1) Notations:

- X - Input
- Y - True Output
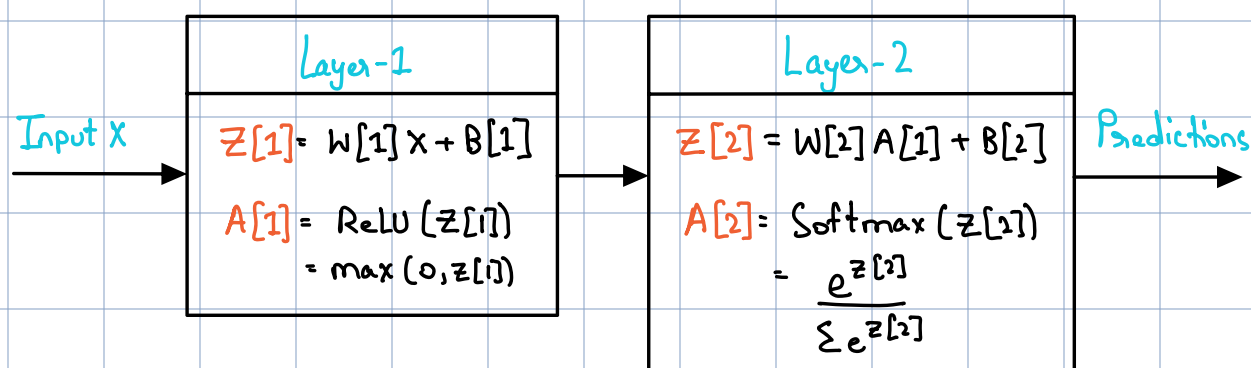- $W[i], B[i]$ - Weights & bias for layer i
- $Z[i], A[i]$ - Linear & Activation for layer i
- α - Learning rate
- m - Number of Examples

ANN from Scratch

2) Forward Propagation

Input X →

Layer-1
$Z[1] = W[1]X + B[1]$
$A[1] = ReLU(Z[1])$
$= \max(0, Z[1])$

Layer-2
$Z[2] = W[2]A[1] + B[2]$
$A[2] = Softmax(Z[2])$
$= \dfrac{e^{Z[2]}}{\sum e^{Z[2]}}$

→ Predictions

① Layer-2:

Weight $\partial W[2]$:

$Z[2] = W[2]A[1] + B[2]$ [from chain rule]

Chain Rule:

$\partial W[2] = \dfrac{\partial J}{\partial Z[2]} \cdot \boxed{\dfrac{\partial Z[2]}{\partial W[2]}}$

$\hookrightarrow A[1]$

Bias $\partial B[2]$:

$Z[2] = W[2]A[1] + B[2]$

Chain Rule:

$\partial B[2] = \dfrac{\partial J}{\partial Z[2]} \cdot \boxed{\dfrac{\partial Z[2]}{\partial B[2]}}$

$\hookrightarrow 1$

$$\partial W[2] = \frac{1}{m} \partial Z[2] A[1]^T$$

$$\partial B[2] = \frac{1}{m} \sum_{i=1} \partial Z[2]$$

② Layer-1:

$\partial Z[1] = \dfrac{\partial J}{\partial Z[2]} \cdot \boxed{\dfrac{\partial Z[2]}{\partial A[1]}} \cdot \boxed{\dfrac{\partial A[1]}{\partial Z[1]}}$

$\qquad W[2] \hookleftarrow \qquad \hookrightarrow g'(Z[1])$

$\partial Z[1] = W[2]^T \partial Z[2] \odot g'(Z[1])$

Weight $\partial W[1]$:

$Z[1] = W[1]A[0] + B[1]$

Chain Rule:

$\partial W[1] = \dfrac{\partial J}{\partial Z[1]} \cdot \boxed{\dfrac{\partial Z[1]}{\partial W[2]}}$

$\hookrightarrow A[0]$

Bias $\partial B[1]$:

$Z[1] = W[1]A[0] + B[1]$

Chain Rule:

$\partial B[1] = \dfrac{\partial J}{\partial Z[1]} \cdot \boxed{\dfrac{\partial Z[1]}{\partial B[2]}}$

$\hookrightarrow 1$

$$\partial W[1] = \frac{1}{m} \partial Z[1] A[0]^T$$

$$\partial B[1] = \frac{1}{m} \sum_{i=1} \partial Z[1]$$

4) Gradient Descent

$W[i] = W[i] - \alpha \cdot \partial W[i]$

$B[i] = B[i] - \alpha \cdot \partial B[i]$

3) Backward Propagation

Cost function $J = \dfrac{1}{m} \sum_{i=1}^{m} (\hat{y} - y)$

Chain Rule $\dfrac{\partial J}{\partial w} = \dfrac{\partial J}{\partial z} \cdot \dfrac{\partial z}{\partial w}$

Starting point $\partial Z[2] = A[2] - Y$

by
P. Toshith