Chương 6
Logistic regression (2)

TS Phạm Tuấn

ptuan@ute.udn.vn

Yếu điểm của Gradient Descent

```
def grad_w1(w1,w0,x,y):
    return -2*np.sum(np.multiply(x,(y - (w1*x + w0))))
def grad_w0(w1,w0,x,y):
    return -2*np.sum(y - (w1*x + w0))
```

Ở hình trên là hàm đạo hàm của Hồi quy tuyến tính, hàm đạo hàm đó chứa 2 phép tính chính là tổng và nhân. Do đó:

1. Giá trị đạo hàm sẽ càng **lớn** nếu điểm dữ liệu càng **nhiều**.
2. Thời gian tính toán sẽ **lâu** nếu điểm dữ liệu càng **nhiều**.

Với 2 nhược điểm trên, và chúng ta sẽ chạy thuật toán qua nhiều vòng lặp sẽ làm cho quá trình tìm điểm hội tụ càng lâu và nặng nề.

Stochastic Gradient Descent

Thay vì lấy đạo hàm trên tất cả các điểm dữ liệu thì chúng ta chỉ cần tính trên **1 điểm ngẫu nhiên**. Và Stochastic nghĩa là ngẫu nhiên.

Stochastic Gradient Descent (SGD) lấy 1 điểm dữ liệu từ tập data để tính giá trị của đạo hàm ở mỗi vòng lặp.

Vì lấy 1 điểm dữ liệu nên thời gian tính đạo hàm trên toàn tập dữ liệu sẽ rất lâu. Do đó, chúng ta lấy ***N* điểm dữ liệu, và *N* này thường là số nhỏ**. Và *N* được gọi là batch size.

Phương pháp lấy *N* điểm dữ liệu gọi là “mini-batch” gradient descent. Mini-batch sẽ cân bằng giữa sức mạnh của gradient descent và tốc độ của SGD.

Biến đổi data

Như ở bài tập trước, việc tính đạo hàm có thể gặp những lỗi số học không mong muốn, do giá trị của những điểm dữ liệu là lớn. Cho nên chúng ta phải biến đổi giá trị của điểm dữ liệu về 1 khoảng cố định.

```
X_std = (X - X.min(axis=0)) / (X.max(axis=0) - X.min(axis=0))
X_scaled = X_std * (max - min) + min
```

Trong đó max, min là khoảng giá trị

Nếu khoảng cố định đó là [0, 1] thì công thức được viết như sau:

```
X_scaled = (X - X.min(axis=0)) / (X.max(axis=0) - X.min(axis=0))
```

Bài tập nhóm số 6

Phân biệt giới tính dựa trên chiều cao và cân nặng:

1. Download file tại đây.
2. Chuyển chiều cao và cân nặng về khoảng [0, 1]. (1điểm)
3. Chia tập dữ liệu thành 2 phần: 80% cho tập training và 20% cho tập testing
4. Giải bài toán phân loại bằng thuật toán Logistic regression và thư viện sklearn. (Tập training) (2 điểm)
5. Giải bài toán phân loại bằng cách xây dựng thuật toán Logistic regression dùng Stochastic Gradient descent. (Tập training). (3 điểm)
6. Viết hàm dự đoán giới tính dùng mô hình Logistic regression đã giải bằng Stochastic Gradient descent. (1 điểm)
7. In ra Confusion matrix với mô hình Logistic regression đã giải bằng Sklearn, với tập testing (1 điểm).