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

Thay vì đi thẳng vào giải thích lý thuyết và công thức, **mục tiêu** của mình lần này sẽ là đi cùng bạn hiểu từng thành phần từ ground up bắt đầu từ trực giác và giả thuyết hình thành nên Linear Regression rồi phát triển lên Logistic Regression.

1.1 Phần 1.1: Vấn đề với Linear Regression

Trong bài tóan dự đoán giá cổ phiếu có 1 biến x, y, 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 mô tả được bằng 1 đường thẳng** (tuyến tính - đỏ). (Note: xu hướng cục bộ - Cam)

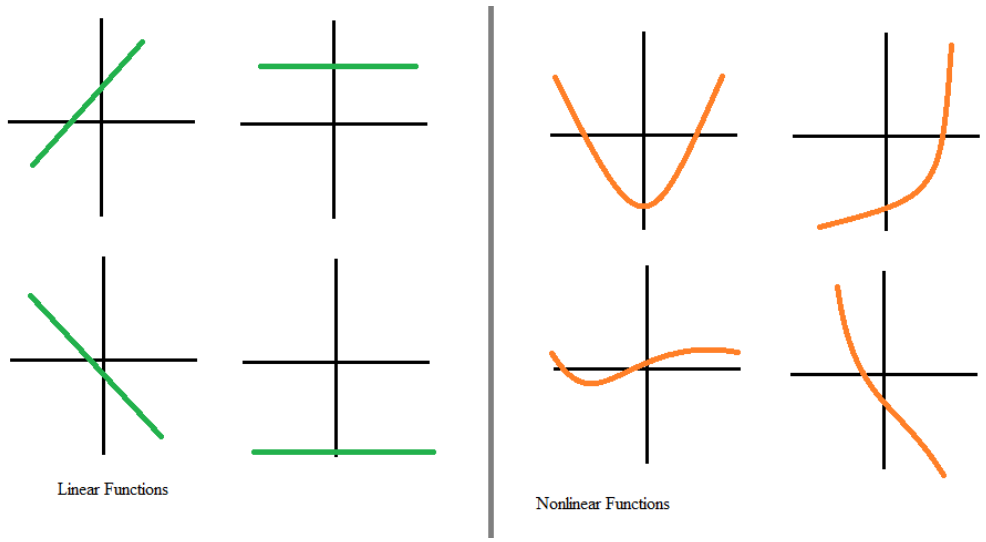


Figure 1: linear function vs nonlinear function

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 $f(x) \approx f(a) + f'(a)(x-a) == \theta_i + \theta_i'(x-a)$.

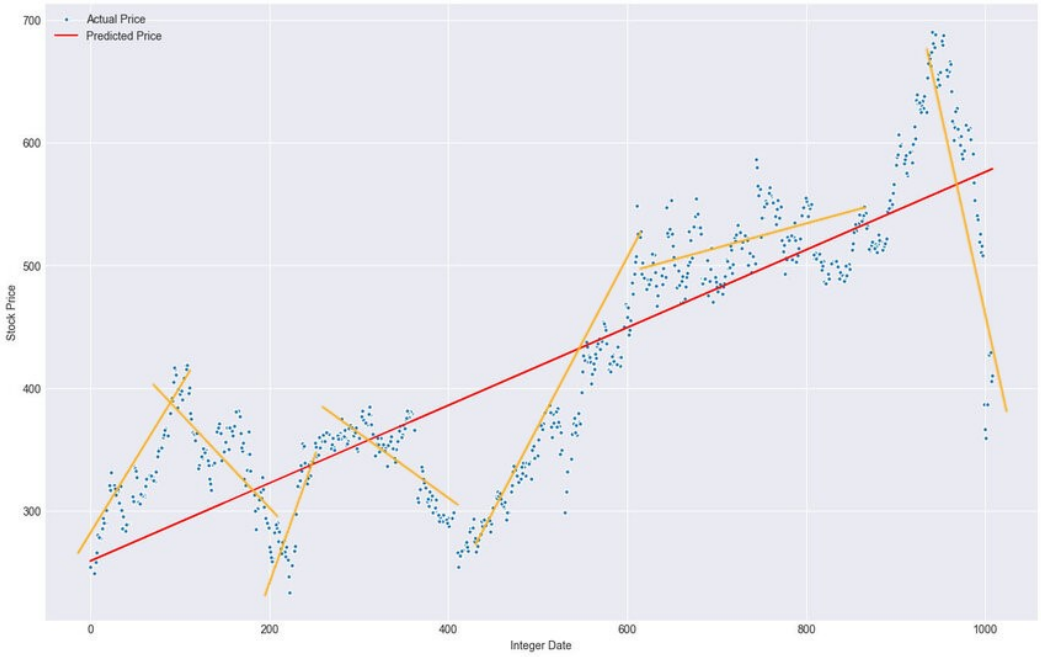


Figure 2: Linear Regression

- (2) **Linear Regression (LR) là mô hình dự đoán đơn giản**, nó chỉ cần 2 tham số: Intercept θ_0 là giá trị ban đầu ($y_0 = \theta_0$) và Slope θ_1 mô tả tốc độ thay đổi (độ dốc) của dữ liệu), còn ϵ_i là sai số, khoảng cách giữa giá trị dự đoán và giá trị thực tế (nếu $\epsilon_i = 0$ thì các điểm dữ liệu sẽ là 1 đường thẳng so với LR.
- (3) **Có thể dùng toán để truy hồi ngược các tham số:** sử dụng hàm số bậc 2 đơn giản để

tính loss (eg. Square Loss), dễ tính tối thiểu sử dụng đạo hàm.

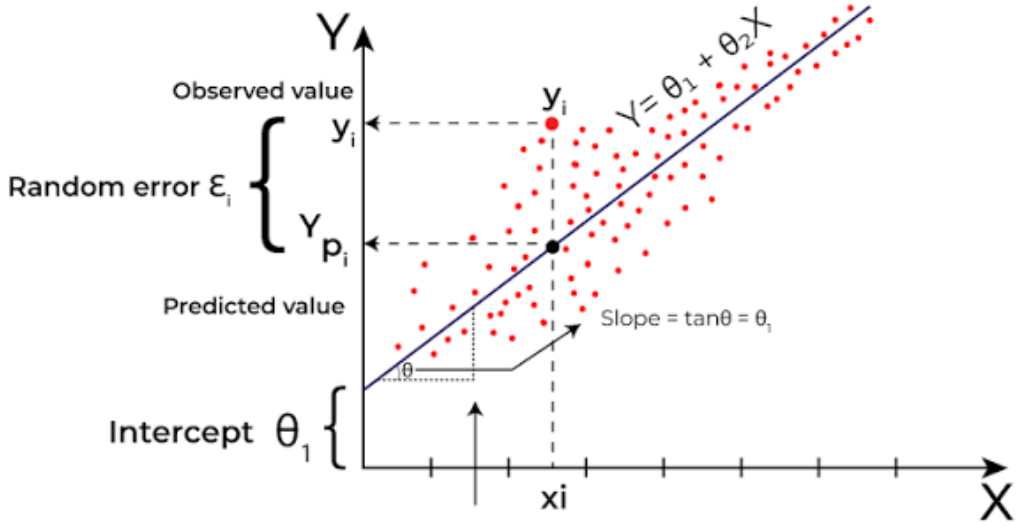
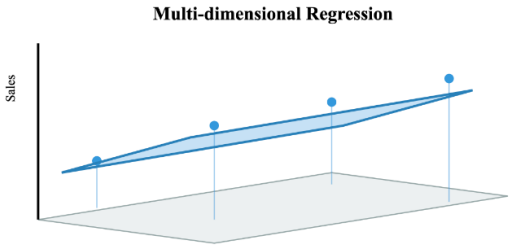


Figure 3: Linear Regression

Cách Linear Regression chọn hàm Loss

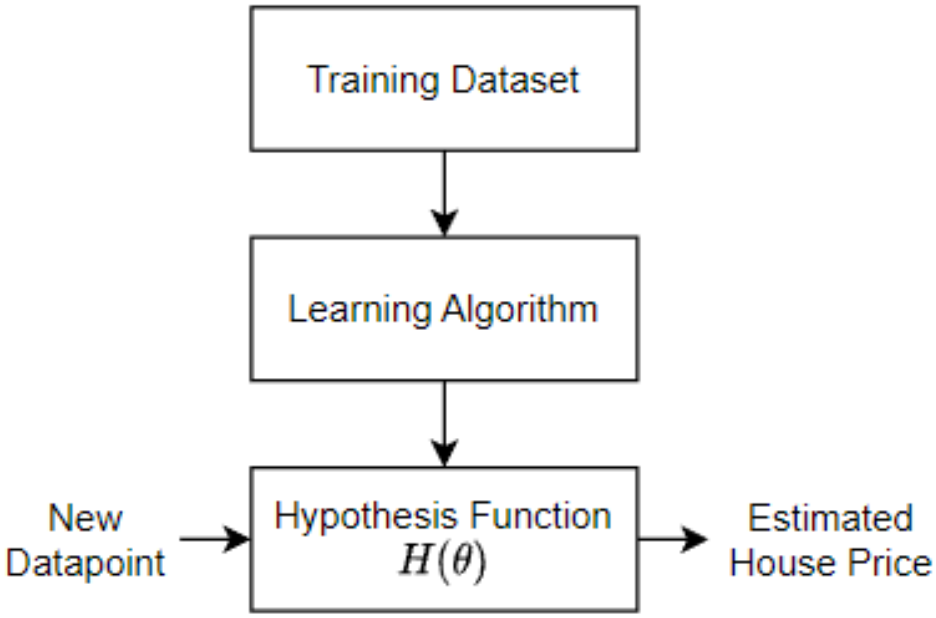
Trong phần này, mình sẽ lấy ví dụ với Linear Regression trong bài toán đa biến có dataset dự đoán giá nhà gồm 3 features (ie. 3 đặc trưng x1, x2, x3) và 1 label (ie. 1 nhãn y) với mục tiêu là thiết kế 1 thuật toán học tập (learning algorithm) để dự đoán giá nhà.

Model: Price = $\theta_0 + \theta_1$(Size) + θ_2(Bedrooms) + θ_3(Age)



Size (x_1)	Bedrooms (x_2)	Age (x_3)	Price (y)
1200	2	10	250k
1800	3	5	450k
2400	4	2	580k
2600	4	15	400k

Mình biết là Linear Regression là Supervised Learning vì nó học nhờ dữ liệu đã được label nên workflow thiết kế là: **Training Dataset → Learning Algorithm (vd. Gradient Descent) → Hàm giả thuyết** $H(\theta)/\hat{y}$ trong đó hàm giả thuyết là đầu ra của Learning Algorithm. Và đầu vào của hàm giả thuyết là dữ liệu mới, đầu ra là giá nhà nó dự đoán. Để dễ tính toán mình sẽ gộp các biến và tham số thành 2 vector: $\theta_i = [\theta_1, \theta_2, \theta_3]$, $X_i = [x_1, x_2, x_3]$. Vậy hàm Linear Regression sẽ thành $\hat{y} = H(X) = \sum_{i=0}^{i} \theta_i X_i$



(1) Để biến đổi 3 features thành giá nhà, mình sẽ nhân mỗi tham số θ với 1 features tương ứng rồi cập nhật nó sử dụng 1 learning algorithm làm sao cho giá nhà dự đoán ($\hat{y}$) gần với giá nhà thực tế nhất, nói cách khác mình cần tối thiểu hàm $H()$.

-> **Xác định mục tiêu mới cho Blog: Làm sao để thiết kế 1 Learning Algorithm/Hàm Loss** (Gradient Descent is a learning algorithm) Hàm Loss -> để minimize -> vector -> Gradient Descent -> visualize hàm Loss vì sao có hình như thế kia a bowl -> Kết luận hàm các bước để tối ưu hàm loss -> thay số với Matrix

Để đưa dự đóan về khoảng [0, 1] mình cần dùng Sigmoid cho Logistic 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 hơi sau khi xây dựng xong BCE)

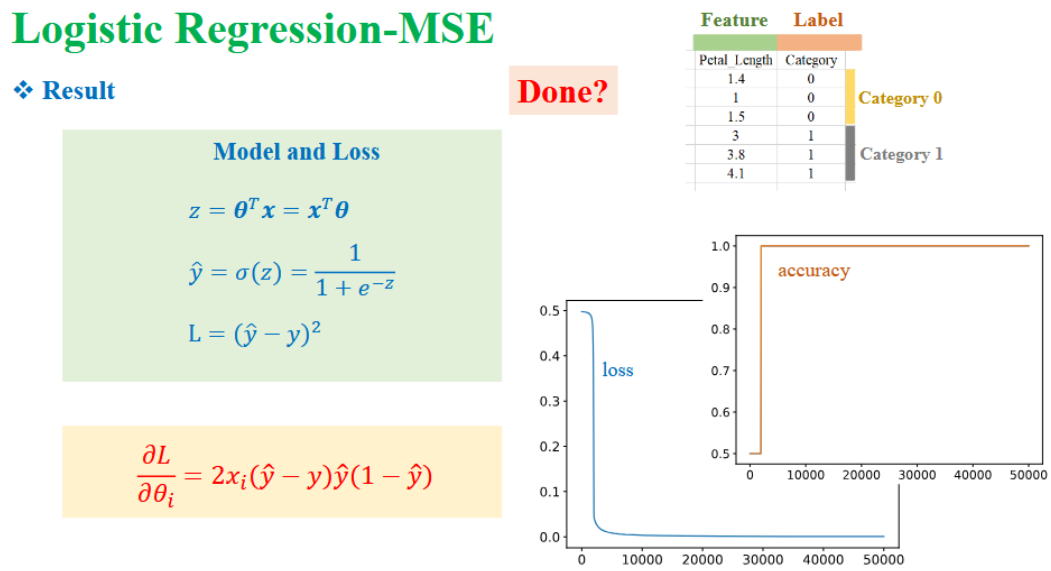


Figure 4: Problem with Non-Convex Loss Function

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

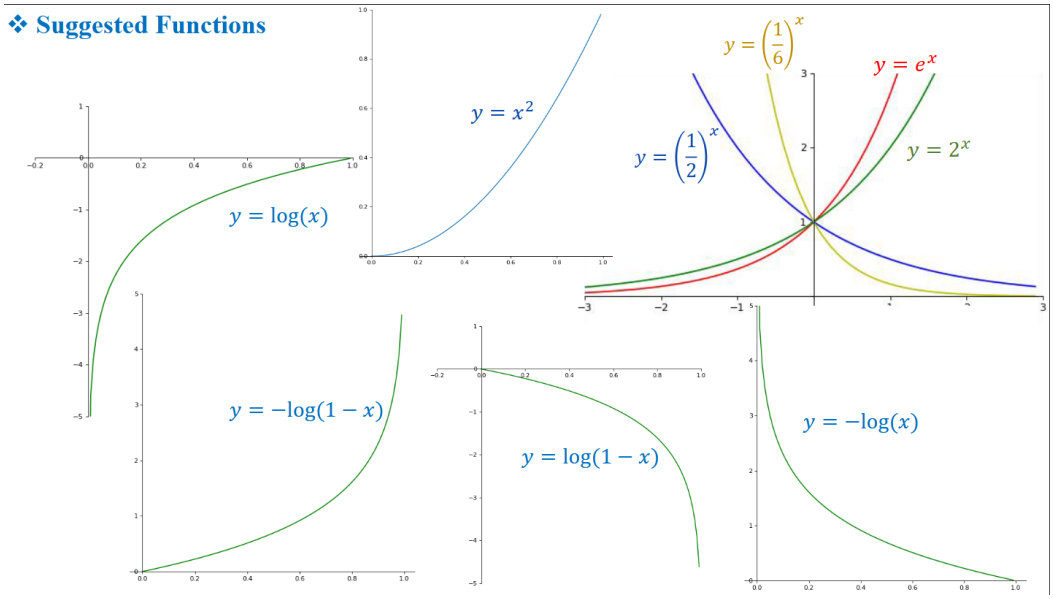


Figure 5: Suggesst Loss Function

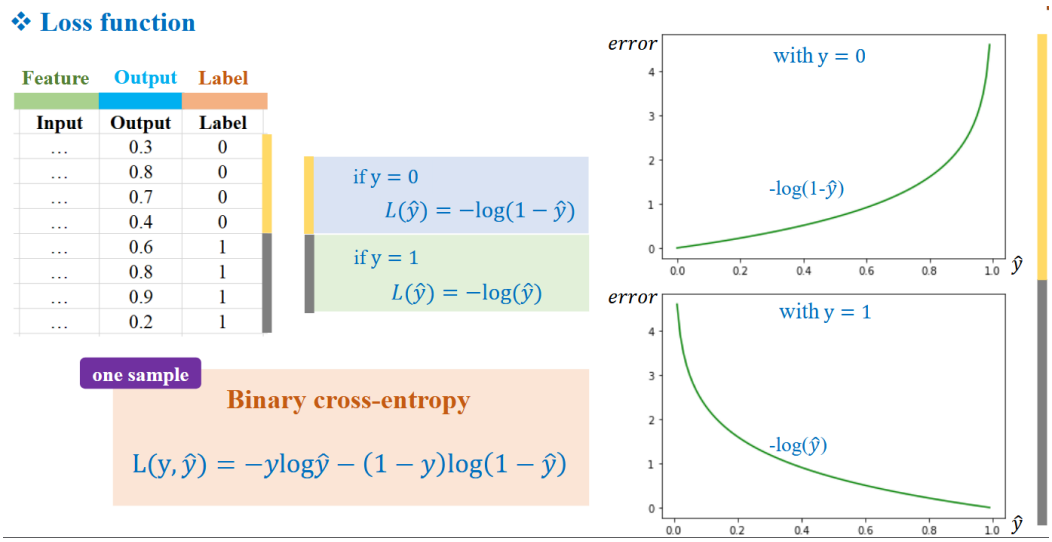


Figure 6: Binary Cross Entropy Loss

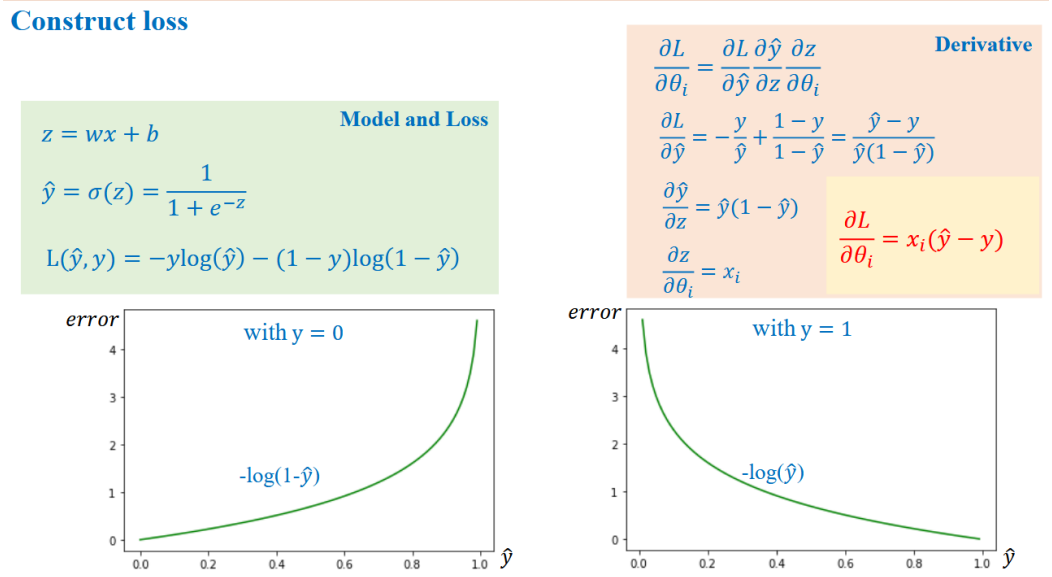


Figure 7: BCE Summarize

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

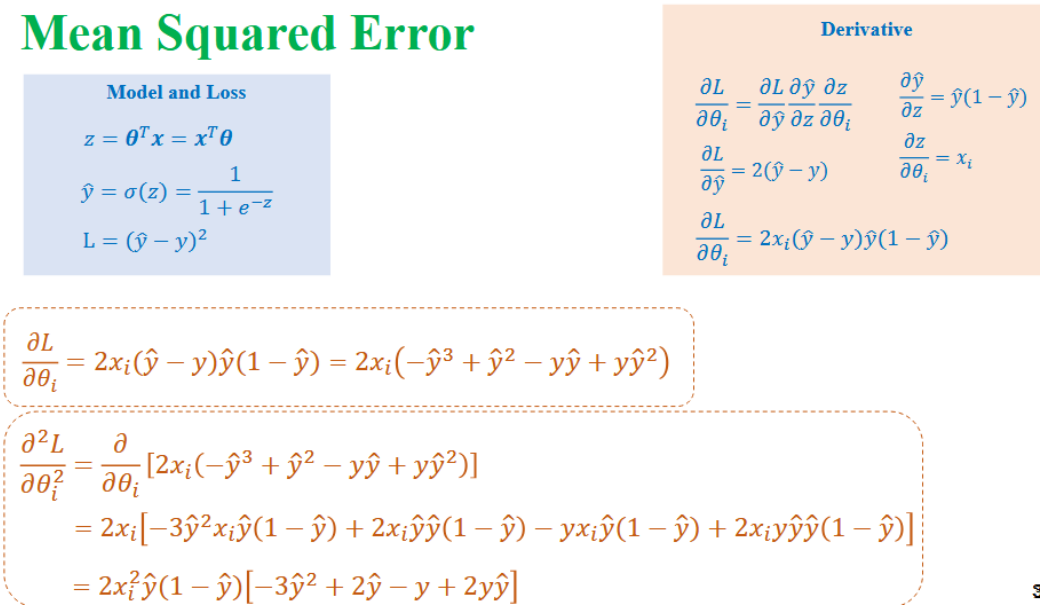


Figure 8: MSE vs BCE

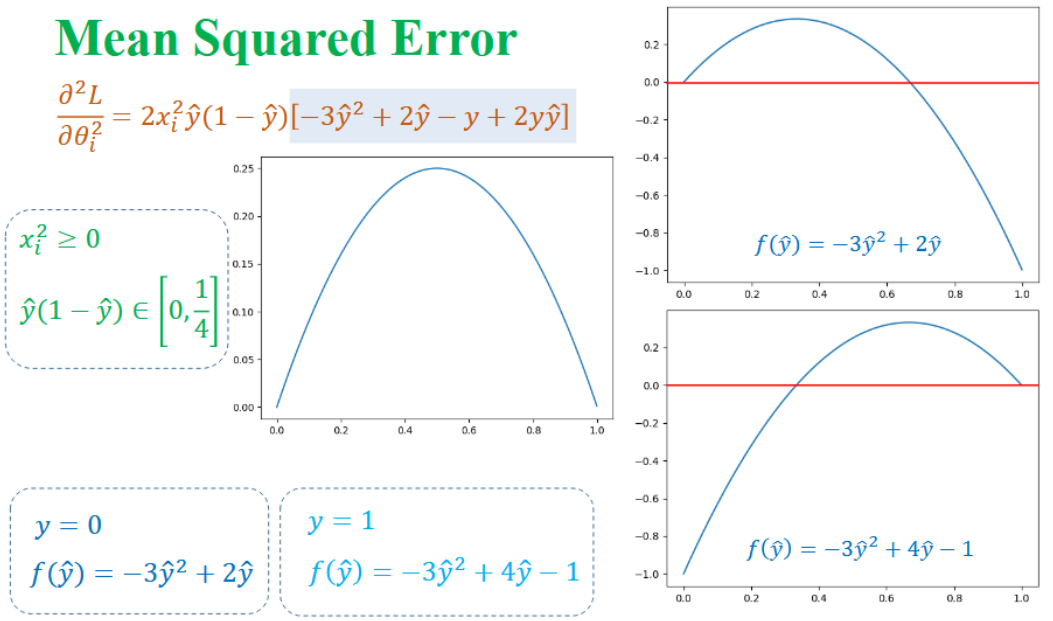


Figure 9: MSE vs BCE

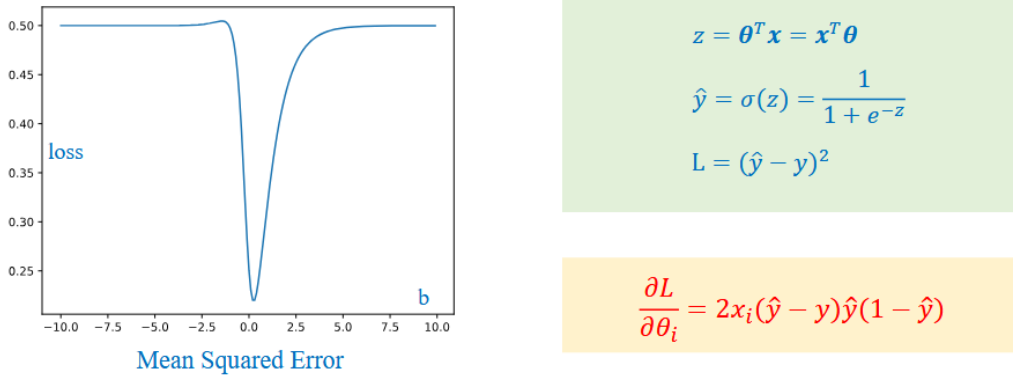


Figure 10: MSE vs BCE

Chứng Minh Convex (độ lồi) của BCE

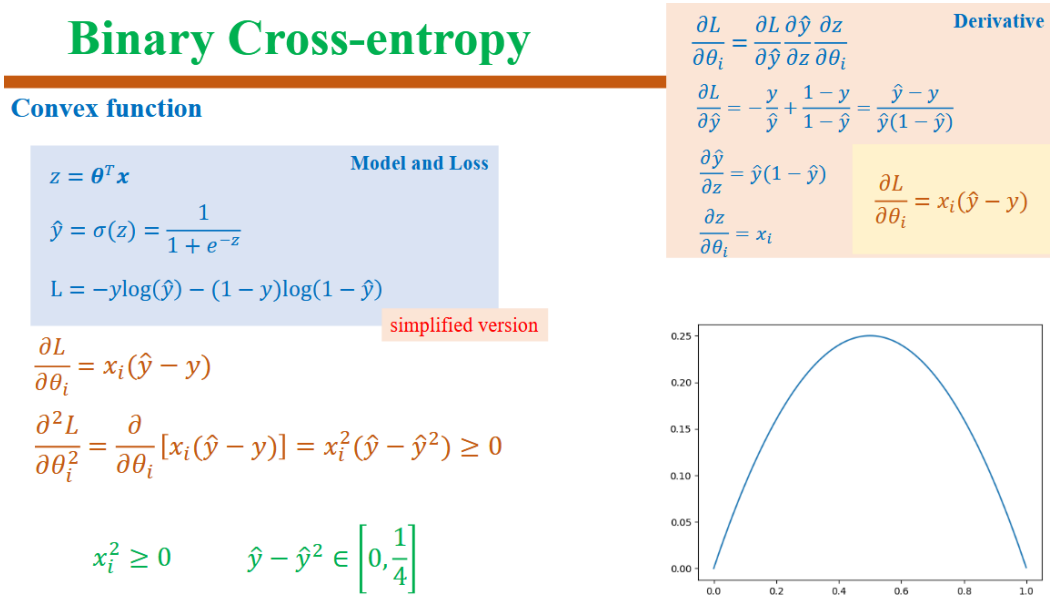


Figure 11: BCE Convex

1.7 So sánh MSE và BCE

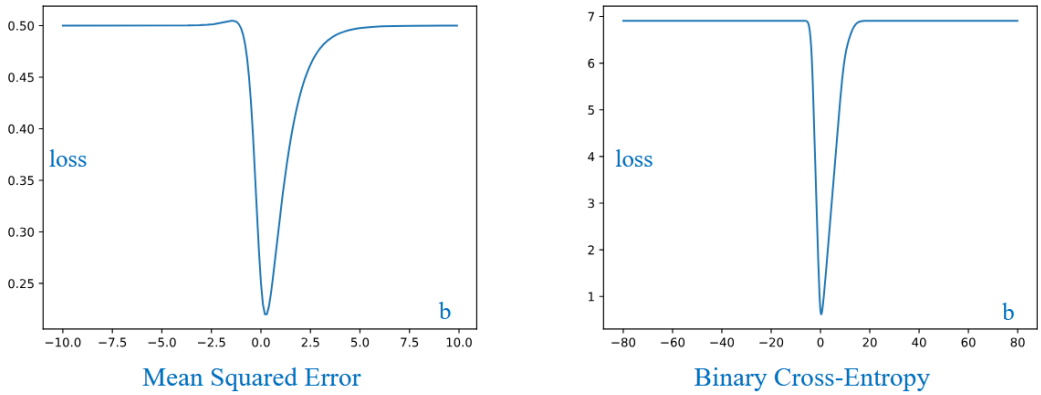


Figure 12: MSE vs BCE

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 dâu 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. -> 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