

Hồi quy tuyến tính

Trình bày: PGS.TS Nguyễn Hữu Quỳnh

Giới thiệu

Bài toán:

Cho: 1000 căn nhà trong thành phố, mỗi căn nhà:

- rộng x_1 m²,
- có x_2 phòng ngủ
- cách trung tâm thành phố x_3 km
- giá của mỗi căn nhà

Hỏi: giá của căn nhà thứ 1001 là bao nhiêu?

Giới thiệu

- Hàm dự đoán $y = f(x)$ có dạng thế nào? Ở đây $x = [x_1, x_2, x_3]$

$$f(x) = w_1x_1 + w_2x_2 + w_3x_3 + w_0$$

- Một hàm số có thể mô tả mối quan hệ giữa giá nhà và 3 đại lượng đầu vào:

$$y \approx f(x) = \hat{y}$$

- Mối quan hệ $y \approx f(x)$ là một mối quan hệ tuyến tính (linear).
- Bài toán trên là bài toán thuộc loại regression. Do đó, bài toán đi tìm các hệ số tối ưu $\{w_1, w_2, w_3, w_0\}$ được gọi là bài toán Linear Regression.

Giới thiệu

- y và \hat{y} là hai giá trị khác nhau do có sai số mô hình, tuy nhiên, chúng ta mong muốn rằng sự khác nhau này rất nhỏ.
- *Linear* là *thẳng, phẳng*:
 - trong không gian hai chiều, một hàm số được gọi là tuyến tính nếu đồ thị của nó có dạng một đường thẳng.
 - trong không gian ba chiều, một hàm số được gọi là tuyến tính nếu đồ thị của nó có dạng một mặt phẳng.
 - trong không gian nhiều hơn 3 chiều, là *siêu mặt phẳng (hyperplane)*.

Dạng của Linear Regression

- Trong phương trình: $f(x) = w_1x_1 + w_2x_2 + w_3x_3 + w_0$, nếu chúng ta đặt
 - $w = [w_1, w_2, w_3, w_0]^T$ là vector hệ số cần phải tối ưu
 - $x = [1, x_1, x_2, x_3]$ là vector (hàng) dữ liệu đầu vào
- Phương trình trên có thể được viết lại dưới dạng:

$$y \approx xw = \hat{y}$$

Sai số dự đoán

- Với một cặp (x_i, y_i) $i = 1, 2 \dots N$, chúng ta muốn sự sai khác e_i giữa giá trị thực y_i và giá trị dự đoán \hat{y}_i là nhỏ nhất:

$$\frac{1}{2} e_i^2 = \frac{1}{2} (y_i - \hat{y}_i)^2 = \frac{1}{2} (y_i - x_i w)^2$$

- Hệ số $\frac{1}{2}$ là để thuận tiện cho việc tính toán (khi tính đạo hàm thì số $\frac{1}{2}$ sẽ bị triệt tiêu).
- Chúng ta cần e_i^2 vì $e_i = y_i - \hat{y}_i$ có thể là một số âm, việc nói e_i nhỏ nhất sẽ không đúng vì khi $e_i = -\infty$ là rất nhỏ nhưng sự sai lệch là rất lớn.

Hàm mất mát

- Chúng ta muốn, tổng sai số là nhỏ nhất, tương đương với việc tìm w để hàm số sau đạt giá trị nhỏ nhất:

$$\mathcal{L}(w) = \frac{1}{2} \sum_{i=1}^N (y_i - x_i w)^2$$

- Chúng ta cần tìm vector hệ số w sao cho giá trị của hàm mất mát này càng nhỏ càng tốt

$$w^* = \underset{w}{\operatorname{argmin}} \mathcal{L}(w)$$

Hàm mất mát

- Trước khi đi tìm lời giải, chúng ta đơn giản hóa phép toán trong phương trình hàm mất mát:

$$\mathcal{L}(w) = \frac{1}{2} \sum_{i=1}^N (y_i - x_i w)^2$$

- $y = [y_1; y_2; \dots y_N]$ là một véc tơ cột chứa tất cả nhãn của dữ liệu huấn luyện
- $X = [x_1; x_2; \dots x_N]$ là ma trận dữ liệu đầu vào mà mỗi hàng là một điểm dữ liệu
- Hàm $\mathcal{L}(w)$ được viết thành:

$$\begin{aligned} \mathcal{L}(\mathbf{w}) &= \frac{1}{2} \sum_{i=1}^N (y_i - \bar{\mathbf{x}}_i \mathbf{w})^2 \\ &= \frac{1}{2} \|\mathbf{y} - \bar{\mathbf{X}} \mathbf{w}\|_2^2 \end{aligned}$$

Nghiem cho bài toán Linear Regression

- Cách phổ biến nhất để tìm nghiệm cho một bài toán tối ưu là giải phương trình đạo hàm bằng 0! Nhưng chỉ trường hợp
 - tính đạo hàm và
 - việc giải phương trình đạo hàm bằng 0 không quá phức tạp.
- Với các mô hình tuyến tính, hai việc này là khả thi
- Đạo hàm theo \mathbf{w} của hàm mất mát là:

$$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}} = \mathbf{X}^T (\mathbf{X}\mathbf{w} - \mathbf{y})$$

Thực hành cho bài toán Linear Regression

- Phương trình đạo hàm bằng 0 tương đương với:

$$\mathbf{X}^T \mathbf{X} \mathbf{w} = \mathbf{X}^T \mathbf{y} \quad (*)$$

Đặt $\mathbf{A} \triangleq \mathbf{X}^T \mathbf{X}$ và $\mathbf{X}^T \mathbf{y} \triangleq \mathbf{b}$

- Nếu ma trận vuông \mathbf{A} khả nghịch thì phương trình (*) có nghiệm duy nhất

$$\mathbf{w} = \mathbf{A}^{-1} \mathbf{b}.$$

Bài toán

- Chúng ta có 1 bảng dữ liệu về chiều cao và cân nặng của 15 người:

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	165	59

STT	Chiều cao (cm)	Cân nặng (kg)
9	168	60
10	170	72
11	173	63
12	175	64
13	178	66
14	180	67
15	183	68

Bài toán

- Bài toán đặt ra là: dự đoán cân nặng của một người dựa vào chiều cao của họ
- Ta nhận thấy: cân nặng tỉ lệ thuận với chiều cao (càng cao càng nặng)
- Để kiểm tra độ chính xác của model tìm được, ta sẽ giữ lại dòng 4 và 6 để kiểm thử

$$w = \begin{bmatrix} -33.73541021 \end{bmatrix}$$

$$\begin{bmatrix} 0.55920496 \end{bmatrix}$$

$$y1 = w_1 * 155 + w_0$$

$$y2 = w_1 * 160 + w_0$$