Exercise 1 - Linear regression

Đặng Linh Anh

December 2020

1. Đọc hiểu code (file `Lesson D3 - One Sample_Implementation (Naive).ipynb` và `Lesson D4 - One Sample_Implementation (Vectorization).ipynb`) về cách train bài toán linear regression theo cách thông thường và vectorization.

2. Trong phần xây dựng công thức, chúng ta tính loss $L = (output - label)^2$. Các bạn hãy xây dựng công thức cho bài toán linear regression với cách tính loss $(label - output)^2$.

Lời giải

Loss:

$$L = (y - \hat{y})^2 \tag{1}$$
$$\tag{2}$$

Đạo hàm:

$$\frac{dL}{dw_j} = \frac{dL}{d\hat{y}}\frac{d\hat{y}}{dw_j} \tag{3}$$
$$\frac{dL}{dw_j} = -2(y - \hat{y})x_j \tag{4}$$
$$\frac{dL}{db_j} = \frac{dL}{d\hat{y}}\frac{d\hat{y}}{db_j} \tag{5}$$
$$\frac{dL}{db_j} = -2(y - \hat{y}) \tag{6}$$

Cập nhật tham số:

$$w_j := w_j - \eta \times \frac{dL}{dw_j} \tag{7}$$
$$\Leftrightarrow w_j := w_j + \eta \times 2(y - \hat{y})x_j \tag{8}$$
$$b_j := b_j - \eta \times \frac{dL}{db_j} \tag{9}$$
$$\Leftrightarrow b_j := b_j + \eta \times 2(y - \hat{y}) \tag{10}$$

Nhận xét:
Kết quả sau cùng khi cập nhật các tham số vẫn không thay đổi so với khi loss $L = (output - label)^2$

3. Trong cách cài đặt vectorization, vector tham số $\theta = [w, b]$. Các bạn hãy chỉnh lại code với cách chọn $\theta = [b, w]$.

Lời giải

Listing 1: Python code

```
# load data
import numpy as np
from numpy import genfromtxt
import matplotlib.pyplot as plt

data = genfromtxt('data.csv', delimiter=',')
areas = data[:, 0]
prices = data[:, 1]
data_size = areas.size

# forward
def predict(x,theta):
    return x.dot(theta)

# compute gradient
def gradient(z,y,x):
    dtheta = 2*x*(z-y)

    return dtheta

# update weights
def update_weight(theta,n,dtheta):
    dtheta_new = theta - n*dtheta

    return dtheta_new

# vector [b, x]
data = np.c_[np.ones((data_size, 1)), areas]

# init weight
n = 0.01
theta = np.array([0.04, -0.34]) #[b, w]

# how long
epoch_max = 10

losses = [] # for debug
for epoch in range(epoch_max):
    for i in range(data_size):
        # get a sample
        x = data[i]
        y = prices[i:i+1]

        # predict z
        z = predict(x, theta)

        # compute loss
        loss = (z-y)*(z-y)
        losses.append(loss[0])
```

```
# compute gradient
dtheta = gradient(z,y,x)

# update weights
theta = update_weight(theta,n,dtheta)
```

---

4. Cài đặt linear regression cho bài toán advertising theo 2 cách (cách thông thường và vectorization) dùng 1 sample (stochastic gradient descent).

Lời giải

Dataset `avertising.csv` (shape 200×4), bao gồm:

- Feature: TV, Radio, Newspaper
- Label: Sales

Loss:

$$L = (\hat{y} - y)^2 \tag{11}$$

i. **Thông thường**:
Lặp cho từng sample của dataset: $(i = 1 \to 200)$
Gradient:

$$\frac{dL}{dw_j} = -2(\hat{y} - y)x_j \quad , 1 \leq j \leq 3 \tag{12}$$

$$\frac{dL}{db} = -2(\hat{y} - y) \tag{13}$$

Cập nhật tham số:

$$w_j := w_j + \eta \times 2(y - \hat{y})x_j \quad , 1 \leq j \leq 3 \tag{14}$$

$$b_j := b_j + \eta \times 2(y - \hat{y}) \tag{15}$$

Code: LINK

ii. **Vectorization**:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{bmatrix} \tag{16}$$

$$\theta = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ b \end{bmatrix} \tag{17}$$

Forward:

$$\hat{y} = \theta^T \mathbf{x} \tag{18}$$

Gradient:

$$\mathbf{L}'_\theta = \begin{bmatrix} 2(\hat{y} - y) \\ 2(\hat{y} - y) \\ 2(\hat{y} - y) \\ 2(\hat{y} - y) \end{bmatrix} \odot \mathbf{x} \tag{19}$$

Cập nhật tham số:

$$\theta \;:=\; \theta - \eta \times \mathbf{L}'_\theta \tag{20}$$

Code: LINK

5. Accuracy có dùng làm hàm loss được không? Tại sao?

Lời giải

Accuracy được tính từ độ chính xác của từng sample.

Trong linear regression, việc dự đoán chính xác nhiều sample cùng lúc là bất khả thi và không thực sự cần thiết. Thay vào đó, chúng ta quan tâm kết quả dự đoán từ model phải đảm bảo một độ chính xác nhất định ở tất cả các sample, tức là giảm sai số tối đa tới mức có thể. Accuracy mang tính đánh giá theo từng sample, vị vậy khi gặp các outlier, accuracy thay đổi rất nhiều, mất tính thống kê, làm model không thể train được. (thường xảy ra ngay từ lúc đầu của quá trình training).

6. Xây dựng hàm linear regression với cách tính loss $L = |output - label|$.

Lời giải

Ta có:

$$L = |\hat{y} - y| \tag{21}$$

$$\Rightarrow L = \sqrt{(\hat{y} - y)^2} \tag{22}$$

$$\Rightarrow \frac{dL}{d\hat{y}} = \frac{\hat{y} - y}{\sqrt{(\hat{y} - y)^2}} \tag{23}$$

$$\Rightarrow \frac{dL}{d\hat{y}} = \frac{\hat{y} - y}{|\hat{y} - y|} \tag{24}$$

$$\Rightarrow \frac{dL}{d\hat{y}} = \begin{cases} 1 & ,\text{nếu } \hat{y} \geq y \\ -1 & ,\text{nếu } \hat{y} < y \end{cases} \tag{25}$$

Đạo hàm:

$$\frac{dL}{dw_j} = \frac{dL}{d\hat{y}} \frac{d\hat{y}}{dw_j} \tag{26}$$

$$\Leftrightarrow \frac{dL}{dw_j} = \frac{\hat{y} - y}{|\hat{y} - y|} x_j \tag{27}$$

$$\frac{dL}{db_j} = \frac{dL}{d\hat{y}} \frac{d\hat{y}}{db_j} \tag{28}$$

$$\Leftrightarrow \frac{dL}{d_bj} = \frac{\hat{y} - y}{|\hat{y} - y|} \tag{29}$$

Cập nhật tham số:

$$w_j \;:=\; w_j - \eta \times \frac{dL}{dw_j} \tag{30}$$

$$\Leftrightarrow w_j \;:=\; w_j + \eta \times \frac{\hat{y} - y}{|\hat{y} - y|} x_j \tag{31}$$

$$b_j \;:=\; b_j - \eta \times \frac{dL}{db_j} \tag{32}$$

$$\Leftrightarrow b_j \;:=\; b_j + \eta \times \frac{\hat{y} - y}{|\hat{y} - y|} \tag{33}$$