

Logistic Regression from the Grounth Up

Nhóm MLOps

Ngày 18 tháng 10 năm 2025

Nội dung về Logistic Regression được chia thành 2 phần chính, phần 1 giải thích vấn đề của Linear Regression và vì sao lại cần Logistic Regression cũng như cách xây dựng nó. Phần 2 tập trung giải thích sâu hơn về việc tại sao lại sử dụng những công thức này và ý nghĩa của chúng.

- **Phần 1: Từ Linear Regression đến Logistic Regression**
- **Phần 2: Mở Rộng: Nhân ma trận với Logistic Regression**
- **Phần 3: Tại Sao**

Phần 1: Từ Linear Regression đến Logistic Regression

1.1 Phần 1.1: Vấn đề với Linear Regression

Ý tưởng của Linear Regression là làm thế nào để mô tả tất cả các biến bằng 1 đường thẳng $f(x)$, để với 1 đầu vào x_i mới ta có thể dự đoán đc đầu ra y_i . (i là số thứ tự). Ý tưởng này đơn giản nhưng thực chất rất hiệu quả vì những lý do sau:

- (1) Kể cả khi dữ liệu thực tế có phức tạp, xu hướng chung (global trend) của nó thường có thể mô tả bằng 1 đường thẳng (tuyến tính - đồ). (Note: xu hướng cục bộ - Cam)

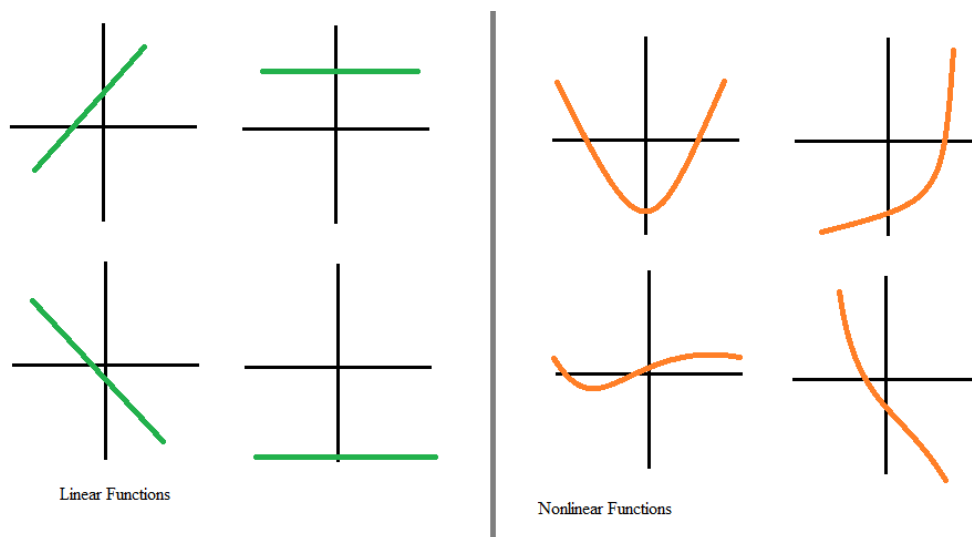


Figure 1: linear function vs nonlinear function

Vì nó nắm bắt xu hướng bậc nhất/lớn nhất (First Order Trend, Màu Đỏ) trong dữ liệu. Nói 1 cách toán học thì nó nắm bắt đc mối quan hệ bậc nhất giữa các biến trong không gian 1D và bỏ lơ các mối quan hệ trong không gian khác (2D, 3D, etc..), Bậc nhất trong đây là bậc nhất (first term) trong khai triển Taylor nhe $f(x) \approx f(a) + f'(a)(x - a)$.

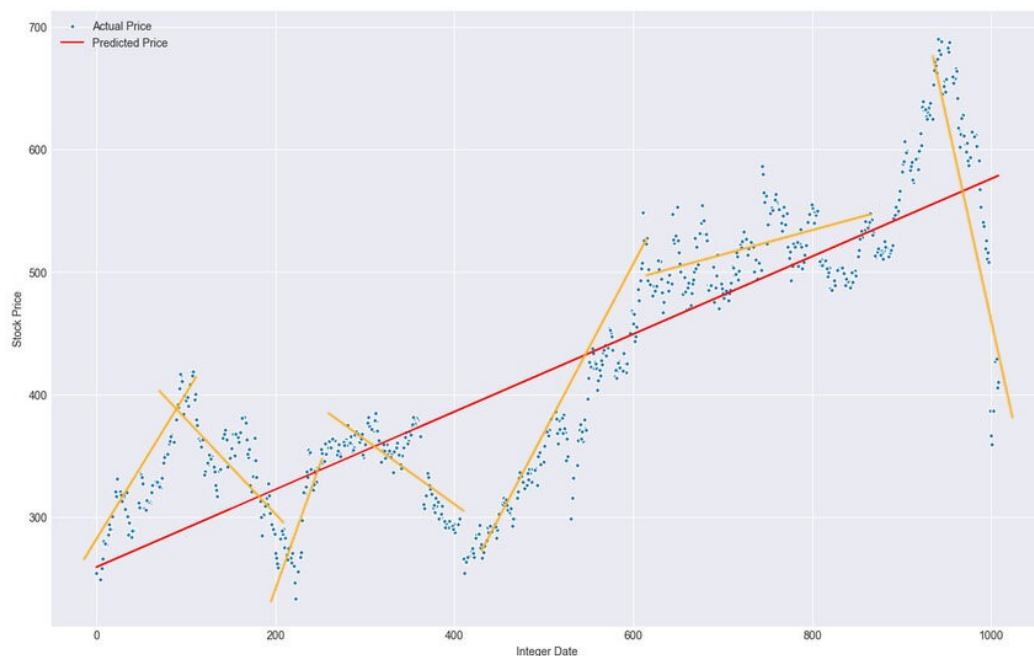


Figure 2: Linear Regression

- Linear (tuyến tính) nghĩa là mối quan hệ giữa các biến có thể được mô tả bởi 1 đường thẳng.
- Regression (hồi quy) là 1 phương pháp dùng để tìm mối quan hệ giữa các biến và đường thẳng dự đoán.

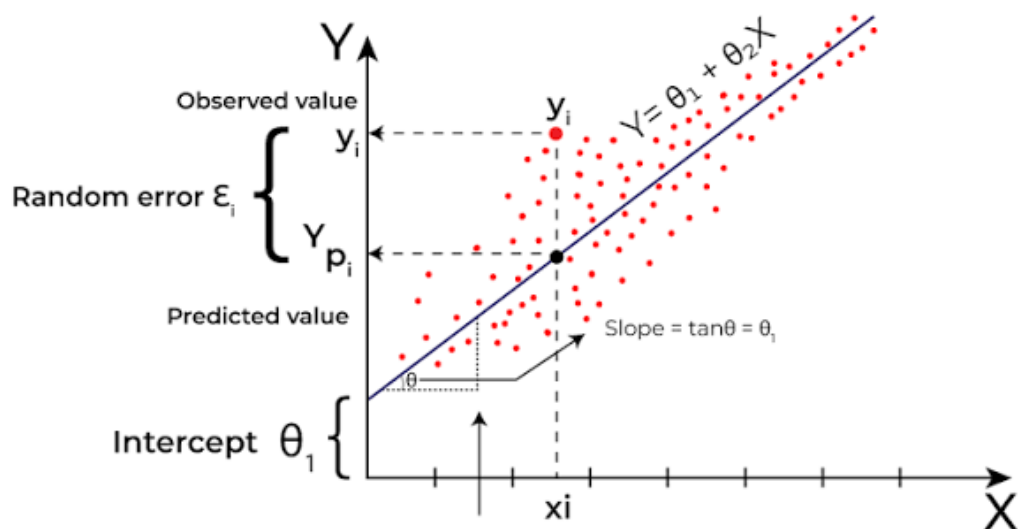


Figure 3: Linear Regression

Tuy nhiên với dữ liệu không tuyến tính và liên tục, thì Linear Regression không thể khớp được. Vậy làm thế nào để chọn được 1 hàm Loss giúp phân loại nhiều lớp ? Đầu tiên mình cần hiểu điều kiện của 1 hàm Loss là gì trước.

1.2 Phần 1.2: Logistic Regression

- Liên kết với Linear Regression và vấn đề Logistic Regression giải quyết - giải thích thuật ngữ - giải thích công thức qua 1 ví dụ từ đầu đến cuối 1. Initialize Parameters 2. Calculating y -> Calculate Sigmoid z 3. Calculate Loss base on \hat{y} and y 4. Calculate Parameters Gradient 5. Update Parameters Gradient (w and b)

- giải thích ý nghĩa công thức - code lại sử dụng ví dụ.

1.3 Phần 1.3: Xây dựng hàm Loss từ Suy Luận

1.4 Nhược Điểm của MSE

+ Tích tiến đến 0. + Non-Convex -> khó hội tụ. (Giải thích sâu hơn sau khi xây dựng xong BCE)

Logistic Regression-MSE

❖ Result

Done?

Feature	Label
Petal Length	Category
1.4	0
1	0
1.5	0
3	1
3.8	1
4.1	1

Category 0

Category 1

Model and Loss

$$z = \theta^T x = x^T \theta$$

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}$$

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

$$\frac{\partial L}{\partial \theta_i} = 2x_i(\hat{y} - y)\hat{y}(1 - \hat{y})$$

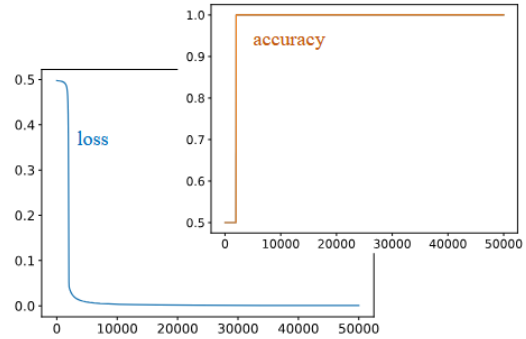


Figure 4: Problem with Non-Convex Loss Function

1.5 Lựa chọn và Xây dựng hàm Loss cho bài toán phân loại

❖ Suggested Functions

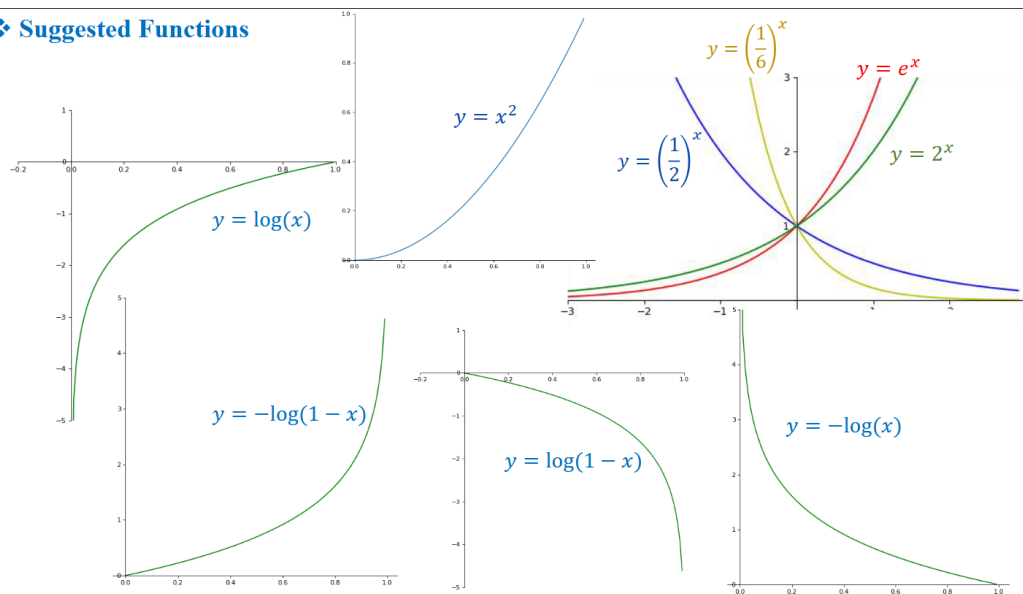


Figure 5: Suggested Loss Function

❖ Loss function

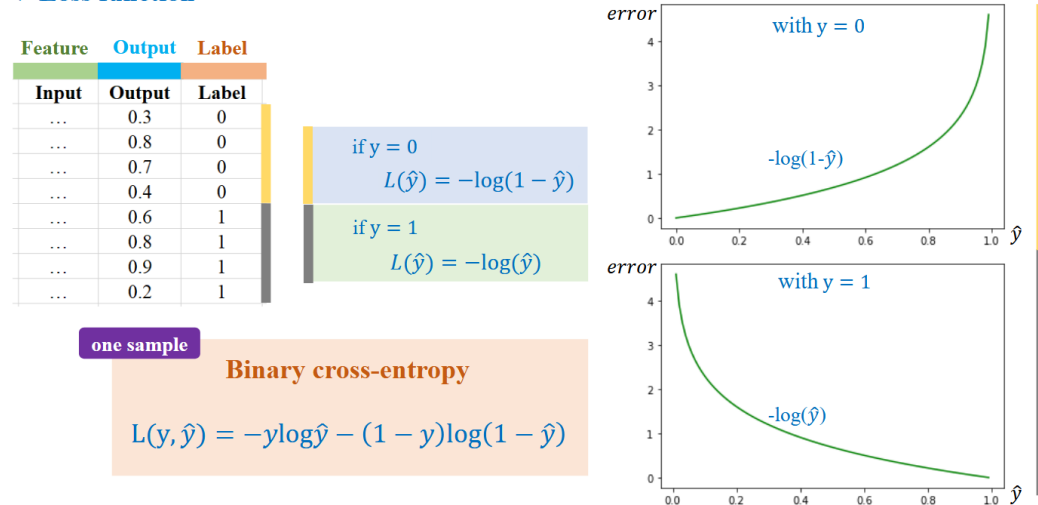


Figure 6: Binary Cross Entropy Loss

Construct loss

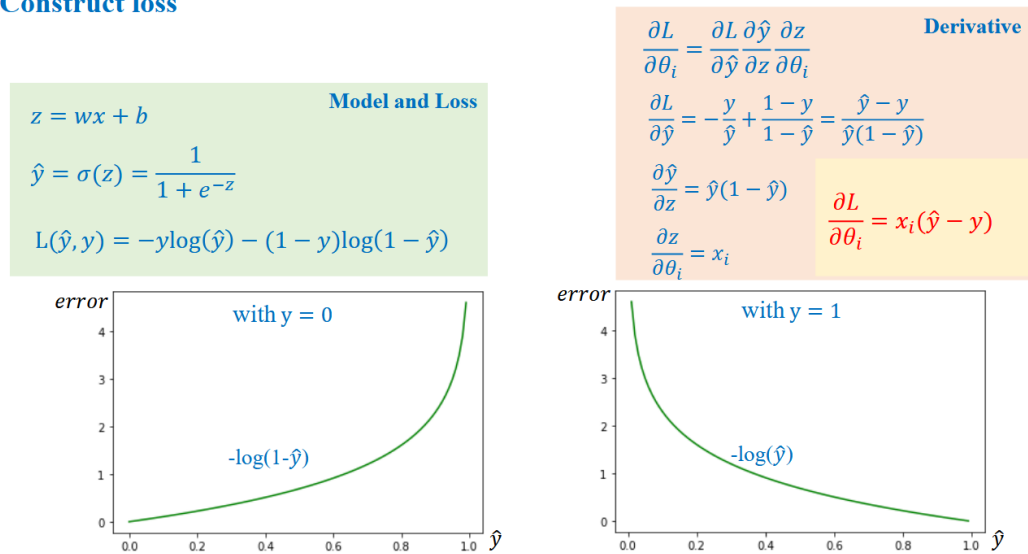


Figure 7: BCE Summarize

Explain Convexity using that Coursera Linear Algebra Course

1.6 Chứng Minh Convex (độ lồi) của MSE và BCE

Chứng Minh Convex (độ lồi) của MSE

Mean Squared Error

Model and Loss

$$z = \theta^T x = x^T \theta$$

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}$$

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

Derivative

$$\frac{\partial L}{\partial \theta_i} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial \theta_i} \quad \frac{\partial \hat{y}}{\partial z} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial \hat{y}} = 2(\hat{y} - y) \quad \frac{\partial z}{\partial \theta_i} = x_i$$

$$\frac{\partial L}{\partial \theta_i} = 2x_i(\hat{y} - y)\hat{y}(1 - \hat{y})$$

$$\frac{\partial L}{\partial \theta_i} = 2x_i(\hat{y} - y)\hat{y}(1 - \hat{y}) = 2x_i(-\hat{y}^3 + \hat{y}^2 - y\hat{y} + y\hat{y}^2)$$

$$\begin{aligned} \frac{\partial^2 L}{\partial \theta_i^2} &= \frac{\partial}{\partial \theta_i} [2x_i(-\hat{y}^3 + \hat{y}^2 - y\hat{y} + y\hat{y}^2)] \\ &= 2x_i[-3\hat{y}^2 x_i \hat{y}(1 - \hat{y}) + 2x_i \hat{y} \hat{y}(1 - \hat{y}) - y x_i \hat{y}(1 - \hat{y}) + 2x_i y \hat{y} \hat{y}(1 - \hat{y})] \\ &= 2x_i^2 \hat{y}(1 - \hat{y})[-3\hat{y}^2 + 2\hat{y} - y + 2y\hat{y}] \end{aligned}$$

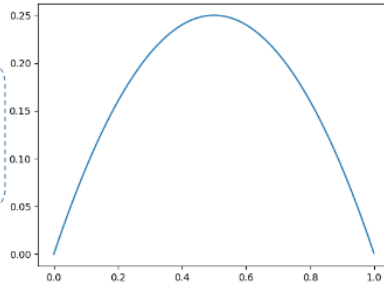
Figure 8: MSE vs BCE

Mean Squared Error

$$\frac{\partial^2 L}{\partial \theta_i^2} = 2x_i^2 \hat{y}(1 - \hat{y})[-3\hat{y}^2 + 2\hat{y} - y + 2y\hat{y}]$$

$$x_i^2 \geq 0$$

$$\hat{y}(1 - \hat{y}) \in \left[0, \frac{1}{4}\right]$$



$$y = 0$$

$$f(\hat{y}) = -3\hat{y}^2 + 2\hat{y}$$

$$y = 1$$

$$f(\hat{y}) = -3\hat{y}^2 + 4\hat{y} - 1$$

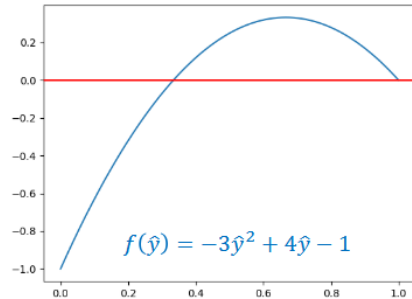
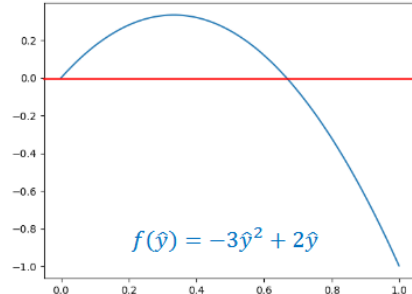
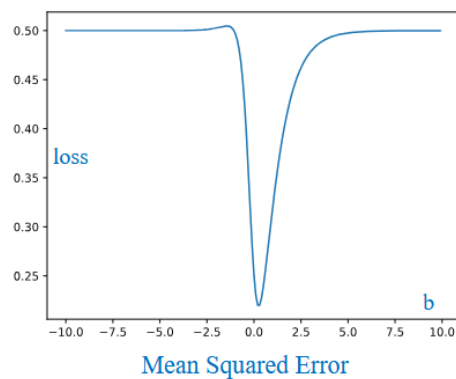


Figure 9: MSE vs BCE



Model and Loss

$$z = \theta^T x = x^T \theta$$

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}$$

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

$$\frac{\partial L}{\partial \theta_i} = 2x_i(\hat{y} - y)\hat{y}(1 - \hat{y})$$

Figure 10: MSE vs BCE

Chứng Minh Convex (độ lồi) của BCE

Binary Cross-entropy

Convex function

$$z = \theta^T x$$

Model and Loss

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}$$

$$L = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

simplified version

$$\frac{\partial L}{\partial \theta_i} = x_i(\hat{y} - y)$$

$$\frac{\partial^2 L}{\partial \theta_i^2} = \frac{\partial}{\partial \theta_i} [x_i(\hat{y} - y)] = x_i^2(\hat{y} - \hat{y}^2) \geq 0$$

$$x_i^2 \geq 0 \quad \hat{y} - \hat{y}^2 \in \left[0, \frac{1}{4}\right]$$

Derivative

$$\frac{\partial L}{\partial \theta_i} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial \theta_i}$$

$$\frac{\partial L}{\partial \hat{y}} = -\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}} = \frac{\hat{y} - y}{\hat{y}(1-\hat{y})}$$

$$\frac{\partial \hat{y}}{\partial z} = \hat{y}(1 - \hat{y})$$

$$\frac{\partial z}{\partial \theta_i} = x_i$$

$$\frac{\partial L}{\partial \theta_i} = x_i(\hat{y} - y)$$

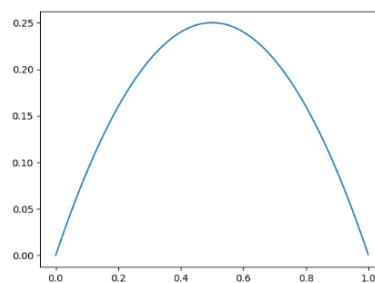


Figure 11: BCE Convex

1.7 So sánh MSE và BCE

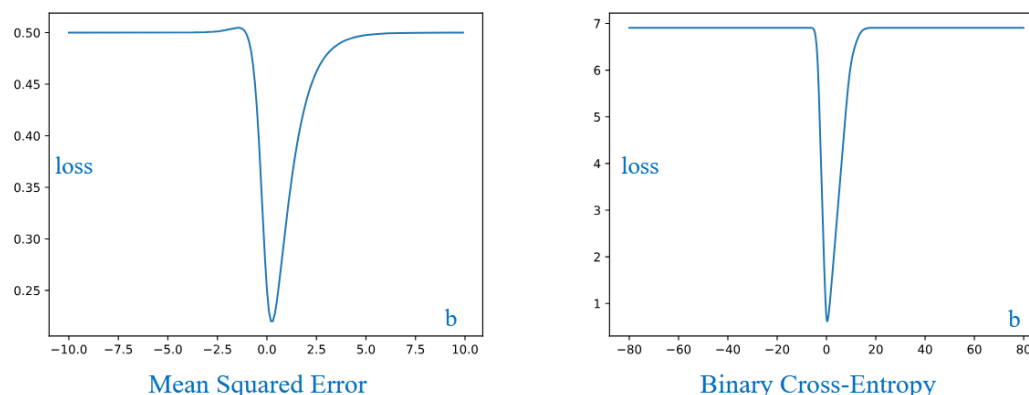


Figure 12: MSE vs BCE

MSE: gentle curve, Phạt nhẹ các dự đoán sai. **BCE:** shallow curve, Phạt nặng các dự đoán sai. Vậy BCE là lựa chọn nhất để tính Loss cho Logistic Regression.

1.8 Phần 1.4: Xây dựng hàm Loss từ Xác Suất Thống Kê

1.9 Giải thích Binary-Cross Entropy từ góc nhìn của Xác Suất Thống Kê

Entropy là gì? Entropy measure information, blah blah

Entropy thể hiện ở đâu trong Logistic Regression ?

Phần 2: Mở Rộng: Nhân ma trận với Logistic Regression

- Giải thích 1 số lý thuyết đại số tuyến tính trong nhân ma trận. + Nhân nghịch đảo - Ví dụ với ma trận X và w . Nhân ma trận. \rightarrow tăng tính tương tác. Gợi ý thay biến để ng đọc học qua tương tác.

2.1 Convexity of BCE for Multiple-Variable using Hessian Matrix

Phần 3: Tại sao

3.1 Tại sao lại nhân nghịch đảo ?

3.2 Tại sao lại sử dụng hàm tăng trưởng Sigmoid ?

3.3 Tại sao lại sử dụng $\ln()$ thay vì $\log()$

3.4 Mối liên hệ giữa $\ln()$ và Sigmoid ?

3.5 Tại sao lại sử dụng loglikelihood thay vì probability

3.6 Convex là gì ?

3.7 Convex trong Logistic Regression cho tính toán ma trận