www.menti.com Code : 178406

CS230: Lecture 3

The mathematics of deep learning

Backpropagation, Initializations, Regularization

Kian Katanforoosh

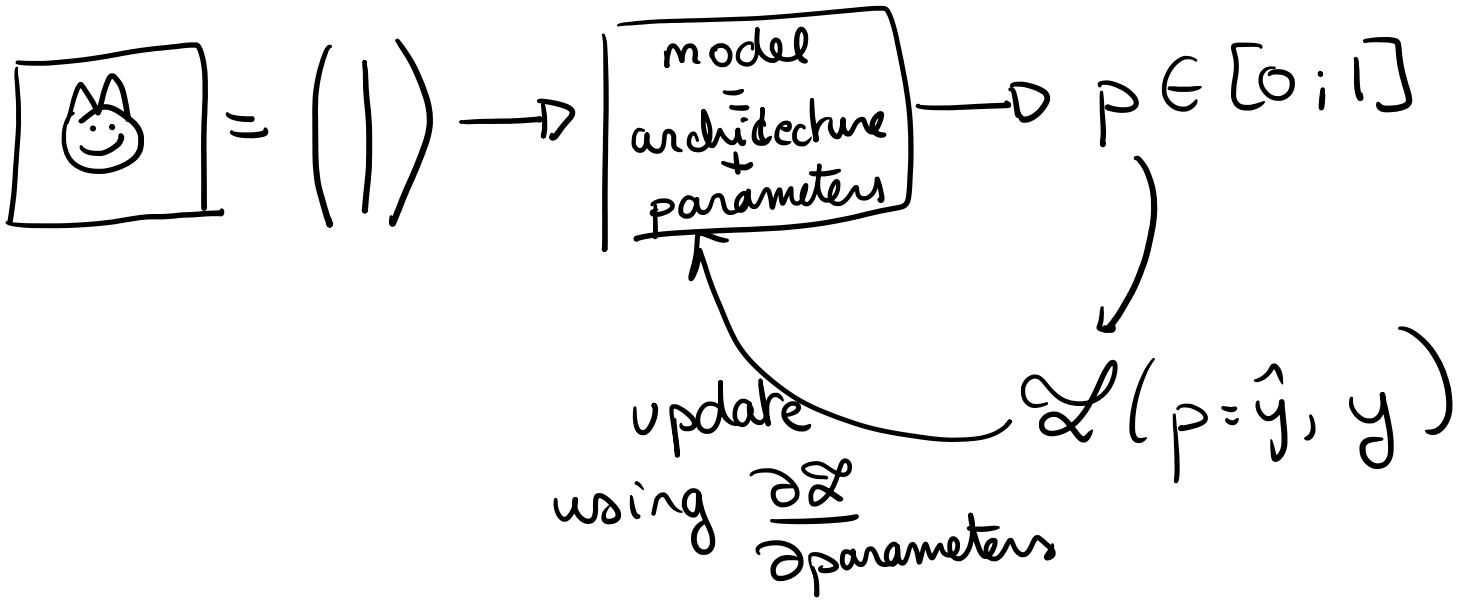
I – Backpropagation - derivatives

II – Initializations - Variance
Mean

III – Regularization

I - Backpropagation

Problem statement

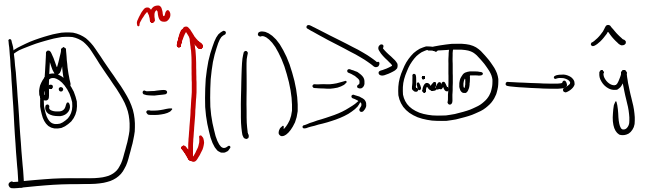


GD: $w = w - \alpha \frac{\partial \mathcal{L}}{\partial w}$

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

A – Logistic Regression backpropagation for one training example

$$\frac{\partial \mathcal{L}}{\partial z} = \frac{\partial \mathcal{L}}{\partial a} \cdot \frac{\partial a}{\partial z}$$



forward propagation :

x $(n,1)$

$z = w x + b$ $(1,n)$ $(1,1)$

$\hat{y} = a = \sigma(z)$

Loss : $\mathcal{L} = -\left(y \log \hat{y} + (1-y) \log(1-\hat{y})\right)$

Backprop : $\frac{\partial \mathcal{L}}{\partial w}$ $\frac{\partial \mathcal{L}}{\partial b}$

$z = wx + b = (w_1 \dots w_n) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + b = \sum_{i=1}^{n} w_i x_i$

$\frac{\partial z}{\partial w_1} = x_1$

$\frac{\partial z}{\partial w_2} = x_2$

$\frac{\partial \mathcal{L}}{\partial w}$?

$$\frac{\partial \mathcal{L}}{\partial a} = -\left(y \frac{\partial \log a}{\partial a} + (1-y) \frac{\partial \log(1-a)}{\partial a}\right)$$

$$= -\left(y \frac{1}{a} + (1-y) \frac{1}{1-a} \cdot (-1)\right)$$

$$\frac{\partial \mathcal{L}}{\partial z} = -\left(y \frac{1}{a} \frac{\partial a}{\partial z} + (1-y) \frac{1}{a-1} \frac{\partial a}{\partial z}\right)$$

$$= -\left(y \frac{1}{a} a(1-a) + (1-y) \frac{-1}{a-1} a(1-a)\right)$$

$$= -y(1-a) - a(1-y) = \boxed{a - y}$$

$$\underbrace{\frac{\partial \mathcal{L}}{\partial w}}_{(1,n)} = \frac{\partial \mathcal{L}}{\partial z} \cdot \underbrace{\frac{\partial z}{\partial w}}_{(1,n)} = \underbrace{(a-y)}_{(1,1)} \underbrace{x^T}_{(1,n)}$$

x, x^T ?

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial z} \cdot \frac{\partial z}{\partial b} = (a-y) \cdot 1$$

$$\begin{cases} w = w - \alpha (a-y) x^T \\ b = b - \alpha (a-y) 1 \end{cases}$$

B – Logistic Regression backpropagation for a batch of m examples

* Forward propagation :

$X \leftarrow (n,m)$

$Z = WX + b$ — $W \leftarrow (1,n)$, $WX \leftarrow (n,m)$, $b \leftarrow (1,1)$ broadcasts to $(1,m)$, $b = (b\ b \dots b)$

$A = \sigma(Z)$ — $Z \rightarrow (1,m)$, $A \rightarrow (1,m)$

* Loss : $J = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}^{(i)}$ ← binary cross entropy

* Backprop : $\frac{\partial J}{\partial w}$, $\frac{\partial J}{\partial b}$

* Notations :

$w = (w_1, w_2, \dots, w_n)$

$Z = (z^{(1)}, z^{(2)}, \dots, z^{(m)})$

$A = (a^{(1)}, \dots \quad a^{(m)})$

$$z^{(i)} = \sum_{k=1}^{n} w_k x_k^{(i)} + b$$

$\forall k \in [\![1, n]\!]$ ← $k = 1 \dots n$

$$\frac{\partial J}{\partial w_k}_{b} = \frac{1}{m} \sum_{i=1}^{m} \frac{\partial \mathcal{L}^{(i)}}{\partial w_k} = \frac{-1}{m} \sum_{i=1}^{m} y^{(i)} \frac{\partial \log a^{(i)}}{\partial w_k}_{b} + (1-y^{(i)}) \frac{\partial \log(1-a^{(i)})}{\partial w_k}_{b}$$

$$= \frac{-1}{m} \sum_{i=1}^{m} y^{(i)} \frac{\partial \log a^{(i)}}{\partial z^{(i)}} \cdot \frac{\partial z^{(i)}}{\partial w_k}_{b} + (1-y^{(i)}) \frac{\partial \log(1-a^{(i)})}{\partial z^{(i)}} \cdot \frac{\partial z^{(i)}}{\partial w_k}_{b}$$

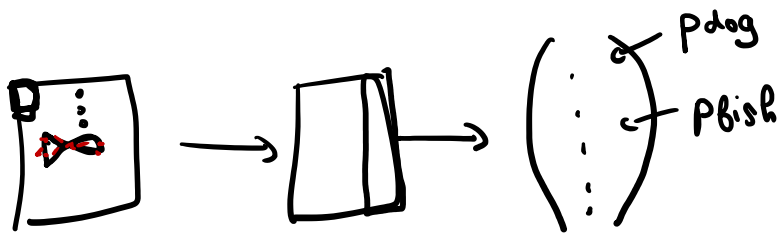
$$= \frac{1}{m} \sum_{i=1}^{m} (a^{(i)} - y^{(i)}) \cdot \frac{\partial z^{(i)}}{\partial w_k}_{b}$$

$$= \frac{1}{m} \sum_{i=1}^{m} \underbrace{(a^{(i)} - y^{(i)})}_{(1,1)} \cdot \underbrace{x_k^{(i)}}_{(1,1)} \cdot 1$$

$$\frac{\partial J}{\partial w} = \left(\frac{\partial J}{\partial w_1}, \dots, \frac{\partial J}{\partial w_n} \right) = \frac{1}{m} (A-Y) \cdot X^T$$

$$\underbrace{\frac{\partial J}{\partial b}}_{(1,1)} = \frac{1}{m} \underbrace{(A-Y)}_{(1,m)} \cdot \underbrace{1}_{(m,1)} \rightarrow \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Question: You have trained an animal classifier. Can you tell what part of the input led to this prediction?



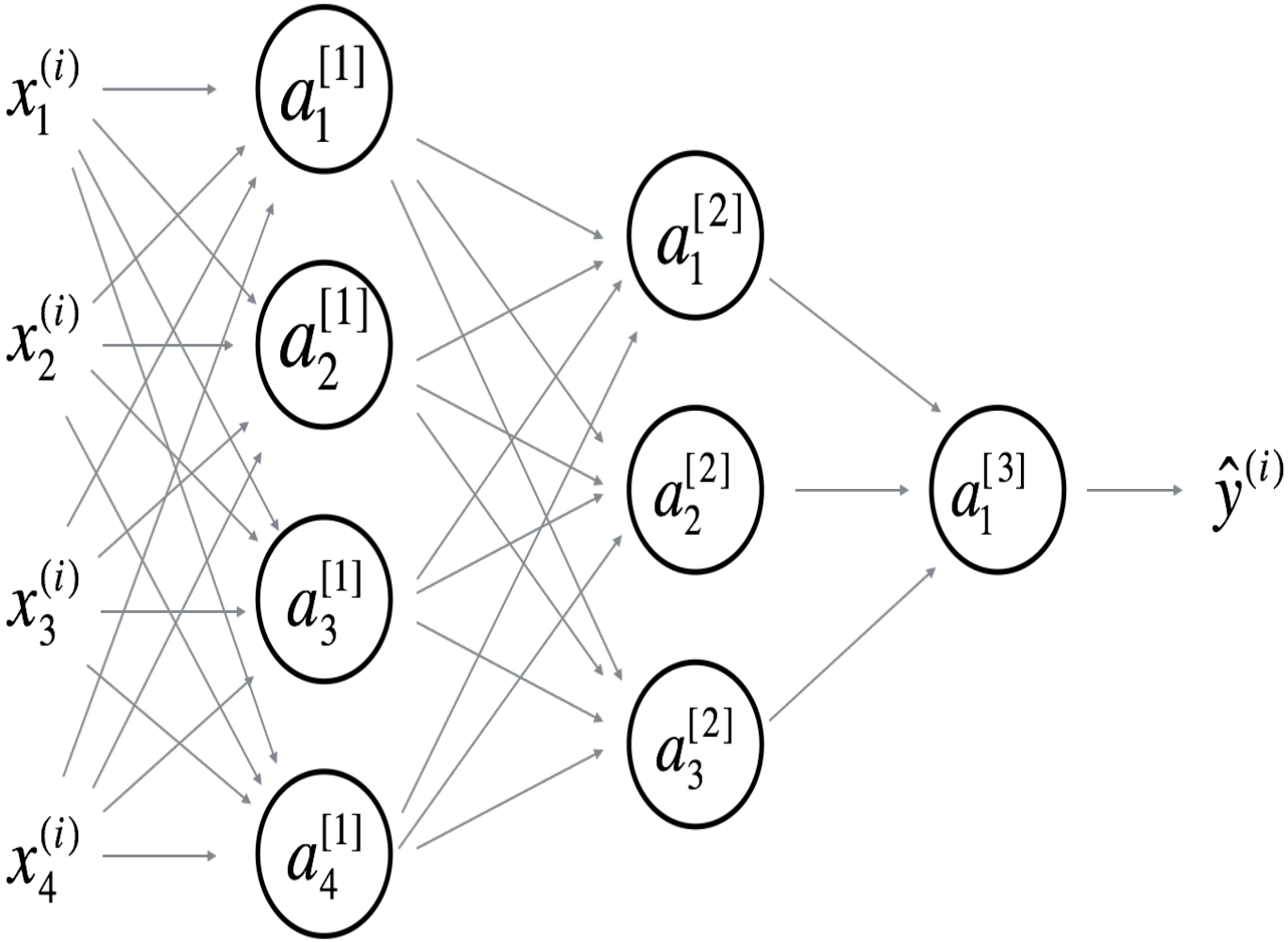
$\frac{\partial \mathcal{L}}{\partial w}, \frac{\partial \mathcal{L}}{\partial b}$

$\frac{\partial \mathcal{L}}{\partial x}$ = what is the value of the impact of shifting x, on $\mathcal{L}$

$\frac{\partial \mathcal{L}}{\partial x_{11}}$

II - Initializations

Problem statement



Forward propagation (L layers)

$$z^{[1]} = W^{[1]}x + b^{[1]}$$
$$a^{[1]} = g^{[1]}(z^{[1]})$$
$$z^{[2]} = W^{[2]}a^{[1]} + b^{[2]}$$
$$a^{[2]} = g^{[2]}(z^{[2]})$$
$$\ldots$$
$$z^{[L]} = W^{[L]}a^{[L-1]} + b^{[L]}$$
$$a^{[L]} = g^{[L]}(z^{[L]})$$

In class, you've seen that:

(i) - not too high/low

(ii) - breaking the symmetry

for a tanh activat° at layer l:

$$\begin{cases} W^{[l]} = np.random.randn(shape) \cdot np.sqrt\left(\frac{1}{n^{[l-1]}}\right) \\ b^{[l]} = 0 \end{cases}$$

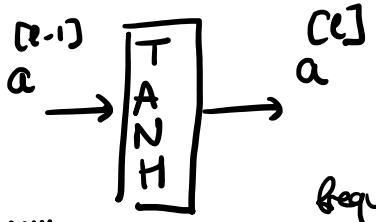
$\Leftrightarrow$

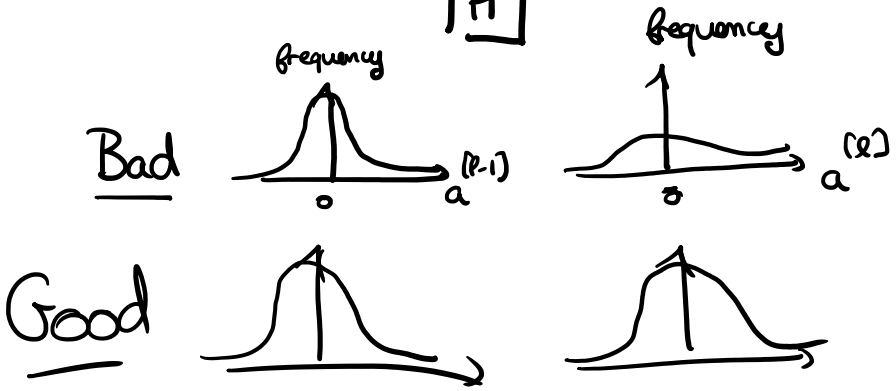
$$W^{[l]} \sim \underbrace{\mathcal{N}}_{\text{(i)}}\left(\underbrace{\mu = 0}_{\text{(ii)}}, \underbrace{\sigma^2 = \frac{1}{n^{[l-1]}}}_{\text{(ii)}}\right)$$

The goal: Given a layer l, we'd like $Var(a^{[l-1]}) = Var(a^{[l]})$

$$\Downarrow$$

$$Var(w^{[l]}) = \frac{1}{n^{[l-1]}}$$



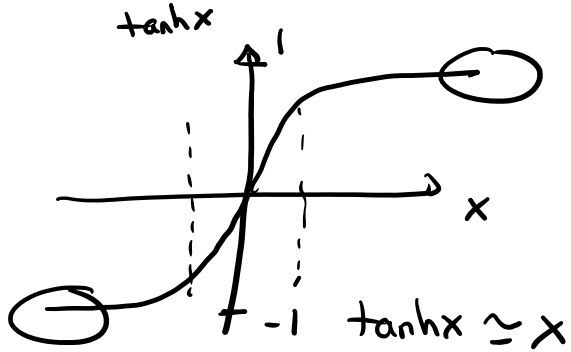


vanishing and exploding gradient

$$Var(a) = Var(a_1) = Var(a_2) = \dots$$

$$a = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

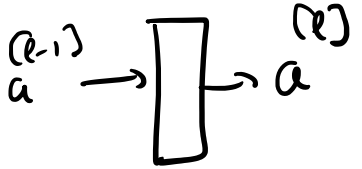
Let's prove that: $Var(a^{[l-1]}) = Var(a^{[l]}) \rightarrow Var(W^{[l]}) = \frac{1}{n^{[l-1]}}$



Forward prop $z^{[l]} = W^{[l]} a^{[l-1]} + b^{[l]} \Rightarrow ?$

$a^{[l]} = \tanh(z^{[l]}) \Rightarrow a^{[l]} = z^{[l]}$

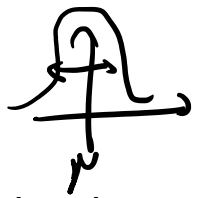
$Var(a^{[l]}) = Var(z^{[l]})$



$$z^{[l]} = W^{[l]} a^{[l-1]} + \underset{b=0}{0} = \begin{pmatrix} w_{11} & \cdots & w_{1n^{[l-1]}} \\ \vdots & \ddots & \vdots \\ w_{n^{[l]}1} & \cdots & w_{n^{[l]}n^{[l-1]}} \end{pmatrix} \begin{pmatrix} a_1^{[l-1]} \\ \vdots \\ a_{n^{[l-1]}}^{[l-1]} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n^{[l-1]}} w_{1i}^{[l]} \cdot a_i^{[l-1]} \\ \vdots \\ \sum_{i=1}^{n^{[l-1]}} w_{n^{[l]}i}^{[l]} a_i^{[l-1]} \end{pmatrix}$$

$$Var(a^{[l]}) = Var(z^{[l]}) = Var\left(\sum_{i=1}^{n^{[l-1]}} w_{1i}^{[l]} a_i^{[l-1]} \right) \longleftrightarrow \underbrace{\cdots}_{=1} Var(a_i^{[l-1]})$$

Let's prove that: $Var(a^{[l-1]}) = Var(a^{[l]}) \rightarrow \; Var(W^{[l]}) = \frac{1}{n^{[l-1]}}$



Checkpoint: $Var(a^{[l]}) = Var(z^{[l]}) = Var\left(\sum_{i=1}^{n^{[l-1]}} W_{1,i}^{[l]} a_i^{[l-1]}\right)$

Assumptions:
- weights are i.i.d
- inputs are i.i.d
- weights and inputs are mutually independant

$$= \sum_{i=1}^{n^{[l-1]}} Var\left(W_{1i}^{[l]} . a_i^{[l-1]}\right)$$

$Var(XY) = E[X]^2 Var Y + E[Y]^2 Var X + Var X Var Y$

$$= \sum_{i=1}^{n^{[l-1]}} \left[\underbrace{E[W_{1i}^{[l]}]}_{=0}^2 Var(a_i^{[l-1]}) + \underbrace{E[a_i^{[l-1]}]}_{=0}^2 Var(W_{1i}^{[l]}) + Var(W_{1i}^{[l]}) Var(a_i^{[l-1]}) \right]$$

$$= \sum_{i=1}^{n^{[l-1]}} \underbrace{Var(W_{1i}^{[l]})}_{\perp \text{ of } i} \underbrace{Var(a_i^{[l-1]})}_{\perp \text{ of } i}$$

$$Var(a^{[l]}) = n^{[l-1]} Var(W^{[l]}) Var(a^{[l-1]}) \Longrightarrow Var(W^{[l]}) = \frac{1}{n^{[l-1]}}$$

We've shown that for every layer l: $\boxed{Var(a^{[l]}) = n^{[l-1]} Var(W^{[l]}) Var(a^{[l-1]})}$

$$Var(a^{[L]}) = n^{[L-1]} Var(W^{[L]}) Var(a^{[L-1]})$$

$$= n^{[L-1]} Var(W^{[L]}) \left(n^{[L-2]} \cdot Var(W^{[L-1]}) \cdot Var(a^{[L-2]}) \right)$$

$$= \dots$$

$$= \left[\prod_{l=1}^{L} n^{[l-1]} \cdot Var(W^{[l]}) \right] Var(\underbrace{a^{[0]}}_{=x}) = \beta^{L} \cdot Var(X)$$

$Var(W^{[l]}) = 1 / n^{[l-1]} \cdot \beta$ $\qquad \beta > 1$