AI VIET NAM – AI COURSE 2024

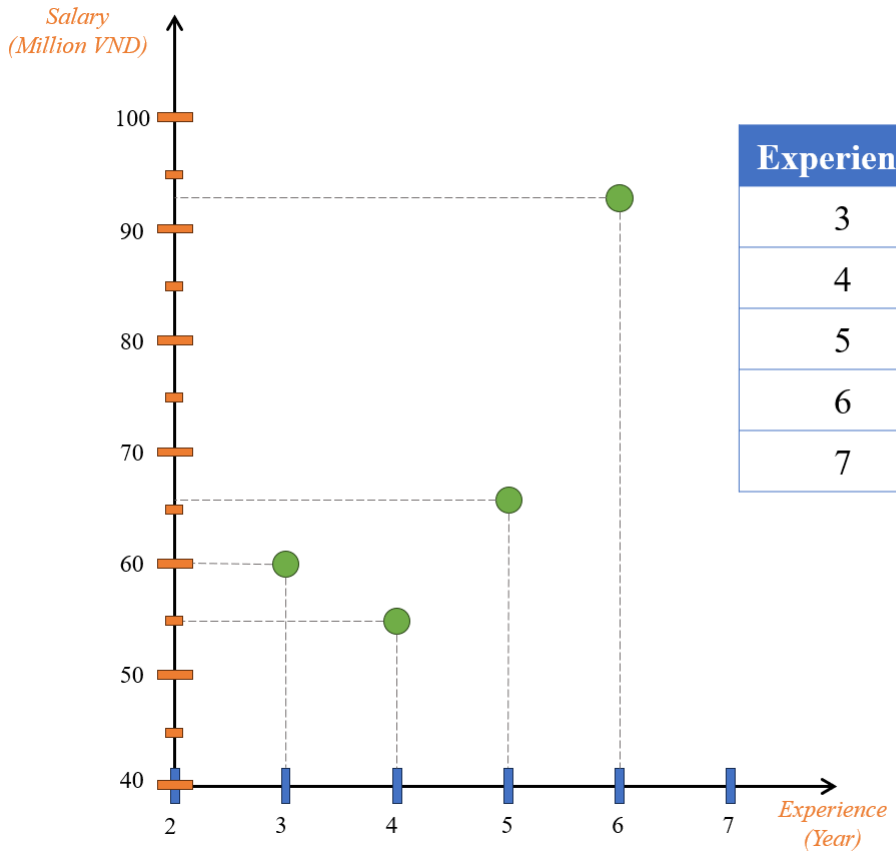
Step-by-Step Simple Linear Regression

Anh-Khoi Nguyen, Dinh-Thang Duong, Quang-Vinh Dinh

Ngày 29 tháng 9 năm 2024

A. Dẫn nhập

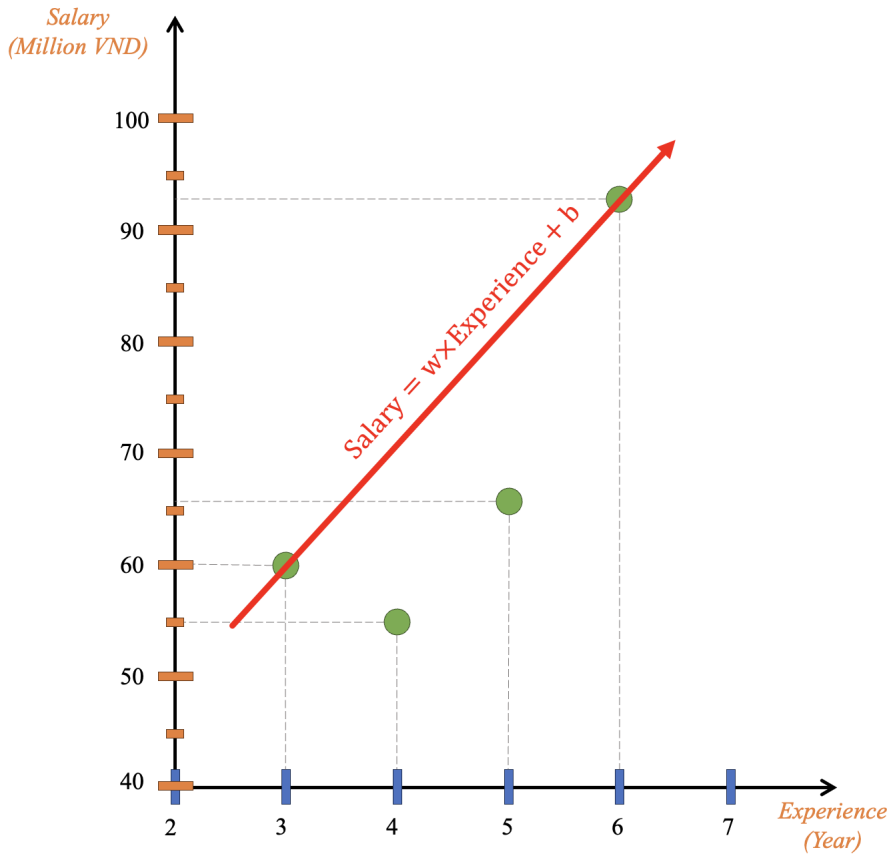
Hãy tưởng tượng chúng ta có một bài toán như sau: *Xây dựng một chương trình tự động dự đoán tiền lương của nhân viên dựa trên số năm kinh nghiệm của họ*. Để trực quan hóa, ta gọi số năm kinh nghiệm là (x) và số tiền lương là (y). Khi đó, với một vài thông tin cho trước của một vài nhân viên, ta có thể sử dụng trục đồ thị Oxy để trực quan hóa dữ liệu như hình sau:



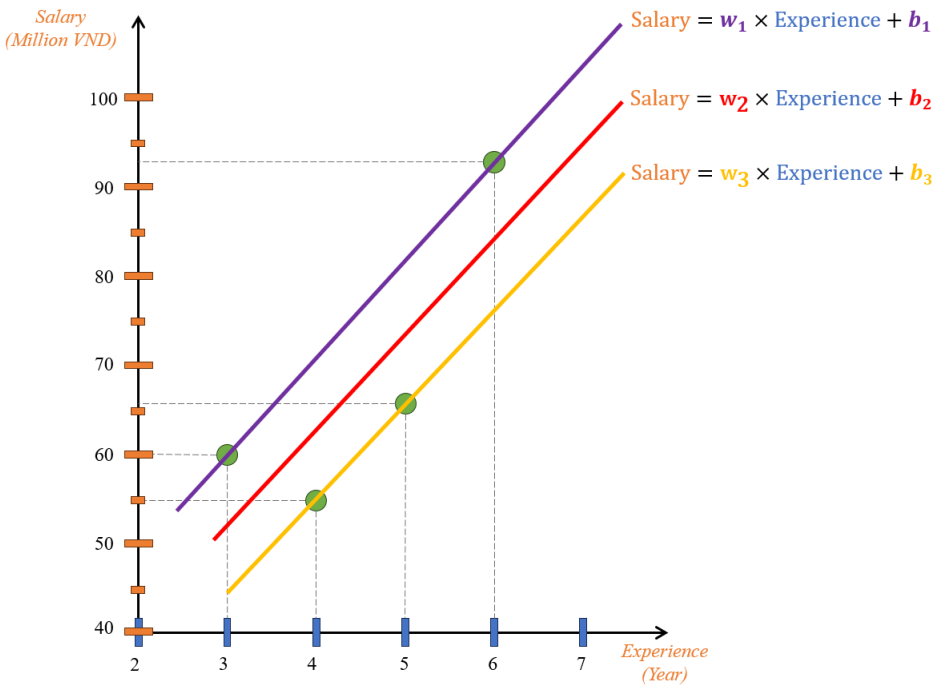
Experience	Salary
3	60
4	55
5	66
6	93
7	?

Câu hỏi đặt ra là, làm cách nào để ta có thể code một chương trình có thể tự động dự đoán tiền lương dựa vào số năm kinh nghiệm? Quan sát thấy rằng, khi ta dùng một đường vẽ ngang

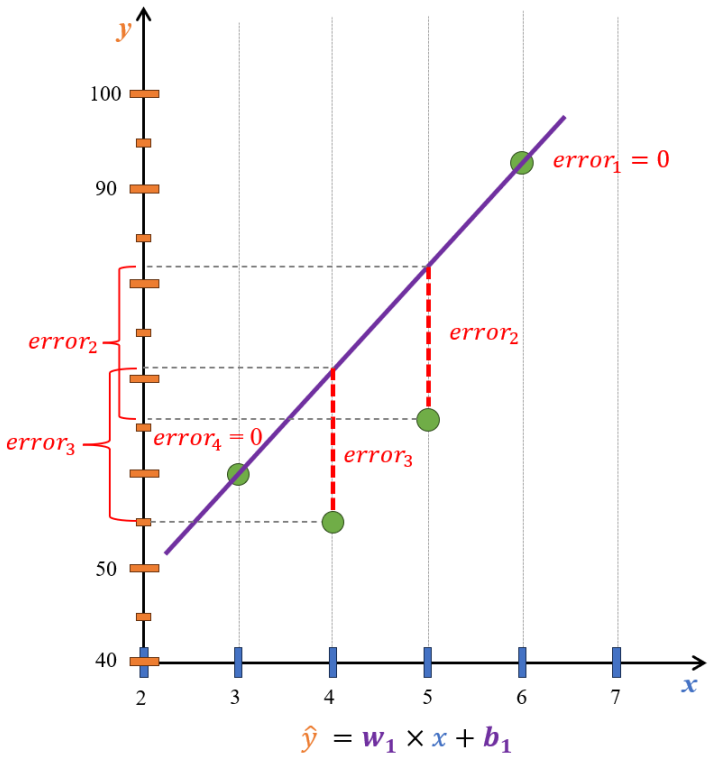
qua một điểm dữ liệu, ta thực chất có thể dùng phương trình đường thẳng này để dự đoán được chính xác tiền lương của nhân viên ứng với số năm kinh nghiệm của họ.



Như vậy, với một đường thẳng bất kỳ, ta hoàn toàn có thể dùng nó cho "chương trình"dự đoán tiền lương. Khi xét một phương trình đường thẳng $f(x) = wx + b$, có thể thấy mỗi đường thẳng khác nhau sẽ có số w và b khác nhau **(1)**.



Mặc dù không thể nào vẽ được một đường thẳng có thể dự đoán chính xác toàn bộ các điểm cho trước. Tuy nhiên, từ quan sát này, với các điểm dữ liệu cho trước như vậy, liệu rằng chúng ta có thể vẽ được một đường thẳng mà có thể xấp xỉ được gần đúng các điểm dữ liệu hiện có hay không?

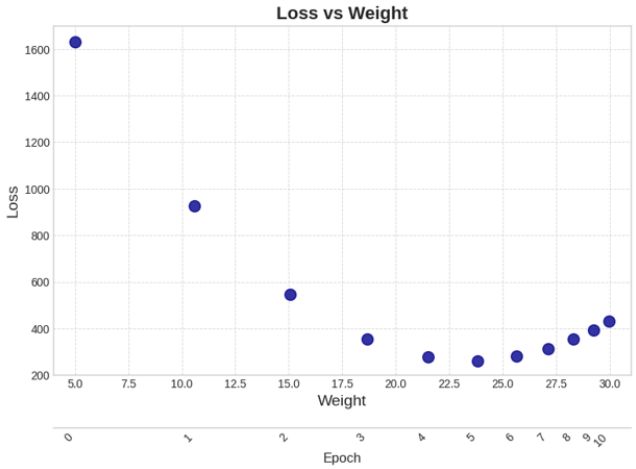


Để tìm được một đường thẳng xấp xỉ tối ưu nhất, ta có thể nghĩ theo hướng rằng, khoảng

cách từ các điểm trong đồ thị đến đường thẳng đó phải có khoảng cách là ngắn nhất **(2)**.

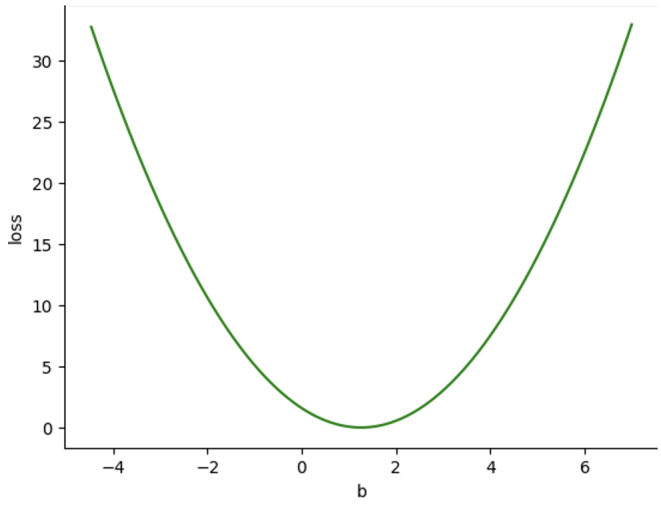
Từ **(1)** và **(2)**, ta có một bài toán mới được phát biểu như sau: *Tìm w và b để tổng các khoảng cách giữa các điểm và đường thẳng là nhỏ nhất*. Ta sẽ tạm gọi giá trị tổng này với một cái tên là **loss**. Để dễ hình dung, các bạn có thể nhìn hình dưới đây:

(a) Các đường thẳng với giá trị b khác nhau.



(b) Tổng khoảng cách giữa các điểm đến đường thẳng tương ứng của mỗi đường thẳng.

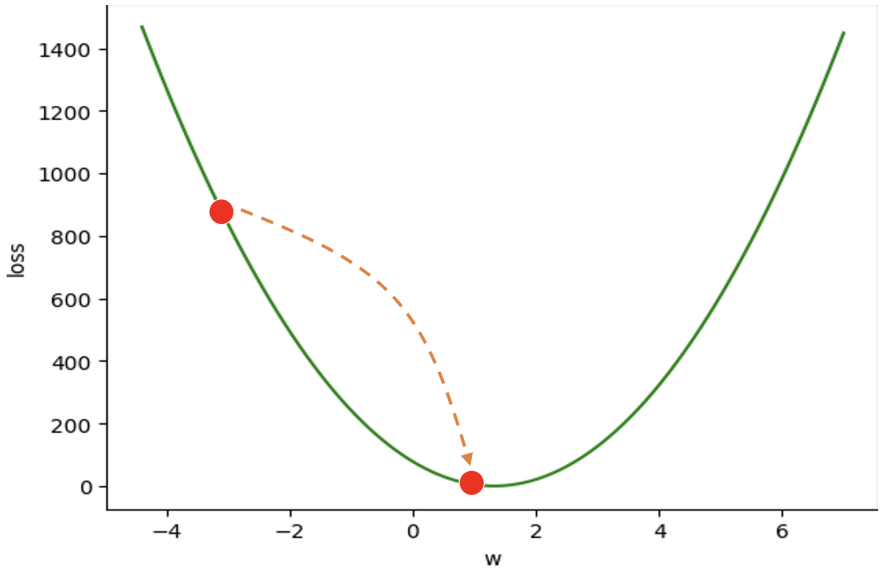
Xét cụ thể theo ví dụ trong hình, làm cách nào để ta có thể tìm được giá trị b tối ưu sao cho loss là nhỏ nhất? Giả sử ta có phương trình khoảng cách giữa y_i và $f(x_i)$ là $loss = (f(x_i) - y_i)^2$, ta có thể trực quan mối quan hệ giữa loss và b như sau:



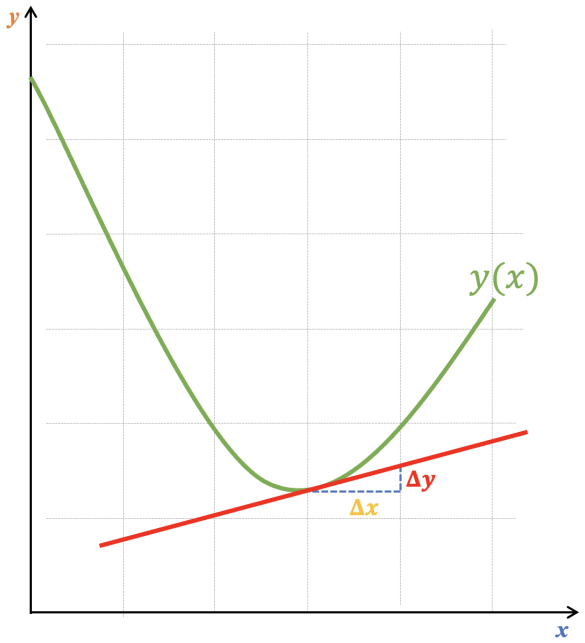
Different b values with a fixed w value

Quan sát thấy, hàm loss của chúng ta lúc này là một đường parabol, và vị trí cực tiểu biểu trưng cho giá trị loss nhỏ nhất. Vậy chẳng phải lúc này nếu ta bằng một cách nào đó có thể tìm được giá trị b tại điểm cực tiểu này, ta sẽ tìm được đường thẳng có loss là thấp nhất không? Nhưng làm cách nào để ta có thể tự động tìm ra được giá trị b tối ưu này?

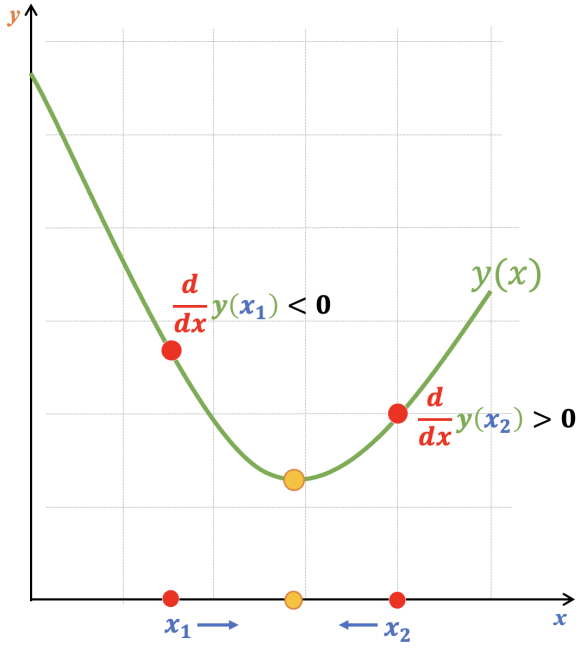
Nói một cách khác, nếu coi giá trị bias ta có ban đầu là một điểm nằm trên hàm loss, làm cách nào để ta có thể di chuyển điểm đó tiến về phía cực tiểu?



Để hiện thực được ý tưởng này, chúng ta có thể sử dụng công cụ đạo hàm. Nhắc qua một chút về khái niệm đạo hàm, khi chúng ta có giá trị đạo hàm tại một vị trí trên hàm số, chúng ta sẽ biết được thông tin về độ dốc của hàm tại điểm đó, biểu thị bằng đường tiếp tuyến tại một điểm, như hình dưới đây:



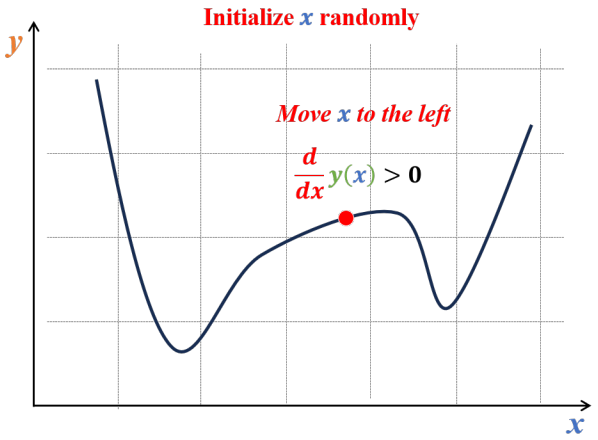
(a) Đường tiếp tuyến của một hàm tại một điểm.



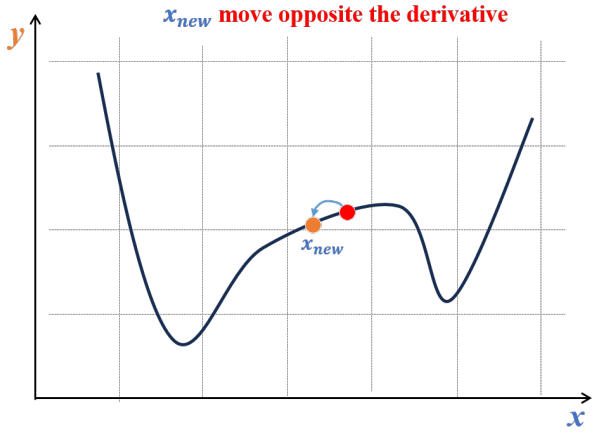
(b) Giá trị đạo hàm (độ dốc) tại các điểm khác nhau.

Nếu giá trị đạo hàm dương, ta biết rằng hàm số đang tăng tại điểm đó. Ngược lại, nếu giá trị đạo hàm âm, ta biết rằng hàm số đang giảm tại điểm đó. Tận dụng tính chất này, chúng ta

có thể dùng thông tin đạo hàm để giúp đẩy điểm b đang ở một vị trí bất kỳ tiến về điểm cực tiểu bằng cách **đi ngược về hướng đạo hàm**:



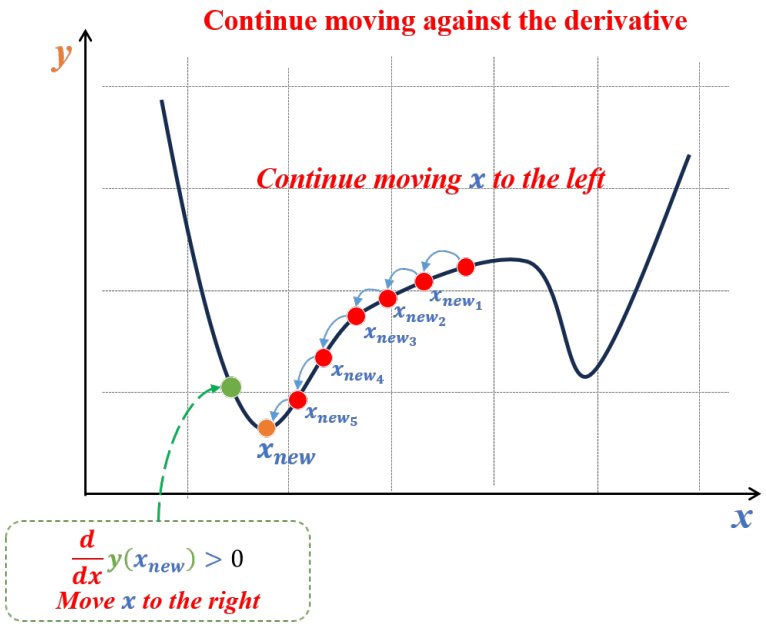
(a) Khởi tạo một x bất kỳ và tính đạo hàm tại điểm đó.



(b) Áp dụng tính chất đi ngược hướng đạo hàm để tiến về cực tiểu.

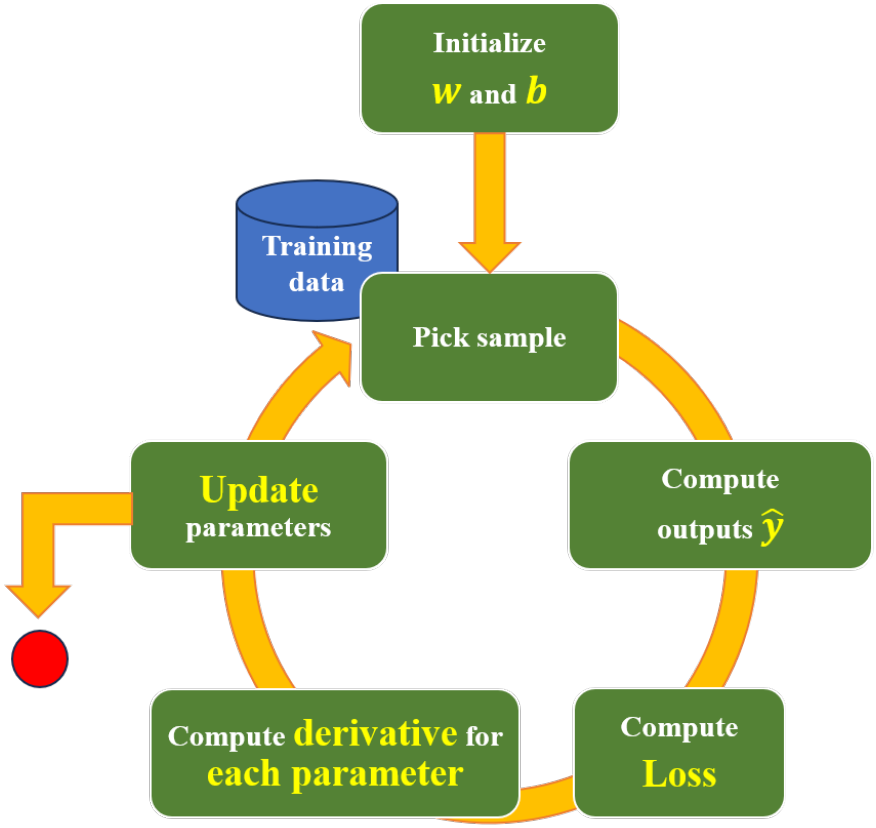
Hãy thử lấy ví dụ một hàm bất kỳ có hình dạng như hình (a), nếu ta khởi tạo một điểm x bất kỳ và có đạo hàm tại điểm x này lớn hơn 0 (hoặc có thể quan sát thấy hàm đang tăng thì nghĩa là đạo hàm đang dương). Để tìm được điểm cực tiểu, ta sẽ đi ngược hướng đạo hàm, đó là đi sang trái (xem hình (b)).

Và nếu hành động này cứ được tiếp diễn, dần dần ta sẽ có thể đưa được điểm b tiến gần đến điểm cực tiểu như hình sau:



Từ đây, ý tưởng từ đây được hình thành như sau: Với lần lượt các điểm dữ liệu có sẵn, ta thực hiện tính đạo hàm b tại điểm đang xét và cập nhật lại giá trị b nhằm đưa b tiến gần đến điểm cực tiểu của hàm loss, dẫn đến ta có một đường thẳng tối ưu cho bài toán dự đoán tiền lương.

Tổng hợp lại toàn bộ các thông tin trên, ta sẽ có một quy trình các bước tính toán (pipeline) để tìm được đường thẳng tối ưu như sau:



Vì đường thẳng được xác định từ hai biến là w và b, ta xét trường hợp tổng quát trong pipeline này bao gồm đi tìm giá trị tối ưu cho hai biến w và b. Như vậy, các bước thực hiện trong pipeline trên có thể diễn giải như sau:

1. Khởi tạo ngẫu nhiên giá trị cho hai tham số là w và b.
2. Với mỗi mẫu dữ liệu thứ i trong bộ dữ liệu, ta áp dụng các bước tính toán sau:
   (a) Thực hiện dự đoán output $\hat{y}_i$ với input là x_i. Ta có công thức sau:
   $$\hat{y}_i = f(x_i) = wx_i + b$$
   (b) Để tính toán sự chênh lệch giữa giá trị dự đoán so với giá trị thực tế, ta đưa y_i và $\hat{y}_i$ vào hàm tính loss. Trong trường hợp này, hàm loss của chúng ta là hàm Squared Error có công thức như sau:
   $$L(\hat{y}_i, y_i) = (\hat{y}_i - y_i)^2$$
   (c) Ta tìm giá trị đạo hàm tại mẫu dữ liệu thứ i cho hai tham số w và b để từ đó cập nhật lại đường tuyến tính với công thức sau:
   $$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}_i}\frac{\partial \hat{y}_i}{\partial b} = 2(\hat{y}_i - y_i)$$

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial \hat{y}_i}\frac{\partial \hat{y}_i}{\partial w} = 2x_i(\hat{y}_i - y_i)$$

(d) Với giá trị đạo hàm riêng vừa tìm được cho hai tham số w và b trên mẫu dữ liệu i, ta cập nhật lại giá trị mới cho hai tham số này với công thức sau:

$$w = w - \eta\frac{\partial L}{\partial w}$$

$$b = b - \eta\frac{\partial L}{\partial b}$$

Với η là một hằng số không đổi. Theo thuật ngữ, ta gọi η là **learning rate**.

Sau đó, như đã đề cập, ta lặp lại quy trình ở bước 2 cho đến khi xử lý hết tất cả các mẫu dữ liệu trong bộ dữ liệu.

B. Thực hành

Dựa theo cách triển khai đã đề cập ở phía trên, chúng ta sẽ thử áp dụng để xây dựng một hàm có khả năng dự đoán số tiền lương của nhân viên dựa theo số năm kinh nghiệm của họ. Đầu tiên, chúng ta xem qua một bộ dữ liệu nhỏ theo bảng dưới đây:

Index	Experience	Salary (.million VND)
0	3	60
1	4	55
2	5	66
3	6	93

Bảng 1: Bộ dữ liệu về thông tin tiền lương của nhân viên theo số năm kinh nghiệm tương ứng tại một công ty X.

Bộ dữ liệu trên gồm có 4 mẫu (sample), ứng với số năm kinh nghiệm (Experience) sẽ có số tiền lương tương ứng (Salary). Ví dụ, tại mẫu dữ liệu có Index bằng 1, một nhân viên có 4 năm kinh nghiệm sẽ có tiền lương tương ứng là 55 triệu VND. Với đề bài trên, ta sẽ tiến hành áp dụng thuật toán mô tả ở phần Dẫn nhập theo từng bước như sau:

1. Khởi tạo giá trị ngẫu nhiên cho hai tham số w và b. Ở đây, mình giả sử $w = 10$ và $b = 5$. Chúng ta cũng sẽ chọn giá trị $\eta = 0.01$ để sử dụng cho phần cập nhật tham số.

   Với hai giá trị này, ta thử áp dụng các công thức dự đoán và tính loss để dùng cho việc so sánh giữa giá trị trước và sau khi áp dụng thuật toán xem mức độ hiệu quả đến đâu. Đầu tiên, ta sẽ dự đoán mức lương của một nhân viên có 7 năm kinh nghiệm với công thức sau:

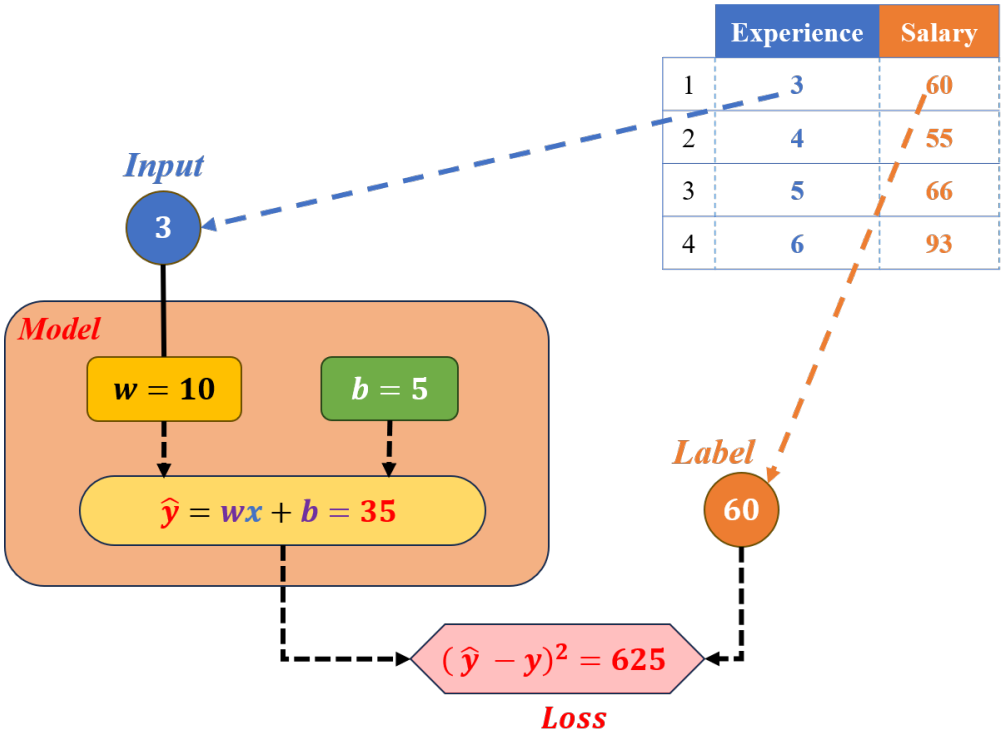
$$\hat{y} = wx + b = 10 \times 7 + 5 = 75$$

Tiếp theo là tính loss, mình sẽ dự đoán dựa trên trực giác dựa theo dữ liệu có được, nhìn chung số năm kinh nghiệm càng lớn thì mức lương càng cao. Mình nghĩ mức lương cho một nhân viên 7 năm sẽ là 100 triệu. Vậy loss sẽ có giá trị bằng:

$$L(\hat{y}, y) = (\hat{y} - y)^2 = (75 - 100)^2 = 625$$

Như có thể thấy, với giá trị loss tương đối cao, do đó, kết quả dự đoán của mô hình không được như mong đợi. Để cải thiện mô hình Linear Regression, chúng ta sẽ từng bước điều chỉnh các tham số w và b sao cho phù hợp với bộ dữ liệu trên.

2. Ta tiến hành duyệt qua lần lượt các mẫu trong bộ dữ liệu: Trước tiên là mẫu dữ liệu thứ 0, gồm có $x_0 = 3$, $y_0 = 60$ và thực hiện các bước tính toán:



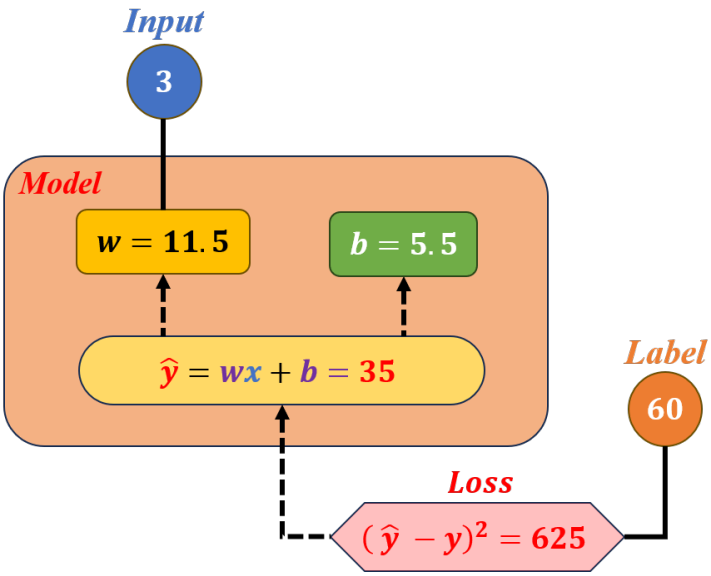
Hình 4: Minh hoạ cho các bước 2.(a) và 2.(b).

(a) Output:

$$\hat{y}_0 = x_0 w + b = 3 \times 10 + 5 = 35$$

(b) Loss:

$$L(\hat{y}_0, y_0) = (\hat{y}_0 - y_0)^2 = (35 - 60)^2 = 625$$



Hình 5: Minh hoạ cho các bước 2.(c) và 2.(d).

(c) Tính đạo hàm:

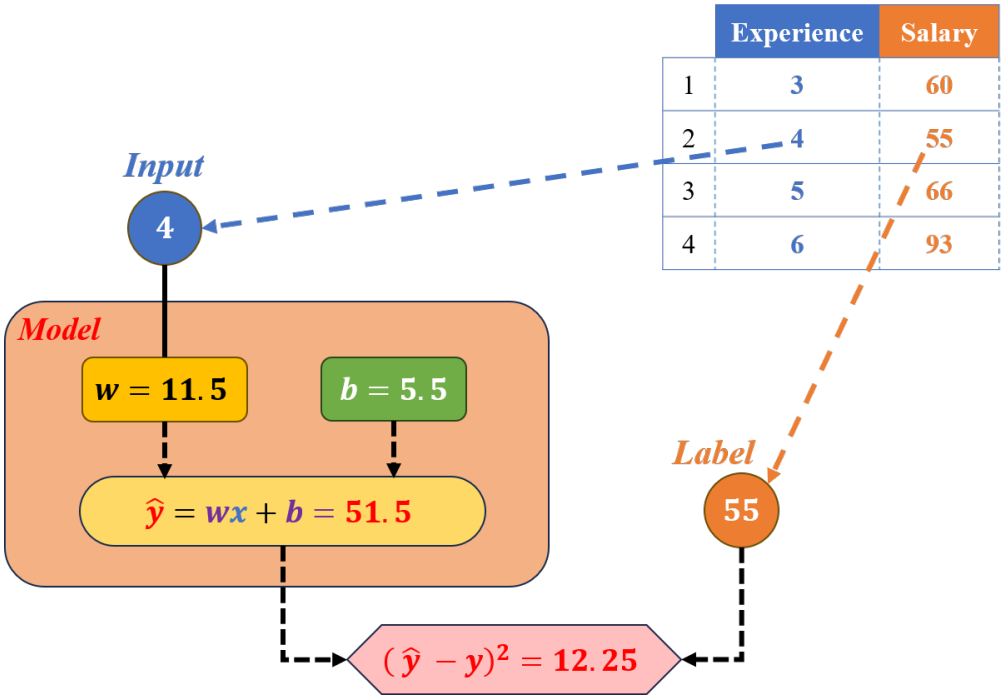
- Tham số w:
$$\frac{\partial L}{\partial w} = 2x_0(\hat{y}_0 - y_0) = 2 \times 3 \times (35 - 60) = -150$$
- Tham số b:
$$\frac{\partial L}{\partial b} = 2(\hat{y}_0 - y_0) = 2 \times (35 - 60) = -50$$

(d) Cập nhật tham số:

- Tham số w:
$$w = w - \eta \frac{\partial L}{\partial w} = 10 - 0.01 \times (-150) = 11.5$$
- Tham số b:
$$b = b - \eta \frac{\partial L}{\partial b} = 5 - 0.01 \times (-50) = 5.5$$

Sau khi cập nhật, ta thu được giá trị tham số mới là $w = 11.5$ và $b = 5.5$.

3. Lặp lại như bước (2) cho mẫu dữ liệu tiếp theo gồm có $x_1 = 4$, $y_1 = 55$ và thực hiện các bước tính toán:



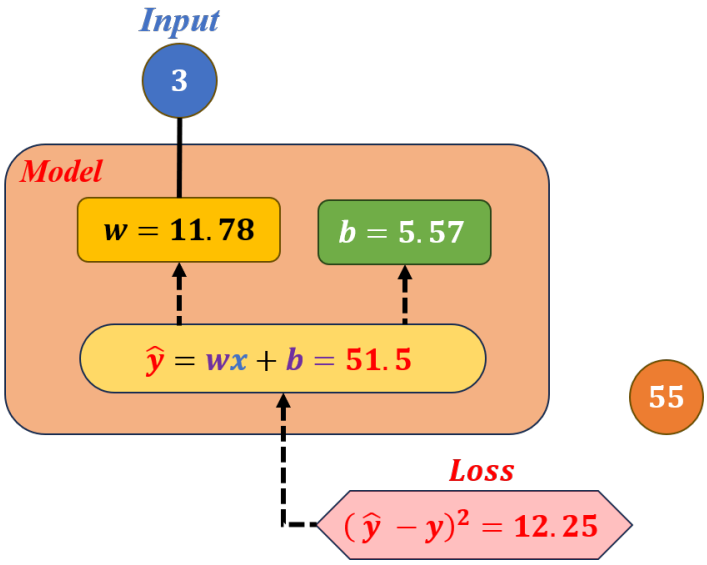
Hình 6: Minh hoạ cho các bước 3.(a) và 3.(b).

(a) Output:

$$\hat{y}_1 = x_1 w + b = 4 \times 11.5 + 5.5 = 51.5$$

(b) Loss:

$$L(\hat{y}_1, y_1) = (\hat{y}_1 - y_1)^2 = (51.5 - 55)^2 = 12.25$$



Hình 7: Minh hoạ cho các bước 3.(c) và 3.(d).

(c) Tính đạo hàm:

- Tham số w:

$$\frac{\partial L}{\partial w} = 2x_1(\hat{y}_1 - y_1) = 2 \times 4 \times (51.5 - 55) = -28$$

- Tham số b:

$$\frac{\partial L}{\partial b} = 2(\hat{y}_1 - y_1) = 2 \times (51.5 - 55) = -7$$

(d) Cập nhật tham số:

- Tham số w:

$$w = w - \eta\frac{\partial L}{\partial w} = 11.5 - 0.01 \times (-28) = 11.78$$

- Tham số b:

$$b = b - \eta\frac{\partial L}{\partial b} = 5.5 - 0.01 \times (-7) = 5.57$$

Sau khi cập nhật, ta thu được giá trị tham số mới là $w = 11.78$ và $b = 5.57$.

4. Thực hiện tương tự như đã làm với mẫu dữ liệu 0 và 1, ta lấy mẫu dữ liệu gồm có $x_2 = 5$, $y_2 = 66$ và thực hiện các bước tính toán:

(a) Output:

$$\hat{y}_2 = x_2 w + b = 5 \times 11.78 + 5.57 = 64.47$$

(b) Loss:

$$L(\hat{y}_2, y_2) = (\hat{y}_2 - y_2)^2 = (64.47 - 66)^2 = 2.3409$$

(c) Tính đạo hàm:

- Tham số w:

$$\frac{\partial L}{\partial w} = 2x_2(\hat{y}_2 - y_2) = 2 \times 5 \times (64.47 - 66) = -15.3$$

- Tham số b:

$$\frac{\partial L}{\partial b} = 2(\hat{y}_2 - y_2) = 2 \times (64.47 - 66) = -3.06$$

(d) Cập nhật tham số:

- Tham số w:

$$w = w - \eta\frac{\partial L}{\partial w} = 11.78 - 0.01 \times (-15.3) = 11.933$$

- Tham số b:

$$b = b - \eta\frac{\partial L}{\partial b} = 5.57 - 0.01 \times (-3.06) = 5.6006$$

Sau khi cập nhật, ta thu được giá trị tham số mới là $w = 11.933$ và $b = 5.6006$.

5. Với mẫu dữ liệu cuối cùng gồm có $x_3 = 6$, $y_3 = 93$ và thực hiện các bước tính toán tương tự:

(a) Output:

$$\hat{y}_3 = x_3 w + b = 6 \times 11.933 + 5.6006 = 77.1986$$

(b) Loss:

$$L(\hat{y}_3, y_3) = (\hat{y}_3 - y_3)^2 = (77.1986 - 93)^2 = 249.6842$$

(c) Tính đạo hàm:

- Tham số w:

$$\frac{\partial L}{\partial w} = 2x_3(\hat{y}_3 - y_3) = 2 \times 6 \times (77.1986 - 93) = -189.6168$$

- Tham số b:

$$\frac{\partial L}{\partial b} = 2(\hat{y}_3 - y_3) = 2 \times (77.1986 - 93) = -31.6028$$

(d) Cập nhật tham số:

- Tham số w:

$$w = w - \eta \frac{\partial L}{\partial w} = 11.933 - 0.01 \times (-189.6168) = 13.82916$$

- Tham số b:

$$b = b - \eta \frac{\partial L}{\partial b} = 5.6006 - 0.01 \times (-31.6028) = 5.916628$$

Sau khi cập nhật, ta thu được giá trị tham số mới là $w = 13.82916$ và $b = 5.916628$.

6. Sau khi duyệt qua tất cả các mẫu dữ liệu và các tham số w và b cũng đã được điều chỉnh lại cho phù hợp hơn với bộ dữ liệu của chúng ta. Để kiểm chứng điều này, ta sẽ thử dự đoán mức lương và tính loss với giá trị w và b mới này cho nhân viên có 7 năm kinh nghiệm. Trước tiên, mức lương của người nhân viên này là:

$$\hat{y} = wx + b = 13.82916 \times 7 + 5.916628 \approx 102.72$$

Kết quả dự đoán có vẻ tốt hơn ban đầu và đúng với mong đợi hơn. Để chắc chắn hơn, sau đây là giá trị loss với dự đoán này:

$$L(\hat{y}, y) = (\hat{y} - y)^2 = (102.72 - 100)^2 = 7.3984$$

Với kết quả loss nhỏ hơn so với ban đầu ($7.3984 < 625$) cho thấy mô hình đã được học và điều chỉnh lại các tham số bên trong sao cho phù hợp hơn với dữ liệu huấn luyện và có thể được sử dụng để dự đoán các dữ liệu mới từ bên ngoài mà vẫn đảm bảo độ chính xác cao.

Tóm lại, Simple Linear Regression được khởi tạo với các tham số w và b có giá trị ngẫu nhiên, các tham số này sau đó sẽ được điều chỉnh lại dựa trên dữ liệu được học, sao cho tối ưu nhất có thể. Dẫu vậy, đây cũng chỉ là một phiên bản đơn giản, trong các bài học tiếp theo, chúng ta sẽ đi sâu hơn vào từng bước "học"của mô hình và khái quát hơn cho các trường hợp tổng quát.

- Hết -