

Linear Regression - Exercise 2

1) Đọc hiểu code (file D5, D6, D7, D8) về cách train bài toán linear regression theo cách thông thường và vectorization cho m and N sample.

Câu (2) và (3) So sánh L1 loss (absolute error) và L2 loss (squared error). Nếu ưu nhược điểm của chúng. Tìm hiểu về Huber loss và cách Huber loss khắc phục nhược điểm của 2 hàm loss trên. Cài đặt Huber loss cho các chương trình ở file D5, D6, D7, D8. (Câu trả lời từ bạn Thân ở nhóm TA. Ad và nhóm TA đã đọc qua. Bạn Thân trình bày đủ ý và chặt chẽ).

Trả lời: Các bạn xem file “Loss Functions.pdf”.

4) Xây dựng công thức linear regression cho N sample khi vector \mathbf{x} có dạng sau:

$$\mathbf{x} = \begin{bmatrix} sample1 \\ \dots \\ sampleN \end{bmatrix} \quad (\text{sắp xếp theo chiều ngang, khác với chiều dọc trong bài học}).$$

X theo chiều ngang

Feature	Label
area	price
6.7	9.1
4.6	5.9
3.5	4.6
5.5	6.7

$$x = \begin{bmatrix} 6.7 & 1 \\ 4.6 & 1 \end{bmatrix} \quad y = \begin{bmatrix} 9.1 \\ 5.9 \end{bmatrix}$$

$$\theta = \begin{bmatrix} w \\ b \end{bmatrix} = \begin{bmatrix} -0.34 \\ 0.049 \end{bmatrix}$$

1) Pick all the N samples from training data

2) Tính output o

$$o = x\theta$$

3) Tính loss

$$L = (o - y)^2$$

4) Tính đạo hàm

$$k = 2(o - y)$$

\odot is element-wise multiplication

$$L'_\theta = \left(\begin{bmatrix} k \\ k \end{bmatrix} \odot x \right) \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

5) Cập nhật tham số

$$\theta = \theta - \eta \frac{L'_\theta}{N}$$

η is learning rate

X theo chiều ngang

Feature	Label
area	price
6.7	9.1
4.6	5.9
3.5	4.6
5.5	6.7

$$x = \begin{bmatrix} 6.7 & 1 \\ 4.6 & 1 \end{bmatrix} \quad y = \begin{bmatrix} 9.1 \\ 5.9 \end{bmatrix}$$

$$\theta = \begin{bmatrix} w \\ b \end{bmatrix} = \begin{bmatrix} -0.34 \\ 0.049 \end{bmatrix}$$

1) Pick all the N samples from training data

2) Tính output o

$$o = x\theta$$

3) Tính loss

$$L = (o - y)^2$$

4) Tính đạo hàm

$$k = 2(o - y)$$

\odot is element-wise multiplication

$$L'_\theta = ([k \quad k] \odot x) \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

5) Cập nhật tham số

$$\theta = \theta - \eta \frac{L'_\theta}{N}$$

η is learning rate

5) Cài đặt linear regression cho bài toán advertising theo 2 cách (cách thông thường và vectorization) dùng m sample (mini-batch gradient descent).

Data advertising:

AI Insight 2020

AI VIETNAM

<https://www.dropbox.com/s/ve57nlqwikgzrh/advertising.csv?dl=0>

Trả lời: Các bạn xem file “Linear regression _ advertising.ipynb”.