước lượng hồi quy là không duy nhất.
  - Tồn tại một phụ thuộc tuyến tính giữa các cột của $\mathbf{X}$.
  - Ta cần tìm cách giảm số chiều của $\mathbf{x}$ bằng việc loại bỏ các đặc trưng phụ thuộc sao cho ma trận thiết kế là đủ hạng.

Một số tính chất của ước lượng hồi quy tuyến tính

Ta thấy $\widehat{\theta}$ là ước lượng không chệch của θ vì

$$\begin{aligned}
\mathbb{E}(\widehat{\theta}|\mathbf{X}) &= \mathbb{E}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}|\mathbf{X}) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\mathbb{E}(\mathbf{y}|\mathbf{X}) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\theta \\
&= \theta.
\end{aligned}$$

Một số tính chất của ước lượng hồi quy tuyến tính

Sử dụng công thức $\text{var}(\mathbf{a} + \mathbf{A}\mathbf{y}) = \mathbf{A}\,\text{var}(\mathbf{y})\mathbf{A}^T$ với $\mathbf{a}, \mathbf{y}$ là các véc-tơ và $\mathbf{A}$ là ma trận hằng số, ta có phương sai của $\widehat{\theta}$ là

$$\begin{aligned}
\text{var}(\widehat{\theta}|\mathbf{X}) &= \text{var}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}|\mathbf{X}) \\
&= [(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\,\text{var}(\mathbf{y}|\mathbf{X})[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}] \\
&= [(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\,[\sigma^2\mathbf{I}]\,[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}] \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.
\end{aligned}$$

Ta thấy phương sai của $\widehat{\theta}$ chỉ phụ thuộc vào $\mathbf{X}$ mà không phụ thuộc vào $\mathbf{y}$.

Phân tích QR

- Trong thực tế, ta ít khi sử dụng trực tiếp phương trình chuẩn để tính toán $\widehat{\theta}$.
  - Lí do: phép lấy nghịch đảo của ma trận $(\mathbf{X}^T\mathbf{X})$ có thể dẫn tới các sai số làm tròn lớn trong quá trình tính toán.
- Hầu hết các phần mềm thống kê sử dụng phương pháp QR để tính toán.

Phân tích QR

Xuất phát từ ma trận $\mathbf{X}$ cỡ $N \times (D+1)$. Giả sử ta có thể tìm được một ma trận $\mathbf{Q}$ cỡ $N \times (D+1)$ và một ma trận $\mathbf{R}$ cỡ $(D+1) \times (D+1)$ sao cho

- $\mathbf{X} = \mathbf{QR}$;
- $\mathbf{Q}$ có các cột trực chuẩn, tức là $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$;
- $\mathbf{R}$ là ma trận tam giác trên, tức là mọi phần tử nằm dưới đường chéo chính đều bằng 0.

Phân tích QR

Sử dụng các tính chất căn bản của đại số ma trận, ta có

$$
\begin{aligned}
\mathbf{X} &= \mathbf{QR} \\
\mathbf{X}^T\mathbf{X} &= (\mathbf{QR})^T(\mathbf{QR}) = \mathbf{R}^T\mathbf{R} \\
(\mathbf{X}^T\mathbf{X})^{-1} &= (\mathbf{R}^T\mathbf{R})^{-1} = \mathbf{R}^{-1}(\mathbf{R}^T)^{-1} \\
\widehat{\theta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \mathbf{R}^{-1}(\mathbf{R}^T)^{-1}(\mathbf{QR})^T\mathbf{y} \\
&= \mathbf{R}^{-1}(\mathbf{R}^T)^{-1}\mathbf{R}^T\mathbf{Q}^T\mathbf{y} \\
&= \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y}.
\end{aligned}
$$

Phân tích QR

Từ đó ta có

$$\mathbf{R}\widehat{\theta} = \mathbf{Q}^T\mathbf{y}.$$

Phương trình này là dễ giải vì $\mathbf{R}$ là ma trận tam giác trên nên ta có thể giải ngược để tìm các tham số. Ví dụ:

$$\begin{pmatrix} 7 & 4 & 2 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix} \widehat{\theta} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$$

- Ta giải phương trình cuối, có $\widehat{\theta}_3 = 1$;
- Thay vào phương trình trước nó có $2\widehat{\theta}_2 + 1 = 2$, từ đó $\widehat{\theta}_2 = 1/2$.
- Phương trình cuối $7\widehat{\theta}_1 + 4\frac{1}{2} + 2 = 3$, từ đó $\widehat{\theta}_1 = -1/7$.

Hồi quy tuyến tính đơn giản

Mô hình hồi quy tuyến tính đơn giản:

$$\mathbb{E}(y|X = x) = h_\theta(x) = \theta_0 + \theta_1 x$$
$$\text{var}(y|X = x) = \sigma^2.$$

Tham số θ_0 gọi là số hạng tự do (số chắn, *intercept*), tham số θ_1 gọi là độ dốc (*slope*). Kí hiệu:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i; \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$
$$\texttt{SXX} = \sum_{i=1}^{N}(x_i - \bar{x})^2$$
$$\texttt{SXY} = \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

Hồi quy tuyến tính đơn giản

Ta có

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

Từ đó

$$(\mathbf{X}^T\mathbf{X}) = \begin{pmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \qquad \mathbf{X}^T\mathbf{y} = \begin{pmatrix} \sum y_i \\ \sum y_i^2 \end{pmatrix}$$

Có thể kiểm tra rằng:

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{\texttt{SXX}} \begin{pmatrix} \sum x_i^2/N & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

Hồi quy tuyến tính đơn giản

Từ đó

$$\widehat{\theta} = \begin{pmatrix} \widehat{\theta}_0 \\ \widehat{\theta}_1 \end{pmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{pmatrix} \bar{y} - \widehat{\theta}_1\bar{x} \\ \texttt{SXY}/\texttt{SXX} \end{pmatrix}$$

Nội dung

Thuật toán giảm gradient

To be continued...

Nội dung

Nhiệt độ và áp suất

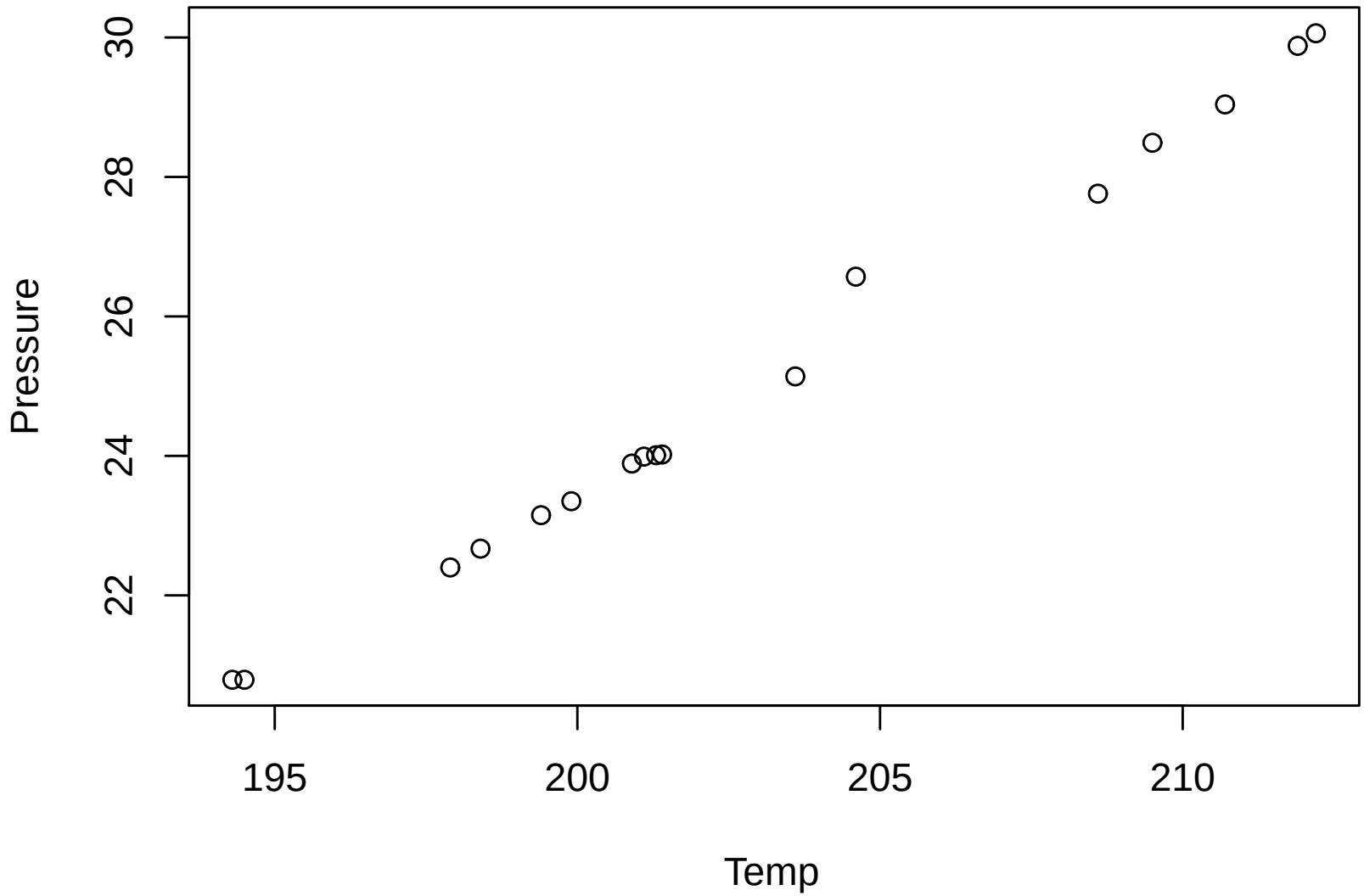
- Vào năm 1857, nhà vật lí học người Scotland James D. Forbes thực hiện một số thí nghiệm nhằm tìm hiểu mối liên hệ giữa áp suất và nhiệt độ sôi của nước.
- Ông biết rằng có thể xác định độ cao từ áp suất không khí đo bằng áp kế: càng lên cao áp suất càng thấp.
- Vào thời gian đó, áp kế thường có độ chính xác không cao. Forbes đề xuất thay thế áp kế bằng nhiệt độ sôi của nước.
- Forbes thu thập dữ liệu ở các dãy Alps và ở Scotland. Tại mỗi điểm, ông đo áp suất (theo inch thủy ngân) bằng áp kế và đo độ sôi của nước (theo độ Fahrenheit) bằng nhiệt kế.

Nhiệt độ và áp suất

Dữ liệu đo ở 17 điểm đo được cho trong bảng sau:

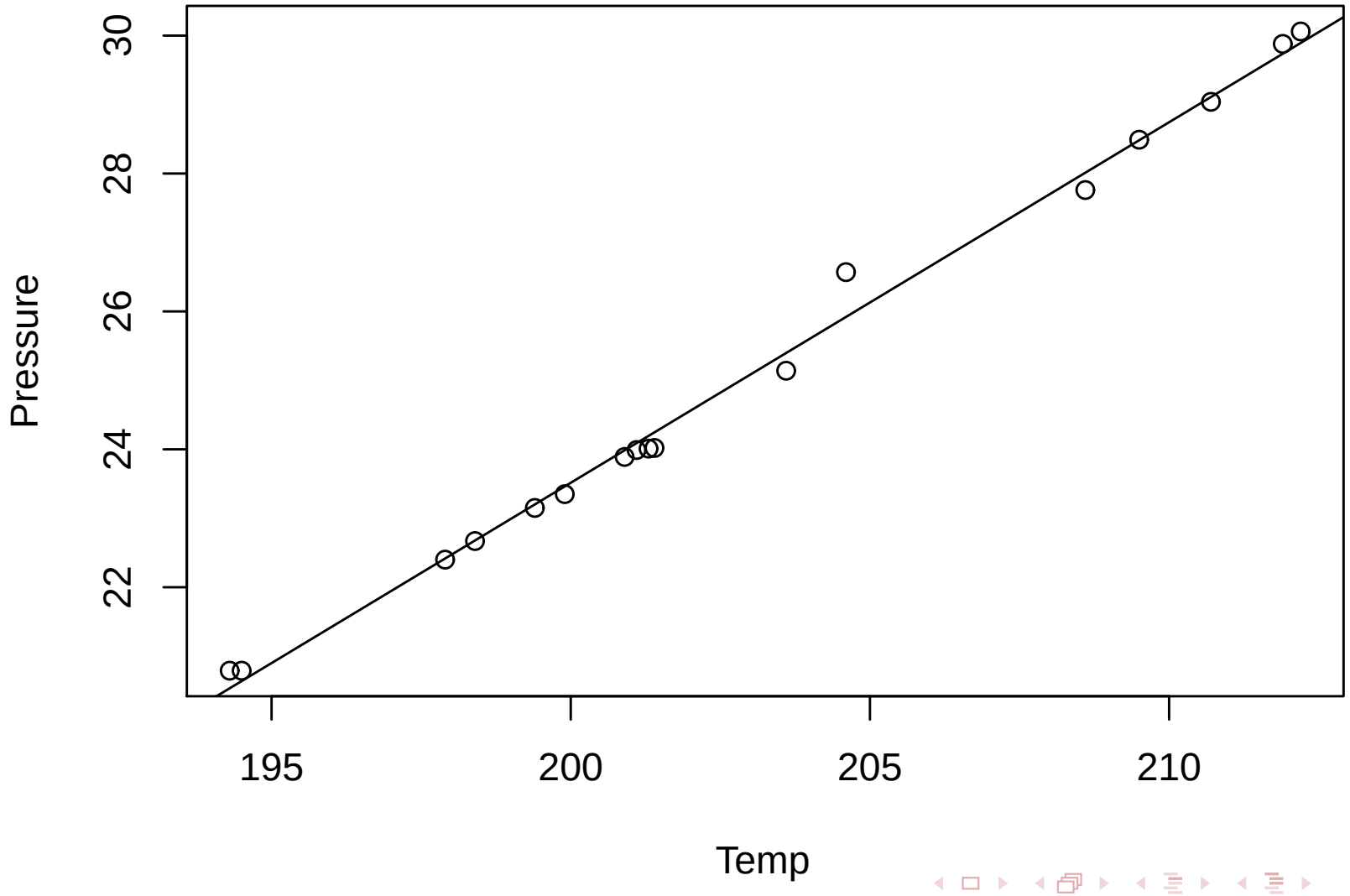
	Nhiệt độ	Áp suất
1	194.5	20.79
2	194.3	20.79
3	197.9	22.40
4	198.4	22.67
5	199.4	23.15
6	199.9	23.35
7	200.9	23.89
8	201.1	23.99
9	201.4	24.02
10	201.3	24.01
11	203.6	25.14
12	204.6	26.57
13	209.5	28.49
14	208.6	27.76
15	210.7	29.04
16	211.9	29.88
17	212.2	30.06

Nhiệt độ và áp suất



Nhiệt độ và áp suất

Ước lượng hồi quy cho kết quả: $h_\theta(x) = -81.06373 + 0.52289x$



Nội dung

Tiêu thụ nhiên liệu

Trong ví dụ này, ta sử dụng mô hình hồi quy tuyến tính để

- Dự báo mức độ tiêu thụ nhiên liệu trong 50 bang của Hoa Kỳ và quận Columbia.
- Tìm hiểu hiệu ứng của tiêu thụ nhiên liệu đối với thuế xăng của các bang.

Tiêu thụ nhiên liệu

Các biến dự báo được sử dụng trong ví dụ[1]. Dữ liệu được thu thập bởi Cục Đường bộ Hoa Kỳ vào năm 2001.

`Drivers`	Số bằng lái được cấp phép trong bang
`FuelC`	Lượng xăng sử dụng cho giao thông đường bộ, theo ngàn gallons
`Income`	Thu nhập bình quân đầu người năm 2000, theo ngàn đôla
`Miles`	Số dặm đường cao tốc của bang được hỗ trợ từ liên bang
`Pop`	Dân số lớn hơn hoặc bằng 16 tuổi
`Tax`	Thuế xăng của bang, theo cents trên một gallon
`State`	Tên bang
`Fuel`	$1000 \times$ `FuelC/Pop`
`Dlic`	$1000 \times$ `Drivers/Pop`
log(`Miles`)	Loga cơ số 2 của `Miles`

[1] Applied Linear Regression, 3rd edition, Sanford Weisberg, Wiley-Interscience, 2005.

Tiêu thụ nhiên liệu

Một số thống kê mô tả dữ liệu tiêu thụ nhiên liệu:

Biến	N	Trung bình	Độ lệch chuẩn	Nhỏ nhất	Trung vị	Lớn nhất
`Tax`	51	20.15	4.5447	7.5	20.0	29.0
`Dlic`	51	903.7	72.858	700.2	909.1	1075.3
`Income`	51	28404	4451.637	20993	27871	40640
`logMiles`	51	15.75	1.4867	10.58	16.27	18.20
`Fuel`	51	613.1	88.96	317.5	626.0	842.8

$$\mathbb{E}(\texttt{Fuel}|X) = \theta_0 + \theta_1\texttt{Tax} + \theta_2\texttt{Dlic} + \theta_3\texttt{Income} + \theta_4\texttt{logMiles}.$$

Tiêu thụ nhiên liệu

Ma trận thiết kế $\mathbf{X}$ và véc-tơ $\mathbf{y}$ là

$$\mathbf{X} = \begin{pmatrix} 1 & 18.00 & 1031.38 & 23471 & 16.5271 \\ 1 & 8.00 & 1031.641 & 30064 & 13.7343 \\ 1 & 18.00 & 908.597 & 25578 & 15.7536 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 25.65 & 904.8936 & 21915 & 15.1751 \\ 1 & 27.30 & 882.329 & 28232 & 16.7817 \\ 1 & 14.00 & 970.7526 & 27230 & 14.7362 \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} 690.264 \\ 514.279 \\ 621.475 \\ \vdots \\ 562.411 \\ 571.794 \\ 842.792 \end{pmatrix}$$

Các cột của ma trận thiết kế tương ứng với hệ số chắn, `Tax`, `Dlic`, `Income` và `logMiles`. Ma trận $\mathbf{X}$ có cỡ 51×5, còn $\mathbf{y}$ có cỡ 51×1.

Tiêu thụ nhiên liệu

Ước lượng hồi quy cho kết quả như sau:

Biến	Hệ số	Sai số chuẩn
`(Intercept)`	154.192845	194.906161
`Tax`	−4.227983	2.030121
`Dlic`	0.471871	0.128513
`Income`	−0.006135	0.002194
`logMiles`	18.545275	6.472174

Nội dung

Hồi quy tuyến tính trong bài toán phân loại

- Mô hình hồi quy tuyến tính được sử dụng để phân loại theo quy tắc:
  - Sau khi ước lượng được tham số θ của mô hình hồi quy, với mỗi $\mathbf{x}$ ta tính
$$\widehat{y} = h_\theta(\mathbf{x}) = \sum_{j=1}^{D} \theta_j x_j.$$
  - Sau đó tùy thuộc vào giá trị của $\widehat{y}$ mà ta quyết định phân $\mathbf{x}$ vào lớp nào.
- Với bài toán phân loại nhị phân, vì $y \in \{0, 1\}$ nên ta có thể sử dụng quy tắc phân loại sau: xếp $\mathbf{x}$ vào lớp 0 nếu $\widehat{y} < 0.5$ và xếp $\mathbf{x}$ vào lớp 1 nếu $\widehat{y} \geq 0.5$.

Chẩn đoán ung thư

- Bộ dữ liệu về bệnh nhân ung thư vú của Đại học Wisconsin–Madison, Hoa Kỳ.[2].
- Để dự báo một bệnh nhân có mắc phải bệnh ung thư vú hay không, người ta lấy mẫu sinh thiết (FNA–Fine Needle Aspiration) của khối u và phân tích mẫu này.
- Lấy một lát cắt của mẫu để soi dưới kính hiển vi, quét và ghi lại mẫu dưới dạng các khung ảnh số của các nhân tế bào.
- Sau khi đã cô lập các nhân tế bào, ta tiến hành tính toán 10 đặc tính của nhân, đo kích thước, hình dạng, kết cấu của chúng.
- Với mỗi đặc trưng này, ta tính toán *giá trị trung bình*, *độ lệch chuẩn*, *các giá trị cực trị*.

[2]Wisconsin Diagnostic Breast Cancer (WDBC), `http://www.cs.wisc.edu/~olvi/uwmp/cancer.html`

Chẩn đoán ung thư

- Có 30 đặc trưng giá trị thực cho mỗi mẫu.
- Tập huấn luyện có 569 mẫu dữ liệu, trong đó có 357 mẫu u lành tính, 212 mẫu u ác tính.
- Với mỗi nhân tế bào, người ta tính 10 đặc trưng sau:

STT.	Đặc trưng	Giải thích
0.	radius	trung bình các khoảng cách từ trung tâm tới các điểm trên chu vi
1.	texture	kết cấu, là độ lệch chuẩn của các giá trị thang xám
2.	perimeter	chu vi
3.	area	diện tích
4.	smoothness	độ trơn, là độ biến đổi cục bộ theo các độ dài bán kính
5.	compactness	độ đặc, tính bởi perimeter^2 / area - 1.0
6.	concavity	độ lõm
7.	concave points	số phần lõm của các đường viền
8.	symmetry	độ đối xứng
9.	fractal dimension	số chiều fractal

Chẩn đoán ung thư

- Mỗi mẫu cần chẩn đoán có 30 đặc trưng:
  - Đặc trưng thứ nhất là trung bình của radius
  - Đặc trưng thứ 11 là độ lệch chuẩn của radius
  - Đặc trưng thứ 21 là radius lớn nhất, được tính bởi giá trị trung bình của 3 giá trị lớn nhất.
- Hai mẫu ví dụ được gán các lớp tương ứng là *ác tính* (M–malignant) và *lành tính* (B–benign).
  1. M, 17.99, 10.38, 122.8, 1001, 0.1184, 0.2776, 0.3001, 0.1471, 0.2419, 0.07871, 1.095, 0.9053, 8.589, 153.4, 0.006399, 0.04904, 0.05373, 0.01587, 0.03003, 0.006193, 25.38, 17.33, 184.6, 2019, 0.1622, 0.6656, 0.7119, 0.2654, 0.4601, 0.1189
  2. B, 7.76, 24.54, 47.92, 181, 0.05263, 0.04362, 0, 0, 0.1587, 0.05884, 0.3857, 1.428, 2.548, 19.15, 0.007189, 0.00466, 0, 0, 0.02676, 0.002783, 9.456, 30.37, 59.16, 268.6, 0.08996, 0.06444, 0, 0, 0.2871, 0.07039

Chẩn đoán ung thư

Mô hình hồi quy tuyến tính sử dụng 10 đặc trưng:

$$\begin{array}{llll} \theta_0 = & 3.0521 & & \\ \theta_1 = & -0.49 & \theta_2 = & -0.022 \\ \theta_3 = & 0.055 & \theta_4 = & 0.001 \\ \theta_5 = & -1.9409 & \theta_6 = & -0.0973 \\ \theta_7 = & -0.8098 & \theta_8 = & -6.431 \\ \theta_9 = & -1.0119 & \theta_{10} = & 0.1193 \end{array}$$

- Mô hình cho độ chính xác trên tập dữ liệu là 93.14%.
- Nếu sử dụng toàn bộ 30 đặc trưng thì độ chính xác của mô hình là 96.48%.

Bài tập

1. Cài đặt thuật toán ước lượng tham số của mô hình hồi quy tuyến tính sử dụng phương trình chuẩn.
2. Cài đặt thuật toán ước lượng tham số của mô hình hồi quy tuyến tính bằng phương pháp giảm gradient.
3. Kiểm tra thực nghiệm kết quả của các thuật toán trên các bộ dữ liệu mẫu (Forbes, Fuel, Breast Cancer).