ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

TOÁN CHO KHOA HỌC MÁY TÍNH

MẠNG NƠ-RƠN NHÂN TẠO

TS. Lương Ngọc Hoàng

Nội dung

1. Giới thiệu mạng nơ-rơn nhân tạo
2. Tính toán gradient với lan truyền ngược
3. Phân tích mạng nơ-rơn

MẠNG NƠ-RƠN NHÂN TẠO

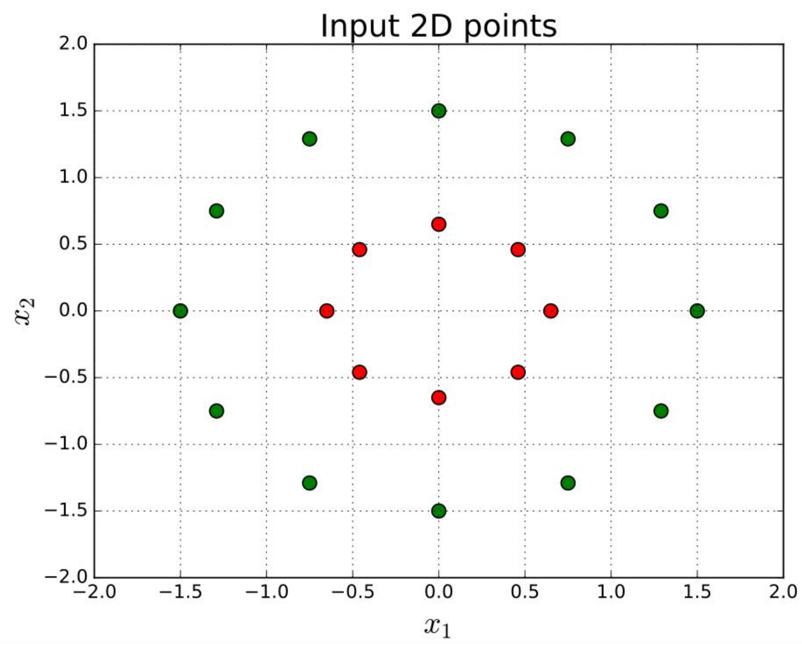
ARTIFICIAL NEURAL NETWORK

Thực hiện bởi Trường Đại học Công nghệ Thông tin, ĐHQG-HCM

Bài toán phân lớp

Bài toán phân lớp (classification):

Với mỗi điểm dữ liệu đầu vào $\boldsymbol{x} = (x_1, x_2)$, ta muốn đầu ra của mô hình dự đoán nhãn (label) của $\boldsymbol{x}$ là nhãn 0 (lớp đỏ), hay nhãn 1 (lớp xanh lá).

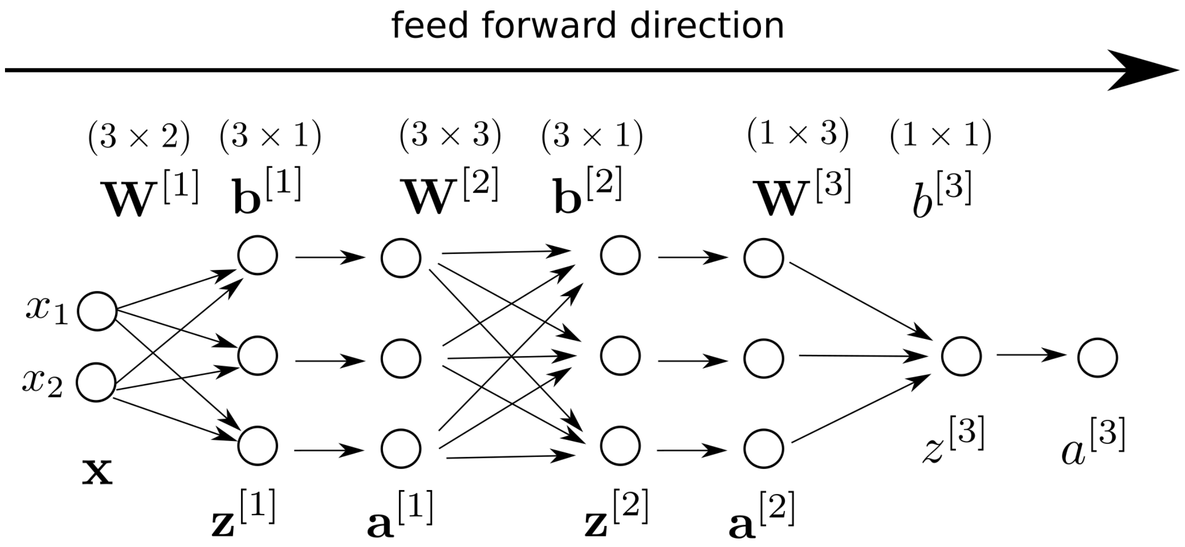


Ta muốn đầu ra của mô hình dự đoán xác suất điểm dữ liệu $\boldsymbol{x} = (x_1, x_2)$ thuộc về lớp đỏ (nhãn 0), và lớp xanh lá (nhãn 1) là bao nhiêu.

Hồi quy logistic có phù hợp với tập dữ liệu này?

- Biên quyết định của hồi quy logistic có dạng tuyến tính (đường thẳng).
- Tập dữ liệu trên không thể phân chia tuyến tính được.

Mạng nơ-rơn (neural network)



$$\boldsymbol{z}^{[1]} = W^{[1]}\boldsymbol{x} + \boldsymbol{b}^{[1]}$$
$$\boldsymbol{a}^{[1]} = g(\boldsymbol{z}^{[1]})$$
$$\boldsymbol{z}^{[2]} = W^{[2]}\boldsymbol{a}^{[1]} + \boldsymbol{b}^{[2]}$$
$$\boldsymbol{a}^{[2]} = g\left(\boldsymbol{z}^{[2]}\right)$$
$$z^{[3]} = W^{[3]}\boldsymbol{a}^{[2]} + b^{[3]}$$
$$a^{[3]} = g\left(\boldsymbol{z}^{[3]}\right) \in (0,1)$$

$g()$ là hàm kích hoạt (activation function) logistic sigmoid. $g(z) = \frac{1}{1+e^{-z}}$

Lớp ẩn (hidden layer) thứ nhất:

$$W^{[1]} = \begin{bmatrix} w_{1,1}^{[1]} & w_{1,2}^{[1]} \\ w_{2,1}^{[1]} & w_{2,2}^{[1]} \\ w_{3,1}^{[1]} & w_{3,2}^{[1]} \end{bmatrix}, \boldsymbol{b}^{[1]} = \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \end{bmatrix}$$

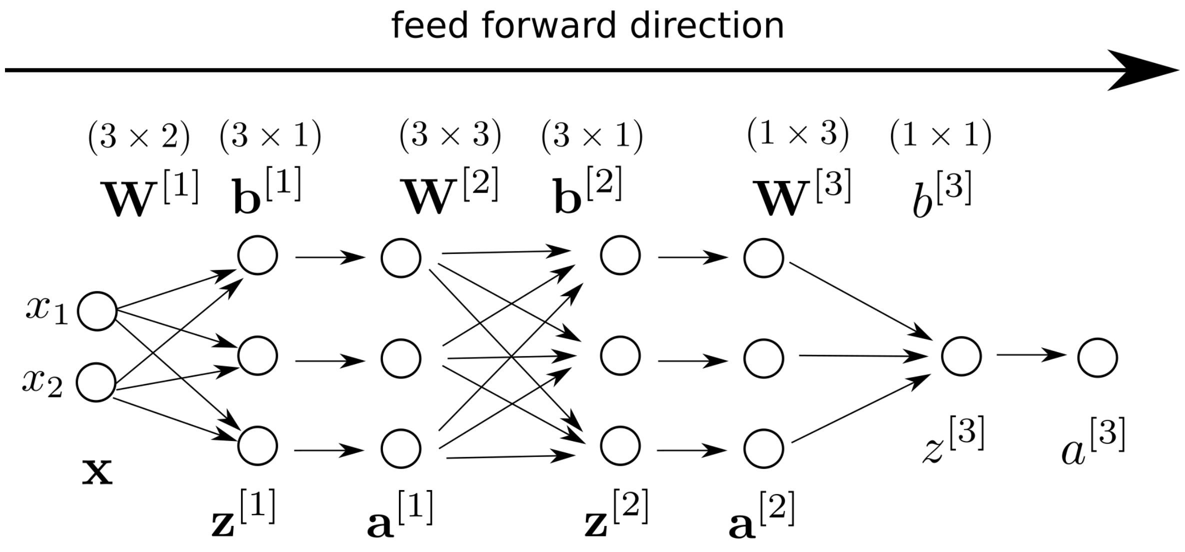
Lớp ẩn (hidden layer) :

$$W^{[2]} = \begin{bmatrix} w_{1,1}^{[2]} & w_{1,2}^{[2]} & w_{1,3}^{[2]} \\ w_{2,1}^{[2]} & w_{2,2}^{[2]} & w_{2,3}^{[2]} \\ w_{3,1}^{[2]} & w_{3,2}^{[2]} & w_{3,3}^{[2]} \end{bmatrix}, \boldsymbol{b}^{[2]} = \begin{bmatrix} b_1^{[2]} \\ b_2^{[2]} \\ b_3^{[2]} \end{bmatrix}$$

Lớp đầu ra (output layer):

$$W^{[3]} = \begin{bmatrix} w_{1,1}^{[3]} & w_{1,2}^{[3]} & w_{1,3}^{[3]} \end{bmatrix}, b^{[3]} = b_1^{[3]}$$

Mạng nơ-rơn (neural network)



$$a^{[3]} = g\left(\boldsymbol{z}^{[3]}\right) = g\left(W^{[3]}\boldsymbol{a}^{[2]} + b^{[3]}\right) \qquad g(z) = \frac{1}{1+e^{-z}}$$

$$= g\left(W^{[3]} g\left(\boldsymbol{z}^{[2]}\right) + b^{[3]}\right)$$

$$= g\left(W^{[3]} g\left(W^{[2]}\boldsymbol{a}^{[1]} + \boldsymbol{b}^{[2]}\right) + b^{[3]}\right)$$

$$= g\left(W^{[3]} g\left(W^{[2]} g(\boldsymbol{z}^{[1]}) + \boldsymbol{b}^{[2]}\right) + b^{[3]}\right)$$

$$= g\left(W^{[3]} g\left(W^{[2]} g(W^{[1]}\boldsymbol{x} + \boldsymbol{b}^{[1]}) + \boldsymbol{b}^{[2]}\right) + b^{[3]}\right)$$

Ta có cần các hàm kích hoạt tại các nơ-rơn của lớp ẩn?

$$a^{[3]} = g\left(W^{[3]}\left(W^{[2]}\,(W^{[1]}\boldsymbol{x} + \boldsymbol{b}^{[1]}) + \boldsymbol{b}^{[2]}\right) + b^{[3]}\right)$$

Lớp ẩn (hidden layer) thứ nhất:

$$W^{[1]} = \begin{bmatrix} w_{1,1}^{[1]} & w_{1,2}^{[1]} \\ w_{2,1}^{[1]} & w_{2,2}^{[1]} \\ w_{3,1}^{[1]} & w_{3,2}^{[1]} \end{bmatrix}, \ \boldsymbol{b}^{[1]} = \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \end{bmatrix}$$

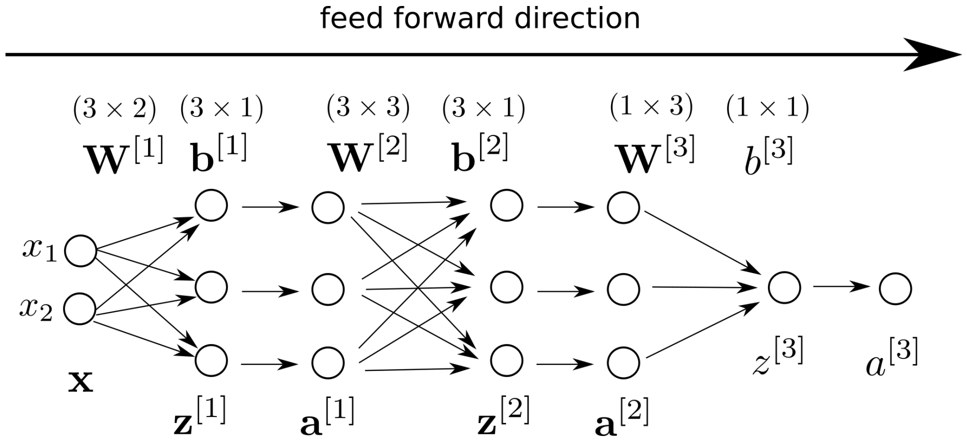
Lớp ẩn (hidden layer) :

$$W^{[2]} = \begin{bmatrix} w_{1,1}^{[2]} & w_{1,2}^{[2]} & w_{1,3}^{[2]} \\ w_{2,1}^{[2]} & w_{2,2}^{[2]} & w_{2,3}^{[2]} \\ w_{3,1}^{[2]} & w_{3,2}^{[2]} & w_{3,3}^{[2]} \end{bmatrix}, \ \boldsymbol{b}^{[2]} = \begin{bmatrix} b_1^{[2]} \\ b_2^{[2]} \\ b_3^{[2]} \end{bmatrix}$$

Lớp đầu ra (output layer):

$$W^{[3]} = \begin{bmatrix} w_{1,1}^{[3]} & w_{1,2}^{[3]} & w_{1,3}^{[3]} \end{bmatrix}, \ b^{[3]} = b_1^{[3]}$$

Mạng nơ-rơn (neural network)



$$g(z) = \frac{1}{1+e^{-z}}$$

Nếu bỏ các hàm kích hoạt tại các nơ-rơn của lớp ẩn:

$$
\begin{aligned}
a^{[3]} &= g\Big(W^{[3]}\big(W^{[2]}\,(W^{[1]}\boldsymbol{x} + \boldsymbol{b}^{[1]}) + \boldsymbol{b}^{[2]}\big) + b^{[3]}\Big) \\
&= g\Big(W^{[3]}\big(W^{[2]}W^{[1]}\boldsymbol{x} + W^{[2]}\boldsymbol{b}^{[1]} + \boldsymbol{b}^{[2]}\big) + b^{[3]}\Big) \\
&= g\big(\underbrace{W^{[3]}W^{[2]}W^{[1]}}_{(1\times3).(3\times3).(3\times2)}\underbrace{\boldsymbol{x}}_{2\times1} + \underbrace{W^{[3]}W^{[2]}}_{(1\times3).(3\times3)}\underbrace{\boldsymbol{b}^{[1]}}_{3\times1} + \underbrace{W^{[3]}}_{(1\times3)}\underbrace{\boldsymbol{b}^{[2]}}_{3\times1} + \underbrace{b^{[3]}}_{1\times1}\big) \\
&= g(\underbrace{W}_{(1\times2)}\underbrace{\boldsymbol{x}}_{2\times1} + \underbrace{b}_{1\times1})
\end{aligned}
$$

$a^{[3]} = g(\boldsymbol{w}^T\boldsymbol{x} + b)$ là hồi quy logistic (chỉ có khả năng tạo các đường biên quyết định tuyến tính). → Hàm kích hoạt sử dụng phải là các hàm phi tuyến tính (non-linear).

Lớp ẩn (hidden layer) thứ nhất:

$$W^{[1]} = \begin{bmatrix} w_{1,1}^{[1]} & w_{1,2}^{[1]} \\ w_{2,1}^{[1]} & w_{2,2}^{[1]} \\ w_{3,1}^{[1]} & w_{3,2}^{[1]} \end{bmatrix}, \quad \boldsymbol{b}^{[1]} = \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \end{bmatrix}$$

Lớp ẩn (hidden layer) :

$$W^{[2]} = \begin{bmatrix} w_{1,1}^{[2]} & w_{1,2}^{[2]} & w_{1,3}^{[2]} \\ w_{2,1}^{[2]} & w_{2,2}^{[2]} & w_{2,3}^{[2]} \\ w_{3,1}^{[2]} & w_{3,2}^{[2]} & w_{3,3}^{[2]} \end{bmatrix}, \quad \boldsymbol{b}^{[2]} = \begin{bmatrix} b_1^{[2]} \\ b_2^{[2]} \\ b_3^{[2]} \end{bmatrix}$$

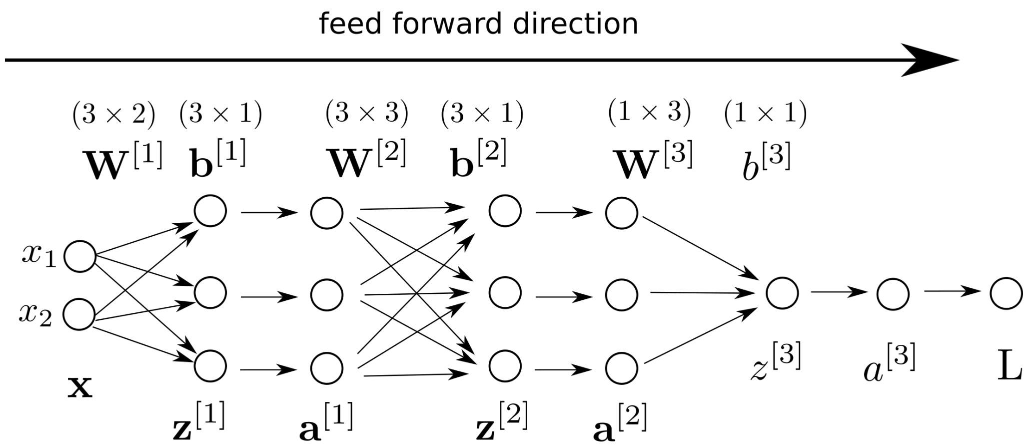
Lớp đầu ra (output layer):

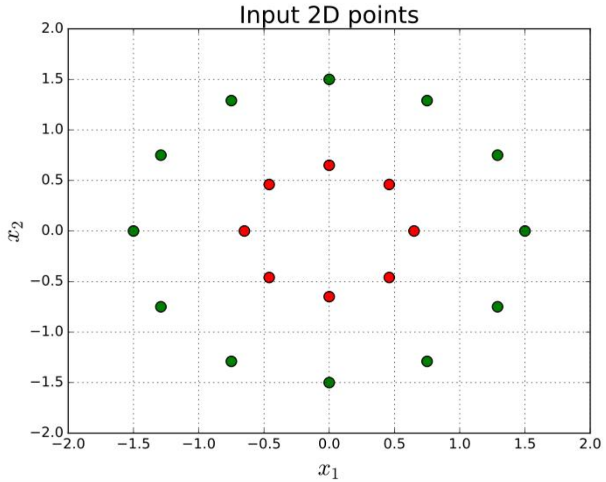
$$W^{[3]} = \begin{bmatrix} w_{1,1}^{[3]} & w_{1,2}^{[3]} & w_{1,3}^{[3]} \end{bmatrix}, \; b^{[3]} = b_1^{[3]}$$

TÍNH TOÁN GRADIENT VỚI LAN TRUYỀN NGƯỢC

GRADIENT COMPUTATION WITH BACKPROPAGATION

Hàm mất mát





Cho bài toán phân lớp, ta có thể sử dụng hàm Binary Cross Entropy (BCE) Loss như trong hồi quy logistic.

- Trên một điểm dữ liệu i:

$$L = -\left(y^{(i)}\log\left(a^{[3](i)}\right) + \left(1 - y^{(i)}\right)\log\left(1 - a^{[3](i)}\right)\right)$$

- Trên toàn bộ tập dữ liệu:

$$J = -\sum_{i=1}^{N}\left(y^{(i)}\log\left(a^{[3](i)}\right) + \left(1 - y^{(i)}\right)\log\left(1 - a^{[3](i)}\right)\right)$$

- Nếu ta có nhiều hơn 2 lớp, ta dùng hàm Cross Entropy (CE) Loss:

$$J = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_j^{(i)}\log\left(a_j^{[3](i)}\right)$$

Ta muốn đầu ra của mô hình dự đoán xác suất điểm dữ liệu $\boldsymbol{x} = (x_1, x_2)$ thuộc về lớp đỏ (nhãn 0), và lớp xanh lá (nhãn 1) là bao nhiêu.

Lan truyền ngược (backpropagation)

Trong hồi quy logistic, ta có:

$$\frac{\partial L}{\partial z^{[3]}} = a^{[3]} - y$$

$$z^{[3]} = W^{[3]}\boldsymbol{a}^{[2]} + b^{[3]}$$

$$= \begin{bmatrix} w_{1,1}^{[3]} & w_{1,2}^{[3]} & w_{1,3}^{[3]} \end{bmatrix} \begin{bmatrix} a_1^{[2]} \\ a_2^{[2]} \\ a_3^{[2]} \end{bmatrix} + b^{[3]}$$

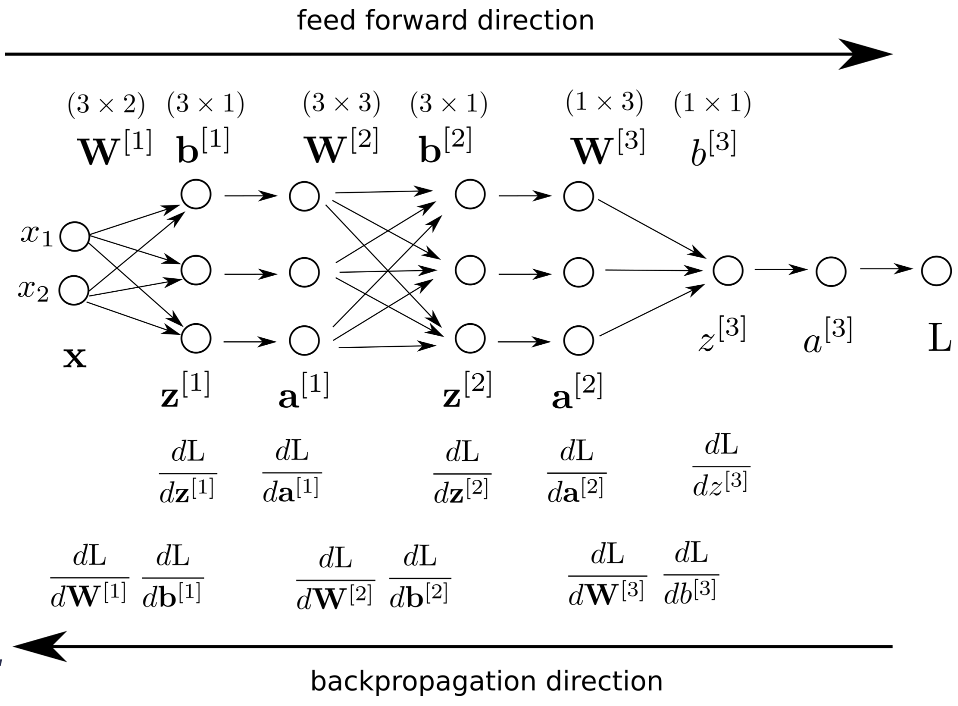
$$= w_{1,1}^{[3]} a_1^{[2]} + w_{1,2}^{[3]} a_2^{[2]} + w_{1,3}^{[3]} a_3^{[2]} + b^{[3]}$$

$$\frac{\partial z^{[3]}}{\partial W^{[3]}} = \begin{bmatrix} \frac{\partial z^{[3]}}{\partial w_{1,1}^{[3]}} & \frac{\partial z^{[3]}}{\partial w_{1,2}^{[3]}} & \frac{\partial z^{[3]}}{\partial w_{1,3}^{[3]}} \end{bmatrix} = \begin{bmatrix} a_1^{[2]} & a_2^{[2]} & a_3^{[2]} \end{bmatrix} = \left(\boldsymbol{a}^{[2]}\right)^T$$

$$\frac{\partial z^{[3]}}{\partial b^{[3]}} = 1$$

$$\frac{\partial L}{\partial W^{[3]}} = \frac{\partial L}{\partial z^{[3]}} \frac{\partial z^{[3]}}{\partial W^{[3]}} = \left(a^{[3]} - y\right)\left(\boldsymbol{a}^{[2]}\right)^T$$

$$\frac{\partial L}{\partial b^{[3]}} = \frac{\partial L}{\partial z^{[3]}} \frac{\partial z^{[3]}}{\partial b^{[3]}} = \left(a^{[3]} - y\right)$$



Lan truyền ngược (backpropagation)

$$\frac{\partial L}{\partial \boldsymbol{a}^{[2]}} = \frac{\partial L}{\partial z^{[3]}} \frac{\partial z^{[3]}}{\partial \boldsymbol{a}^{[2]}}$$

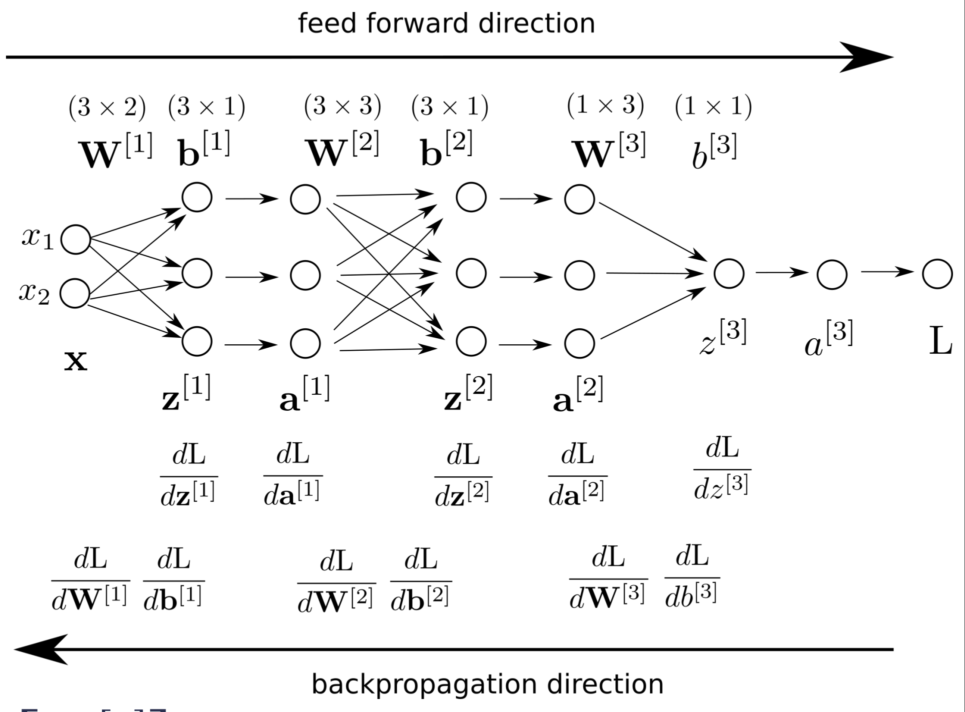
$$z^{[3]} = W^{[3]}\boldsymbol{a}^{[2]} + b^{[3]}$$

$$= \begin{bmatrix} w_{1,1}^{[3]} & w_{1,2}^{[3]} & w_{1,3}^{[3]} \end{bmatrix} \begin{bmatrix} a_1^{[2]} \\ a_2^{[2]} \\ a_3^{[2]} \end{bmatrix} + b^{[3]}$$

$$= w_{1,1}^{[3]} a_1^{[2]} + w_{1,2}^{[3]} a_2^{[2]} + w_{1,3}^{[3]} a_3^{[2]} + b^{[3]}$$

$$\frac{\partial z^{[3]}}{\partial \boldsymbol{a}^{[2]}} = \begin{bmatrix} \frac{\partial z^{[3]}}{\partial a_1^{[2]}} \\ \frac{\partial z^{[3]}}{\partial a_2^{[2]}} \\ \frac{\partial z^{[3]}}{\partial a_3^{[2]}} \end{bmatrix} = \begin{bmatrix} w_{1,1}^{[3]} \\ w_{1,2}^{[3]} \\ w_{1,3}^{[3]} \end{bmatrix} = \left(W^{[3]}\right)^T$$

$$\frac{\partial L}{\partial \boldsymbol{a}^{[2]}} = \frac{\partial L}{\partial z^{[3]}} \frac{\partial z^{[3]}}{\partial \boldsymbol{a}^{[2]}} = \left(a^{[3]} - y\right)\left(W^{[3]}\right)^T$$



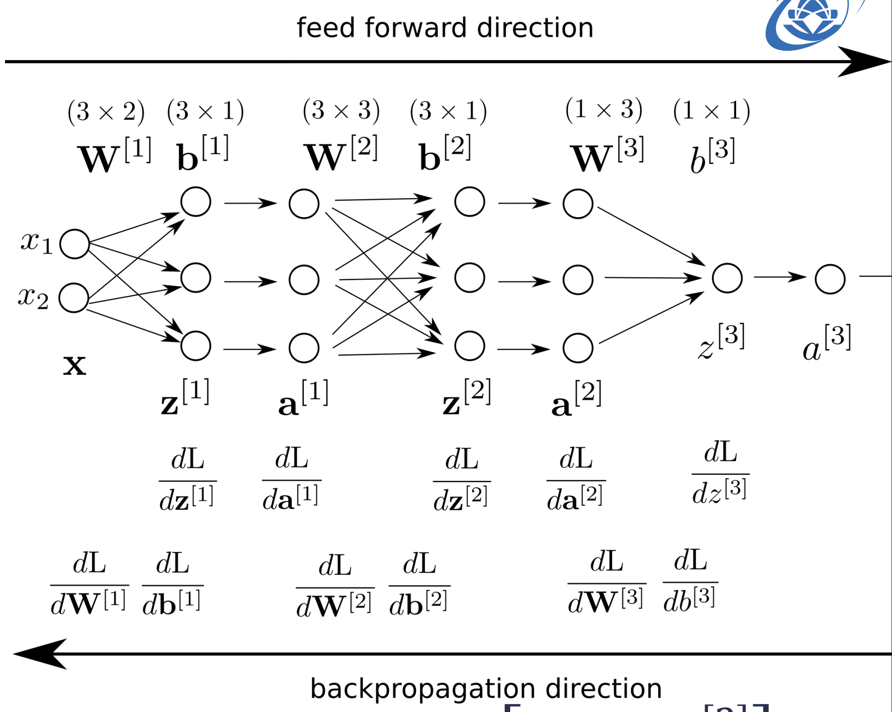
$$\frac{\partial \boldsymbol{a}^{[2]}}{\partial \boldsymbol{z}^{[2]}} = \begin{bmatrix} \frac{\partial a^{[2]}}{\partial z_1^{[2]}} \\ \frac{\partial a^{[2]}}{\partial z_2^{[2]}} \\ \frac{\partial a^{[2]}}{\partial z_3^{[2]}} \end{bmatrix} = \begin{bmatrix} a_1^{[2]}\left(1 - a_1^{[2]}\right) \\ a_2^{[2]}\left(1 - a_2^{[2]}\right) \\ a_3^{[2]}\left(1 - a_3^{[2]}\right) \end{bmatrix}$$

$$= \boldsymbol{a}^{[2]} \circ \left(1 - \boldsymbol{a}^{[2]}\right)$$

Lan truyền ngược

$$\boldsymbol{z}^{[2]} = W^{[2]}\boldsymbol{a}^{[1]} + \boldsymbol{b}^{[2]}$$

$$\begin{bmatrix} z_1^{[2]} \\ z_2^{[2]} \\ z_3^{[2]} \end{bmatrix} = \begin{bmatrix} w_{1,1}^{[2]} & w_{1,2}^{[2]} & w_{1,3}^{[2]} \\ w_{2,1}^{[2]} & w_{2,2}^{[2]} & w_{2,3}^{[2]} \\ w_{3,1}^{[2]} & w_{3,2}^{[2]} & w_{3,3}^{[2]} \end{bmatrix} \begin{bmatrix} a_1^{[1]} \\ a_2^{[1]} \\ a_3^{[1]} \end{bmatrix} + \begin{bmatrix} b_1^{[2]} \\ b_2^{[2]} \\ b_3^{[2]} \end{bmatrix}$$



$$\frac{\partial L}{\partial W^{[2]}} = \begin{bmatrix} \frac{\partial L}{\partial w_{1,1}^{[2]}} & \frac{\partial L}{\partial w_{1,2}^{[2]}} & \frac{\partial L}{\partial w_{1,3}^{[2]}} \\ \frac{\partial L}{\partial w_{2,1}^{[2]}} & \frac{\partial L}{\partial w_{2,2}^{[2]}} & \frac{\partial L}{\partial w_{2,3}^{[2]}} \\ \frac{\partial L}{\partial w_{3,1}^{[2]}} & \frac{\partial L}{\partial w_{3,2}^{[2]}} & \frac{\partial L}{\partial w_{3,3}^{[2]}} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial z_1^{[2]}}\frac{\partial z_1^{[2]}}{\partial w_{1,1}^{[2]}} & \frac{\partial L}{\partial z_1^{[2]}}\frac{\partial z_1^{[2]}}{\partial w_{1,2}^{[2]}} & \frac{\partial L}{\partial z_1^{[2]}}\frac{\partial z_1^{[2]}}{\partial w_{1,3}^{[2]}} \\ \frac{\partial L}{\partial z_2^{[2]}}\frac{\partial z_2^{[2]}}{\partial w_{2,1}^{[2]}} & \frac{\partial L}{\partial z_2^{[2]}}\frac{\partial z_2^{[2]}}{\partial w_{2,2}^{[2]}} & \frac{\partial L}{\partial z_2^{[2]}}\frac{\partial z_2^{[2]}}{\partial w_{2,3}^{[2]}} \\ \frac{\partial L}{\partial z_3^{[2]}}\frac{\partial z_3^{[2]}}{\partial w_{3,1}^{[2]}} & \frac{\partial L}{\partial z_3^{[2]}}\frac{\partial z_3^{[2]}}{\partial w_{3,2}^{[2]}} & \frac{\partial L}{\partial z_3^{[2]}}\frac{\partial z_3^{[2]}}{\partial w_{3,3}^{[2]}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial L}{\partial z_1^{[2]}} a_1^{[1]} & \frac{\partial L}{\partial z_1^{[2]}} a_2^{[1]} & \frac{\partial L}{\partial z_1^{[2]}} a_3^{[1]} \\ \frac{\partial L}{\partial z_2^{[2]}} a_1^{[1]} & \frac{\partial L}{\partial z_2^{[2]}} a_2^{[1]} & \frac{\partial L}{\partial z_2^{[2]}} a_3^{[1]} \\ \frac{\partial L}{\partial z_3^{[2]}} a_1^{[1]} & \frac{\partial L}{\partial z_3^{[2]}} a_2^{[1]} & \frac{\partial L}{\partial z_3^{[2]}} a_3^{[1]} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial z_1^{[2]}} \\ \frac{\partial L}{\partial z_2^{[2]}} \\ \frac{\partial L}{\partial z_3^{[2]}} \end{bmatrix} \begin{bmatrix} a_1^{[1]} & a_2^{[1]} & a_3^{[1]} \end{bmatrix}$$

$$= \frac{\partial L}{\partial \boldsymbol{z}^{[2]}} \left(\boldsymbol{a}^{[1]}\right)^T$$

$$\frac{\partial L}{\partial \boldsymbol{z}^{[2]}} = \begin{bmatrix} \frac{\partial L}{\partial z_1^{[2]}} \\ \frac{\partial L}{\partial z_2^{[2]}} \\ \frac{\partial L}{\partial z_3^{[2]}} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial a_1^{[2]}}\frac{\partial a_1^{[2]}}{\partial z_1^{[2]}} \\ \frac{\partial L}{\partial a_2^{[2]}}\frac{\partial a_2^{[2]}}{\partial z_2^{[2]}} \\ \frac{\partial L}{\partial a_3^{[2]}}\frac{\partial a_3^{[2]}}{\partial z_3^{[2]}} \end{bmatrix}$$

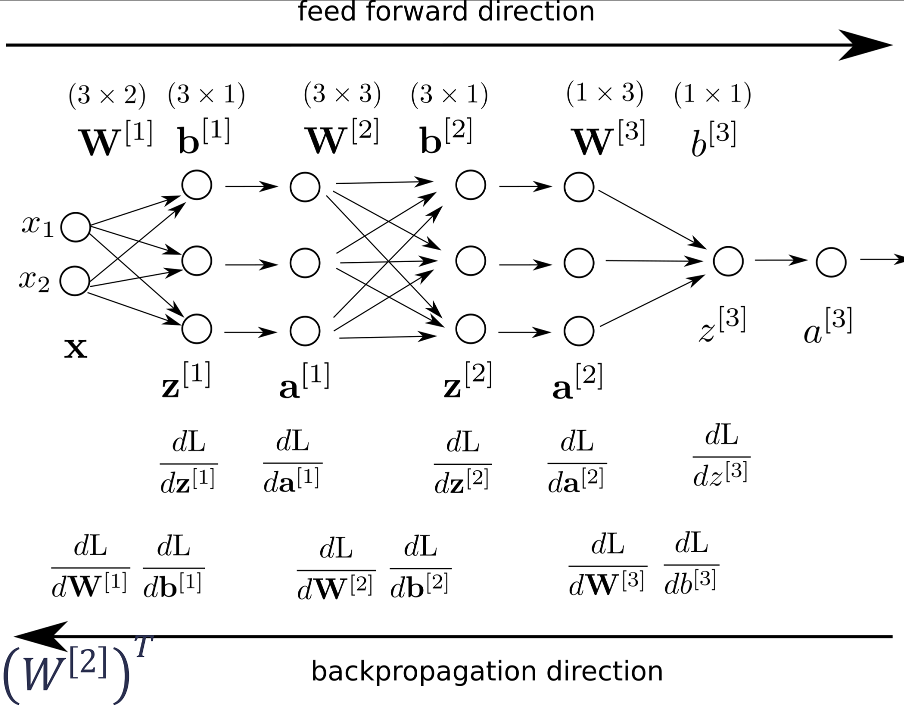
$$\frac{\partial L}{\partial \boldsymbol{b}^{[2]}} = \frac{\partial L}{\partial \boldsymbol{z}^{[2]}}$$

$$= \frac{\partial L}{\partial \boldsymbol{a}^{[2]}} \circ \left(\boldsymbol{a}^{[2]} \circ \left(1 - \boldsymbol{a}^{[2]}\right)\right)$$

Lan truyền ngược

$$\boldsymbol{z}^{[2]} = W^{[2]}\boldsymbol{a}^{[1]} + \boldsymbol{b}^{[2]}$$

$$\begin{bmatrix} z_1^{[2]} \\ z_2^{[2]} \\ z_3^{[2]} \end{bmatrix} = \begin{bmatrix} w_{1,1}^{[2]} & w_{1,2}^{[2]} & w_{1,3}^{[2]} \\ w_{2,1}^{[2]} & w_{2,2}^{[2]} & w_{2,3}^{[2]} \\ w_{3,1}^{[2]} & w_{3,2}^{[2]} & w_{3,3}^{[2]} \end{bmatrix} \begin{bmatrix} a_1^{[1]} \\ a_2^{[1]} \\ a_3^{[1]} \end{bmatrix} + \begin{bmatrix} b_1^{[2]} \\ b_2^{[2]} \\ b_3^{[2]} \end{bmatrix}$$



$$\frac{\partial \boldsymbol{z}^{[2]}}{\partial \boldsymbol{a}^{[1]}} = \begin{bmatrix} \frac{\partial \boldsymbol{z}^{[2]}}{\partial a_1^{[1]}} \\ \frac{\partial \boldsymbol{z}^{[2]}}{\partial a_2^{[1]}} \\ \frac{\partial \boldsymbol{z}^{[2]}}{\partial a_3^{[1]}} \end{bmatrix} = \begin{bmatrix} \frac{\partial z_1^{[2]}}{\partial a_1^{[1]}} & \frac{\partial z_2^{[2]}}{\partial a_1^{[1]}} & \frac{\partial z_3^{[2]}}{\partial a_1^{[1]}} \\ \frac{\partial z_1^{[2]}}{\partial a_2^{[1]}} & \frac{\partial z_2^{[2]}}{\partial a_2^{[1]}} & \frac{\partial z_3^{[2]}}{\partial a_2^{[1]}} \\ \frac{\partial z_1^{[2]}}{\partial a_3^{[1]}} & \frac{\partial z_2^{[2]}}{\partial a_3^{[1]}} & \frac{\partial z_3^{[2]}}{\partial a_3^{[1]}} \end{bmatrix} = \begin{bmatrix} w_{1,1}^{[2]} & w_{2,1}^{[2]} & w_{3,1}^{[2]} \\ w_{1,2}^{[2]} & w_{2,2}^{[2]} & w_{3,2}^{[2]} \\ w_{1,3}^{[2]} & w_{2,3}^{[2]} & w_{3,3}^{[2]} \end{bmatrix} = \left(W^{[2]}\right)^T$$

$$\frac{\partial L}{\partial \boldsymbol{a}^{[1]}} = \frac{\partial \boldsymbol{z}^{[2]}}{\partial \boldsymbol{a}^{[1]}} \frac{\partial L}{\partial \boldsymbol{z}^{[2]}} = \left(W^{[2]}\right)^T \frac{\partial L}{\partial \boldsymbol{z}^{[2]}}$$

$$\boldsymbol{z}^{[1]} = W^{[1]}\boldsymbol{x} + \boldsymbol{b}^{[1]}$$

$$\begin{bmatrix} z_1^{[1]} \\ z_2^{[1]} \\ z_3^{[1]} \end{bmatrix} = \begin{bmatrix} w_{1,1}^{[1]} & w_{1,2}^{[1]} \\ w_{2,1}^{[1]} & w_{2,2}^{[1]} \\ w_{3,1}^{[1]} & w_{3,2}^{[1]} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \end{bmatrix}$$

Tương tự, ta cũng tính được

$$\frac{\partial \mathrm{L}}{\partial \boldsymbol{z}^{[1]}} = \frac{\partial L}{\partial \boldsymbol{a}^{[1]}} \circ \left(\boldsymbol{a}^{[1]} \circ \left(1 - \boldsymbol{a}^{[1]}\right)\right)$$

$$\frac{\partial L}{\partial W^{[1]}} = \frac{\partial L}{\partial \boldsymbol{z}^{[1]}} \left(\boldsymbol{a}^{[0]}\right)^T = \frac{\partial L}{\partial \boldsymbol{z}^{[1]}} \boldsymbol{x}^T$$

$$\frac{\partial L}{\partial \boldsymbol{b}^{[1]}} = \frac{\partial L}{\partial \boldsymbol{z}^{[1]}}$$

CÀI ĐẶT & PHÂN TÍCH

IMPLEMENTATION & ANALYSIS

Gradient Descent

Hàm mất mát:

$$L = -\left(y^{(i)} \log\left(a^{[3](i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - a^{[3](i)}\right)\right)$$

Đạo hàm:

$$\frac{\partial L}{\partial z^{[3]}} = a^{[3]} - y$$

$$\frac{\partial L}{\partial b^{[3]}} = \frac{\partial L}{\partial z^{[3]}}$$

$$\frac{\partial L}{\partial W^{[3]}} = \frac{\partial L}{\partial z^{[3]}} \left(\boldsymbol{a}^{[2]}\right)^T$$

$$\frac{\partial L}{\partial \boldsymbol{a}^{[2]}} = \frac{\partial L}{\partial z^{[3]}} \frac{\partial z^{[3]}}{\partial \boldsymbol{a}^{[2]}} = \left(a^{[3]} - y\right)\left(W^{[3]}\right)^T$$

$$\frac{\partial L}{\partial \boldsymbol{z}^{[2]}} = \frac{\partial L}{\partial \boldsymbol{a}^{[2]}} \circ \left(\boldsymbol{a}^{[2]} \circ \left(1 - \boldsymbol{a}^{[2]}\right)\right)$$

$$\frac{\partial L}{\partial \boldsymbol{b}^{[2]}} = \frac{\partial L}{\partial \boldsymbol{z}^{[2]}}$$

$$\frac{\partial L}{\partial W^{[2]}} = \frac{\partial L}{\partial \boldsymbol{z}^{[2]}} \left(\boldsymbol{a}^{[1]}\right)^T$$

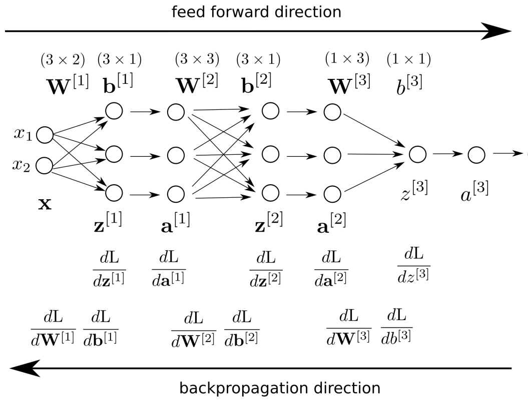
$$\frac{\partial L}{\partial \boldsymbol{a}^{[1]}} = \frac{\partial \boldsymbol{z}^{[2]}}{\partial \boldsymbol{a}^{[1]}} \frac{\partial L}{\partial \boldsymbol{z}^{[2]}} = \left(W^{[2]}\right)^T \frac{\partial L}{\partial \boldsymbol{z}^{[2]}}$$

$$\frac{\partial L}{\partial \boldsymbol{z}^{[1]}} = \frac{\partial L}{\partial \boldsymbol{a}^{[1]}} \circ \left(\boldsymbol{a}^{[1]} \circ \left(1 - \boldsymbol{a}^{[1]}\right)\right)$$

$$\frac{\partial L}{\partial \boldsymbol{b}^{[1]}} = \frac{\partial L}{\partial \boldsymbol{z}^{[1]}}$$

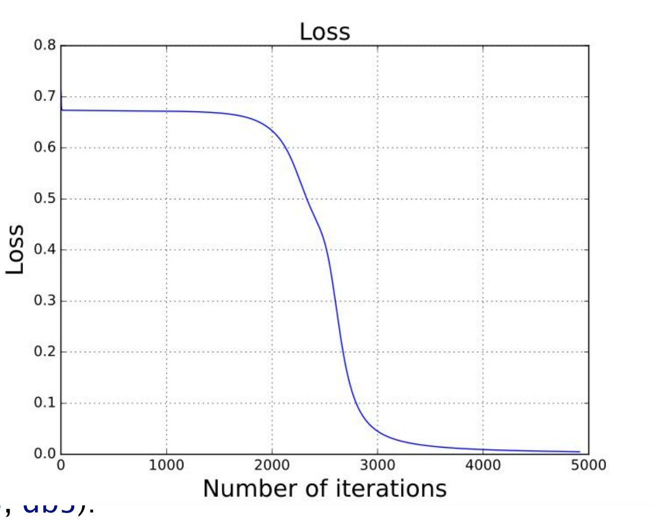
$$\frac{\partial L}{\partial W^{[1]}} = \frac{\partial L}{\partial \boldsymbol{z}^{[1]}} \boldsymbol{x}^T$$

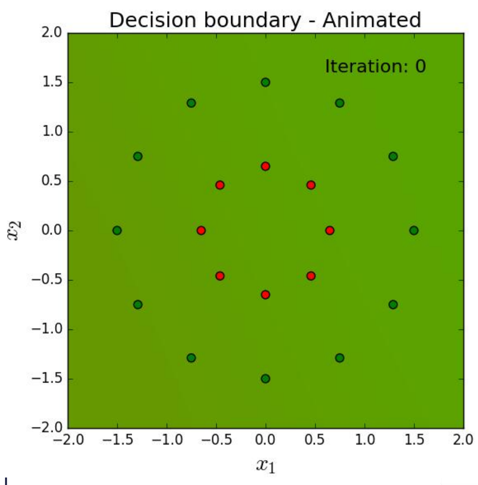
```python
def get_loss(y, a):
  return -1 * (y * np.log(a) + (1-y) * np.log(1-a))

def get_gradients(z1, a1, z2, a2, z3, a3, x, y, W1, b1, W2, b2, W3, b3):
  dz3 = a3 - y # dL/dz_3
  db3 = dz3    # dL/db_3

  dW3 = dz3 * a2.T  # dL/dW_3
  da2 = dz3 * W3.T

  dz2 = da2 * (a2 * (1-a2))
  db2 = dz2

  dW2 = np.matmul(dz2, a1.T)
  da1 = np.matmul(W2.T, dz2)

  dz1 = da1 * (a1 * (1-a1))
  db1 = dz1
  dW1 = np.matmul(dz1, x.T)

  return dW1, db1, dW2, db2, dW3, db3
```

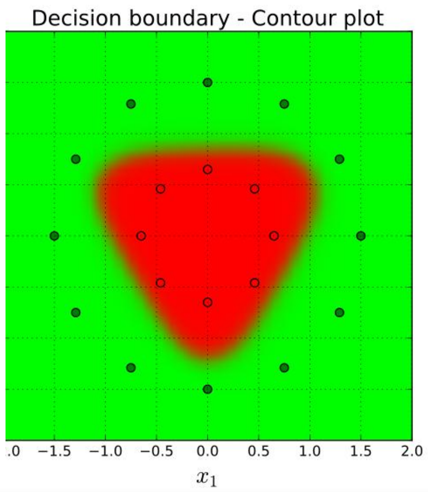


Gradient Descent

```
def gradient_descent(W1, b1, W2, b2, W3, b3, dW1, db1, dW2, db2, dW3, db3, alpha):
  W1 -= alpha
  b1 -= alpha
  W2 -= alpha
  b2 -= alpha
  W3 -= alpha
  b3 -= alpha

  return W1, b

def add_grad                                          W1, db1,
dW2, db2, dW
  tdW1 += dW1
```



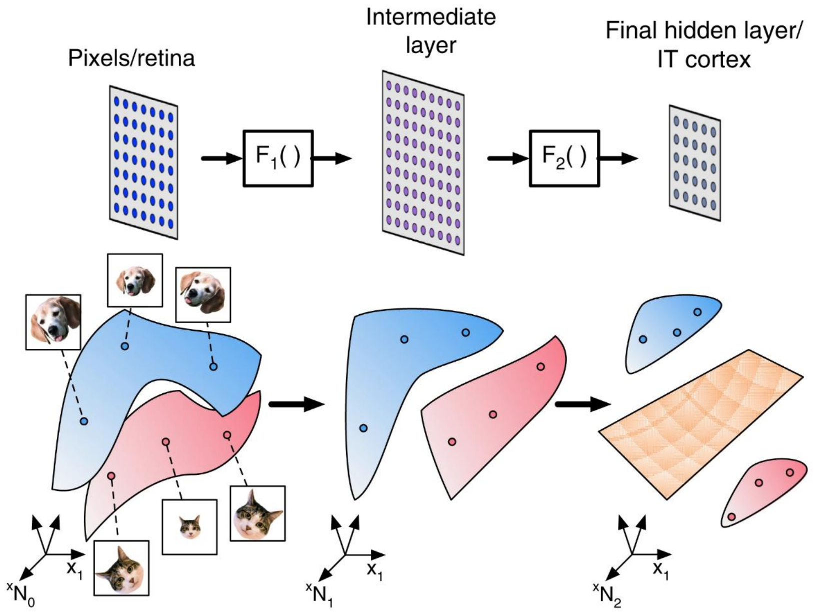




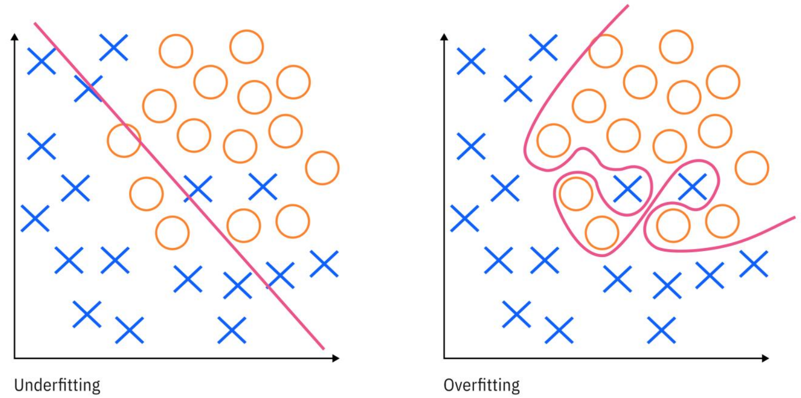

```
alpha = 0.4
for i in range(20000):
  totalL = 0
  tdW1, tdb1, tdW2, tdb2, tdW3, tdb3 = get_zero_gradients(W1, b1, W2, b2, W3, b3)
  for j in range(X.shape[0]):
      x = X[j, :].reshape(2,1)
      z1, a1, z2, a2, z3, a3 = forward(x, W1, b1, W2, b2, W3, b3)
      L = (1.0 / 20) * get_loss(y[j], a3)
      totalL += L
    dW1, db1, dW2, db2, dW3, db3 = get_gradients (z1, a1, z2, a2, z3, a3, x, y[j],
W1, b1, W2, b2, W3, b3)
    tdW1, tdb1, tdW2, tdb2, tdW3, tdb3 = add_gradients(tdW1, tdb1, tdW2, tdb2,
tdW3, tdb3, dW1, db1, dW2, db2, dW3, db3)

  W1, b1, W2, b2, W3, b3 = gradient_descent(W1, b1, W2, b2, W3, b3, tdW1, tdb1,
tdW2, tdb2, tdW3, tdb3, alpha)

  if totalL[0,0] < 0.005:
   break
```

Phân tích mạng nơ-rơn

- Mạng nơ-rơn nhân tạo (artificial neural network – ANN) với đủ độ sâu (số lớp) và đủ số lượng nơ-rơn ở mỗi lớp có thể tạo ra những biên quyết định (decision boundary) có độ phức tạp cao.



Cohen, U., Chung, S., Lee, D.D. *et al.* Separability and geometry of object manifolds in deep neural networks. *Nat Commun* **11**, 746 (2020). https://doi.org/10.1038/s41467-020-14578-5



Tim Mucci. https://www.ibm.com/think/topics/overfitting-vs-underfitting

- Mạng nơ-rơn có khả năng khớp rất tốt vào tập dữ liệu huấn luyện → rủi ro quá khớp (overfitting).
- Để khắc phục rủi ro quá khớp của mạng nơ-rơn, ta cần có tập dữ liệu huấn luyện đủ lớn.