# Logistic regression

1. **Giới thiệu:**

**Hồi quy Logistic** là một thuật toán học máy được sử dụng để phân loại dữ liệu. Nó là một mô hình thống kê dự đoán xác suất của một sự kiện xảy ra dựa trên một tập hợp các biến độc lập.

**Những điểm chính**:

* Hồi quy logistic dự đoán đầu ra của một biến phụ thuộc phân loại. Do đó, kết quả phải là một giá trị phân loại hoặc rời rạc.
* Nó có thể là Có hoặc Không, 0 hoặc 1, đúng hoặc Sai, v.v. nhưng thay vì đưa ra giá trị chính xác là 0 và 1, nó đưa ra các giá trị xác suất nằm trong khoảng từ 0 đến 1.
* Trong hồi quy logistic, thay vì khớp đường hồi quy, chúng tôi khớp hàm logistic hình chữ “S”, dự đoán hai giá trị tối đa (0 hoặc 1).

**Các loại hồi quy logistic:**

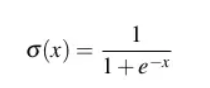
* **Nhị thức**: Trong hồi quy logistic nhị thức, chỉ có thể có hai loại biến phụ thuộc, chẳng hạn như 0 hoặc 1, Đạt hoặc Thất bại, v.v.
* **Đa thức**: Trong hồi quy logistic đa thức, có thể có 3 hoặc nhiều loại biến phụ thuộc không có thứ tự, chẳng hạn như “mèo”, “chó” hoặc “cừu”
* **Thứ tự**: Trong hồi quy logistic thứ tự, có thể có 3 loại biến phụ thuộc được sắp xếp theo thứ tự trở lên, chẳng hạn như “thấp”, “Trung bình” hoặc “Cao”.

1. **Thiết lập thuật toán**:
   1. Thiết lập mô hình:

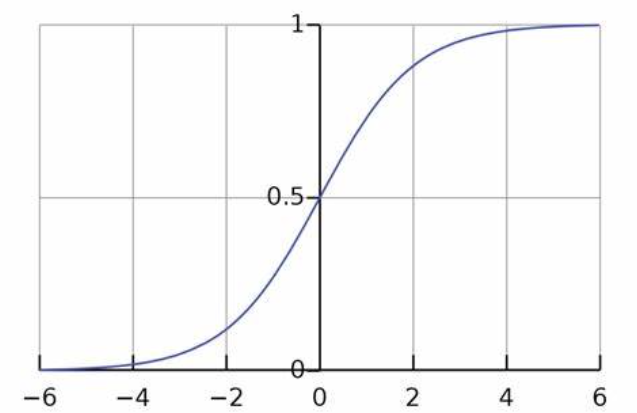
Data X -> Mô hình Logistic Regression -> Giá trị dự đoán wps wps Giá trị thực tế Y

Linear regression: 

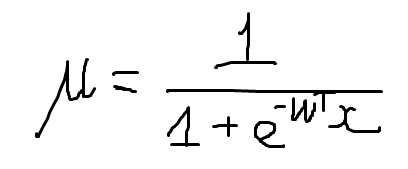
Giá trị của hàm f(x) khi này cho ra là từ (-∞, ∞) để mô hình Logistic Regression cho ra xác suất của một sự kiện [0, 1] người ta phải sử dụng hàm sigmoid:



Vì hàm sigmoid xác định từ (-∞, ∞) và giới hạn giá trị từ [0,1]:

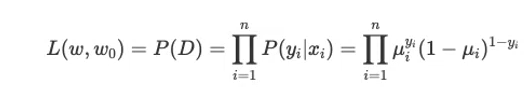


Nên ta có công thức của logistic regression:



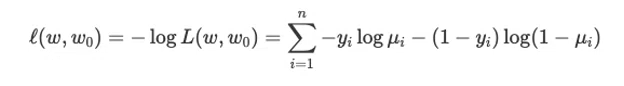
* 1. Hàm mất mát (Loss function):

Nguyên lí MLE:



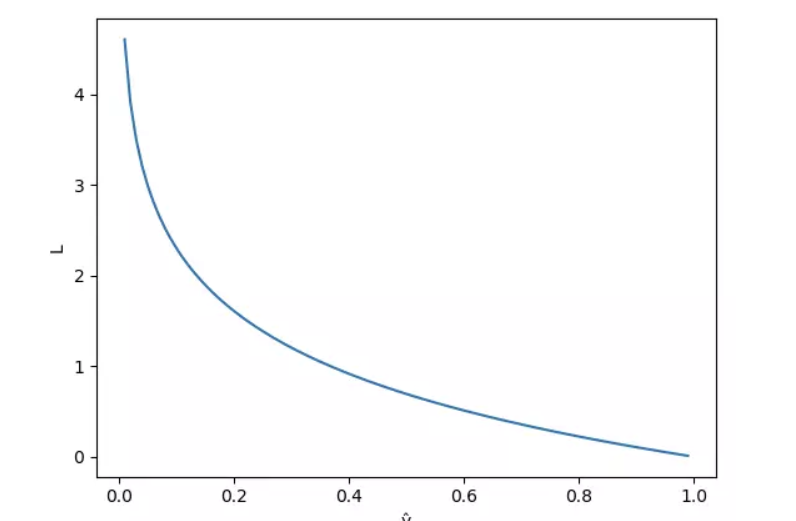
Với mu\_i là xác suất dự đoán của Yi

Ta lấy nagative-loglikelihood (NLL):



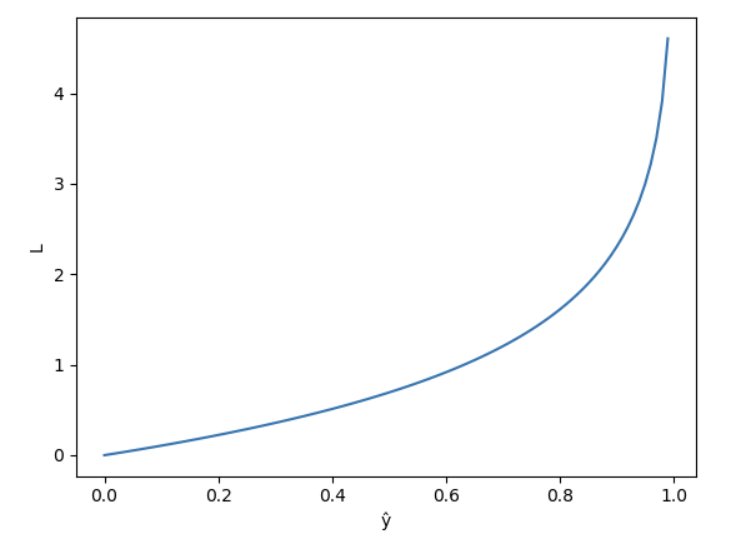
Đánh giá hàm mất mát này:

- Giả sử Yi = 1 thì L = -log(mu\_i) và có đồ thị như hình:



Ta dễ thấy khi giá trị mu\_i càng gần về 1 thì giá trị của L càng gần về ) tức là mất mát nhỏ và ngược lại.

- Giả sử Yi = 0 thì L = -log(1-mu\_i) và có đồ thị như hình sau:

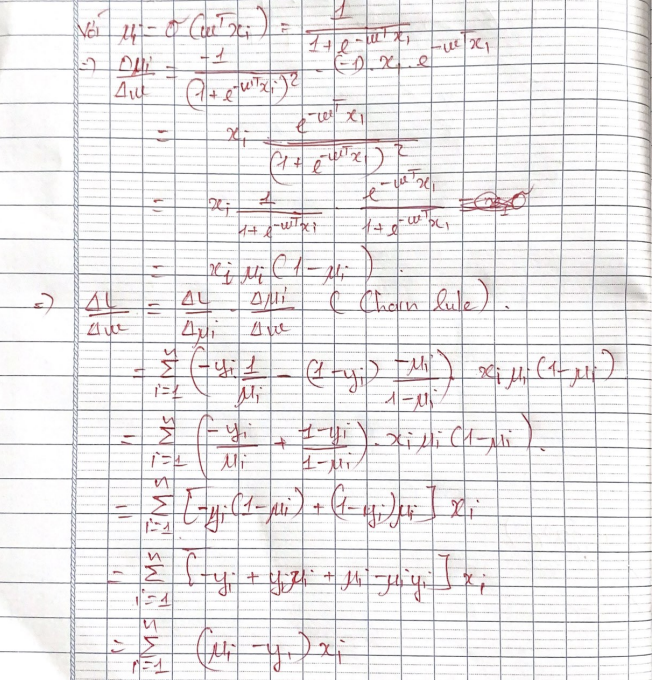


Ta thấy khi giá trị của mu\_i càng gần về 0 thì giá tri của L cũng càng nhỏ cho thấy mất mát là rất nhỏ và khi mu\_i càng lớn thì L càng lớn (giá trị dự đoán khác giá trị thực) nên L càng lớn tức là mát mát càng lớn.

* 1. Tối ưu hàm mất mát:

Để tối ưu hàm mất mát thì phải tìm cực tiểu của hàm này.

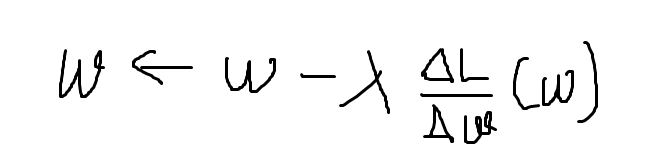
Sau đây là đạo hàm của hàm mất mát:



Đạo hàm bằng tổng của các sai lệch mu\_i - Yi với trọng số chính là dữ liệu Xi.

Vì đạo hàm của hàm này rất khó tìm nghiệm, nên người ta sử dụng một phương pháp gọi là gradent descent để tìm cực trị:

- Chọn một w ngẫu nhiên, wps > 0

- Tính giá trị đạo hàm của hàm L dựa trên w đã chọn, nếu đạo hàm lớn hơn 0 thì ta giảm bớt w, ngược lại nếu đạo hàm bé hơn 0 thì ta tăng w theo công thức sau: 

- Lặp lại quá trình đến khi đạo hảm đủ nhỏ, hoặc đạt tới một số vòng lặp nhất định ta thu được w và có thể bắt đầu dự đoán.

* 1. Đánh giá mô hình Logistic regression:

1. **Ứng dụng**:

* Dự đoán email có phải spam hay không
* Dự đoán giao dịch ngân hàng là gian lận hay không
* Dự đoán khối u lành hay ác tính
* Dự đoán khoản vay có trả được không
* Dự đoán khoản đầu tư vào start-up có sinh lãi hay không.
* ...

1. **Demo:**