13·9·21

Neural Networks - HW03

① Prove that,

$$D_{h^{(K)}} C = D_{\hat{y}} L_y(\hat{y})$$

From forward propagation, we know that

$$C(\theta) = L_y(\hat{y}) + \lambda \cancel{\Omega(\theta)}^{\,0} \quad (\text{since, } \lambda = 0)$$

therefore $C = L_y(\hat{y})$

We need to prove that

$$D_{h^{(K)}} L_y(\hat{y}) = D_{\hat{y}} L_y(\hat{y})$$

Again by definition, from forward propagation

$$a^{(K)} = b^{(K)} + W^{(K)} h^{(K-1)}$$

$$h^{(K)} \leftarrow g(a^{(K)})$$

$$\hat{y} \leftarrow h^{(K)}$$

then,

$$D_{h^{(K)}} C = D_{h^{(K)}} L_y(\hat{y}) = D_{\hat{y}} L_y(\hat{y})$$

where the first equality follows from (1),

and the second follows from (2).

$C \circ L_y$

② $D_{a^{(k)}} C = D_{h^{(k)}} C \odot g'(a^{(k)})^T$.

$C = L_y(\hat{y}) + \lambda \Omega$

Since $\lambda = 0$

$C = L_y(\hat{y})$, $\hat{y} = h(k)$

$g(a^{(k)}) = h^{(k)} = \sigma a^{(k)}$

By applying chain rule,

$$D_{a^{(k)}}(L_y(\hat{y})) = \frac{d\, L_y(\hat{y})}{d a^{(k)}} = \frac{d\, L_y(h^{(k)})}{d a^{(k)}}$$

$$= D_{\hat{y}} L_y(\hat{y}) \cdot D_\sigma a^{(k)} \cdot \cancel{D_{a^{(k)}} a^{(k)}}^{\,1}$$

$$= D_{\hat{y}} L_y(\hat{y}) \cdot D_\sigma a^{(k)} = D_{\hat{y}} L_y(\hat{y}) \cdot g'(a^{(k)}) \quad \text{— ①}$$

$g' = D\sigma$ (given)

$$\frac{\partial L_y(\hat{y})}{\partial a_i^{(k)}} = \sum_i \frac{d L_y(\hat{y})}{\partial h_j^{(k)}} \cdot \frac{\partial h^{(k)}}{\partial a_i^{(k)}}$$

$$= \frac{\partial L_y(\hat{y})}{\partial h_j^{(k)}} \sum_i \frac{\partial h^{(k)}}{\partial a_i^{(k)}}$$

We also know that, $\forall\ i \neq j$, $\frac{\partial h^{(k)}}{\partial a_i^{(k)}} = 0$.

$$\frac{\partial L_y(\hat{y})}{\partial a_i^{(k)}} = \frac{\partial L_y}{\partial h_j^{(k)}} \cdot \sum_i \frac{\partial h_i^{(k)}}{\partial a_i}$$

From what i understand, $\frac{\partial h_i^{(k)}}{\partial a_i^{(k)}}$ is a diagnol element matrix which is nothing but the activation function thats being multiplied to $D_{h^{(k)}} \boxed{C}$

$$\Rightarrow \frac{\partial L_y}{\partial h_j^{(k)}} \cdot \begin{bmatrix} \frac{\partial h_0^{(k)}}{\partial a_0^{(k)}} & \cdots & \frac{\partial h_n}{\partial a_0} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_0^{(k)}}{\partial a_n} & \cdots & \frac{\partial h_n^{(k)}}{\partial a_n^{(k)}} \end{bmatrix}.$$

Only diagnol elements will be having values.

Since for a symmetrix matrix, $A = A^T$,

we can write this as a transpose.

we also know that $\text{diag}\left[\frac{\partial h_0^{(k)}}{\partial a_0} \cdots \frac{\partial h_n^{(k)}}{\partial a_n^{(k)}}\right] = g(a^{(k)})$

we use this result now in 1

which gives us;

$$\frac{\partial L_y(\hat{y})}{\partial a_i^{(k)}} = [g'(a^k)]^T \cdot \frac{\partial L_y(\hat{y})}{\partial h_j^{(k)}}$$

$$= D_{h^{(k)}} \cdot L_y(\hat{y}) \cdot \odot [g'(a^{(k)})]^T$$

$$= D_{h^{(k)}} \cdot C \odot [g'(a^{(k)})]^T$$

Hence proved.

3 Given, $D_{W^{(k)}} C = h^{(k-1)} \cdot D_{a^{(k)}} \cdot C$

Also given, $[D_{W^{(k)}} C]_{ij} = \frac{\partial C}{\partial w_{ji}^{(k)}}$

This means that the matrix $\frac{\partial C}{\partial W^{(k)}}$ is a symmetric matrix.

We know, $\hat{y} = h^{(k)}$

$c = L_y(\hat{y})$

$$\frac{\partial c}{\partial w^{(k)}} = \frac{\partial L_y(\hat{y})}{\partial w^{(k)}} = \frac{\partial L_y(h^{(k)})}{\partial w^{(k)}} \quad \text{—— (i)}$$

$$\frac{\partial L_y(\hat{y})}{\partial w_{ij}(k)} = \sum_m \frac{\partial L_y(\hat{y})}{\partial a_m^k} \cdot \frac{\partial a_m^{(k)}}{\partial w_{ij}^{(k)}} \quad \text{—— (1)}$$

$$a^k = b^k + \sum w^k h^{k-1}$$

for m^{th} column of vector a^k

$$a_m^k = b_m^k + \sum_i w_{im}^k h_i^{k-1}$$

from (1), we can understand that all terms except $\frac{\partial a_m}{\partial w_{ij}} = 0$

$$\frac{\partial L_y(\hat{y})}{\partial w_{ij}^{(k)}} = \sum_m \frac{\partial L_y(\hat{y})}{\partial a_m^k} \cdot \sum_l^m h_i$$

$\times$ $\cancel{\nabla_w^{(k)} L_y(\hat{y}) = h^{(k-1)} \cdot \nabla_a^{(k)} L_y(\hat{y})}$.

$h^{(k)} = \sigma\, a^{(k)}$. where $\sigma \rightarrow$ activation function.

$$\frac{\partial L_y(\sigma(a^{(k)}))}{\partial w^{(k)}}$$

We know, $a^{(k)} = w^{k} h^{(k-1)} + b^{(k)}$

$$\frac{\partial c}{\partial w^{(k)}} = \frac{\partial L_y}{\partial w^{(k)}} \left[\sigma \left(w^{(k)} h^{(k-1)} + b^{(k)} \right] \right.$$

$$= D_{\hat{y}} L_y(\hat{y}) \cdot D\sigma\, a^{(k)} \cdot D_{w^{(k)}} \left[w^{(k)} \cdot h^{(k-1)} + b^{(k)} \right]$$

$$= D_{h^{(k)}} C \cdot D\sigma\, a^{(k)} \cdot h^{(k-1)}$$

$g' = D\sigma$

Symmetric matrix, as from previous question.

$$= D_{h^{(k)}} C \cdot \boxed{g'(a^{(k)})} (h^{(k-1)})$$

$$\frac{dc}{dw^{(k)}} = D_{h^{(k)}} C \cdot \times\ g'(a^{(k)})^{T} \cdot h^{(k-1)}$$

from previous question again,

$$D_{a^{(k)}} \cdot C = D_{h^{(k)}} C \odot g'(a^{(k)})^{T} \quad \text{--- (3)}$$

$$\frac{dc}{\partial w^{(k)}} = D_{a^{(k)}} C \cdot h^{(k-1)}$$

Hence proved

for (3),

$$\underbrace{\frac{\partial L_y(\hat{y})}{\partial h^{(k)}} \odot g'(a^{(k)})^{T}}$$

The above multiplication has to result in a scalar quantity. since it is a gradient.

from the results in previous question,

$[g'(a^{(k)})]^T$. will yield a square matrix with

$\frac{\partial h_i^{(k)}}{\partial a_i^{(k)}} \rightarrow$ as activation.

④ $D_{h^{(k-1)}} \cdot C = (D_{a^{(k)}} C) \, w^{(k)}$

$$D_{H^{(k-1)}} C = \frac{dC}{dh^{(k-1)}} = \frac{d\, L_y(\hat{y})}{dh^{(k-1)}}$$

$$\hat{y} = h^{(k)} \; ; \; D_{h^{(k-1)}} C = \frac{dL_y(h^{(k)})}{dh^{(k-1)}}$$

then $h^{(k)} = \sigma \, a^{(k)}$

$$D_{h^{(k-1)}} C = \frac{dL_y}{dh^{(k-1)}} (\sigma (a^{(k)})]$$

$$a^{(k)} = w^{(k)} h^{(k-1)} + b^{(k)}$$

$$D_{h^{(k-1)}} C = \frac{dL_y}{dh^{(k-1)}} \cdot (\sigma [w^k h^{(k-1)} + b^{(k)}])$$

$$= D_{\hat{y}} L_y(\hat{y}) \, D\sigma(a^{(k)}) . D_{h^{(k-1)}} a^{(k)} .$$

$$= D_{\hat{y}} L_y(\hat{y}) . D\sigma(a^{(k)}) \, w^{(k)}$$

$$= D_{h^{(k)}} C \cdot D\sigma (a^{(k)}) \, w^{(k)}$$

from the previous results,

$$D\sigma = g' \text{ applied element wise.}$$

therefore,

$$= D_{h^{(k)}} C \odot [g'(a^{(k)})]^T \cdot w^{(k)}$$

$$= [D_{a^{(k)}} C]\, w^{(k)}$$