DEEP LEARNING

7/21/2023

$x^{(i)} \in \mathbb{R}^d$ TRAIN DATA
$a^{(0)} = x$

for l in $1, \dots, L$
$\quad z^{[l]} = W^{[l]} a^{[l-1]} + b^{[l]}$
$\quad a^{[l]} = g(z^{[l]})$

$\hat{y} = a^{[L]}$
$\mathcal{L} = \text{Loss}(y, \hat{y})$

TRAIN DATA

MODEL WITH L LAYERS

PREDICTION AND LOSS

IN ORDER TO TRAIN THE MODEL, THE APPROACH IS TO MAXIMIZE THE LIKELIHOOD $\mathcal{L}$ (MINIMIZE THE $-\log \mathcal{L}$)

EX (*)

$x \in \mathbb{R}^d$, $y \in \{0, 1\}$ (CLASSIFICATION PROBLEM)

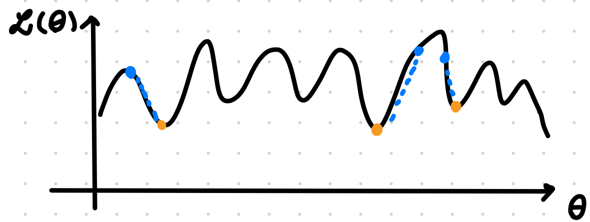
$$\mathcal{L} = -[y \log \hat{y} + (1-y) \log 1-y] \longrightarrow \boxed{\hat{y} = \text{Model}_\theta(x)}$$

$\mathcal{L}$ OPTIMIZATION : BACKPROPAGATION

NOTE THE PARAMETERS ARE UPDATED $\forall_l$, BUT $\mathcal{L}$ IS APPLIED ONLY AT L (OUTPUT LAYER)

WE NEED TO COMPUTE $\nabla\mathcal{L}$ IN ORDER TO MAX THE LIKELIHOOD AND THEN APPLY GRADIENT DESCENT $\theta = \theta - \alpha\nabla\mathcal{L}$. IN NEURALNETS, THE OPTIMIZATION ALG. IS BACKPROP

NOTE DEEP LEARNING MODELS ARE <u>NOT</u> CONVEX! HOW DO WE FIND θ?



• CONVERGE
• RAND INIT

LOCAL MINIMA? GLOBAL?
LOCAL MOST LIKELY!

LET'S DERIVE BACKPROP STARTING FROM THE EXAMPLE (*)

ALGORITHM

→ INIT

$$\begin{cases} W^{[l]} \sim N\left(\vec{0}, \sqrt{\dfrac{2}{n^{[l]} + n^{[l-1]}}}\right) \\ \vec{b} = 0 \end{cases}$$

RANDOM INIT OF W

BIAS INIT AT Ø

WE DEFINE Θ AS

$$\Theta = \left\{ W^{[1]}, b^{[1]}, W^{[2]}, b^{[2]}, W^{[3]}, b^{[3]} \right\} \quad (3 \text{ LAYERS})$$

$$x^{(i)} \in \mathbb{R}^{d_0}$$

$$W^{[1]} \in \mathbb{R}^{d_1 \times d_0}$$

$$b^{[1]} \in \mathbb{R}^{d_1}$$

$$z^{[1]} = W^{[1]} \overbrace{a^{[0]}}^{x^{(i)}} + b^{[1]} \in \mathbb{R}^{d_1} \quad (\text{LOGIT})$$

$$a^{[1]} = g(z^{[1]}) \in \mathbb{R}^{d_1}$$

$$W^{[2]} = \mathbb{R}^{d_2 \times d_1}$$

MODEL PARAMS

WE DEFINE $\frac{\partial \mathcal{L}}{\partial \Theta}$ AS THE FOLLOWING TENSOR (CONTAINS THE INFOS REGARDING DERIVATIVE OF $\mathcal{L}$ W.R.T. EACH WEIGHT & BIAS)

$$\frac{\partial \mathcal{L}}{\partial \Theta} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial W^{[1]}}, & \frac{\partial \mathcal{L}}{\partial b^{[1]}} \\ \frac{\partial \mathcal{L}}{\partial W^{[2]}}, & \frac{\partial \mathcal{L}}{\partial b^{[2]}} \\ \frac{\partial \mathcal{L}}{\partial W^{[3]}}, & \frac{\partial \mathcal{L}}{\partial b^{[3]}} \end{bmatrix}$$

LET'S ANALYZE EACH MEMBER $\frac{\partial \mathcal{L}}{\partial W^{[i]}}$

EX

$$\frac{\partial \mathcal{L}}{\partial W^{[2]}} \longrightarrow \begin{cases} \mathcal{L} \in \mathbb{R} \\ W^{[2]} \in \mathbb{R}^{d_2 \times d_1} \\ \frac{\partial \mathcal{L}}{\partial W^{[2]}} \in \mathbb{R}^{d_2 \times d_1} \end{cases}$$

THE OPTIMIZATION OPERATION, IN GENERAL, LOOKS LIKE THIS (G.D.)

$$\theta = \theta - \alpha \frac{\partial \mathcal{L}}{\partial \theta}$$

WE SPLIT THIS OPERATION INTO SUBPROBLEMS

$$\begin{cases} W^{[1]} = W^{[1]} - \alpha \frac{\partial \mathcal{L}}{\partial W^{[1]}} \\ \vdots \\ b^{[3]} = b^{[3]} - \alpha \frac{\partial \mathcal{L}}{\partial b^{[3]}} \end{cases} \rightarrow \quad W^{[2]} = \begin{bmatrix} W_{11}^{[2]} & \cdots & W_{1d_1} \\ \vdots & \ddots & \vdots \\ W_{d_2 1}^{[2]} & \cdots & W_{d_2 d_1} \end{bmatrix} \longrightarrow \frac{\partial \mathcal{L}}{\partial W^{[2]}} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial W_{11}^{[2]}} & \cdots & \frac{\partial \mathcal{L}}{\partial W_{1d_1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{L}}{\partial W_{d_2 1}^{[2]}} & \cdots & \frac{\partial \mathcal{L}}{\partial W_{d_2 d_1}} \end{bmatrix}$$

WE DO THIS $\forall W, b$ AND GET A SET OF PARTIAL DERIVATIVES

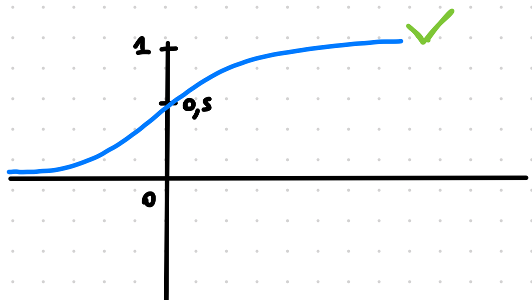
DEPENDING ON THE NATURE OF $\hat{y}$, THE LAST NON-LINEARITY FUNCTION SHOULD BE CHOOSED ACCORDINGLY

Ex

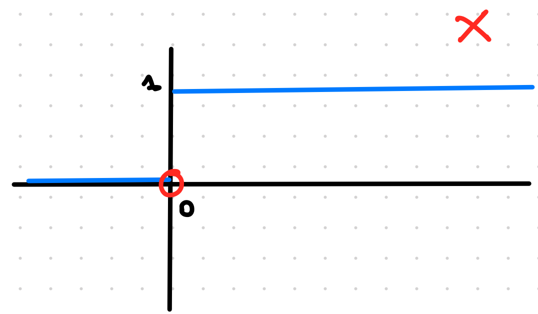
$y \in \{0, 1\}$

$$\hat{y} \in \begin{cases} 0 & z < 0 \\ z & z \geq 0 \end{cases}$$

NOTE HAS TO BE SOME NON-LINEARITY DIFFERENTIABLE IN EACH POINT



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



$$\hat{y} = \mathbb{1}[z > 0]$$

BY APPLYING THE SIGMOID $\sigma(z)$ FUNCTION AS NON LINEARITY AND CHOOSE A BINARY CLASSIFICATION PROBLEM WE GET THE FOLLOWING RESULT

$$\begin{cases} \hat{y} = \dfrac{1}{1+e^{-z}} \\ \mathcal{L} = y \log \hat{y} + (1-y)\log(1-\hat{y}) \end{cases} \longrightarrow \boxed{\frac{\partial \mathcal{L}}{\partial z} = \hat{y} - y}$$

WHICH IS EXTREMELY IMPORTANT IN THE DEFINITION PROCESS OF BACKPROP

FROM THE DEFINITIONS OF THE MODEL, WE CAN APPLY THE CHAIN RULE OF DERIVATION IN ORDER TO OBTAIN $\frac{\partial \mathcal{L}}{\partial W_{ij}^{[2]}}$. THE CHAIN RULE CONSISTS OF BREAKING DOWN DERIVATIVES INTO JACOBIANS

EX

$$\frac{\partial \mathcal{L}}{\partial W_{ij}^{[2]}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial W_{ij}^{[2]}} =$$

$$= \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial W_{ij}^{[2]}} = \frac{\partial \mathcal{L}}{\partial a^{[3]}} \frac{\partial a^{[3]}}{\partial W_{ij}^{[2]}} =$$

$$= \frac{\partial \mathcal{L}}{\partial a^{[3]}} \frac{\partial a^{[3]}}{\partial W_{ij}^{[2]}} = \frac{\partial \mathcal{L}}{\partial a^{[3]}} \frac{\partial a^{[3]}}{\partial z^{[3]}} \frac{\partial z^{[3]}}{\partial W_{ij}^{[2]}}$$

$$\frac{\partial \mathcal{L}}{\partial a^{[3]}} \frac{\partial a^{[3]}}{\partial z^{[3]}} \frac{\partial z^{[3]}}{\partial W_{ij}^{[2]}} = \overbrace{\frac{\partial \mathcal{L}}{\partial a^{[3]}} \frac{\partial a^{[3]}}{\partial z^{[3]}}}^{\frac{\partial \mathcal{L}}{\partial z^{3}}} \frac{\partial z^{[3]}}{\partial a^{[2]}} \frac{\partial a^{[2]}}{\partial W_{ij}^{[2]}} =$$

$$= \underbrace{\frac{\partial \mathcal{L}}{\partial z^{[3]}}}_{(a^{[3]}-y)} \frac{\partial z^{[3]}}{\partial a^{[2]}} \frac{\partial a^{[2]}}{\partial W_{ij}^{[2]}} = \left(a^{[3]}-y\right) \frac{\partial z^{[3]}}{\partial a^{[2]}} \frac{\partial a^{[2]}}{\partial W_{ij}^{[2]}}$$

AND FINALLY OBTAINING OUR GOAL EXPRESSION

DEPENDS ON THE FINAL N.L. FUNCTION

BLOCK: REPEAT

GOAL

$$\boxed{\left(a^{[3]}-y\right)} \boxed{\frac{\partial z^{[3]}}{\partial a^{[2]}} \frac{\partial a^{[2]}}{\partial z^{[2]}}} \boxed{\frac{\partial z^{[2]}}{\partial W_{ij}^{[2]}}}$$

IN GENERAL

- GOAL: CALCULATE $\frac{\partial \mathcal{L}}{\partial W_{ij}^{[l]}}$
- BLOCKS: $\frac{\partial z^{[L]}}{\partial a^{[L-1]}} \cdots \frac{\partial z^{[l+1]}}{\partial a^{[l]}} \frac{\partial a^{[l]}}{\partial z^{[l]}} \cdot \frac{\partial z^{[l]}}{\partial W_{ij}^{[l]}}$

THE OBJECTIVE IS TO REACH THE z OF THE CORRESPONDING LAYER (OF THE W, b W.R.T. WE ARE DERIVATING) $\longrightarrow$ GO BACKWARDS!

LET'S DO SOME ANALYSIS OF THE DIMENSIONS

$$a^{[2]} = g(z^{[2]})$$

$$\frac{\partial a^{[2]}}{\partial z^{[2]}} = \mathrm{diag}\left(g'(z^{[2]})\right)$$

→ JACOBIAN MATRIX (DERIVATIVE APPLIED TO EACH ELEMENT)

→ IT IS A DIAGONAL MATRIX

→ g IS APPLIED ELEMENT-WISE

$$\boxed{(a^{[3]} - y)}\ \boxed{\frac{\partial z^{[3]}}{\partial a^{[2]}}}\ \boxed{\frac{\partial a^{[2]}}{\partial z^{[2]}}}\ \boxed{\frac{\partial z^{[2]}}{\partial W_{ij}^{[2]}}}$$

$\mathbb{R}$

$$z^{[3]} = W^{[3]} a^{[2]} + b^{[3]} \in \mathbb{R}^{d_3}$$

$$\frac{\partial z^{[3]}}{\partial a^{[2]}} = W^{[3]} \in \mathbb{R}^{d_3 \times d_2}$$

NOTE WE CAN THINK IT IN THIS WAY

$$\begin{tikzcd} & b^{[l]} \\ z^{[l]} \arrow[ur] \arrow[dr] & \\ & W^{[l]} \end{tikzcd}$$

$$\cdots \leftarrow a^{[l-1]} \leftarrow z^{[l]} \leftarrow a^{[l]} \leftarrow \hat{y} \leftarrow \mathcal{L}$$

WE OBTAIN THE FOLLOWING FORM

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

$$\begin{bmatrix} \\ z_i^{[2]} \\ \\ \end{bmatrix} = \begin{bmatrix} \longleftrightarrow \\ \leftarrow W_{ij}^{[2]} \rightarrow \\ \longleftrightarrow \end{bmatrix} \begin{bmatrix} \\ a_j^{[1]} \\ \\ \end{bmatrix} + \begin{bmatrix} \\ b^{[2]} \\ \\ \end{bmatrix}$$

$$z_i^{[2]} = \sum_j W_{ij}^{[2]} \underbrace{a_j^{[1]}}_{\in \mathbb{R}^{d_1}} + b^{[2]}$$

!

$$\frac{\partial z^{[2]}}{\partial W_{ij}^{[2]}} = \boxed{a_j^{[i]} \cdot C_j} = a_j^{[1]} \cdot e_j$$

WHERE e_j IS THE j^{TH} BASIS VECTOR

$$a_j^{[i]} e_i = a_j^{[i]} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \begin{array}{cl} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ a_j^{[1]} \\ \vdots \\ 0 \end{bmatrix} & \begin{array}{l} \\ \\ \\ \rightarrow j^{TH} \\ \\ \\ \end{array} \end{array}$$

NOW, WE CAN FINALLY WORK OUR WAY TO A CLOSED FORM EXPRESSION

$$\frac{\partial \mathcal{L}}{\partial W_{ij}^{[2]}} = (a^{[3]} - y) W^{[3]} \operatorname{diag}(g'(z^{[2]})) \frac{\partial z^{[2]}}{\partial W_{ij}^{[2]}}$$

$$= (a^{[3]} - y) W^{[3]} \underline{\odot} g'(z^{[2]}) \cdot a_j^{[1]} e_i$$

HADAMART PRODUCT
(ELEMENT-WISE MATRIX
MULTIPLICATION)

$$\frac{\partial \mathcal{L}}{\partial W_{ij}^{[2]}} = \left[(a^{[3]} - y) W^{[3]} \odot g'(z^{[2]}) \right]_i a_j^{[1]}$$

IN GENERAL

$$\boxed{\frac{\partial \mathcal{L}}{\partial W_{ij}^{[l]}} = \left[(a^{[L]} - y) W^{[l+1]} \odot g'(z^{[l]}) \right]_i {a_j^{[l-1]}}^T}$$

WE ARE EXTRACTING THE i-TH TERM AND MULTIPLYING IT BY a_j

ONCE WE HAVE THE z OF THE LAST LAYER WE CAN WORK OUR WAY BACK TO OUR z OF INTEREST.

NOTE COMPUTATIONALLY, THE EXPRESSION WE FOUND IS MUCH MORE CONVENIENT THAN CALCULATING THE CHAIN BY MAT MULT ALL THE L BLOCKS

$$(a^{[L]}-y)\, W^{[L]} \operatorname{diag}(\ )\, W^{[L-1]} \operatorname{diag}(\ )\, W^{[\ ]} \operatorname{diag}(\ ) \cdots \times \quad :($$