linear regesssion

Giới thiệu sơ qua về Linear Regression

Linear regression là một hàm dự đoán, trong đó `training data` sẽ được sử dụng để xây dựng một linear model[1] để dự đoán. Đối với linear model thì mối quan hệ giữa mục tiêu(dependent variable)[2] và các thuộc tính(independent variable)[3] là tuyến tính.

> 🖊 mối quan hệ tuyến tính:
>
> Mối quan hệ tuyến tính(mối quan hệ bậc nhất) là mối quan hệ mà trong đó quan hệ giữa 2 biến số có thể được biểu diễn dưới dạng đường thẳng. Khi một biến thay đổi thì biến còn lại sẽ được thay đổi theo tỉ lệ cố định.
>
> Đối với Linear regression thì ta có công thức tổng quát như sau:
>
> $$y = b + \sum_{i=1}^{N} w_i x_i$$
>
> trong đó:
> y là biến target
> b là bias dùng để nâng hoặc hạ đường thẳng dự đoán.
> w_i là các thuộc tính đầu vào.

Đối với Linear regression thì có thể có một hoặc nhiều independent variable. Trong trường hợp chỉ có một independent variable thì công thức tổng quát sẽ có dạng $y = mx + b$, trong đó y là target, m sẽ là độ dốc của đường thẳng dự đoán và b là bias. Nhưng trong hầu hết các trường hợp thì các independent variable sẽ có nhiều hơn một, khi đó công thức sẽ có dạng $y = b + \sum_{i=1}^{N} w_i x_i$ như ở trên. Với N independent variables và w_i là trọng số[4] tương ứng với mỗi independent variable.

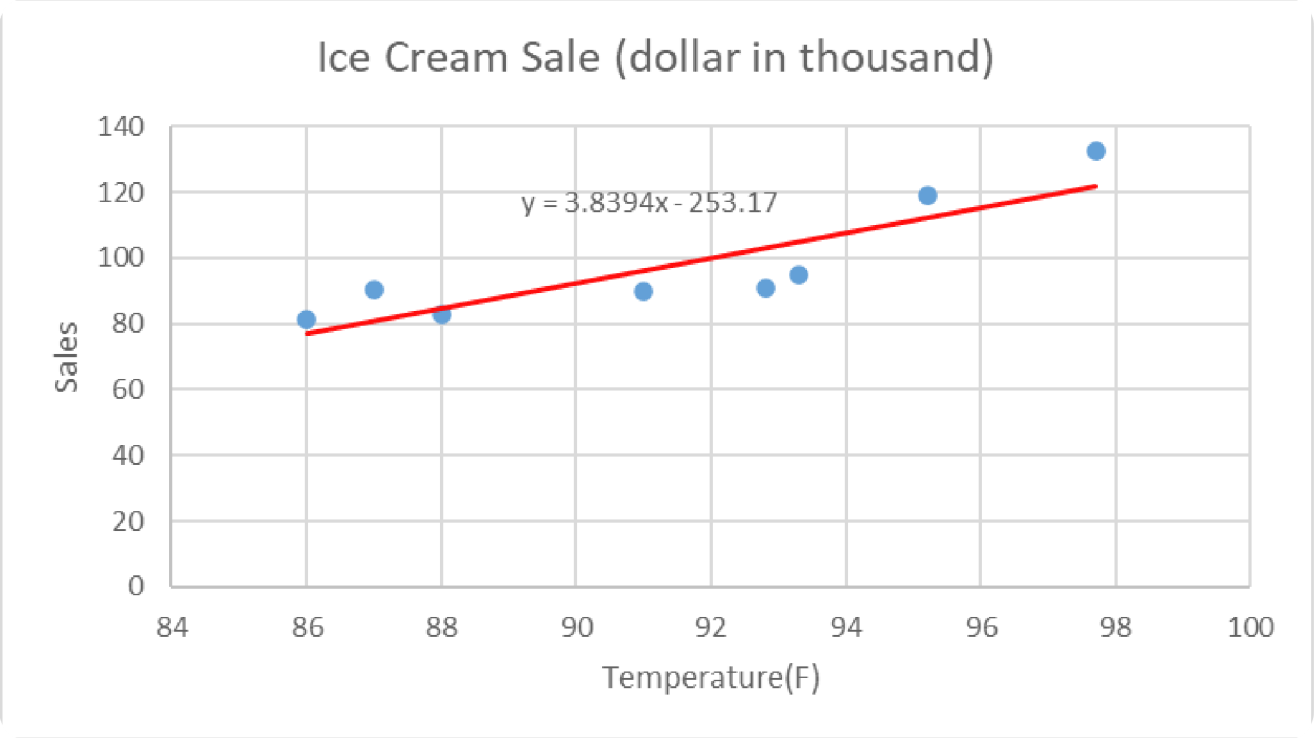
Để xây dựng linear model, chúng ta phải tìm ra được các giá trị của w_i và b dựa trên các giá trị đã biết của target và các independent variable trong `training data`.

Lấy ví dụ như sau:
In a southern beach town, Tommy, the manager of a supermarket is thinking about predicting ice cream sales based on the weather forecast. He has collected some data that relate weekly averaged daily high temperatures to ice cream sales during a summer season. The training data are presented in Table 2-1

Table 2-1. *Ice cream sale vs. temperature*

Temperature (F)	Ice Cream Sale (Thousands of Dollars)
91	89.8
87	90.2
86	81.1
88	83.0
92.8	90.9
95.2	119.0
93.3	94.9
97.7	132.4

Với phần dữ liệu này: ta có thể tạo đường dự đoán dựa trên các điểm dữ liệu đã biết như hình dưới đây:



Khi đó: $m = 3.8394$ và $b = -253.17$. Vậy công thức sẽ là $y = 3.8394x - 253.17$

Với giá trị m và b ta có thể dự đoán được giá trị của y khi biết giá trị của x. Ví dụ: nếu nhiệt độ trung bình hàng ngày là 88.8 độ F thì ta có thể dự đoán doanh số bán kem như sau:

$$y = 3.8394 * 88.8 - 253.17 = 87.8$$

Vậy m và b được tính như thế nào?

Least Squares

Để tìm được các giá trị của m và b thì ta sẽ sử dụng phương pháp Least Squares(phương pháp bình phương tối thiểu). Ý tưởng chính của phương pháp này là tìm ra đường thẳng sao cho tổng bình phương khoảng cách từ các điểm dữ liệu đến đường thẳng là nhỏ nhất.

Vị trí của tất cả các điểm dữ liệu so với đường dự đoán đều có một khoảng cách nhất định (được gọi là residuals). Mục tiêu của chúng ta là tìm ra đường thẳng sao cho tổng bình phương của tất cả các khoảng cách này là nhỏ nhất.

Tổng của bình phương các khoảng cách này được tính như sau:

$$E = \sum_{i=1}^{n} (y_i - (mx_i + b))^2$$

Trong đó:
y_i là giá trị thực tế của điểm dữ liệu thứ i trong `training data`.

Chúng ta có thể lấy **partial derivative** của E đối với m và b để tìm ra giá trị tối ưu của chúng.

$$\frac{\partial E}{\partial m} = -2\sum_{i=1}^{n} x_i(y_i - (mx_i + b)) = 0$$

$$\frac{\partial E}{\partial b} = -2\sum_{i=1}^{n} (y_i - (mx_i + b)) = 0$$

Giải hệ phương trình này ta sẽ tìm được giá trị của m và b.

$$m = \frac{n\left(\sum_{i=1}^{N} x_i y_i\right) - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{n\left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum x_i\right)^2}$$

$$b = \frac{\left(\sum_{i=1}^{N} x_i^2\right)\left(\sum_{i=1}^{N} y_i\right) - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} x_i y_i\right)}{n\left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)^2} = \frac{\left(\sum_{i=1}^{N} y_i\right) - m\left(\sum_{i=1}^{N} x_i\right)}{n}$$

Linear regression though excel

tính tay

Đối với ví dụ ban đầu:

	A	B	C	D
1	x	y		
2	91	89.8		
3	87	90.2	SUM(X)	=SUM(A:A)
4	86	81.1	SUM(Y)	=SUM(B:B)
5	88	83	SUM(XY)	=SUMPRODUCT(A:A, B:B)
6	92.8	90.9	SUM(X^2)	=SUMPRODUCT(A:A,A:A)
7	95.2	119	n	=COUNT(A:A)
8	93.3	94.9	m	=(D7*D5-D3*D4)/(D7*D6-D3^2)
9	97.7	132.4	b	=(D6*D4-D3*D5)/(D7*D6-D3^2)

Figure 2-4. *Computing m and b via least square approximation*

như ta có thể thấy:
$Sum(x) = Sum(A:A)$ đây là tổng các giá trị nhiệt độ, cũng là $\sum_{i=1}^{N} x_i$.

$Sum(y) = Sum(B:B)$ đây là tổng các giá trị doanh số bán kem, cũng là $\sum_{i=1}^{N} y_i$.

$Sum(xy) = SumProduct(A:A, B:B)$ đây là tổng tích của từng cặp giá trị nhiệt độ và doanh số bán kem, cũng là $\sum_{i=1}^{N} x_i y_i$.

$Sum(x^2) = SumProduct(A:A, A:A)$ đây là tổng bình phương của các giá trị nhiệt độ, cũng là $\sum_{i=1}^{N} x_i^2$.

Khi đó ta có thể tính m và b bằng:
$m = (n * Sum(xy) - Sum(x) * Sum(y))/(n * Sum(x^2) - Sum(x)^2)$ hay
$m = (D7 * D5 - D3 * D4)/(D7 * D6 - D3^2)$

$b = (Sum(x^2) * Sum(y) - Sum(x) * Sum(xy))/(n * Sum(x^2) - Sum(x)^2)$ hay
$b = (D6 * D4 - D3 * D5)/(D7 * D6 - D3^2)$)

Khi đó ta có thể dự đoán các giá trị dựa vào công thức $y = mx + b$.

Scoring Data	Prediction
88.8	87.8
96.9	118.9
94.7	110.4

sử dụng hàm tích hợp sẵn của excel

Ngoài cách tính tay như trên, chúng ta cũng có thể sử dụng hàm tích hợp sẵn của excel để tính m và b.

Hàm để tính m là `SLOPE(known_y's, known_x's)`.
Hàm để tính b là `INTERCEPT(known_y's, known_x's)`.

Với ví dụ ở trên thì công thức là:

$m = SLOPE(B2:B9, A2:A9)$
$b = INTERCEPT(B2:B9, A2:A9)$

> ⚠ Lưu ý:
>
> Chúng ta cũng có thể tính m và b bằng Data Analysis tool của excel. Nhưng các giá trị nhận được này là tĩnh và sẽ không thay đổi khi dữ liệu thay đổi. Còn sử dụng hàm tích hợp sẵn thì các giá trị này sẽ tự động cập nhật khi dữ liệu thay đổi.

bài tập mở rộng

Bảng dưới đây có thêm thuộc tính: Tourist và Sunny Days:

Temperature (F)	Tourists	Sunny Days	Ice Cream Sale (dollar in thousand)
91	998	4	89.8
87	1256	7	90.2
86	791	6	81.1
88	705	5	83
92.8	1089	3	90.9
95.2	1135	6	119
93.3	1076	4	94.9
97.7	1198	7	132.4

Khi đó công thức sẽ là:

$$y = b + w_1x_1 + w_2x_2 + w_3x_3$$

với x_1 là Temperature, x_2 là Tourist, x_3 là Sunny Days.

> Note:
>
> Ta cũng có thể sử dụng hàm Linest và Index để tính các trọng số w_1, w_2, w_3 và b. Khi đó công thức sẽ là:
>
> INDEX(LINEST(known_y's, known_x's, const, stats), row_num, [column_num])
> Trong đó:
>
> - known_y's: là giá trị của biến target (doanh số bán kem)
> - known_x's: là các giá trị của các biến independent variable (Temperature, Tourist, Sunny Days)
> - const: là giá trị logic xác định xem hằng số b có được đặt thành 0 hay không. Nếu const là TRUE hoặc bị bỏ qua thì b được tính toán bình thường. Nếu const là FALSE thì b được đặt thành 0 và các trọng số được điều chỉnh cho phù hợp.
> - stats: là giá trị logic xác định xem các thống kê bổ sung có được trả về hay không. Nếu stats là TRUE thì hàm trả về các thống kê bổ sung về độ phù hợp của mô hình. Nếu stats là FALSE hoặc bị bỏ qua thì chỉ trả về các trọng số.

- row_num: là số hàng trong mảng kết quả mà bạn muốn trả về. Trong trường hợp const là TRUE và stats là TRUE thì hàng đầu tiên sẽ chứa các trọng số cho các biến independent variable và hằng số b. Hàng thứ hai sẽ chứa các thống kê bổ sung như R-squared, standard error, v.v.
- column_num: là số cột trong mảng kết quả mà bạn muốn trả về. Nếu bạn bỏ qua thì toàn bộ hàng được trả về dưới dạng mảng.

Vậy công thức trong excel sẽ là:

- Sunny Days = INDEX(LINEST(D2:D9,A2:C9,TRUE,TRUE),1,1)
- Tourist = INDEX(LINEST(D2:D9,A2:C9,TRUE,TRUE),1,2)
- Temperature = INDEX(LINEST(D2:D9,A2:C9,TRUE,TRUE),1,3)
- Y-in-intercept = INDEX(LINEST(D2:D9,A2:C9,TRUE,TRUE),1,4)

Khi đó ta có thể dự đoán doanh số bán kem dựa trên công thức:

$$y = b + w_1x_1 + w_2x_2 + w_3x_3$$

Hay:

$$y = Y - intercept + Temperature * w_1 + Tourist * w_2 + SunnyDays * w_3$$

Với hàng 2 thì công thức cụ thể sẽ là:

$$A2 * C\$13 + B2 * C\$12 + C2 * C\$11 + C\$14$$

Và Error =$POWER(D2 - E2, 2)$
và Sum of Error hay E sẽ được tính bằng tổng của các Error.

	A	B	C	D	E	F
1	Temperature	Tourists	Sunny Days	Ice Cream Sale (dollar in thousand)	Predicted	Error
2	91	998	4	89.8	88.7706	1.059665
3	87	1256	7	90.2	90.390726	0.036376
4	86	791	6	81.1	81.085358	0.000214
5	88	705	5	83	83.195593	0.038256
6	92.8	1089	3	90.9	89.847272	1.108235
7	95.2	1135	6	119	117.19569	3.25555
8	93.3	1076	4	94.9	97.80896	8.462051
9	97.7	1198	7	132.4	133.0058	0.367
10						
11	Sunny Days	1	5.95647	Sum of Errors	14.327347	
12	Tourists	2	-0.0013			
13	Temperature	3	3.97541			
14	Y-intercept	4	-295.47			

---

1. mô hình tuyến tính
2. Gọi là dependent bởi vì thuộc tính cần dự đoán sẽ phụ thuộc vào các giá trị input
3. Được gọi là independent vì chúng ta giả định rằng các thuộc tính này không phụ thuộc vào giá trị target
4. Một hệ số nhân với mỗi independent variable để biểu thị mức độ ảnh hưởng của thuộc tính đó đến giá trị của target. Nếu trọng số càng lớn thì mức độ ảnh hưởng của thuộc tính đó đến giá trị của target càng cao và ngược lại.