Mục lục

Chương 1

Giới thiệu

1.1 Giới thiệu

1.1.1 Logistic Regression

Logistic Regression (LR) là một kỹ thuật vay mượn từ lĩnh vực xác suất thống kê của Machine Learning (ML) áp dụng cho bài toán phân lớp (classification). Classification là bài toán mà nhãn (label) của dữ liệu được thể hiện dưới dạng từng đối tượng được chỉ đinh bằng tên của đối tượng đó hay các con số mã hóa thay cho tên đối tượng đó. Để có thể hiểu hơn về khái niệm của bài toán phân lớp ta xét ví dụ sau: "phân biệt học sinh cấp 2 và học sinh cấp 3". Ở đây ta hoàn toàn có thể sử dụng độ tuổi như một thuộc tính (feature) để phân biệt đâu là học sinh cấp 2 và đâu là học sinh cấp 3 bằng cách tạo một điều kiện để phân biệt học sinh cấp 2 là học sinh dưới 16 tuổi và ngược lại học sinh cấp 3 trên 16 tuổi. Từ đó ta xây dựng và huấn luyện một mô hình phân lớp cơ bản dựa vào điều kiện trên. Mô hình (model) phân lớp của ta sẽ dự đoán được học sinh cấp 2 hay cấp 3 dựa vào độ tuổi đã cho của học sinh đó.

1.2 Một số kiến thức toán

1.2.1 Maximum Likelihood

Ý tưởng

Maximum Likelihood Estimation (MLE) liên quan đến việc coi vấn đề như một vấn đề tối ưu hóa hoặc tìm kiếm, theo đó tìm kiếm một tập hợp các tham số

dẫn đến kết quả phù hợp nhất với xác suất chung của mẫu dữ liệu

Giả sử có các điểm dữ liệu $x_1, x_2, ..., x_n$ đã tuân theo một phân phối A nào đó được mô tả bởi một bộ tham số θ. Từ đó đi tìm một bộ tham số θ sao cho xác suất sau đạt giá trị lớn nhất:

$$\theta = \max p(x_1, x_2, ..., x_n|\theta) \tag{1.1}$$

Trong đó, $p(x_1, x_2, ..., x_n|\theta)$ chính là xác suất để các điều kiện $x_1, x_2, ..., x_2$ đồng thời xảy ra (xác suất có điều kiện) và xác suất đồng thời này được gọi là *likelihood*.

Ta cần tìm giá trị xác suất cao nhất cuả một likelihood bởi vì sự việc này đã được xảy ra và đã có kết quả nên ta cần phải đi tìm nguyên nhân sao cho xác suất để xảy ra kết quả đó càng cao càng tốt.

Vấn đề của MLE và log-likelihood

Việc giải trực tiếp bài toán 1.1 là một vấn đề khó khăn vì tìm kiếm một mô hình xác suất đồng thời cho toàn bộ dữ liệu là một điều bất khả thi. Hơn hết khi áp dụng MLE cho bài toán tối ưu thì việc giải phương trình đạo hàm với hàng ngàn điểm dữ liệu để tìm nghiệm tối ưu cũng là điều bất khả thi. Vì thế một cách tiếp cận phổ biến là đơn giản hóa bài toán bằng cách xem như rằng tất cả các điểm dữ liệu $x_1, x_2, ..., x_n$ là độc lập với nhau. Hay nói cách khác, likelihood được xấp xỉ bởi:

$$p(x_1, x_2, ..., x_n|\theta) \approx \prod_{n=1}^{N} p(X_n|\theta) \tag{1.2}$$

Vì các sự kiện là độc lập với nhau nên xác xuất đồng thời của chúng sẽ bằng tích xác suất của từng sự kiện.

$$p(x_1, x_2, ..., x_n) = p(x_1)p(x_2)...p(x_n) \tag{1.3}$$

Và khi là xác suất có điều kiện thì 1.3 trở thành:

$$p(x_1, x_2, ..., x_n|\theta) = p(x_1|\theta)p(x_2|\theta)...p(x_n|\theta) \tag{1.4}$$

Từ 1.4 thì 1.1 có thể giải quyết bằng cách giải bài toán sau:

$$\theta = \max \approx \prod_{n=1}^{N} p(X_n|\theta) \tag{1.5}$$

Ở đây việc tối ưu một tích sẽ thường phức tạp hơn khi làm việc với những số lớn và cần một lượng lớn về bộ nhớ và chi phí tính toán cao cho việc thực hiện tích các số lớn này. Vì vậy để đảm bảo vẫn giữ nguyên tính chất của dữ liệu và giảm chi phí tính toán cũng như việc giải bài toán này trở nên dễ dàng hơn thì hàm mục tiêu thường được chuyển về việc tối đa log của hàm mục tiêu. Lúc này 1.5 trở thành:

$$\theta = \max \approx \prod_{n=1}^{N} log(p(X_n|\theta)) \tag{1.6}$$

1.2.2 Phân phối Bernoulli

Một phép thử Bernoulli có kết quả nhận được là một trong hai giá trị hoặc "thành công" hoặc "thất bại". "Thành công" xảy ra với xác suất là p, "thất bại" với xác suất là $q = 1 - p$. Tham số p là số thực nằm giữa 0 và 1. Một biến ngẫu nhiên nhịn phân rời rạc X có phân phối Bernoulli nhận một trong 2 giá trị: 1 (thành công) hoặc 0 (thất bại). Xác suất thành công $P(X = 1) = p$ và xác suất thất bại $P(X = 0) = q = 1 - p$.

$$p(x_i|\lambda) = \lambda^{x_i}(1 - \lambda)^{1-x_i} \tag{1.7}$$

Chương 2

Logistic Regression

2.1 Mô hình Logistic Regression

Logistic Regression - mặc dù tên của nó không phải là thuật toán cho các vấn đề về hồi quy mà là một trong những thuật toán Machine Learning được sử dụng rộng rãi cho các vấn đề phân lớp. Logistic Regression là một thuật toán mô hình dự đoán để mô hình hóa các biến phân loại. Logistic Regression thuộc nhóm các thuật toán học có giám sát (supervised learning). Mục tiêu chính của nó để tìm xác suất có điều kiện, đầu ra là y = 1 cho rằng đầu vào là X.

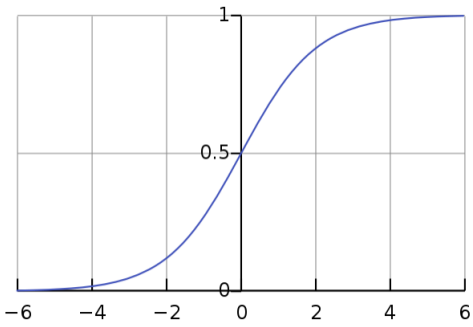
Đưa ra một tập hợp các biến độc lập, Logistic Regression dự đoán kết quả nhị phân (TRUE / FALSE, 0/1, có / không, hạnh phúc / buồn, ...). Logistic Regression được sử dụng khi có một trong 3 trường hợp sau:

- Biến nhị phân hoặc nhị phân Y.
- Các biến X giải thích có liên quan đến biến y.
- Giá trị của các biến y phụ thuộc vào các biến giải thích.

Đầu ra của thuật toán không phải là một lớp rời rạc hoàn toàn mà nó còn chứa thông tin về xác suất của một điểm dữ liệu khi rơi vào một lớp.

2.2 Hàm Sigmoid

Sigmoid là một hàm tăng liên tục, đơn điệu với đường cong đặc trưng hình chữ "S" đối xứng qua điểm (0, 0.5) nhờ sở hữu một số tính chất đặc biệt nên sigmoid



HÌNH 2.1: Ví dụ minh họa về hàm sigmoid (Nguồn Internet).

thường được lựa chọn là hàm kích hoạt sử dụng trong mạng nơ-ron nhân tạo đặc biệt đối với bài toán phân lớp nhị phân (binary classification).

$$f(s) = \frac{1}{1+e^{-s}} \tag{2.1}$$

Hàm sigmoid thường được sử dụng nhiều trong mạng nơ-ron nhân tạo bởi vì giá trị đầu ra của nó bị chặn trong khoảng (0, 1). Hơn hết, lim của f(s) bằng 0 khi tiến đến âm vô cùng và bằng 1 khi tiến đến dương vô cùng.

Vì sigmoid có dạng đường cong đối xứng qua điểm (0, 0.5) nên khi cộng kết quả của hàm sigmoid cho hai giá trị x đối xứng thì sẽ cho ra ra được kết quả bằng 1

$$\sigma(x) + \sigma(-x) = 1 \tag{2.2}$$

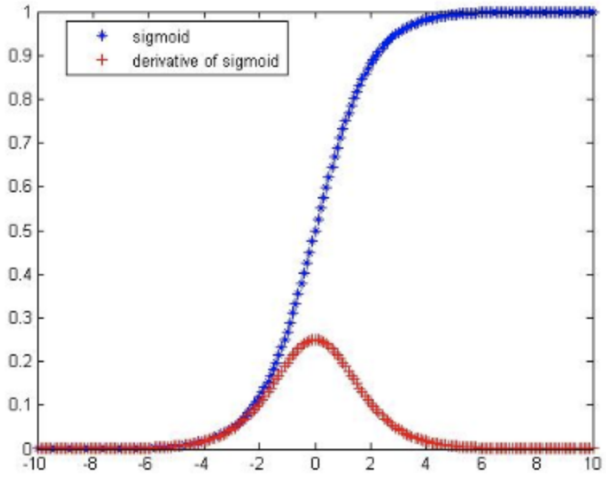
Ta sẽ chứng minh 2.2 như sau:

$$\sigma(x) + \sigma(-x) = \frac{1}{1+e^{-x}} + \frac{1}{1+e^{-(-x)}} = \frac{1+e^{x}}{(1+e^{-x})(1+e^{x})} + \frac{1+e^{-x}}{(1+e^{-x})(1+e^{x})} \tag{2.3}$$

Từ 2.3 ta có:

$$\frac{2+e^{x}+e^{-x}}{1+e^{x}+e^{-x}+e^{x-x}} = \frac{2+e^{x}+e^{-x}}{2+e^{x}+e^{-x}} = 1 \tag{2.4}$$

Trong khi giá trị của hàm sigmoid bị chặn trọng khoảng (0, 1) thì giá trị đạo hàm của nó bị chặn trong khoảng (0, 0.25)



HÌNH 2.2: Ví dụ minh họa về giá trị của sigmoid và đạo hàm của nó (Nguồn Internet).

Ta sẽ tính đạo hảm của sigmoid như sau:

$$\frac{d\sigma(x)}{dx} = \frac{d}{dx}(\frac{1}{1+e^{-x}}) = \frac{d}{dx}(1+e^{-x})^{-1} = -(1+e^{-x})^{-2}(-e^{-x}) \tag{2.5}$$

Từ 2.5 ta có

$$\frac{e^{-x}}{(1+e^{-x})(-2)} = \sigma(x)\frac{(e^{-x})}{1+e^{-x}} = \sigma(x)\sigma(-x) \tag{2.6}$$

2.3 Xây dựng hàm mất mát

ta có thể giả sử rằng xác suất để một điểm dữ liệu x rơi vào lớp thứ nhất là $f(W^T x)$ và rơi và lớp còn lại là $1 - f(W^T x)$

$$p(y_i = 1|x_i; W) = f(W^T x_i) \tag{2.7}$$

$$p(y_i = 0|x_i; W) = 1 - f(W^T x_i) \tag{2.8}$$

Trong đó $p(y_i = 1|x_i; W)$ là xác suất xảy ra sự kiện đầu ra $y_i = 1$ khi biết tham số mô hình w với dữ liệu đầu vào x_i. Mục đích cuối cùng là tìm các hệ số w sao cho với các điểm dữ liệu ứng với $y_i = 1$, $f(W^T x_i)$ gần với 1 và ngược lại. Gọi $z_i = f(W^T x_i)$, áp dụng phân phối Bernoulli ta có:

$$p(y_i|x_i; W) = z_i^{y_i}(1 - z_i)^{1-y_i} \tag{2.9}$$

Biểu thức 2.9 cho ta thấy khi $y_i = 1$ thì 2.9 sẽ bằng z_i, khi $y_i = 0$ phần thứ nhất sẽ bằng $(1 - z_i)$. Ta muốn mô hình gần với dữ liệu đã cho nhất tức xác suất này phải đạt giá trị cao nhất. Áp dụng MLE cho 2.9 ta có:

$$p(y|X; W) = \prod_{i=1}^{N} p(y_i|X_i; W) = \prod_{i=1}^{N} z_i^{y_i}(1 - z_i)^{1-y_i} \tag{2.10}$$

Áp dụng Log-Likelihood cho 2.10 ta có

$$\prod_{i=1}^{N} z_i^{y_i}(1 - z_i)^{1-y_i} = \prod_{i=1}^{N} log(z_i^{y_i}(1 - z_i)^{1-y_i}) \tag{2.11}$$

Áp dụng tính chất của logarit tự nhiên 2.11 trở thành

$$-\frac{1}{N}\sum_{i=1}^{N}(y_i log(z_i) + (1 - y_i)log(1 - z_i)) \tag{2.12}$$

Xét hàm 2.12 trên 1 điểm dữ liệu ta có:

$$-y_i log(z_i) - (1 - y_i)log(1 - z_i) \tag{2.13}$$

Khi $y = 1$, thì $loss = -ylog(z)$ ngược lại khi $y = 0$ thì $loss = -log(1 - z)$. Và hàm mất mát này thường được gọi là Binary CrossEntropy.

2.4 Tối ưu hàm mất mát

Ở phần này ta sẽ tìm hiểu mối liên hệ giữa Logistic Regression với hàm sigmoid bằng cách giải bài toán tối ưu hàm loss của nó sau đây. Từ 2.13 ta giải phương trình đạo hàm theo biến w như sau:

$$\bigtriangledown_w J = -(\frac{y_i}{z_i} - \frac{1-y_i}{1-z_i})(\bigtriangledown_w z_I) = -(\frac{y_i - z_i}{z_i(1-z_i)}) = \frac{z_i - y_i}{z_i(1-z_i)} \tag{2.14}$$

Đến bước này để có thể giải tiếp phương trình 2.14 ta sẽ đặt $s_i = w^T x_i$ và giải được kết quả $\bigtriangledown_w z_i$ như sau:

$$\bigtriangledown_w z_i = \frac{\partial z_i}{\partial s_i} \bigtriangledown_w z_i = \frac{\partial z_i}{\partial s_i} x_i \tag{2.15}$$

Để có thể dễ dàng tính toán đạo hàm thì ta sẽ khử mẫu của 2.14 bằng cách đặt hàm số $z = f(s)$ sao cho $\frac{\partial z_i}{\partial s_i} = z_i(1-z_i)$ và giải phương trình đã đặt như sau:

$$\frac{\partial z_i}{z_i(1-z_i)} = \partial s_i \iff (\frac{1}{z_i} + \frac{1}{1-z_i})\partial z_i = \partial s_i \tag{2.16}$$

Nguyên hàm hai vế ta được:

$$log(z_i) - log(1-z_i) = s \iff log(\frac{z_i}{1-z_i}) = s_i + C \iff \frac{z_i}{1-z_i} = e^{s_i + C} \tag{2.17}$$

Từ 2.17 ta có

$$z(1 + e^{s_i+C}) = e^{s_i+C} \iff z_i = \frac{e^{s_i+C}}{1+e^{s_i+C}} \iff z = \frac{1}{1+e^{-(s_i+C)}} \tag{2.18}$$

Do C là hằng số nên ta có:

$$z = \frac{1}{1+e^{-(s_i)}} \iff z_i = \sigma(s_i) \tag{2.19}$$

Từ 2.19 ta thấy được rằng sigmoid chính là kết quả tối ưu hàm mất mát của logistic regression.

2.5 Tính chất của Logistic Regresssion

Sau khi có được mô hình, việc xác định lớp cho một điểm dữ liệu được xác định bằng việc so sánh hai biểu thức xác suất $P(y = 1|X; W)$ và $P(y = 0|X; W)$. Nếu biểu thức thứ nhất lớn hơn, ta kết luận điểm dữ liệu thuộc lớp thứ nhất và ngược lại. Vì tổng hai biểu thức này luôn bằng 1 (theo phân phối Bernoulli) nên một cách ngắn gọn hơn, ta chỉ xác định xem $P(y = 1|X; W)$ có lớn hơn 0.5 hay không.

Ngưỡng đầu ra quyết định đều được lấy tại 0.5. Trong nhiều trường hợp, ngưỡng này có thể được thay đổi. Ta hoàn toàn có thể lấy ngưỡng cao hơn 0.5 hay thấp hơn 0.5 tùy thuộc vào bài toán mà ta cần quan tâm. Ví dụ, đối với bài toán dò mìn, việc xác định kim loại nằm dưới đất có phải là mìn hay không là cực kỳ quan trọng. Chúng ta chấp nhận việc dự đoán nhầm còn hơn bỏ sót trường hợp này gây nên những hậu quả nghiêm trọng khó lường, ta hoàn toàn có thể điều chỉnh ngưỡng đầu ra này thấp hơn như 0.2 để đảm bảo rằng tối thiểu việc bỏ sót trường hợp.

2.6 Hạn chế của Logistic Regression

Logistic regression được áp dụng cho các bài toán phân lớp nhị phân. Nhưng trong thực tế số lớp nhiều hơn 2 được gọi là bài toán phân lớp đa lớp. Ta có thể áp dụng một vài kỹ thuật hỗ trợ để áp dụng cho logistic regression khi làm việc với nhiều lớp như: one-vs-one, phân tầng, binary encoding, one-vs-rest.