

HÍK - KIỂM TRA

$$h(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

⇒ Viết thuật toán Gradient Descent

ÔN TẬP THI GIỮA KÌ

⌚ Các chủ đề ôn tập thi giữa kỳ

1. Linear Regression
2. Logistic Regression
3. Model Validation (Overfitting and Regularization)
4. Neural Networks

I> Linear Regression:

Tập huấn luyện: $(x^{(i)}, y^{(i)})$, $i = 1, m$

Hypothesis: $h(x) = \theta_0 + \theta_1 x \rightarrow x: \text{Input}$
 \downarrow
 $h(x): \text{Output}$

Mục tiêu: Xác định θ_0, θ_1 để mô hình $h(x)$ khớp với dữ liệu huấn luyện nhất.

Cho x , xác định $h(x)$ sao cho $h(x)$ gần y nhất!

* Về cốt bản thì các thuật ngữ sau có ý nghĩa giống nhau:

Cost function = Loss function = Error function.

Lỗi của mô hình ứng với 1 input:

$$= \frac{1}{2} (h(x) - y)^2 = \frac{1}{2} (\theta_0 + \theta_1 x - y)^2$$

Mãm chi phí

$$J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^m \frac{1}{2} (h(x^{(i)}) - y^{(i)})^2$$

$$\Rightarrow J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (\theta_0 + \theta_1 x^{(i)} - y^{(i)})^2$$

Mục tiêu: Tìm giá trị của θ_0, θ_1 sao cho $J(\theta_0, \theta_1)$ đạt mức thấp.

Minimize $\theta_0, \theta_1 \ J(\theta_0, \theta_1)$

Thực hiện hóa đơn (Gradient Descent)

$$\text{Với } J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (\theta_0 + \theta_1 x^{(i)} - y^{(i)})^2 \Rightarrow \text{Tìm } \theta_0, \theta_1 \text{ để } J(\theta_0, \theta_1) \text{ nhỏ nhất.}$$

Bắt đầu từ 1 điểm θ_0, θ_1 nào đó (VD: $\theta_0 = 0, \theta_1 = 0$)

Thay đổi θ_0, θ_1 cho đến khi $J(\theta_0, \theta_1)$ đạt mức thấp.

\Rightarrow Thay đổi θ_0, θ_1 như thế nào?

\Rightarrow Sử dụng Gradient Descent.

Đạo hàm riêng phần:

$$\frac{dJ}{d\theta_0} = \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})$$

$$\frac{dJ}{d\theta_1} = \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) x^{(i)}$$

$$\Rightarrow \begin{cases} \frac{dJ}{d\theta_0} = \frac{1}{m} \sum_{i=1}^m (\theta_0 + \theta_1 x^{(i)} - y^{(i)}) \\ \frac{dJ}{d\theta_1} = \frac{1}{m} \sum_{i=1}^m (\theta_0 + \theta_1 x^{(i)} - y^{(i)}) x^{(i)} \end{cases}$$

\Rightarrow Vecto gradient $\left(\frac{dJ}{d\theta_0}, \frac{dJ}{d\theta_1} \right)$

Trường ngược với vecto gradient làm cho hàm số giảm dần.

Thuật toán hạ độe:

Lặp cho đến khi hội tụ:

{

$$\theta_0 = \theta_0 - \alpha \frac{dJ}{d\theta_0}$$

+) α là hệ số học

+) Các bước thực hiện:

1) Tính vecto gradient

2) Cập nhật các thành phần của vecto θ

3) Tính lại chi phí và thông báo

{ α quá nhỏ \rightarrow hội tụ chậm

α quá lớn \rightarrow ko hội tụ

}

Hai quy trình tính đa biến.

{ Đam biến: $h_\theta(x) = \theta_0 + \theta_1 x$: giá trị số

Đa biến:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n : \text{veeto.}$$

Hypothesis

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

$$x \in \mathbb{R}^n, \theta \in \mathbb{R}^{n+1}$$

Đặt $x_0 = 1$.

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \dots \\ \theta_n \end{bmatrix} \Rightarrow h_{\theta}(x) = \theta^T x$$

$\theta, x \in \mathbb{R}^{n+1}$

Hàm chi phí

$$\text{Hypothesis } h_{\theta}(x) = \theta^T x$$

$$\text{Cố tham số } \theta = [\theta_0, \theta_1, \theta_2, \dots, \theta_n]^T$$

$$\text{Hàm chi phí: } J(\theta) = \frac{1}{2m} \sum_{i=1}^m (\theta^T x^{(i)} - y^{(i)})^2$$

Vecto gradient:

$$\textcircled{R} \quad \frac{\partial J}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^m (\theta^T x^{(i)} - y^{(i)}) x_j^{(i)}$$

$$j = 0, 1, 2, \dots, n$$

Repeat

{

until

$$\theta_j = \theta_j - \alpha \frac{\partial J}{\partial \theta_j}$$

convergence

}

}

Chuẩn hóa đặc trưng

$$x_j = \frac{x_j - \mu_j}{s_j} \Rightarrow x_j \in [-1, 1]$$

Tổ công thức.

$$\theta_j = \theta_j - \alpha \frac{\partial J}{\partial \theta_j} = \theta_j - \frac{\alpha}{m} \sum_{i=1}^m (\theta^T x^{(i)} - y^{(i)}) x_j^{(i)}$$

$$= \theta_j - \frac{\alpha}{m} (\theta^T x^{(i)} x_j^{(i)} - y^{(i)} x_j^{(i)}) = \dots \text{ (tính nhau)}$$

II > Logistic Regression:

Sử dụng để phục vụ cho bài toán phân loại.

Thường là Yes / No $\{1, 0\}$

Đường phân loại $h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$

Luật phân loại $\begin{cases} \text{Nếu } h_{\theta}(x) \geq 0, y = 1 \\ \text{Nếu } h_{\theta}(x) < 0, y = 0 \end{cases}$

Đường phân loại: $h_{\theta}(x) = 0 \Leftrightarrow \theta^T x = 0$

$\begin{cases} \text{Nếu } \theta^T x \geq 0, y = 1 \\ \text{Nếu } \theta^T x < 0, y = 0 \end{cases}$

Chúng ta cần: $\begin{cases} y = 1, h_{\theta}(x) \rightarrow 1 \\ y = 0, h_{\theta}(x) \rightarrow 0 \end{cases}$

Mô hình mới: $h_{\theta}(x) = g(\theta^T x)$

$\begin{cases} \theta^T x \text{ càng lớn hơn } 0 \text{ thì } g(\theta^T x) \text{ càng tiến tới } 1. \\ \theta^T x \text{ càng nhỏ hơn } 0 \text{ thì } g(\theta^T x) \text{ càng tiến tới } 0 \end{cases}$

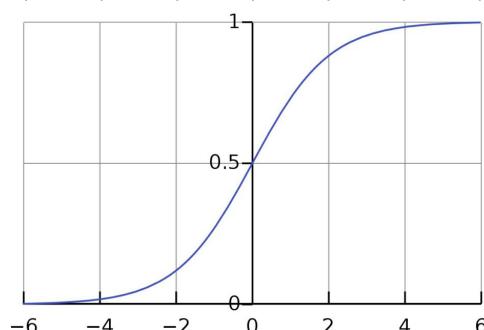
Với g là hàm sigmoid $g(z) = \frac{1}{1 + e^{-z}}$

$$\Rightarrow h_{\theta}(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

Chú ý: $y \neq h_{\theta}(x)$

Nếu $h_{\theta}(x) \geq 0,5, y = 1$
 $\Rightarrow \theta^T x \geq 0$

$h_{\theta}(x) < 0,5, y = 0$
 $\Rightarrow \theta^T x < 0$



* $h_{\theta}(x)$ chính là xác suất $y = 1$

$$h_{\theta}(x) = P(y = 1 | x, \theta)$$

$$\left. \begin{array}{l} y = 1, h_{\theta}(x) \geq 0,5 \\ \theta^T x \geq 0 \end{array} \right\} \quad \left. \begin{array}{l} y = 0, h_{\theta}(x) < 0,5 \\ \theta^T x < 0 \end{array} \right\}$$

$$\text{Hypothesis : } h_{\theta}(x_i) = \frac{1}{1 + e^{-\theta^T x}}$$

$$0 \leq h_{\theta}(x_i) \leq 1$$

Hàm Chi phí:

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

Tùy hàm chi phí, ta thấy:

Với $y = 1$, nếu $h_{\theta}(x)$ càng gần 1 thì chi phí $\rightarrow 0$

nếu $h_{\theta}(x)$ càng gần 0 thì chi phí $\rightarrow \infty$

Với $y = 0$, nếu $h_{\theta}(x)$ càng gần 0 thì chi phí $\rightarrow 0$

nếu $h_{\theta}(x)$ càng gần 1 thì chi phí $\rightarrow \infty$

Tùy ý thức của bạn chi phí, ta có thể viết lại thành 1 biểu thức:

$$\text{Cost}(h(x), y) = -y^{(i)} \log h_{\theta}(x^{(i)}) - (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)}))$$

Ta suy ra được hàm chi phí áp dụng cho toàn bộ tập dữ liệu:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m y^{(i)} \log h_{\theta}(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)}))$$

$$\frac{\partial J}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$



Gradient Descent:

{

$$\theta_j = \theta_j - \alpha \frac{\partial J}{\partial \theta_j}$$

Tương tự như mô hình

Linear Regression.

}

Hồi quy Logistic với nhiều lớp

$$h_{\theta}^C(x) = P(y = C | x; \theta)$$

Để obtain cho input mới: $y = \max_C h_{\theta}^C(x)$

III > Overfitting, Regularization and Model Validation:

Mô hình quá khớp dữ liệu (Overfitting)

Khi số lượng đặc trưng lớn:

- + Mô hình có "thể" quá khớp với dữ liệu học (chi phí ≈ 0)
- + Không thể tăng quá mức cho dữ liệu mới (chi phí $\gg 0$)

Giai pháp cho mô hình quá khớp dữ liệu:

1. Giảm số đặc trưng: ? Dimensionality Reduction?

+ Lựa chọn đặc trưng qua phân tích dữ liệu.

+ Lựa chọn đặc trưng qua đánh giá mô hình.

2. Bình thường hóa tham số: (Regularization)

+ Giảm mức ảnh hưởng của tham số đối với output.

+ Tốt trong trường hợp các đặc trưng chi phí nhỏ đối với output.

Lựa chọn đặc trưng:

\Rightarrow Tùy hiệu lùi trùi mồi liên hệ giữa input và output, sử dụng cái công cụ như histogram, scatter plot để thử nghiệm.

+ Lựa chọn đặc trưng bằng cách lượng correlation

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \text{Corr}(x, y)$$

x_i, y_i : variable in a sample \bar{x}, \bar{y} : means of variable.

Sau khi lựa chọn đặc trưng:

+ Kiểm tra mô hình với các đặc trưng đã chọn.

+ Lặp lại quá trình lựa chọn cho tới khi mô hình ổn định.

Bình thường hóa tham số: \Rightarrow Giảm overfitting.

* Giá trị tham số nhỏ hơn tạo ra mô hình đơn giản hơn.

VD: $\left\{ \begin{array}{l} \theta_0 + \theta_1 x + \theta_2 x^2 \end{array} \right.$

$\Leftrightarrow \theta_0 + \theta_1 x + \theta_2 x^2 + \cancel{\theta_3 x^3} + \cancel{\theta_4 x^4}$

Hàm chi phí:

$$J(\theta) = \frac{1}{2m} \left[\sum_{i=1}^m (\hat{y}_i - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right]$$

- Lambda (λ) : weight decay (trong số phần rã)

- Lambda càng lớn, mô hình càng đơn giản.

{ Mô hình quá đơn giản \rightarrow Underfitting }

{ Mô hình quá phức tạp \rightarrow Overfitting }

Tử hàm chi phí $J(\theta)$, ta suy ra vector gradient:

$$\frac{\partial J}{\partial \theta_0} = \frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y^{(i)}) x_0^{(i)}$$

$$\frac{\partial J}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y^{(i)}) x_j^{(i)} + \frac{\lambda}{m} \theta_j$$

\Rightarrow Trở về áp dụng thuật toán hạ độ gradient descent như thường?

Kiểm tra mô hình:

+ Quá trình học làm cho mô hình khớp với dữ liệu học.

+ Mô hình học được có tổng quát hóa cho dữ liệu mới?

\Rightarrow Kiểm tra mô hình với dữ liệu chưa nhìn thấy' (test set)

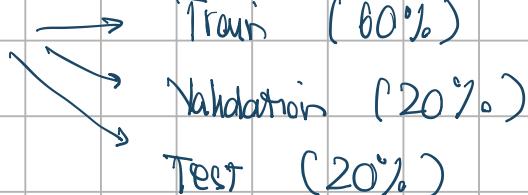
+ Giải pháp:

- Học, kiểm tra và chọn mô hình tốt nhất.

- Kiểm tra mô hình này có tốt cho dữ liệu mới không

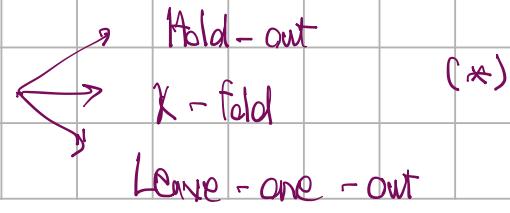


Chia tập dữ liệu:



Cross Validation là phương pháp kiểm tra

để chính xác, sử dụng toàn bộ tập dữ liệu cho quá trình học.



Hàm chi phí:

Train error

$$J_{\text{train}}(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

Cross Validation error: $J_{CV}(\theta) = \frac{1}{2m_{CV}} \sum_{i=1}^{CV} (h_{\theta}(x_{CV}^{(i)}) - y_{CV}^{(i)})^2$

Test error

$$J_{\text{test}}(\theta) = \frac{1}{2m_{\text{test}}} \sum_{i=1}^{\text{test}} (h_{\theta}(x_{\text{test}}^{(i)}) - y_{\text{test}}^{(i)})^2$$

k-fold cross validation:

- + Chia ngẫu nhiên dữ liệu thành k phần.
- + Học và kiểm tra k lần, với k thường là 10.

Generalization Error = Average Error of Each Iteration.

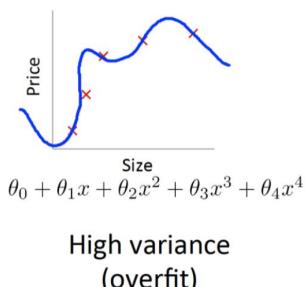
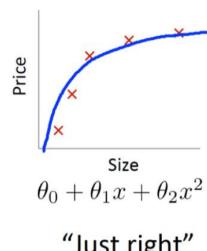
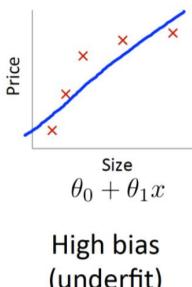
Hiệu chỉnh mô hình:

- * Mô hình học quá nhiều tham số'
- * Thuật toán học có nhiều siêu tham số' (hyper parameters)
- * làm gì khi thuật toán hoạt động không tốt?
 - + Thu thập thêm dữ liệu
 - + Điều chỉnh mô hình
- + Rủi ro tập đặc trưng
- + Giảm weight decay (λ)
- + Thủ với đặc trưng khác
- + Tăng weight decay

Bias và Variance:

- + Bias là lỗi của mô hình trên tập học.
- + Variance là độ lệch giữa lỗi của mô hình trên tập đánh giá và lỗi của mô hình trên tập học.

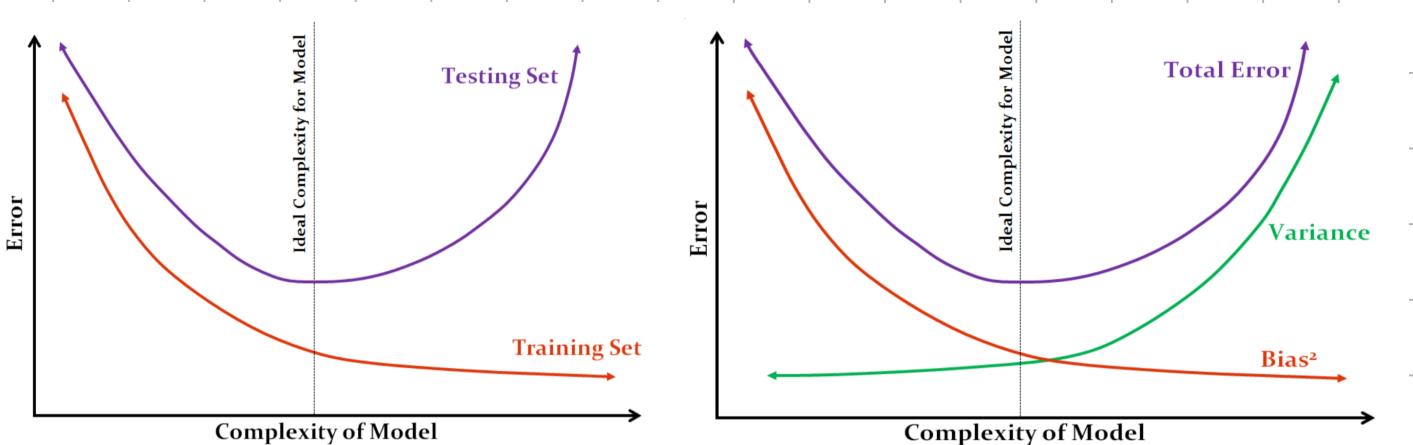
\Rightarrow Hiệu chỉnh cho đến khi cả bias và variance đều đạt mức thấp.



Mô hình đơn giản, bias lớn
 \Rightarrow underfit.

Mô hình phức tạp, variance lớn
 \Rightarrow overfit.

\Rightarrow Cần bằng bias & variance.



Đánh giá mô hình

Mỗi bộ dữ liệu để đánh giá mô hình

$$\text{Precision} = \frac{\text{True Positive}}{\text{Predicted Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Positive}}$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Positive} + \text{Negative}}$$

Tay vào bút toán mà chúng ta cần chọn thuộc về phái hợp.

Bonus: Confusion Matrix and Classification Report in Sklearn.

Confusion Matrix

		ACTUAL	
		Positive	Negative
PREDICTED	Positive	TP	FP
	Negative	FN	TN

⇒ Type I Error

⇒ Classification report

$$\Rightarrow \text{Precision} = \frac{TP}{TP + FP}$$

Support là số lần xuất hiện

thật sự của mỗi lớp trong

tập dữ liệu ⇒ Khuyến khích

$$\Rightarrow \text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Support giữa các lớp

xấp xỉ bằng nhau để

tránh Imbalanced data

Type II

Error

$$\text{Recall} = \frac{TP}{TP + FN}$$

Multiclass for confusion matrix

			gold labels		
			urgent	normal	spam
		urgent	8	10	1
system	urgent	true	11		
system	not	not			
system	urgent	true	60	55	
system	normal	normal	40	212	
system	not	not			
		spam	200	33	
system	spam	true	51	83	
system	not	not			
		yes	268	99	
system	yes	true	99	635	
system	no	not			

$$\text{precision} = \frac{8}{8+11} = .42$$

$$\text{precision} = \frac{60}{60+55} = .52$$

$$\text{precision} = \frac{200}{200+33} = .86$$

$$\text{microaverage precision} = \frac{268}{268+99+51+83} = .73$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = .60$$

system output

			gold labels		
			urgent	normal	spam
		urgent	8	10	1
system	urgent	true	11		
system	not	not			
system	normal	true	60	55	
system	not	not	40	212	
system	spam	true	200	33	
system	not	not	51	83	
		yes	268	99	
system	yes	true	99	635	
system	no	not			

$$\text{precision}_u = \frac{8}{8+10+1}$$

$$\text{precision}_n = \frac{60}{5+60+50}$$

$$\text{precision}_s = \frac{200}{3+30+200}$$

$$\text{recall}_u = \frac{8}{8+5+3}$$

$$\text{recall}_n = \frac{60}{10+60+30}$$

$$\text{recall}_s = \frac{200}{1+50+200}$$

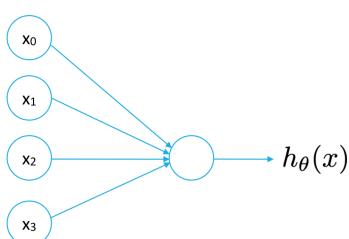
DL > Neural Networks:

Phân lớp phi tuyến (nonlinear): dữ liệu không thể phân lớp bằng màng hình tuyến tính (hyperplane)

Phân lớp phi tuyến với hồi quy logistic cẩn thận nhiều feature

Hồi quy logistic và Mạng Neural

Hồi quy logistic

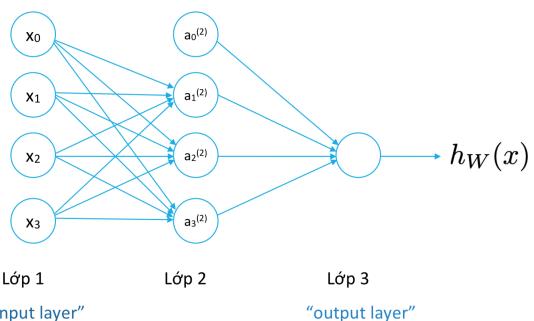


$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Hàm truyền sigmoid

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

Mạng nơ ron



Lớp 1

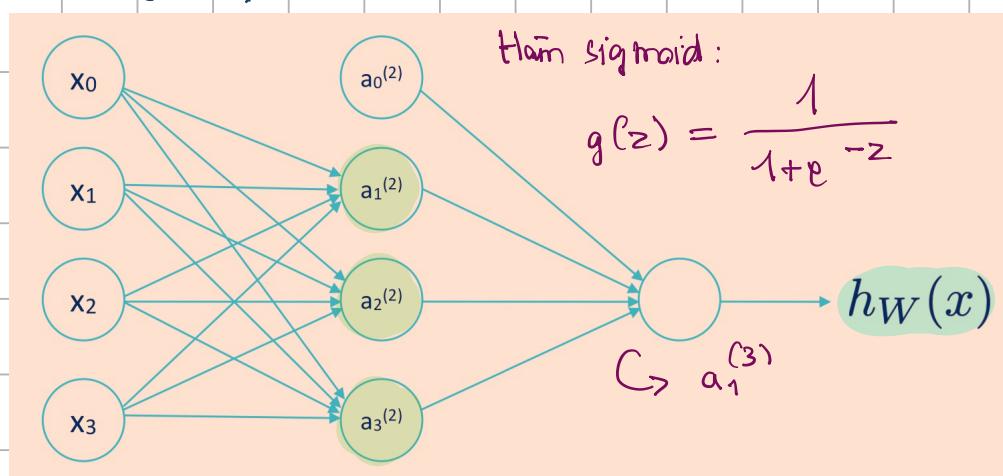
"input layer"

Lớp 2

"output layer"

Hàm truyền tại nút:

Activation / Transfer Function

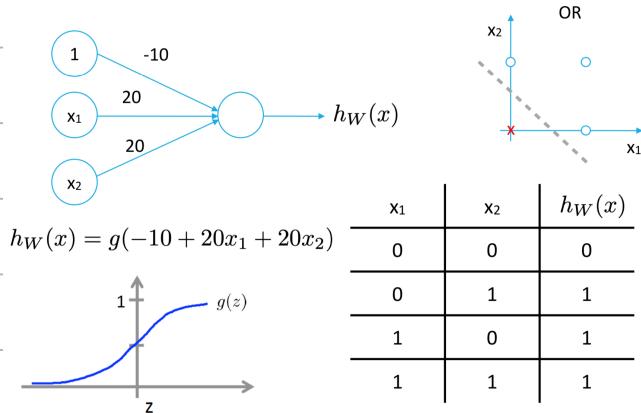


$$\left\{ \begin{array}{l} a_1^{(2)} = g(W_{10}^{(1)} x_0 + W_{11}^{(1)} x_1 + W_{12}^{(1)} x_2 + W_{13}^{(1)} x_3) \\ a_2^{(2)} = g(W_{20}^{(1)} x_0 + W_{21}^{(1)} x_1 + W_{22}^{(1)} x_2 + W_{23}^{(1)} x_3) \\ a_3^{(2)} = g(W_{30}^{(1)} x_0 + W_{31}^{(1)} x_1 + W_{32}^{(1)} x_2 + W_{33}^{(1)} x_3) \end{array} \right.$$

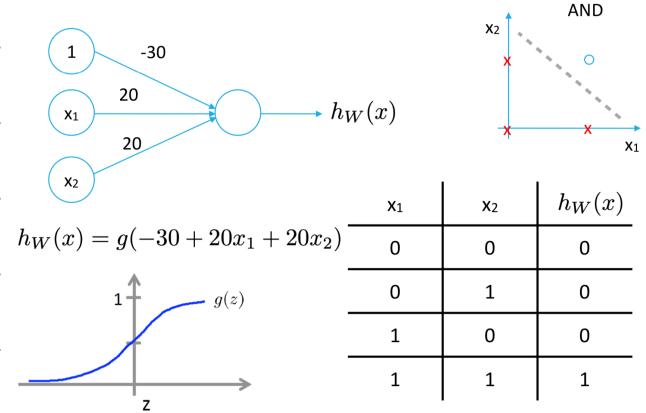
Mạng Neural cẩn` hàm truyền` vì

- +) Tính phi tuyến (nonlinearity) của hàm truyền giúp chuyển đổi dữ liệu sang không gian (representation) mới.
- +) Dữ liệu dạng mới được kỳ vọng phân tách tuyến tính.
 => Không có biến đổi phi tuyến, mạng neural hoạt động như một màng hình tuyến tính

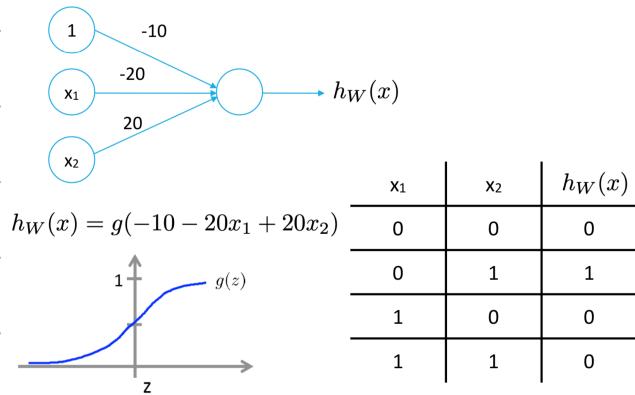
Hàm OR :



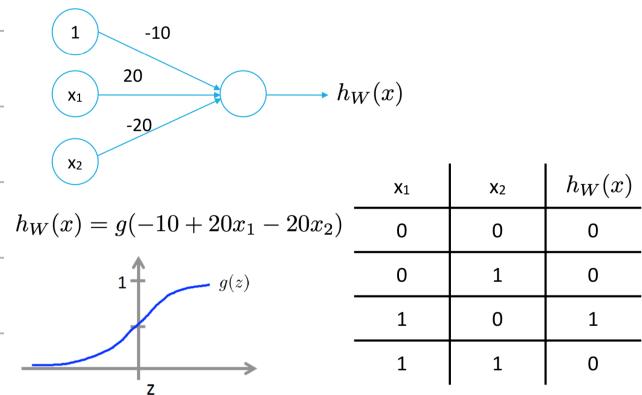
Hàm AND :



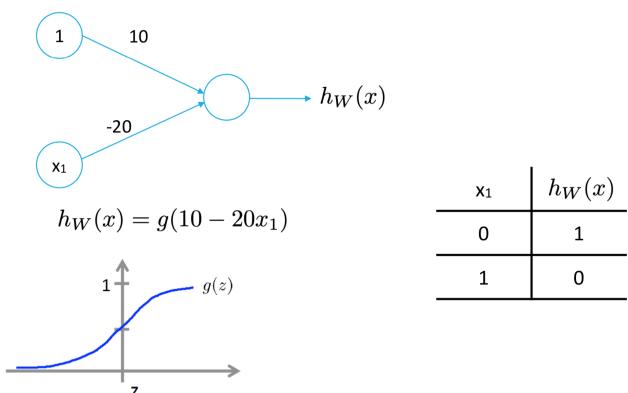
Hàm AND ($\text{NOT}(x_1), x_2$)



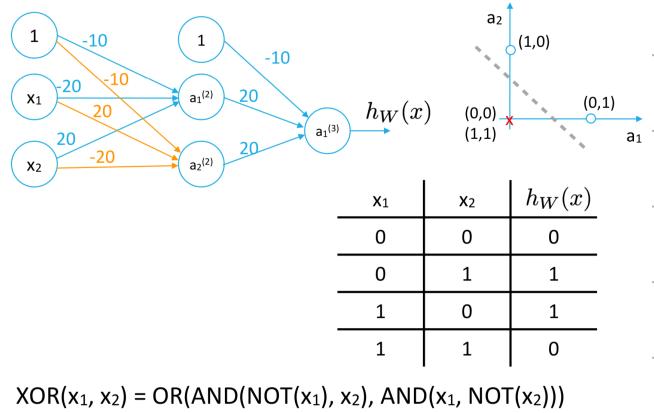
Hàm AND ($x_1, \text{NOT}(x_2)$)



Hàm NOT :



Hàm XOR



Nhắc lại hàm chi phí:

$$J(W) = -\frac{1}{m} \left[\sum_{i=1}^m y^{(i)} \log(h_W(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_W(x^{(i)})) \right]$$

Gradient :

$$\frac{dJ}{dW} = \left[\frac{dJ}{dW_{10}^{(1)}} , \frac{dJ}{dW_{11}^{(1)}} , \dots , \frac{dJ}{dW_{10}^{(L-1)}} , \dots , \frac{dJ}{dW_{S_{L+1}}^{(L-1)}} \right]$$

