

Week 5: Logistic Regression

Ekarat Rattagan

January 29, 2026

Logistic Regression Overview

- Logistic regression is used for binary/multi-class classification
- Logistic regression outputs probability $\in (0, 1)$
- Example: cancer prediction, churn prediction, attrition prediction, etc.

Intuition

- Linear regression cannot bound output between 0 and 1, i.e.
unbounded: $(-\infty, +\infty)$
- Apply sigmoid function $\sigma(z)$ to squash the linear regression output into $(0, 1)$

Given, $z = \theta^T x$

$$\sigma(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(\theta^T x)}}, \quad \text{where } e \approx 2.71828$$

- S-shaped curve centered at $z = 0$

Linear VS Logistic Regression

- Linear regression: $h_{\theta}(x) = \theta^T x$
- Logistic regression: $h_{\theta}(x) = \sigma(\theta^T x)$

Decision Boundary

Suppose we predict:

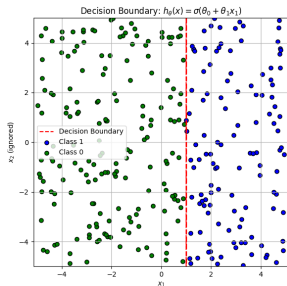
- $y = 1$ if $h_{\theta}(x) \geq 0.5$
- $y = 0$ if $h_{\theta}(x) < 0.5$

$$h_{\theta}(x) = \sigma(\theta_0 + \theta_1 x_1)$$

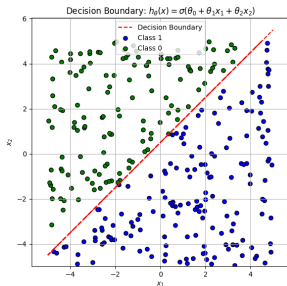
$$h_{\theta}(x) = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

$$h_{\theta}(x) = \sigma(-1 + x_1^2 + x_2^2)$$

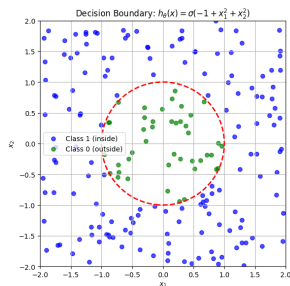
Three Decision Boundaries



$$h_{\theta}(x) = \sigma(\theta_0 + \theta_1 x_1)$$



$$h_{\theta}(x) = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$



$$h_{\theta}(x) = \sigma(-1 + x_1^2 + x_2^2)$$

Ref: Colab

Cost function for Logistic Regression?

- MSE (Mean Squared Error) is used in linear regression:

$$J(\theta) = \frac{1}{N} \sum_{i=1}^N (h_{\theta}(x_i) - y_i)^2$$

- Not ideal for logistic regression because
 - Non-convex curve
 - Poor convergence properties

Negative Log Likelihood Loss (NLLL)

- General form:

$$J(\theta) = \frac{1}{N} \sum_{i=1}^N \text{Cost}(h_{\theta}(x_i), y_i)$$

- Cost function (NLLL):

$$\text{Cost}(h_{\theta}(x), y) = \begin{cases} -\ln(h_{\theta}(x)) & \text{if } y = 1 \\ -\ln(1 - h_{\theta}(x)) & \text{if } y = 0 \end{cases}$$

- Combined:

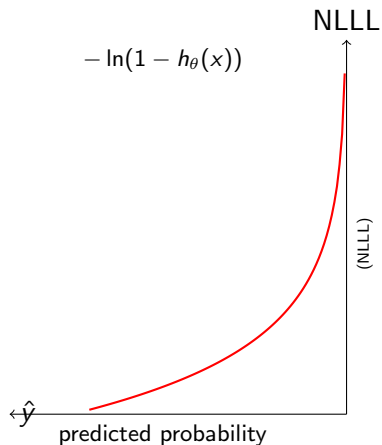
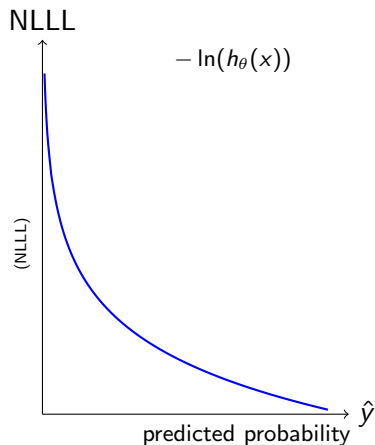
$$J(\theta) = -\frac{1}{N} \sum_{i=1}^N [y_i \ln h_{\theta}(x_i) + (1 - y_i) \ln(1 - h_{\theta}(x_i))]$$

- Based on Bernoulli distribution likelihood

$$P(y_i | x_i; \theta) = (h_{\theta}(x_i))^{y_i} \cdot (1 - h_{\theta}(x_i))^{1-y_i}$$

$$\ln(P(y_i | x_i; \theta)) = \textcolor{red}{y_i} \ln(h_{\theta}(x_i)) + (1 - \textcolor{red}{y_i}) \ln(1 - h_{\theta}(x_i))$$

Cross-Entropy Loss



Derivation of Gradient (1)

$$J(\theta) = -\frac{1}{N} \sum_{i=1}^N (y_i \ln(h_{\theta}(x_i)) + (1 - y_i) \ln(1 - h_{\theta}(x_i)))$$

$$\Rightarrow \frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{N} \sum_{i=1}^N \left(y_i \cdot \frac{1}{h_{\theta}(x_i)} \cdot \frac{\partial h_{\theta}(x_i)}{\partial \theta_j} \right. \\ \left. + (1 - y_i) \cdot \frac{1}{1 - h_{\theta}(x_i)} \cdot \frac{\partial (1 - h_{\theta}(x_i))}{\partial \theta_j} \right)$$

Note: $\frac{\partial \ln(f(x))}{\partial x} = \frac{1}{f(x)} \frac{\partial (f(x))}{\partial x}$

Derivation of Gradient (2)

Note: $\frac{\partial \sigma(f(x))}{\partial x} = \sigma(f(x))(1 - \sigma(f(x))) \frac{\partial(f(x))}{\partial x}$

$$\begin{aligned} \frac{\partial J(\theta)}{\partial \theta_j} = & -\frac{1}{N} \sum_{i=1}^N \left(y_i \cdot \frac{1}{h_{\theta}(x_i)} \cdot \frac{\partial \sigma(\theta^T x)}{\partial \theta_j} \right. \\ & \left. + (1 - y_i) \cdot \frac{1}{1 - h_{\theta}(x_i)} \cdot (-1) \cdot \frac{\partial \sigma(\theta^T x)}{\partial \theta_j} \right) \end{aligned}$$

$$\begin{aligned} = & -\frac{1}{N} \sum_{i=1}^N \left(y_i \cdot \frac{1}{h_{\theta}(x_i)} \cdot \sigma(\theta^T x)(1 - \sigma(\theta^T x))x_{ij} \right. \\ & \left. + (1 - y_i) \cdot \frac{-1}{1 - h_{\theta}(x_i)} \cdot \sigma(\theta^T x)(1 - \sigma(\theta^T x))x_{ij} \right) \end{aligned}$$

Simplified Gradient Expression

$$\begin{aligned} &= -\frac{1}{N} \sum_{i=1}^N (y_i(1 - h_{\theta}(x_i))x_{ij} + (1 - y_i)(-1)h_{\theta}(x_i)x_{ij}) \\ &= -\frac{1}{N} \sum_{i=1}^N (y_ix_{ij} - y_ih_{\theta}(x_i)x_{ij} - h_{\theta}(x_i)x_{ij} + y_ih_{\theta}(x_i)x_{ij}) \\ &= -\frac{1}{N} \sum_{i=1}^N (y_ix_{ij} - h_{\theta}(x_i)x_{ij}) = -\frac{1}{N} \sum_{i=1}^N (y_i - h_{\theta}(x_i))x_{ij} \\ &\Rightarrow \frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{N} \sum_{i=1}^N (h_{\theta}(x_i) - y_i)x_{ij} \quad \text{for } j = 0, \dots, d \end{aligned}$$

*** Same form as BGD of MSE**

Linear vs Logistic Regression

Linear Regression	Logistic Regression
$h_{\theta}(x) = \theta^T x$	$h_{\theta}(x) = \sigma(\theta^T x)$
$\text{MSE} = J(\theta) = \frac{1}{N} \sum_{i=1}^N (h_{\theta}(x_i) - y_i)^2$	$\text{NLLL} = J(\theta) = -\frac{1}{N} \sum_{i=1}^N (y_i \ln(h_{\theta}(x_i)) + (1 - y_i) \ln(1 - h_{\theta}(x_i)))$
MSE, R^2 , MAE	Accuracy, Precision, Recall, F1, AUC
$\theta_j^{(t+1)} = \theta_j^{(t)} - \eta \cdot \frac{1}{N} \sum_{i=1}^N (h_{\theta}(x_i) - y_i) x_{ij}$	$\theta_j^{(t+1)} = \theta_j^{(t)} - \eta \cdot \frac{1}{N} \sum_{i=1}^N (h_{\theta}(x_i) - y_i) x_{ij}$

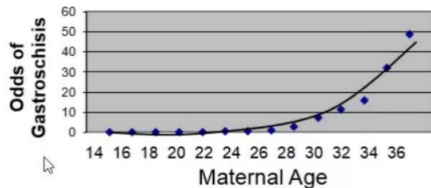
Appendix: Odds

- Odds refer to the probability of an event happening divided by the probability of it not happening.

$$\text{Odds} = \frac{P(\text{event}=\text{success})}{1 - P(\text{event}=\text{success})}$$

- Examples:

$$\frac{0.8}{0.2} = 4, \quad \frac{0.9}{0.1} = 9, \quad \frac{0.5}{0.5} = 1, \quad \frac{0.2}{0.8} = 0.25$$



Maternal age VS Odds of Gastroschisis



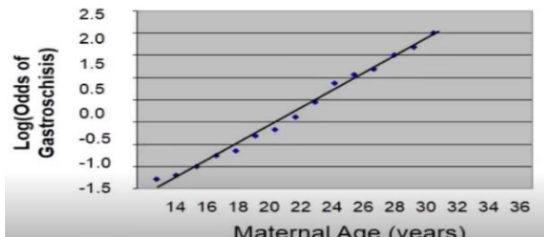
Gastroschisis

Appendix: Logit Function

Take $\ln(\cdot)$ to the chart above, we obtain the linear relationship.

$$\ln(\text{Odds of Gastroschisis}) = \theta_0 + \theta_1 \cdot \text{Age}$$

This is called a logit function.



Appendix: Logistic Regression

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

Exponentiate and take the multiplicative inverse of both sides,

$$\frac{1-p}{p} = \frac{1}{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

Partial out the fraction on the left-hand side of the equation and add one to both sides,

$$\frac{1}{p} = 1 + \frac{1}{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

Change 1 to a common denominator,

$$\frac{1}{p} = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) + 1}{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

Finally, take the multiplicative inverse again to obtain the formula for the probability $P(Y = 1)$,

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

ref: stats.oarc.ucla.edu

- Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemiamedica*, 24(1), 12-18.