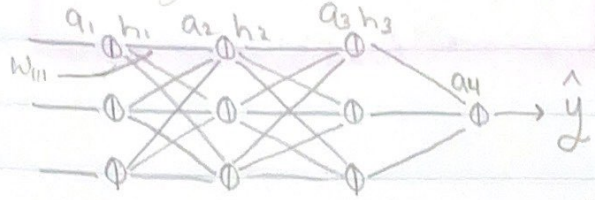
Gradient Descent

04/11/2023

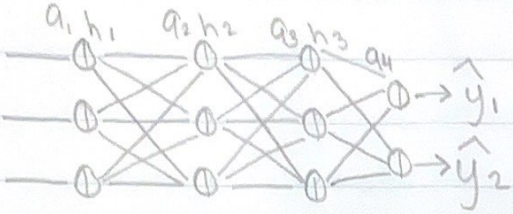
defining the model



Regression

$x \in \mathbb{R}^3$

$y \in \mathbb{R}$



Classifion

$x \in \mathbb{R}^3$

$y \in \mathbb{R}^2$

	Outputs: Real Values	Outputs: Properties
Output function	Linear Layers	Softmax
Loss function	MSE	Cross Entropy: $-\sum_{i=1}^{C} p_i \log q_i$

W_{111} is updated by

$$W_{111} \leftarrow W_{111} - \alpha \frac{dL}{dW_{111}}$$

how do we compute this?

Simplified example

: Loss L is computed using y and $\hat{y}$

$$\frac{dL}{dW_{111}} = \frac{dL}{d\hat{y}} \times \frac{d\hat{y}}{da_L} \times \cdots \times \frac{da_3}{dh_2} \times \frac{dh_2}{da_2} \times \frac{da_2}{da_1} \times \frac{da_1}{dW_{111}}$$

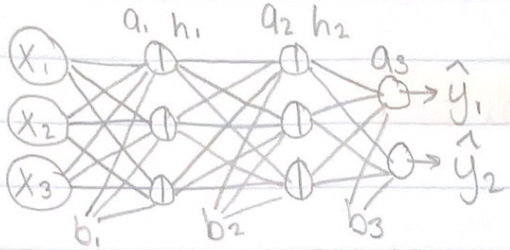
*only include W's that you want to optimize

$$\frac{dL}{dW_{111}} = \frac{dL}{dh_1} \frac{dh_1}{dW_{111}}$$

(break chain between $\frac{da_2}{dh_1}$ and $\frac{dh_1}{da_1}$)

Gradient Descent (Part II)

04/13/2023



★ assuming a binary classification problem ★

Loss = L $\quad \hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix}$

★ $W_{111} \leftarrow W_{111} - \alpha \frac{dL}{dW_{111}}$

$$\frac{dL}{dW_{111}} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{da_3} \cdot \frac{da_3}{dh_2} \cdot \frac{dh_2}{da_2} \cdot \frac{da_2}{dh_1} \cdot \frac{dh_1}{da_1} \cdot \frac{da_1}{dW_{111}}$$

- $\frac{dL}{d\hat{y}} \frac{da_3}{}$: gradient across output layer
- $\frac{dh_2}{} \frac{da_2}{}$: gradient from output layer to last hidden layer
- $\frac{dh_1}{} \frac{da_1}{}$: gradient from 2nd hidden layer to 1st
- $\frac{dW_{111}}{}$: gradient to the weight of interest

Quantities of interest

1) gradient with respect to output layer
2) gradient with respect to hidden layers
3) gradient with respect to weights and biases

For classification:

cost function: cross entropy

output layer: soft max

example (cross entropy):

$\hat{y} = \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix}$ true prob $= \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$CE = -\sum p_i \log q_i$ (q_i: $\hat{y}_i$)

$= -\log(0.8)$

★ Negative log likelihood ★

parameters (for example):

$$\theta = [W_1, W_2, W_3, b_1, b_2, b_3]$$

① $L = -\log(\hat{y}_\ell)$ $\{\ell$ = true class label

$$\frac{dL}{d\hat{y}_i} = \begin{cases} -\frac{1}{\hat{y}_i} & \text{if } i = \ell \\ 0 & \text{otherwise} \end{cases} \qquad -\frac{\mathbb{1}_{i=\ell}}{\hat{y}_\ell} \quad \left\{\mathbb{1}_{i=\ell} = \begin{cases} 1 & i = \ell \\ 0 & \text{otherwise} \end{cases}\right\}$$

(indicator function; $\mathbb{1}_{\ell=1}$: is 1 the true class?)

$$\nabla_{\hat{y}} L = \begin{bmatrix} \frac{dL}{d\hat{y}_1} \\ \frac{dL}{d\hat{y}_2} \\ \vdots \end{bmatrix} = -\frac{1}{\hat{y}_\ell} \begin{bmatrix} \mathbb{1}_{\ell=1} \\ \mathbb{1}_{\ell=2} \\ \vdots \end{bmatrix} = -\frac{1}{\hat{y}_\ell} e(\ell)$$

where $e(\ell)$ = a vector of all zeros with length = len(y), and a one at ℓ^{th} position

$$\boxed{\nabla_{\hat{y}} L = -\frac{1}{\hat{y}_\ell} e(\ell)}$$

$\frac{d\hat{y}}{da_{Last}}$ (in our example, this is $\frac{d\hat{y}}{da_3}$)

simpler: $\frac{d\hat{y}_\ell}{da_{Last\,i}} = \frac{d}{da_{Last\,i}} \left(\frac{\exp(a_{Last\,\ell})}{\sum_j \exp(a_{Last\,j})} \right)$

$$\frac{d\hat{y}_\ell}{da_{Last\,i}} = \hat{y}_\ell (\mathbb{1}_{\ell=i} - \hat{y}_i)$$

$$\frac{d\hat{y}}{da_{Last\,i}} = \hat{y} (\mathbb{1}_{\ell=i} - \hat{y}_i)$$

$$\frac{d\hat{y}}{da_{last\,i}} = \begin{bmatrix} \frac{dy_1}{da_{last\,i}} \\ \frac{dy_2}{da_{last\,i}} \\ \vdots \end{bmatrix} = \begin{bmatrix} \hat{y}_1 (\mathbb{1}_{1=i} - \hat{y}_i) \\ \hat{y}_2 (\mathbb{1}_{2=i} - \hat{y}_i) \\ \vdots \end{bmatrix} = \hat{y}(\mathbb{1}_{\ell=i} - \hat{y}_i)$$

$$\frac{dL}{da_{last}} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{da_{last}}$$

$$\frac{dL}{da_{last\,i}} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{da_{last\,i}}$$

$$= \frac{-1}{\hat{y}_\ell} e(\ell) \cdot \hat{y}(\mathbb{1}_{\ell=i} - \hat{y}_i)$$

$$= -(\mathbb{1}_{\ell=i} - \hat{y}_i)$$

$$\frac{dL}{da_{last}} = \begin{bmatrix} \frac{dL}{da_{last\,1}} \\ \frac{dL}{da_{last\,2}} \\ \vdots \end{bmatrix} = \begin{bmatrix} -(\mathbb{1}_{\ell=1} - \hat{y}_1) \\ -(\mathbb{1}_{\ell=2} - \hat{y}_2) \\ \vdots \end{bmatrix} = -(e(\ell) - \hat{y})$$

$$\boxed{\frac{dL}{da_{last}} = -(e(\ell) - \hat{y})}$$