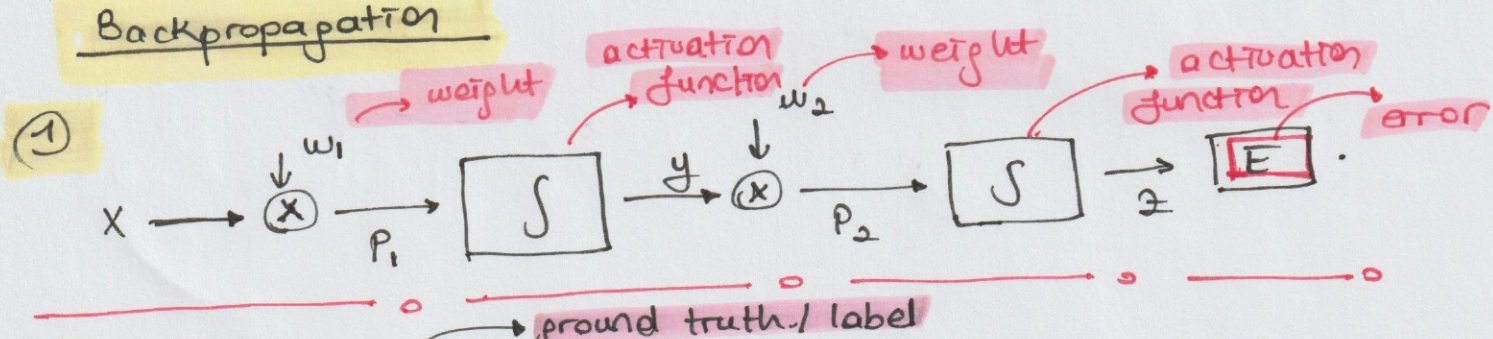
Backpropagation



② $E = \frac{-1}{2}(d - z)^2$

d → ground truth / label

z → prediction

E → error-calculation

③ $\frac{\partial E}{\partial w_1} = ?$ $\frac{\partial E}{\partial w_2} = ?$

① $\frac{\partial E}{\partial w_2} = \frac{\partial E}{\partial z} \cdot \frac{\partial z}{\partial P_2} \cdot \frac{\partial P_2}{\partial w_2}$

$= (d - z) \cdot z(1 - z) \cdot y$

- $(d - z)$: calculated from label and output of the current layer.
- $z(1 - z)$: calculated from output of the current layer
- y: input of the current layer.

* $\frac{\partial E}{\partial z} = \frac{-1}{2} \cdot 1 \cdot 2 (d - z) = (d - z)$

* $\frac{\partial P_2}{\partial w_2} = y$

* Let's say we use sigmoid activation function, so $z = Act(P_2)$ → sigmoid

* $\frac{\partial z}{\partial P_2} = z(1 - z)$

② $\frac{\partial E}{\partial w_1} = \frac{\partial E}{\partial z} \frac{\partial z}{\partial P_2} \frac{\partial P_2}{\partial y} \frac{\partial y}{\partial P_1} \frac{\partial P_1}{\partial w_1}$

$= (d - z) \quad z(1 - z) \quad w_2 \quad y(1 - y) \quad x$

- $(d - z)$: Known (Calculated)
- $z(1 - z)$: Known (Calculated)
- w_2: weight of the next layer (already calculated in fprop)
- $y(1 - y)$: calculated from output of the current layer
- x: input of the current layer

* $\frac{\partial E}{\partial z} = (d - z)$

* $\frac{\partial z}{\partial P_2} = z(1 - z)$

* $\frac{\partial P_2}{\partial y} = w_2$

* $\frac{\partial y}{\partial P_1} = y(1 - y)$

* $\frac{\partial P_1}{\partial w_1} = x$

① ← navigation

Question - 1

⇒ The last layer of the network:

$$\begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \rightarrow \begin{bmatrix} \sigma_1(x) \\ \vdots \\ \sigma_N(x) \end{bmatrix}$$

** Softmax function takes a N dimensional vector, and produces a N dimensional vector → $R^N \rightarrow R^N$

$$x = [x_1, \dots, x_N] \in R^{1 \times N}$$

$$\sigma(x) = \text{softmax}(x) = [\sigma_1(x), \dots, \sigma_N(x)]$$

$$\sigma_i(x) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}$$

(a) $D_x \sigma(x) \in R^{N \times N} = ?$ ⇒

$$\begin{bmatrix} \frac{\partial \sigma_1(x)}{\partial x_1} & \cdots & \frac{\partial \sigma_N(x)}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial \sigma_1(x)}{\partial x_N} & \cdots & \frac{\partial \sigma_N(x)}{\partial x_N} \end{bmatrix}$$

Reminder: $J = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_N} \end{bmatrix}$ Jacobian Matrix

⇒ How we're gonna compute $\frac{\partial \sigma_i(x)}{\partial x_i}$ or $\frac{\partial \sigma_j(x)}{\partial x_j}$?? values.??

⇒ We're gonna use two tricks:

① Calculate the dragonal entries

② Calculate off-dragonal entries.

① Calculate dragonal entries

$$\frac{\partial \sigma_i(x)}{\partial x_i} = \frac{\partial}{\partial x_i}\left[\frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}\right] \Rightarrow$$

Trick-1

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$$

②

$f \to e^{x_i}$

$g \to \sum_{j=1}^{N} e^{x_j}$

$f' \to (e^{x_i})' \to e^{x_i}$

$g' \to \left(\sum_{j=1}^{N} e^{x_j}\right)' \to e^{x_i}$

or

$g' \to \frac{\partial (e^{x_1} + e^{x_2} + \dots + e^{x_i} + \dots + e^{x_N})}{\partial x_i}$

$\to e^{x_i}$

$$\frac{\partial \sigma_i(x)}{\partial x_i} = \frac{f'g - fg'}{g^2}$$

$$= \frac{e^{x_i}\left(\sum_{j=1}^{N} e^{x_j}\right) - e^{x_i} \cdot e^{x_i}}{\left(\sum_{j=1}^{N} e^{x_j}\right)^2}$$

(2) Calculate off-diagonal entries:

$$\frac{\partial \sigma_i(x)}{\partial x_j} = \frac{\partial}{\partial x_j}\left[\frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}\right] \Rightarrow$$

Trick -2

$(c \cdot f^{-1})' = -c \cdot f^{-2} \cdot f'$

where c = constant

$$= \frac{\partial}{\partial x_j}\left[e^{x_i} \frac{1}{\sum_{j=1}^{N} e^{x_j}}\right]$$

$e^{x_i} \to c$

$\frac{1}{\sum_{j=1}^{N} e^{x_j}} \to f^{-1}$ where $f = \sum_{j=1}^{N} e^{x_j}$

$f \to \sum_{j=1}^{N} e^{x_j}$

$c \to e^{x_i}$

$f' \to \left(\sum_{j=1}^{N} e^{x_j}\right)' \to e^{x_j}$

$f' \to \frac{\partial (e^{x_1} + e^{x_2} + \dots + e^{x_j} + \dots + e^{x_N})}{\partial x_j}$

$f' \to e^{x_j}$

$$\frac{\partial \sigma_i(x)}{\partial x_j} = \frac{-e^{x_i} \cdot e^{x_j}}{\left(\sum_{j=1}^{N} e^{x_j}\right)^2}$$

① $$\frac{\partial \sigma_i(x)}{\partial x_i} = \frac{e^{x_i}\left(\sum_{j=1}^{N} e^{x_j}\right) - (e^{x_i})^2}{\left(\sum_{j=1}^{N} e^{x_j}\right)^2}$$

$$= \frac{e^{x_i}\left(\sum_{j=1}^{N} e^{x_j}\right)}{\left(\sum_{j=1}^{N} e^{x_j}\right)^2} - \left(\frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}\right)^2$$

$$= \sigma_i(x) - \sigma_i^2(x)$$

$$= \sigma_i(x)\left(1 - \sigma_i(x)\right)$$ diagonal

② $$\frac{\partial \sigma_i(x)}{\partial x_j} = -\underbrace{\frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}}_{\sigma_i(x)} \cdot \underbrace{\frac{e^{x_j}}{\sum_{j=1}^{N} e^{x_j}}}_{\sigma_j(x)}$$

$$= -\sigma_i(x)\,\sigma_j(x)$$ off-diagonal

Finally, $D_x \sigma(x) =$

$$\begin{bmatrix} \sigma_1(x)(1-\sigma_1(x)) & -\sigma_1(x)\sigma_2(x) & \cdots & -\sigma_1(x)\sigma_N(x) \\ \vdots & \ddots & & \vdots \\ -\sigma_N(x)\sigma_1(x) & -\sigma_N(x)\sigma_2(x) & \cdots & \sigma_N(x)(1-\sigma_N(x)) \end{bmatrix}$$

diagonal: $\sigma_1(x)(1-\sigma_1(x))$ … $\sigma_N(x)(1-\sigma_N(x))$

off-diagonal: $-\sigma_1(x)\sigma_2(x)$, $-\sigma_1(x)\sigma_N(x)$, $-\sigma_N(x)\sigma_1(x)$, $-\sigma_N(x)\sigma_2(x)$

(b) $z = \underbrace{[\vartheta_1, \dots, \vartheta_N]}_{1\times N} \begin{bmatrix} D_x\sigma(x) \end{bmatrix}$

$1\times N \quad N\times N \Rightarrow 1\times N$

$= [z_1, \dots, z_N]_{1\times N}$

⇒ Let's calculate z_1, specifically $i=1$. for z_i

$z_1 = \vartheta \cdot \text{col}_1(D_x\sigma(x))$

$= (\vartheta_1 \sigma_1(x))(1-\sigma_1(x)) + \vartheta_2(-\sigma_1(x)\sigma_2(x)) + \dots + \vartheta_N(-\sigma_1(x)\sigma_N(x))$

$= \vartheta_1\sigma_1(x) - \vartheta_1\sigma_1^2(x) + \vartheta_2(-\sigma_1(x)\sigma_2(x)) + \dots + \vartheta_N(-\sigma_1(x)\sigma_N(x))$

$= \sigma_1(x)(\vartheta_1 - \vartheta_1\sigma_1(x) - \vartheta_2\sigma_2(x) - \dots - \vartheta_N\sigma_N(x))$ ↘ Trick

$= \sigma_1(x)(\vartheta_1 - \vartheta_1\sigma_1(x) - \vartheta_2\sigma_2(x) - \dots - \vartheta_N\sigma_N(x) + \vartheta_1\sigma_1(x) - \vartheta_1\sigma_1(x))$

$-\sum_{J=1}^{N}\vartheta_J\sigma_J(x)$

$= \sigma_1(x)(\vartheta_1 - \vartheta_1\sigma_1(x) - \sum_{J=1}^{N}\vartheta_J\sigma_J(x) + \vartheta_1\sigma_1(x))$

$= \sigma_1(x)(\vartheta_1 - \sum_{J=1}^{N}\vartheta_J\sigma_J(x))$ → $\vartheta\sigma(x)^T$

$= \sigma_1(x)(\vartheta_1 - \vartheta\cdot\sigma(x)^T)$

Generally, $\boxed{z_i = \sigma_i(x)(\vartheta_i - \vartheta\cdot\sigma(x)^T)}$

(c) $l(x,t) = -\sum_{i=1}^{N} t_i \ln(x_i)$ with $t \in [0,1]^{1\times N}$

for example $t = [0\ 0\ 1\ 0]$

$D_x l(x,t) \in R^{1\times N} = ?$

$$D_x l(x,t) = \left[\frac{\partial l(x,t)}{\partial x_1}, \dots, \frac{\partial l(x,t)}{\partial x_N} \right]$$

$$\frac{\partial l(x,t)}{\partial x_i} = ?$$

$$\frac{\partial l(x,t)}{\partial x_i} = \frac{\partial}{\partial x_i}\left(-t_i \ln(x_i)\right)$$

or $\frac{\partial}{\partial x_i}\left(-t_1 \ln(x_1) - t_2 \ln(x_2) - \dots - t_i \ln(t_i) - t_N \ln(\dots)\right)$

$$= \frac{\partial}{\partial x_i}\left(-t_i \ln(x_i)\right)$$

Trick $\ln(x)' = \frac{1}{x}$

$x = x_i$

$$= \frac{\partial}{\partial x_i}\left(-t_i \ln(x_i)\right) = -t_i \frac{1}{x_i}$$

$$D_x l(x,t) = \left[\frac{-t_1}{x_1}, \dots, \frac{-t_N}{x_N} \right]$$

(d) Division by zero when $x_i = 0$ or any class gets 0 probability.