LOGISTIC REGRESSION

Nguồn tham khảo: <https://en.wikipedia.org/wiki/Logistic_regression>

<https://aws.amazon.com/vi/what-is/logistic-regression/>

<https://bvag.com.vn/wp-content/uploads/2013/01/k2_attachments_PHAN-TICH-HOI-QUY-LOGISTIC-DON-VA-DA-BIEN.pdf>

<https://viblo.asia/p/logistic-regression-bai-toan-co-ban-trong-machine-learning-924lJ4rzKPM>

<https://machinelearningcoban.com/2017/01/27/logisticregression/#mo-hinh-logistic-regression>

<https://vi.wikipedia.org/wiki/H%C3%A0m_sigmoid>

1. Giới thiệu

Mô hình Logistic Regression là một kỹ thuật thống kê để xem xét mối liên hệ giữa biến độc lập (biến số hoặc biến phân loại) với biến phụ thuộc là biến nhị phân. Trong thuật toán này, biến phụ thuộc y chỉ có 2 trạng thái 1 (ví dụ tử vong) và 0 (ví dụ sống). Muốn đổi ra biến số liên tục người ta tính xác suất của 2 trạng thái này.

Mô hình Logistic Regression được phát triển bởi nhà toán học người Pháp Pierre-Simon Laplace vào thế kỷ 19.

Bằng cách nói ngắn gọn về định nghĩa của mô hình này như trên, điểm đặc biệt của nó là: mô hình Logistic Regression được thiết kế đặc biệt để phân loại dữ liệu vào hai nhóm khác nhau. Ngoài ra, Mô hình Logistic Regression cung cấp khả năng dự đoán xác suất xảy ra của sự kiện nhị phân và mô tả mối quan hệ tuyến tính giữa biến độc lập và biến phụ thuộc dựa trên hàm sigmoid.

1. Chi tiết thuật toán
2. Xây dựng hàm Sigmoid

Đầu tiên, cần phải chuyển đổi giá trị được đưa vào thành một giá trị nằm trong khoảng từ 0 đến 1

Công thức toán học của hàm Sigmoid là:

f(s)=

Trong đó: s là giá trị đầu vào của hàm

1. Xây dựng mô hình logistic

Mô hình hồi quy logistic sử dụng một hàm tuyến tính để tính toán giá trị z, tổng của các đặc trưng nhân với các tham số tương ứng. Công thức tính z được tính như sau:

z = w0 + w1x1 + w2x2 + … + wnxn

Trong đó:

* w0, w1, w2, …, wn là các tham số mô hình cần tìm kiếm
* x1,x2,...,xn là các đặc trưng của mẫu dữ liệu

1. Xây dựng hàm mất mát

Mô hình hồi quy Logistic sử dụng hàm mất mát để đo lường sự sai lệch giữa dự đoán mô hình và giá trị thực tế. Hàm mất mát phổ biến trong mô hình này là hàm Cross-Entropy (Log Loss)

Trong đó:

* là nhãn thực tế của mẫu dữ liệu (0 hoặc 1)
* là dự đoán được tính bởi mô hình, tức là xác suất dự đoán rằng mẫu dữ liệu thuộc vào lớp dương (1).

1. Tối ưu hàm mất mát

Sử dụng phương pháp Stochastic Gradient Descent (SGD) để tối ưu hàm mất mát.

1. Các siêu tham số của thuật toán

* Penalty: đây là tham số điều chỉnh regularization - kiểm soát mức độc phức tạp của mô hình và giảm hiện tượng overfitting.
* Hệ số điều chỉnh (C) : đây là siêu tham số có liên quan đến regularization trong logistic regression. Giá trị C nhỏ sẽ tạo ra môi trường có regularization mạnh và ngược lại.
* Solver: Logistic Regression sử dụng các phương pháp tối ưu để tìm ra bộ trọng số tốt nhất. Có nhiều phương pháp tối ưu như là ‘newton-cg’, ‘lbfgs’, ‘liblinear’, ‘sag’ và ‘saga’.
* max\_iter: số lần lặp quá trình tối ưu để tìm ra bộ trọng số tốt nhất. Nếu quá trình tối ưu không hội tụ trước khi đạt đến số lượng vòng lặp tối đa, nó sẽ dừng và trả lại kết quả tốt nhất tìm được cho tới thời điểm đó.

1. Đánh giá thuật toán

Ưu điểm:

* Đơn giản và dễ hiểu
* Có thể sử dụng với nhiều loại dữ liệu và phân loại nhãn đầu ra.
* Có thể xử lí đặc trưng liên tục và rời rạc
* Không đòi hỏi tính toán tài nguyên

Nhược điểm:

* Với mối quan hệ phi tuyến, logistic regression hoạt động không hiệu quả.
* Nhạy cảm với nhiễu và giá trị cực đại.
* Khó cho kết quả tốt khi dữ liệu mất cân bằng

1. Tổng kết

Khi nhắc đến Logistic Regression, điều cần nhớ là:

* Nó dựa trên hàm sigmoid để biến đổi giá trị đầu vào.
* Là mô hình tuyến tính
* Dữ liệu huấn luyện được sử dụng để ước lượng các tham số thông qua phương pháp tối ưu như: Gradient Descent