Chương 6 : Hồi qui

Trịnh Anh Phúc [1]

[1] Bộ môn Khoa Học Máy Tính, Viện CNTT & TT,
Trường Đại Học Bách Khoa Hà Nội

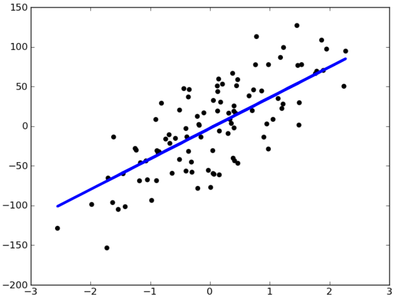
Ngày 23 tháng 5 năm 2014

Giới thiệu

Bài toán hồi qui

Định nghĩa về bài toán hồi qui

Giống bài toán phân loại nhưng không gian giá trị đầu ra $y \in \mathbb{R}$ là giá trị liên tục, vô hạn



Bài toán hồi qui

Ứng dụng trong khai phá dữ liệu

- Sử dụng tập các đối tượng đã được quan sát gồm các cặp $(\mathbf{X}, Y) = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$ trong đó $(\mathbf{x}_i, y_i)$ với $i = 1, \cdots, n$ với $\mathbf{x}_i$ là tập giá trị thuộc tính, còn $y_i \in \mathbb{R}$ là giá trị cần ước lượng
- Sử dụng *lớp các hàm số* $f(\mathbf{X}) \mapsto \mathbb{R}$ trong đó $\mathbf{X}$ là tập giá trị thuộc không gian thuộc tính
- Sử dụng *tiêu chí* xác định tham số của hàm $f(\mathbf{X})$ sao cho ánh xạ có giá trị đầu ra gần y nhất có thể

Bài toán hồi qui

Lớp các hàm số $f(\mathbf{X}) \mapsto \mathbb{R}$

- Lớp các hàm số truyến tính có dạng $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \Leftarrow$ lớp hàm hay dùng
- Lớp các hàm số phi tuyến, e.g. $f(\mathbf{x}) = exp(-||\mathbf{x}||^2)$, $f(\mathbf{x}) = \frac{exp(||\mathbf{x}||^2) - exp(-||\mathbf{x}||^2)}{exp(||\mathbf{x}||^2) + exp(-||\mathbf{x}||^2)} \cdots$

Các tiêu chí

Dựa trên tập các điểm dữ liệu $(\mathbf{X}, Y)$

- Tổng bình phương lỗi $\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2 \Leftarrow$ tiêu chí hay dùng
- Trung bình tổng bình phương lỗi $\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2$
- Tổng trị tuyệt đối lỗi $\sum_{i=1}^{n}|f(\mathbf{x}_i) - y_i|$

Bài toán hồi qui

Nhận xét về bài toán hồi qui áp dụng KPDL

- Áp dụng hoàn toàn thuộc tính có giá trị là số
- Thường dùng chính tiêu chi để đánh giá độ tốt xấu của ánh xạ $f(\mathbf{X}) \mapsto \mathbb{R}$ với tập dự đoán
- Bài toán hồi qui khó hơn bài toán phân loại do không gian đầu ra $y \in \mathbb{R}$ là vô hạn

Dành cho trả lời câu hỏi

Bài toán hồi qui

Mô hình tuyến tính đơn - Simple Linear Regression

Với các thuộc tính $x^1, x^2, \ldots, x^m \equiv \mathbf{x}$ ta sử dụng

- Lớp các hàm tuyến tính một chiều

$$f(x) = ax + b$$

- Tiêu chí đánh giá là tổng bình phương lỗi (sum of squared errors)

$$se = \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2 = \sum_{i=1}^{n} (a x_i^j + b - y_i)^2$$

  để ước lượng a và b sao cho tổng bình phương nhỏ nhất có thể

- Thế ngược để xác định thuộc tính lựa chọn $x^j \in \mathbf{x}$ với $j = \overline{1, m}$

Giải thuật SimpleLinearRegression

- **Đầu vào** : Tập điểm dữ liệu $(\mathbf{X}, Y) = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$
- **Đầu ra** : Ánh xạ $f(x) = ax + b$ và thuộc tính làm cực tiểu hóa bình phương lỗi

Procedure SimpleLinearRegression($\mathbf{X}$, Y)

jmin $\leftarrow$ 1; $(amin, bmin) \leftarrow (0,0)$; $se_{min} \leftarrow \infty$;
for $j \leftarrow 1$ **to** m **do**
    $se \leftarrow 0$
    for $i \leftarrow 1$ **to** n **do** // với x_i^j giá trị thuộc tính j của $\mathbf{x}_i$
        $se \leftarrow se + (ax_i^j + b - y_i)^2$
    endfor
    $(a_j, b_j, se_j) \leftarrow \text{ArgMin}(se)$
    if $(se_{min} > se_j)$ **then** jmin $\leftarrow$ j; $se_{min} \leftarrow se_j$; $(amin, bmin) \leftarrow (a_j, b_j)$ **endif**
endfor
return $(amin, bmin)$ và $jmin$

Giải thuật SimpleLinearRegression

Tổng bình phương lỗi

$$se = \sum_{i=1}^{n}(ax_i^j + b - y_i)^2 = \sum_{i=1}^{n}\Big((ax_i^j)^2 + 2(ax_i^j)(b - y_i) + (b - y_i)^2\Big)$$

$$= \sum_{i=1}^{n}\Big(a^2(x_i^j)^2 + ab(2x_i^j) - a(2x_i^j y_i) + b^2 - b(2y_i) + y_i^2\Big)$$

$$= a^2\underbrace{\sum_{i=1}^{n}(x_i^j)^2}_{A} + ab\underbrace{\sum_{i=1}^{n}2x_i^j}_{B} - a\underbrace{\sum_{i=1}^{n}2x_i^j y_i}_{C} + nb^2 - b\underbrace{\sum_{i=1}^{n}2y_i}_{D} + \underbrace{\sum_{i=1}^{n}y_i^2}_{E}$$

$$= a^2A + abB - aC + nb^2 - bD + E$$

Do hệ số $A > 0$ và $n > 0$ nên bình phương lỗi có cực tiểu toàn cục

$$\partial se/\partial b = aB + 2nb - D = 0$$

$$\partial se/\partial a = 2aA + bB - C = 0$$

Giải thuật SimpleLinearRegression

Giải hệ phương trình

$$aB + b2n - D = 0 \Rightarrow aB + b2n = D$$
$$a2A + bB - C = 0 \Rightarrow a2A + bB = C$$

Các định thức $\Delta = \begin{vmatrix} B & 2n \\ 2A & B \end{vmatrix}$, $\Delta_a = \begin{vmatrix} D & 2n \\ C & B \end{vmatrix}$ và $\Delta_b = \begin{vmatrix} B & D \\ 2A & C \end{vmatrix}$

Vậy

$$a_j = \frac{\Delta_a}{\Delta}; \quad b_j = \frac{\Delta_b}{\Delta}$$

Thế ngược lại phương trình bình phương lỗi, ta có được se_j

Giải thuật SimpleLinearRegression

Thủ tục con ArgMin

- **Đầu vào** : Tổng bình phương lỗi $se = a^2A + abB - aC + nb^2 - bD + E$ với $A > 0$
- **Đầu ra** : Tham số a_j, b_j và giá trị cực tiểu bình phương lỗi se_j

Function ArgMin(se)

1. Tính các giá trị A, B, C, D và E
2. Tính các định thức Δ, Δ_a và Δ_b
3. $a_j \leftarrow \frac{\Delta_a}{\Delta}$; $b_j \leftarrow \frac{\Delta_b}{\Delta}$
4. $se_j \leftarrow se(a_j, b_j)$
5. **return** (a_j, b_j, se_j)

End

Bài toán hồi qui

Xét tập dữ liệu house.arff, để tính thủ tục con ArgMin()

kích thước	đất	phòng	granit	phòng tắm phụ	giá
1076	2801	6	0	0	324500
990	3067	5	1	1	466000
1229	3094	5	0	1	425900
731	4315	4	1	0	387120
671	2926	4	0	1	312100

- Giả sử ta chọn thuộc tính x_j là đất, hay $j = 2$
- Tổng bình phương lỗi $se = a^2A + abB - aC + nb^2 - bD + E$ với $n = 5$ và $m = 5$

Bài toán hồi qui

- Các giá trị tính được gồm

$$A = 54005627, \; B = 32406, \; C = 12479017000,$$
$$D = 3831240, \; E = 751115364400$$

Các định thức

$$\Delta = -29963704, \; \Delta_a = -635006560, \; \Delta_b = -9422011872960$$

Các tham số

$$a_2 = \frac{\Delta_a}{\Delta} = 21.2; \quad b_2 = \frac{\Delta_b}{\Delta} = 314447.5$$

Tổng bình phương lỗi

$$se_2 = 16522497885.39$$

Dành cho trả lời câu hỏi

Bài toán hồi qui

Mô hình hồi qui tuyến tính đơn - Linear Regression Model

Với các thuộc tính $x^1, x^2, \dots, x^m \equiv \mathbf{x} \in \mathbb{R}^m$ ta sử dụng trọng số $w^1, w^2, \dots, w^m \equiv \mathbf{w} \in \mathbb{R}^m$

- Lớp các hàm tuyến tính đa chiều

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$$

  trong đó $\cdot$ là phép nhân vô hướng hai vec tơ $\mathbf{x}$ và $\mathbf{w}$
- Tổng bình phương lỗi $\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2 = \sum_{i=1}^{n}(\mathbf{w} \cdot \mathbf{x}_i + b - y_i)^2$ dùng để ước lượng $\mathbf{w}$ và b

[1]

[1] Các số trên giá trị thuộc tính chỉ thứ tự thuộc tính, không phải số mũ

Bài toán hồi qui

Giải thuật Linear Regression

- **Đầu vào** : Tập điểm dữ liệu $(\mathbf{X}, Y) = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$
- **Đầu ra** : Ánh xạ $f(x) = \mathbf{w} \cdot \mathbf{x} + b$

Function LinearRegresion($(\mathbf{X}, Y)$)

$se \leftarrow 0$

for $i \leftarrow 1$ **to** n **do**

  $se \leftarrow se + (\mathbf{w} \cdot \mathbf{x}_i + b - y_i)^2$

endfor

$(\mathbf{w}, b) \leftarrow \mathrm{ArgMin}(se)$

return $(\mathbf{w}, b)$

End

Bài toán hồi qui

Với phép nhân vô hướng $\mathbf{w} \cdot \mathbf{x} = w^1 x^1 + w^2 x^2 + \cdots w^m x^m$, hệ phương trình đạo hàm bộ phận bậc 1 của bình phương lỗi tạo nên gradient

$$\frac{\partial se}{\partial w^1} = 2\sum_{i=1}^{n}(\mathbf{w} \cdot \mathbf{x}_i + b - y_i)x_i^1 \tag{1}$$

$$\frac{\partial se}{\partial w^2} = 2\sum_{i=1}^{n}(\mathbf{w} \cdot \mathbf{x}_i + b - y_i)x_i^2 \tag{2}$$

$$\vdots \tag{3}$$

$$\frac{\partial se}{\partial w^m} = 2\sum_{i=1}^{n}(\mathbf{w} \cdot \mathbf{x}_i + b - y_i)x_i^m \tag{4}$$

$$\frac{\partial se}{\partial b} = 2\sum_{i=1}^{n}(\mathbf{w} \cdot \mathbf{x}_i + b - y_i) \tag{5}$$

Bài toán hồi qui

Khác với mô hình Simple Linear Regression có thể giải được một cách tường minh, ta thường dùng các kỹ thuật tối ưu - optimization techniques - để ước lượng $\mathbf{w}$ và b

- Bước gradient thấp nhất (Descent gradient)
- Phương pháp gần Newton (Quasi-Newton method)
- Limited Memory-BFGS (Broyden–Fletcher–Goldfarb–Shanno Method)

Vậy thủ tục con ArgMin(se) có thể áp dụng một trong các kỹ thuật tối ưu trên. Nhắc lại các tham số trả lại $(\mathbf{w}, b)$ là không duy nhất cho cùng bộ dữ liệu.

Chương trình Weka

Yêu cầu

- Tải tập dữ liệu cholesterol.arff
- Chạy mô hình Simple Linear Regression và Linear Regression Model [2]
- Giải thích các tham số của các giải thuật

[2] Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference : a practical information-theoretic approach. Springer, New York. Akaike Information Criterion (AIC)