PHÂN TÍCH HỒI QUY

THỰC HÀNH PHƯƠNG PHÁP SỐ CHO KHOA HỌC DỮ LIỆU

Ngày 27 tháng 5 năm 2024

Liên hệ

GOOGLE CLASSROOM: saio6uy

TRỢ GIẢNG:

- Nguyễn Thị Kiều Trang: ntktrang@hcmus.edu.vn
- Lý Như Bình: lnbinh@hcmus.edu.vn

LƯU Ý:

- Email đăng nhập google classroom thể hiện đầy đủ họ và tên, tránh sử dụng email có biệt danh.
- Tiêu đề mail (bắt buộc):
  [2024-HK2-THPPSKHDL] [Tiêu đề thư]
  VD: [2024-HK2-THPPSKHDL] HỎI BÀI
  Vui lòng giới thiệu họ tên, MSSV và tên ca học khi gửi email.

Một vài điều về lớp

Điểm thực hành: Chiếm 30% tổng điểm:

- Điểm danh: 0.5 điểm (Mỗi buổi)
- Bài tập: 2.5 điểm (Nộp bài tập thực hành mỗi tuần)

Cách thức nộp bài:

- Nộp trên google classroom
- Nộp file .txt
- Tên file: Y_MSSV_Hoten_baix.txt,
  - $Y = C204$ nếu bạn học phòng C204.
  - $Y = C203$ nếu bạn học phòng C203.
  - $x \in \{1, 2, 3, 4, ...\}$

Phân tích hồi quy tuyến tính

Bài 1: Cho bảng số liệu sau:

STT	**Diện tích (m^2)**	**Số phòng ngủ**	**Khoảng cách tới TT**	**Giá (tỷ VND)**
1	40	1	30	1.1
2	60	2	32	1.55
3	53	2	30.1	1.68
4	71	2	35.7	1.75
5	80	2	24.5	5.5
6	56	2	27.6	2.3
7	75	2	27.6	3
8	79	2	27.6	3.5
9	56	2	29.7	2.4

STT	Diện tích (m^2)	Số phòng ngủ	Khoảng cách tới TT	Giá (tỷ VND)
10	60	2	29.7	2.9
11	72	2	29.7	3
12	95	3	29.7	4.2
13	47	1	19.3	1.5
14	91	2	18.1	2.2
15	68	1	21.4	1.5
16	69	2	17.5	3.15
17	82	2	25.1	3.4
18	60	2	26.5	2.245
19	68	2	26.5	2.4

Dựa vào bảng số liệu trên, hãy dự đoán giá của một căn nhà có diện tích là $79m^2$, 2 phòng ngủ, khoảng cách tới trung tâm là 26.5 km bằng cách:

a) Giải phương trình đạo hàm mất mát

b) Dùng các thuật toán Gradient descent, Accelerated gradient descent, Stochastic gradient descent.

c) Dùng thư viện scikit-learn

Biết giá trị thực tế của căn nhà trên là 2.5 tỷ VND, hãy so sánh các kết quả trên với nhau.

Thuật toán:

- Đầu vào: Thông tin diện tích, số phòng ngủ và khoảng cách tới trung tâm của một căn nhà.
- Giá tiền dự đoán của căn nhà đó.

Các bước làm bài:
Bài toán tối ưu mà ta cần giải có dạng như sau:

$$\min_{x} L(x) = \min_{x} \frac{1}{2} \sum_{i=1}^{N} (d_i - \overline{w}_i x) \tag{1}$$

trong đó

- $\overline{w}_i = (a_i, b_i, c_i, 1)$ là vector hàng chứa dữ liệu đầu vào của căn nhà thứ i.
- d_i là giá trị của căn nhà thứ i.
- $x = (x_1, x_2, x_3, x_4)^T$ là vector cần phải tối ưu. Nói cách khác, đây là nghiệm của bài toán tối ưu trên.

Câu a: Giải phương trình đạo hàm mất mát.

- Đặt

$$d = \begin{bmatrix} 1.1 \\ 1.55 \\ \vdots \\ 2.4 \end{bmatrix} \text{ là vector cột chứa giá trị các căn nhà.} \tag{2}$$

$$\overline{W} = \begin{bmatrix} 40 & 1 & 30 & 1 \\ 60 & 2 & 32 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 68 & 2 & 26.5 & 1 \end{bmatrix} \tag{3}$$

- Lúc này, bài toán tối ưu sẽ được viết lại dưới dạng ma trận như sau:

$$L(x) = \frac{1}{2}\|d - \overline{W}x\|_2^2 \tag{4}$$

Đạo hàm hàm mất mát ta được:

$$\overline{W}^T\overline{W}x = \overline{W}^T d \tag{5}$$

- Kiểm tra tính khả nghịch của ma trận $\overline{W}^T W$
  - $\overline{W}^T W$ khả nghịch: $x = (\overline{W}^T W)^{-1}\overline{W}^T d$
  - $\overline{W}^T W$ khả nghịch: Áp dụng thuật toán SVD để tìm ma trận giả nghịch đảo của $\overline{W}^T W$. Sau đó tính $x = (\overline{W}^T W)^{+}\overline{W}^T d$

Câu b: Dùng GD, AGD, SGD

Lần lượt giải bài toán (1) bằng các thuật toán GD, AGD và SGD.

Gradien Descent:

- Tính vector gradient $\nabla L(x)$.
- Đặt $i = 0$.
- while $i \leq N$:
  - Tính $x_{t+1} = x_t - \eta \nabla L(x_t)$
  - Nếu $\|\nabla L(x_{t+1})\|_2 < \epsilon$ thì
    Xuất ra màn hình giá trị x_{t+1}
    Dừng lại
  - $x_t = x_{t+1}, i = i + 1$
- Xuất ra màn hình thông báo: thuật toán không thành công sau N bước lặp.

Accelerated gradient descent:

- Tính vector gradient $\nabla L(x)$.
- Đặt $i = 0$ và $x_{i-1} = x_i$
- while $i \leq N - 1$:
  - Tính

$$y_i = x_i + \frac{i-1}{i+2}(x_i - x_{i-1}) \tag{6}$$

$$x_{i+1} = y_i - \eta \nabla L(y_i) \tag{7}$$

  - Nếu $\|\nabla L(x_{t+1})\|_2 < \epsilon$ thì
    Xuất ra màn hình giá trị x_{t+1}
    Dừng lại
  - Cập nhật giá trị $x_{i-1} = x_i, x_i = x_{i+1}$ và $i = i + 1$.
- Xuất ra màn hình thông báo: thuật toán không thành công sau N bước lặp.

Stochastic gradient descent:

- Tính vector gradient $\nabla L_{i_t}(x)$
- Đặt $i = 0, m = len(a)$
- while $i \leq N$:
  - Chọn ngẫu nhiên $i_t \in 1, 2, \ldots, m$
  - Tính $x_{t+1} = x_t - \eta \nabla L_{i_t}(x_t)$
  - Nếu $\|\nabla L(x_{t+1})\|_2 < \epsilon$ thì
    Xuất ra màn hình giá trị x_{t+1}
    Dừng lại
  - $x_t = x_{t+1}, i = i + 1$
- Xuất ra màn hình thông báo: thuật toán không thành công sau N bước lặp.

Câu c: Dùng scikitlearn *Sinh viên đọc về scikit-learn tại đây để làm bài.*

Bài 2: Cho bảng số liệu sau:

STT	**Chiều cao** (cm)	**Cân nặng** (kg)
1	147	49
2	150	50
3	153	51
4	155	52
5	158	54
6	160	56
7	163	58
8	168	60
9	170	72
10	173	63
11	175	64
12	178	66
13	180	67
14	183	68

Bài toán đặt ra là từ bảng số liệu trên, hãy dự đoán cân nặng của một người có chiều cao là 165 *cm* bằng cách:

a) Giải phương trình đạo hàm mất mát.

b) Dùng các thuật toán Gradient descent, Accelerated gradient descent, Stochastic gradient descent.

c) Sử dụng thư viện scikit-learn.

Biết cân nặng thực tế của người đó trên là 59 *kg*, hãy so sánh các kết quả trên với nhau.

Phân tích hồi quy logistic

Bài 3: Cho bảng số liệu sau:

	Lương	Thời gian làm việc	Cho vay
0	10	1.0	1
1	5	2.0	1
2	6	1.8	1
3	7	1.0	1
4	8	2.0	1
5	9	0.5	1
6	4	3.0	1
7	5	2.5	1
8	8	1.0	1
9	4	2.50	1

	Lương	Thời gian làm việc	Cho vay
10	8	0.10	0
11	7	0.15	0
12	4	1.00	0
13	5	0.80	0
14	7	0.30	0
15	4	1.00	0
16	5	0.50	0
17	6	0.30	0
18	7	0.20	0
19	8	0.15	0

a) Từ bảng số liệu trên, áp dụng thuật toán Gradient Descent để viết hàm tính xác suất cho vay của một hồ sơ bất kỳ.

b) Giả sử ngân hàng yêu cầu hồ sơ đạt 80% mới cho vay, hãy vẽ đường phân cách giữa hồ sơ cho vay và không cho vay. Từ đó xác định xem một người có mức lương là 9 triệu và kinh nghiệm làm việc là 0.5 năm thì có được vay hay không?

Thuật toán:

- Đầu vào: Lương và kinh nghiệm làm việc của một bộ hồ sơ.
- Đầu ra: Xác xuất cho vay của hồ sơ đó và quyết định được cho vay hay không.

Các bước làm bài:
Với hồ sơ thứ i mà ta đang xét, ta gọi:

- $x_1^{(i)}$ là lương và $x_2^{(i)}$ là thời gian làm việc của người nộp hồ sơ vay.
- $p(y_i = 1) = \hat{y}_i$ là xác suất mà chúng ta dự đoán hồ sơ được cho vay.
- $p(y_i = 0) = 1 - \hat{y}_i$ là xác suất mà chúng ta dự đoán hồ sơ *không được cho vay*.

Bài toán tối ưu cần giải quyết có dạng như sau:

$$\min_w L(w) = \min_w -\sum_{i=1}^{N}(y_i \ln(\hat{y}_i) + (1-y_i)\ln(1-\hat{y}_i)) \qquad (8)$$

trong đó:

$$\hat{y}_i = \frac{1}{1+e^{-\left(\omega_0+\omega_1 x_1^{(i)}+\omega_2 x_2^{(i)}\right)}} \qquad (9)$$

Bước 1: Tìm vector gradient $\nabla L(w)$
Bước 2: Khởi tạo giá trị ban đầu w_0, dùng Gradient Descent để tìm $w_{t+1} = w_t - \eta \nabla L(w_t)$
Bước 3: Tính phần trăm cho vay $\hat{y}_i$ bằng công thức (9)
Bước 4: Với yêu cầu hồ sơ đạt 80% mới cho vay, hãy khai triển công thức đường phân cách và vẽ nó trên đồ thị.

$$\hat{y}_i > s \Leftrightarrow w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)} > -\ln\left(\frac{1}{s} - 1\right)$$

Bài 4: Với thông số của bài 3, hãy

a) Từ bảng số liệu trên, áp dụng thuật toán Accelerate Gradient Descent để viết hàm tính xác suất cho vay của một hồ sơ bất kỳ.

b) Giả sử ngân hàng yêu cầu hồ sơ đạt 80% mới cho vay, hãy vẽ đường phân cách giữa hồ sơ cho vay và không cho vay. Từ đó xác định xem một người có mức lương là 9 triệu và kinh nghiệm làm việc là 0.5 năm thì có được vay hay không?

Thuật toán: Tương tự bài trên nhưng thay bằng Gradient Descent bằng Accelerate Gradient Descent.