① P.T. $\sigma(x) = \sigma(x+c)$.

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \rightarrow ①$$

$$\sigma(x_i+c) = \frac{e^{x_i+c}}{\sum_j e^{x_j+c}} = \frac{e^{x_i} \cdot e^{c}}{\sum_j e^{x_j} \cdot e^{c}}$$

$$= \frac{e^{c} \cdot e^{x_i}}{e^{c} \sum_j e^{x_j}} = \frac{e^{x_i}}{\sum_j e^{x_j}} \rightarrow ②$$

① = ②. Hence proved.

$\frac{d}{dx}\sigma(x) = ?$ $\quad \frac{d}{dx}\sigma(x) = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right)$

$$= \frac{d}{dx}(1+e^{-x})^{-1} = -1\,(1+e^{-x})^{-2}\left[\frac{d}{dx}(1+e^{-x})\right]$$

$$= -(1+e^{-x})^{-2} \cdot \left[0 + (-e^{-x})\right] = \frac{e^{-x}}{(1+e^{-x})^{+2}}$$

$$= \underbrace{\frac{e^{-x}}{1+e^{-x}}}_{1-\sigma(x)} \cdot \underbrace{\frac{1}{(1+e^{-x})}}_{\sigma(x)}$$

$$= \sigma(x) \cdot [1-\sigma(x)]$$

(2 b)

$$\hat{y} = \text{Softmax}(\theta)$$

$$CE(y, \hat{y}) = -\sum_i^d y_i \cdot \log(\hat{y}_i)$$

$$\frac{d}{d\theta} \cdot CE(y, \hat{y}) = ?$$

Soln:

$$\frac{d}{d\theta}\left[- \sum_i \cdot y_i \cdot \log(\hat{y}_i) \right] = ?$$

$$= -\frac{d}{d\theta}\left[\Sigma_i y_i \log \cdot (\overset{\text{softmax}}{(\text{sigmoid}}(\theta_i)) \right] = ?$$

y is one-hot-encoded vector.

$\therefore$ chose K, s.t. $y_K \neq 0$

$$= -\frac{d}{d\theta}\left[y_K \cdot \log \left(\overset{\text{softmax}}{\text{sigmoid}}(\theta_K) \right) \right] = ?$$

$$= -\frac{d}{d\theta} \cdot \left[y_K \cdot \log \frac{e^{\theta_K}}{\sum_{j=1}^{d} e^{\theta_j}} \right]$$

$$= -\frac{d}{d\theta}\left[y_K \cdot \left(\log e^{\theta_K} - \log \sum_{j=1}^{d} e^{\theta_j} \right) \right]$$

$$= -\frac{d}{d\theta}\left[y_K \cdot \left(\theta_K - \log \sum_{j=1}^{d} e^{\theta_j} \right) \right]$$

chose i, s.t. $i == K$.

$$= -\frac{d}{d\theta_i}\left[y_K \left(\theta_K - \log \sum_{j=1}^{d} e^{\theta_j} \right) \right]$$

$$= -\left[y_k \cdot \left(1 - \frac{1}{\sum_{j=1}^{d} e^{\theta_j}} e^{\theta_k} \right) \right]$$

$$= -\left[y_k \cdot \left(1 - \text{Softmax}(\theta_k) \right) \right]$$

$$= -\left[y_k \left(1 - \hat{y}_k \right) \right] = (\hat{y}_k - 1) \cdot y_k .$$

But $y_k = 1$.

$$\Rightarrow \frac{d}{d\theta_k} CE(y, \hat{y}) = (\hat{y}_k - 1). \rightarrow (1)$$

chose $j \neq k$ s.t $j \neq k$.

$$\therefore -\frac{d}{d\theta_j} \left[y_k \left(\theta_k - \log \sum_{j=1}^{d} e^{\theta_j} \right) \right]$$

$$= -\left[y_k \cdot \left(0 - \frac{1}{\sum_{j=1}^{d} e^{\theta_j}} \cdot e^{\theta_j} \right) \right]$$

$$= -\left[y_k \left(- \text{softmax}(\theta_j) \right) \right]$$

$$= y_k \cdot \text{softmax}(\theta_j) = y_k \cdot \hat{y}_j = \hat{y}_j . \rightarrow (2) .$$

From (1) & (2),

$$\frac{d}{d\theta_k} CE(y, \hat{y}) = \hat{y}_k - 1 \rightarrow (3)$$

$$\frac{d}{d\theta_j} CE(y, \hat{y}) = \hat{y}_j - 0 \rightarrow (4)$$

But, $y_k = 1$ & $y_g = 0$. in 1-hot-encoding format

$\therefore$ (3) & (4) $\Rightarrow$

$$\frac{\partial}{\partial \theta_k} CE(y, \hat{y}) = \hat{y}_k - y_k$$

$$\frac{\partial}{\partial \theta_g} CE(y, \hat{y}) = \hat{y}_g - y_g$$

Generalize,

$$\boxed{\frac{\partial}{\partial \theta} CE(y, \hat{y}) = \hat{Y} - Y}$$

---

2 c) $J = CE = -\sum_{i=1}^{N} y_i \cdot \log(\hat{y}_i)$

$\hat{y}_i = \text{softmax}(h W_2 + b_2)$

$h = \text{sigmoid}(x W_1 + b_1)$

$\frac{dJ}{dx} = ?$

$J = -\sum_{i=1}^{N} y_i \log(\text{softmax}(\theta_i))$

$\therefore J(\theta)$.

$\theta = h W_2 + b_2$

$\therefore \theta(h)$.

$h = \text{sigmoid}(z)$; $z = x W_1 + b_1$

$\therefore h(z)$.

$z = x \cdot w_1 + b_1$

$\therefore z(x)$.

$$\therefore \frac{dJ}{dx} = \frac{dJ(\theta)}{d\theta} \cdot \frac{d\theta(h)}{dh} \cdot \frac{dh(z)}{dz} \cdot \frac{dz(x)}{d(x)}.$$

From previous question, $\frac{dJ(\theta)}{d\theta} = (\hat{y} - y) \rightarrow (1)$.

$$\frac{d\theta(h)}{dh} = w_2 \rightarrow (2).$$

$$\frac{dh(z)}{dz} = \sigma(z) \odot \sigma(1-z) \rightarrow (3)$$

element-wise prod.

$$\frac{dz(x)}{dx} = w_1 \rightarrow (3).$$

$$\frac{dJ}{dx} = (\hat{y} - y) * (w_2) \odot (\sigma(z) \odot \sigma(1-z)) * w_1$$

$1 \times D_y \cdot D_y \times D_h \cdot 1 * D_h \cdot D_H \times D_x$

$1 \times D_h \cdot 1 * D_h \cdot D_h \times D_x$

$= 1 \times D_h \times D_h \times D_x$

$\sim 1 \times D_x$

(2) D_x H D_y

No of Params = $\left(\left((1 + D_x) * H\right)\right) + \left((H + 1) * D_y\right)$

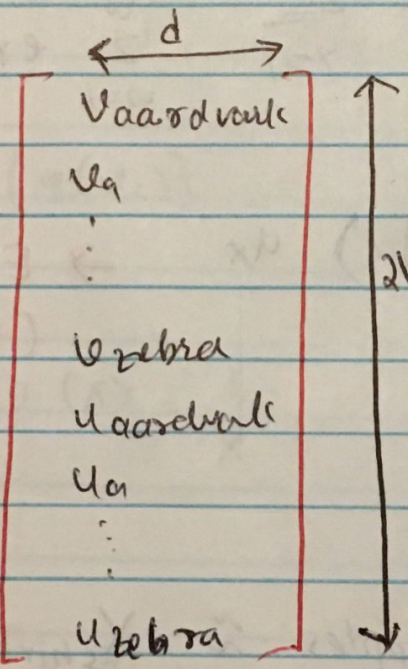
CNF

$\omega = 1$

Tody I am giving a lecture

v_{today} . v_I v_{am}

$$P(o|c) \equiv \frac{\exp(v_{today}^T \cdot v_I)}{\Sigma} \quad P(o|c) \equiv \frac{\exp(v_{am}^I \cdot v_I)}{\Sigma}$$



$\theta \in \mathbb{R}^{2dv}$

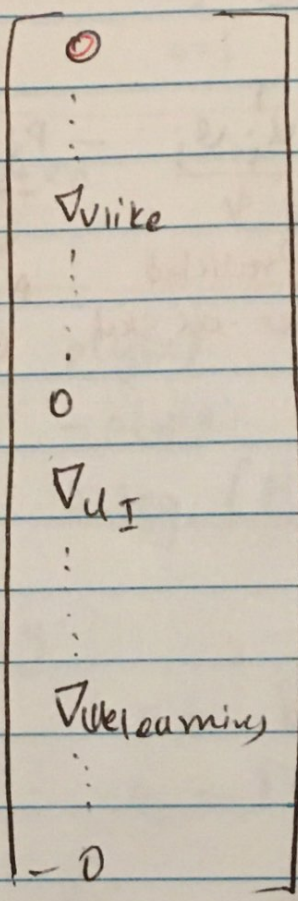
$$\theta^{new} = \theta^{old} - \nabla_\theta J(\theta)$$

→ Use SGD instead of batch gradient descent

→ Use small random numbers as initial values.

Stochastic Gradient Descent with word vectors:

$\nabla_\theta J_t(\theta) =$



⇒ very sparse.

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

I/P: $-\infty \leq x \leq +\infty$

O/P: $0 \leq \sigma(x) \leq 1$

DL CO
DL NLP

PSet 1: Skip-Gram Model
+
Negative Sampling

skip-gram:

CBOW: Avg all outside words & predict center word.

→ Continous Bag of words.

Stochastic Gradient Ascent with word vectors:

$$\nabla_\theta J_t(\theta) = \begin{bmatrix} 0 \\ \vdots \\ \nabla_{v_{like}} \\ \vdots \\ 0 \\ \nabla_{u_I} \\ \vdots \\ \nabla_{u_{learning}} \\ \vdots \\ 0 \end{bmatrix}$$

⇒ very sparse.

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

I/P: $-\infty \le x \le +\infty$

O/P: $0 \le \sigma(x) \le 1$

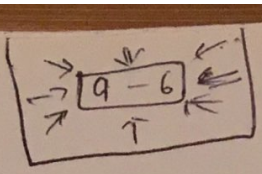
DL CO
DL NLP

PSet 1: Skip-Gram Model
+
Negative sampling

skip-gram:
CBOW: Avg all outside words & predict center word.

→ Continous Bag of words.

SVD + skip-gram.



GloVe:

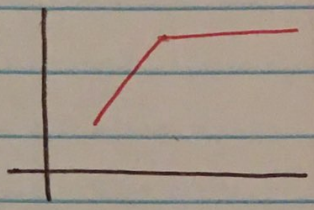
for every pair of words.

$$J(\theta) = \frac{1}{2} \sum_{i,j=1}^{w} f(P_{ij}) \left(u_i^T v_j - P_{ij} \right)^2$$

$u_i^T v_j$ → Predicted co-occ stat.

P_{ij} → Actual co-ocurance stastics

→ scalable.



Hyperparameters

1. $d \approx 300$
2. $w \approx 8$
3. time

~~Softmax:~~

Sigmoid : 2 class classification

Softmax : Multiclass classification.

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\sigma(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

$W = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix}$ $x = [x_1 \; x_1 \; x_{3}]$

[→ prep
→ Advan Normal life.
→]

$$p(y|x) = \frac{\exp(W_y x)}{\sum_{c=1}^{C} \exp(W_c \cdot x)}$$

$\frac{a}{\sum}$ $\frac{b}{\sum}$ $\frac{c}{\sum}$

$w \in \mathbb{R}^{C \times d}$ $w \in \mathbb{R}^{C \times d}$

→ How are you ?
→ prep X ←
→ Busy ?

Max $p(y|x)$
or Min $-p(y|x)$
or Min $-\log(p(y|x))$

→ How
→ kelsa incharan
→ Pari

Cross - entropy:

$p = [0, 1, 0]$
$q = [0.01, 0.9, 0.09]$

$$H(p, q) = -\sum_c p_c * \log(q_c).$$

[1 2]

$\sigma(x) = \sigma(x + c).$ [1 2)

$\sigma[1 \; 2]$ $\sigma[2 \; 3].$

a+6

w = 1,1
2,2

$\frac{\exp(3)}{\exp(3) + \exp(6)}$ $\frac{\exp(5)}{\exp(5) + \exp(10)}$

$\frac{\exp(1)}{\exp(1) + \exp(2)}$ $\frac{\exp(2)}{\exp(2) + \exp(3)}$

1 2

$\frac{1}{1+2}$ $\frac{2}{2+3}$

$\frac{1}{3}$ $\frac{2}{5}$

$2^{2+3} = 2^2 \cdot 2^3$

$2^5 = 4 \cdot 8$

$= 32$

$$\frac{\exp(x_1)}{\exp(x_1)+\exp(x_2)} \qquad \frac{\exp(x_1+c)}{\exp(x_1+c)+\exp(x_2+c)}$$

$$= \frac{\exp(x_1)\cdot\exp(c)}{\exp(c)\left[\exp(x_1)+\exp(x_2)\right]}$$

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \rightarrow (1)$$

$$\sigma(x_i+c) = \frac{e^{(x_i+c)}}{\sum_j e^{(x_j+c)}} = \frac{e^{x_i}\cdot e^{c}}{\sum_j e^{x_j}\cdot e^{c}}$$

$\sum i 3$

$$= \frac{e^{x_i}/e^{c}}{e^{c}\cdot\sum_j x_j} = \frac{e^{x_i}}{\sum_j x_j} \rightarrow (2)$$

(1) = (2) . Hence proved.

$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$

$$\frac{e^{x_i}}{\sum_j x_j}.$$

np.sum

x = [1 2]

1 x 2 .

2 x 2 .

$x = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

a = (1,2)

b = 3 a*b

RNN, CNN, LSTM.

x = np.exp(x)

x.reshape(

(2,2)

[1 2] [[],[]] (1,2)

a b

$e \quad \frac{e^{\wedge}}{e^{a} + e^{b}}$.

-1, len(x)

x(0) [1,2]

[-1, 0] = $[\frac{1}{e}, 1]$

1 2 2 4. -1 0

3 4 4 2 4 -2 0

1 3 2 4

2 4 2 4

1 2 2

3 5 5

1 3 2 4.

2 5 2 4 -1 0

-2 0

1 3 2 5 -1 / -1

2 5 2 5 -2 0. / -2

X.T

-1 -2 -(-2

0 0 -(-2

$$\sigma(x) = \frac{1}{1+e^{-x}} \cdot \frac{i}{e^x}$$

$\frac{d}{dx}(x+y)^2$

$= 2(x+y)\cdot[1+0]$

$$\frac{d}{dx}\sigma(x) = \frac{d}{dx}(1+e^{-x})^{-1}$$

$\frac{d}{dx} e^{\frac{-x}{g(x)}}$

$g(x) = -x$

$\frac{1+e^{-x}-x}{1+e^{-x}}$

$$= -1\,(1+e^{-x})^{-2}\cdot\left[0 + e^{-x}\cdot(-1)\right]$$

$\frac{d}{dx}(e^x)$

$= e^x \cdot 1$

$$= \frac{-1-e^{-x}}{(1+e^{-x})^2} = \frac{-1(1+e^{-x})}{(1+e^{-x})\,1}$$

$y = e^{-x}$

$\ln y = -x$

$$= \frac{-1}{1+e^{-x}} = -\sigma(x).$$

a^{-x}

$\frac{d}{dx} a^x$

$\frac{d}{dx} e^x$

$$\therefore \frac{d}{dx}\sigma(x) = \sigma(x)\,(1-\sigma(x)).$$

$$\frac{d}{dx}\sigma(x) = \frac{d}{dx}(1+e^{-x})^{-1}$$

$\frac{d}{dx} f(g(x))$

$\frac{df(z)}{dz}\cdot\frac{dg(x)}{dx}$

$$= -1\,(1+e^{-x})^{-2}\cdot\frac{d}{dx}(1+e^{-x})$$

$$= -1\,(1+e^{-x})^{-2}\cdot\left[0 + \frac{d}{dx}e^{-x}\right]$$

$$= -1\,(1+e^{-x})^{-2}\,\frac{d}{dx}e^{-x}\,(-e^{-x})$$

$$= \frac{e^x}{(1+e^{-x})^2} \quad \frac{e^x}{1+e^{-x}}\cdot\frac{1}{1+e^{-x}}$$

$$(1-\sigma(x))\,\sigma(x).$$

e^{-x} $f(x) = e^{g(x)}$ $g(x) = -x$

$e^{g(x)}$ $z = g(x) = -x$

$\frac{d}{dx} f(x) = \frac{d}{dz} \cdot e^z \cdot \frac{d}{dx} -x$

$= e^z \cdot (-1) = -e^{-x}$

$\frac{d}{dx} e^{-x} = -e^{-x}$

$f(z) = e^z$ $z = g(x) = -x$

$\therefore [f(g(x))]'$

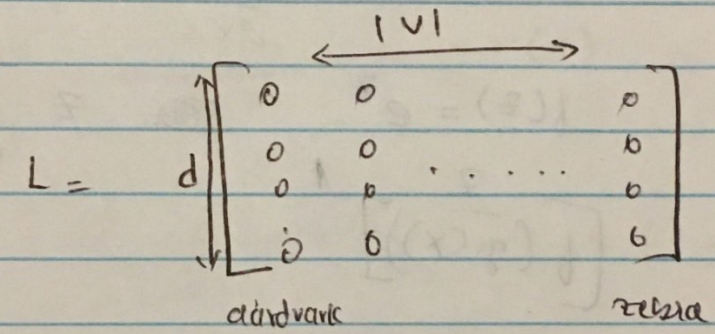
$\Rightarrow \frac{df(z)}{dz} \cdot \frac{d}{dx} \cdot g(x)$

$\Rightarrow e^z \cdot \frac{d}{dx}(-x) = -e^z = -e^{-x}$

$\{x_i, y_i\}_{i=1}^{N}$

$$\theta = \begin{bmatrix} w_{\cdot 1} \\ \vdots \\ w_{\cdot d} \end{bmatrix} = w(:) \in \mathbb{R}^{Cd}$$

$$\nabla_\theta J(\theta) = \begin{bmatrix} \nabla w_{\cdot 1} \\ \vdots \\ \nabla w_{\cdot d} \end{bmatrix} \in \mathbb{R}^{Cd}$$



... museums in Paris are amazing ...

$$X_{window} = \begin{bmatrix} x_m & x_E & x_p & x_a & x_q \end{bmatrix}^T$$

Y = location | Not location

$X \in \mathbb{R}^{5d}$ Col vector.

$0 \le x \le 1$ $\quad x \to 0 \quad \log(x) \to \infty$

$-\infty \le \log(x) \le 0 \quad x \to 1 \quad \log(x) \to 0$

$\hat{y} = \text{softmax}(\theta)$ $\quad y = [0, 1, 0]$

$\hat{y} = [0.01, 0.95, 0.04]$

$CE(y, \hat{y}) = -\sum_i y_i \log(\hat{y}_i)$

$\frac{\partial}{\partial \theta} \text{softmax}(\theta) = ?$ $\quad \sigma(\theta) \cdot [1 - \sigma(\theta)]$

$\neq -\sum_i y_i [\log(\sigma(\theta)$

$= -\sum_i y_i \log($

$CE(y, \hat{y}) = -y_x \log(\hat{y}_x)$

$= -y_x \log(\sigma(\theta_x))$

$\frac{\partial}{\partial \theta} CE(y, \hat{y}) = -\frac{\partial}{\partial \theta_x} y_x \log[\sigma(\theta_x)]$

$= -\left[y_x \left[\log(\sigma(\theta_x))\right]' + \log(\sigma(\theta_x)) \, y_x' \right]$

$= -\left[y_x \frac{1}{\sigma(\theta_x)} \cdot \sigma(\theta_x) \cdot (1 - \sigma(\theta_x)) \right]$

$= -\left[y_x 1 (1 - \sigma(\theta_x)) \right]$

$= \left[\sigma(\theta_x) - 1 \right]$ $\quad x; y_x = 1$

$X = [x_1 \; x_2] \qquad A = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$

$\frac{d}{dx}\sigma(x)$

$\frac{d}{dx}\frac{1}{1+e^{-x}}$

$(1+e^{-x})^{-1}$

$= -1(1+e^{-x})^{-2}[0+0]$

$\frac{1}{1+e^{-x}}$

$X^T . A = [x_1 \; x_2] \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$

$= \sum_i x_i A_i$

$\frac{d}{dX} X^T . A = \begin{bmatrix} \frac{\partial}{\partial x_1} \sum_i x_i A_i = a_1 \\ \\ \frac{\partial}{\partial x_2} \sum_i x_i A_i = a_2 \end{bmatrix}$

$\sigma(\theta^T x)$

$CE(y, \hat{y}) = -\sum_i y_i \cdot \log(\hat{y}_i)$

$\frac{d}{d\theta} CE(y, \hat{y}) = ? \begin{bmatrix} -\frac{\partial}{\partial \theta_1} \sum_i y_i \cdot \log(\hat{y}_i) \\ \vdots \\ \frac{\partial}{\partial \theta_n} \sum_i y_i \cdot \log(\hat{y}_i) \end{bmatrix}$

$[0 \; 1]$

$[\hat{y}_1 \; \hat{y}_2]$

$\sigma(\theta_1) \; \sigma(\theta_2)$

$= -\frac{\partial}{\partial \theta_1} \sum_i y_i \log(\sigma(\theta_i))$

$[0 \; 1]$

$[\sigma(\theta_1) \; \sigma(\theta_2)]$

$y_i \log(\sigma(\theta_i)) + y_2 \log(\sigma(\theta_2)) \quad \sigma(\theta_2)(1-\sigma(\theta_2))$

$y_i \frac{1}{\sigma(\theta_i)}$

$= \begin{bmatrix} -y_1 (1 - \sigma(\theta_1)) \\ \vdots \\ -y_n (1 - \sigma(\theta_n)) \end{bmatrix}$

$$= \begin{bmatrix} -y_i\,(1-\hat{y}_1) \\ \vdots \\ -y_n\,(1-\hat{y}_n) \end{bmatrix}$$

$\hat{y}-y$

$$= \begin{bmatrix} y_1\,(\hat{y}_1-1) \\ \vdots \\ y_n\,(\hat{y}_n-1) \end{bmatrix}$$

$\hat{y}-y.$

$Y * \hat{Y}-1$

$Y*(\hat{Y}-1)$

$= Y\hat{Y} - Y.$

$= \hat{Y} - 1$

$\hat{Y}-Y.$

when ~~$y_i=1$~~ $j=i$

$\Rightarrow [0 \;\; 0 \;\; \cdots \;\; \hat{y}_i - 1 \;\; \cdots \;\; 0]$

$\frac{\partial}{\partial \theta}(E_j = \hat{y}_j - 1$

when $j \neq i$

$\Rightarrow$ ~~[illegible]~~ $Y*(\hat{Y}-1)$

when $i=k$; $\frac{\partial}{\partial \theta_k}(E(y,\hat{y}) = \hat{y}_i - 1$

$i \neq k$; $\frac{\partial}{\partial \theta_i}(E(y,\hat{y}) = 0$

$\ln \theta$

$$\hat{y} = \text{softmax}(\theta)$$

$$CE = -\sum_i y_i \cdot \log(\hat{y}_i).$$

$$\frac{\partial}{\partial \theta} CE = ?$$

$$\frac{\partial}{\partial \theta} CE(y, \hat{y}) = \hat{y} = \text{softmax}(\theta)$$

$$\hat{y}_k = \frac{e^{\theta_k}}{\sum_{i=1}^{d} e^{\theta_i}}$$

$$\hat{y} = \text{softmax}(\theta)$$

$$CE = -\sum_i y_i \log(\hat{y}_i)$$

$$\frac{\partial}{\partial \theta} CE(y, \hat{y}) = ?$$

$$\frac{\partial}{\partial \theta} \left[-\sum_i y_i \log(\hat{y}_i) \right] = ?$$

$$= \frac{\partial}{\partial \theta} \left[-\sum_i y_i \cdot \log\left(\text{softmax}(\theta_i)\right) \right] = ?$$

$$\Rightarrow \frac{\partial}{\partial \theta} \left[-\sum_i y_i \cdot \log\left(\frac{e^{\theta_i}}{\sum_{j=1}^{d} e^{\theta_j}} \right) \right] = ?$$

$\log(\text{cat})$

$$= \begin{bmatrix} \frac{d}{d\theta_1}\left[-\sum_i y_i \cdot \log\left(\frac{e^{\theta_i}}{\sum_{j=1}^{d} e^{\theta_j}}\right)\right] \\ \vdots \\ \frac{d}{d\theta_d}\left[-\sum_i y_i \cdot \log\left(\frac{e^{\theta_i}}{\sum_{j=1}^{d} e^{\theta_j}}\right)\right] \end{bmatrix}$$

$$= \frac{d}{d\theta_1}\left[-\sum_i y_i \cdot \log\left[\frac{e^{\theta_i}}{\sum_{j=1}^{d} e^{\theta_j}}\right]\right]$$

$$\log \frac{e^{\theta_i}}{\sum_{j=1}^{d} e^{\theta_j}} = \left[\log \exp(\theta_i) - \log \sum_{j=1}^{d} \exp(\theta_j)\right] \qquad \frac{d}{dx} e^x = e^x$$

$$= \theta_i - \log \sum_{j=1}^{d} \exp(\theta_j) \qquad \left[1 - \log\right.$$

$$\frac{d}{d\theta_1}\left(\theta_i - \log \sum_{j=1}^{d} \exp(\theta_j)\right) \qquad \frac{d}{d\theta_1}\exp(\theta_j)$$

$$= 1 - \frac{\exp(\theta_1)}{\sum_{j=1}^{d} \exp(\theta_j)} \qquad 1 - \text{Softmax}(\theta_1)$$

$$\therefore \frac{d}{d\theta_x} \log\left(\text{Softmax}(\theta_x)\right) = 1 - s(\theta$$

$\frac{\partial}{\partial o_1}\left(o_2 - \log \sum_{i=1}^{d} \exp(o_i)\right)$

$= 0 - 0$

$\frac{\partial}{\partial o_1}\left[-\sum_i y_i \cdot \log\left[\text{Softmax}(o_i)\right]\right] = ?$

$y_1 \cdot \log(s(o_1)) + y_2 \cdot \log(s(o_2))$

$\Rightarrow \frac{\partial}{\partial o_1}\left[-y_1 \cdot \log\left(\text{Softmax}(o_1)\right)\right] = ?$

$\Rightarrow -\left[y \cdot (1 - \text{Softmax}(o_1))\right] = -\left[y_1 \cdot (1 - \hat{y}_1)\right]$

$$= \begin{bmatrix} y_1 (\hat{y}_1 - 1) \\ y_2 (\hat{y}_2 - 1) \\ \vdots \\ y_d (\hat{y}_d - 1) \end{bmatrix}$$

when $i = k$

$\frac{\partial}{\partial o_i}(E(y_i, \hat{y}_i)) = y_i (\hat{y}_i - 1)$ $\quad \frac{dy}{dx} = 1$

$= 1 \cdot (\hat{y}_i - 1)$

$= \hat{y}_i - 1$

$i \neq k$ $\quad X + Y + Z$

$\frac{\partial}{\partial o_i} =$

$$\hat{y} = \text{softmax}(\theta)$$

$$CE(y, \hat{y}) = -\sum_i y_i \cdot \log(\hat{y}_i),$$

$$\frac{\partial}{\partial \theta} CE(y, \hat{y}) = ?$$

$$\frac{\partial}{\partial \theta} CE(y, \hat{y}) = \frac{\partial}{\partial \theta} \cdot -\sum_i y_i \cdot \log(\text{softmax}(\theta_i))$$

chose k, s.t. $y_k \neq 0 \Rightarrow$

$$= \frac{\partial}{\partial \theta} -\sum_i y_i \cdot \frac{\partial}{\partial \theta} \cdot -y_k \cdot \log(\text{softmax}(\theta_i))$$

$$= \begin{bmatrix} \frac{\partial}{\partial \theta_1} -y_k \cdot \log(\text{softmax}(\theta_k)) \\ \vdots \\ \frac{\partial}{\partial \theta_d} -y_k \cdot \log(\text{softmax}(\theta_k)) \end{bmatrix}$$

when $i = k$; $\quad \frac{\partial}{\partial \theta_i} = ?$

$$\Rightarrow -\frac{\partial}{\partial \theta_k} \left[+ y_k \log(\text{softmax}(\theta_k)) \right]$$

$$= -\left[y_k \cdot (1 - \text{softmax}(\theta_k)) \right] \cdot = -\left[y_k (1 - \hat{y}_k) \right]$$

$$= y_k (\hat{y}_k - 1) = (\hat{y}_k - 1) \hat{y}_k - 1$$

when $i \neq k$

$$-\frac{\partial}{\partial \theta_i} \left[y_k \cdot \log(\text{softmax}(\theta_k)) \right]$$

$=$

$\log(xy) = \log x + \log y$

$$\frac{d}{d\theta_k} \log \left[\frac{e^{\theta_i}}{\sum_{j=1}^{d} e^{\theta_j}} \right] =$$

$$= \frac{d}{d\theta_k} \left[\theta_i - \log \sum_{j=1}^{d} e^{\theta_j} \right]$$

$$= \frac{d}{d\theta_k} 0 - \frac{1}{\sum_{j=1}^{d} e^{\theta_j}} \cdot \frac{d}{d\theta_k} \cdot \left(\sum_{j=1}^{d} e^{\theta_j} \right)$$

$$= - \frac{e^{\theta_k}}{\sum_{j=1}^{d} \theta_j} = - \text{Softmax}(\theta_k)$$

$$= - \frac{d}{d\theta_i} \left[y_k \cdot f(\theta_i) \right]$$

$$= - \left[y_k \cdot (-S(\theta_i)) + S(\theta_k) \cdot 0 \right]$$

$$= y_k \cdot \text{Softmax}(\theta_i)$$

$$= y_k \cdot \hat{y}_i = 1 \cdot \hat{y}_i = y_i$$

$$\therefore \frac{\partial}{\partial \theta_j} CE(y, \hat{y}) = \hat{y}_j - 1 \quad ; \text{ when } j == k$$

$$= \hat{y}_j \quad ; \text{ when } j \neq k$$

$$= [\hat{y}_j, \hat{y}_j, \ldots \hat{y}_j - 1, \hat{y}_j, \ldots \hat{y}_j]$$

$$\therefore \frac{\partial}{\partial \theta_j} CE(y, \hat{y}) = \hat{y}_j - y_j$$

$$= \hat{y} - y$$

$\{x_i, y_i\}_{i=1}^{M}$

x_i → words | sentences | vectors

y_i → sentiment | words | sentences.

$$P(y|x) = \frac{e^{w_y \cdot x}}{\sum_{c=1}^{C} e^{w_c \cdot x}}$$

$\{x_i, y_i\}_{i=1}^{N}$ Actual | class.

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} -\log \left(\frac{e^{f_{y_i}}}{\sum_{c=1}^{C} e^{f_c}} \right)$$

for each $\{x_i, y_i\}$,

$$f_y = f_y(x) = W_y \cdot x = \sum_{j=1}^{d} W_{y_j} \cdot x_j$$

$f = W \cdot x$.

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{e^{f_{y_i}}}{\sum_{c=1}^{C} e^{f_c}}\right)$$

$$f_y = W_y \cdot x$$

$\{x_i, y_i\}$; $f_y' = W_y \cdot x = \sum_{j=1}^{d} W_{y_j} \cdot x_j$

$$J(\theta) \quad \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{e^{f_{y_i}}}{\sum_{k=1}^{C} e^{f_k}}\right)$$

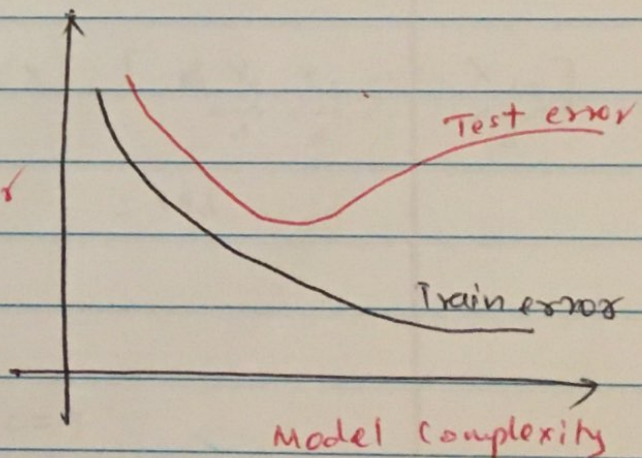
$$\begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \qquad [x_1 \; x_2]$$

→ Why cross entropy over MSOS error.

→ Squared error doesn't work well with neural networks. ∴ use cross-entropy.

Regularized: (prevents over-fitting)

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{e^{f_{y_i}}}{\sum_{c=1}^{C} e^{f_c}}\right) + \lambda \cdot \sum_{k=1}^{d} \theta_k^2$$



$$\theta = \begin{bmatrix} w_{\cdot 1} \\ \vdots \\ w_{\cdot d} \end{bmatrix} \qquad \theta \in \mathbb{R}^{cd}$$

$$\nabla_\theta f(\theta) = \begin{bmatrix} \nabla_{w_{\cdot 1}} \\ \vdots \\ \nabla_{w_{\cdot d}} \end{bmatrix} \qquad \nabla_\theta f(\theta) \in \mathbb{R}^{cd}$$

Deep learning:

Learn both w & x_i.

$$\nabla_\theta f(\theta) = \begin{bmatrix} \nabla_{w_{\cdot 1}} \\ \vdots \\ \nabla_{w_{\cdot d}} \\ \nabla_{x_{\text{aardvark}}} \\ \vdots \\ \nabla_{x_{\text{zebra}}} \end{bmatrix} \qquad \nabla_\theta f(\theta) \in \mathbb{R}^{cd+Vd}$$

Large ↑ $cd+Vd$ ↓ over-fitting danger

word vec = word embeddings = word representations

$$L = d \begin{bmatrix} \circ & \circ & & \circ \\ \circ & \circ & & \circ \\ \circ & \circ & \cdots & \circ \\ \circ & \circ & & \circ \\ \circ & \circ & & \circ \end{bmatrix}$$

$|V|$ columns: $x_{aardvark}$ … x_{zebra}

ML, PGM, NLP

→ Rarely training single words work.
∴ use window classification.
EX: 'in' says something about location.

museums in Paris are amazing

$x_{window} = [\ x_{mu} \quad x_{in} \quad x_{pa} \quad x_{ar} \quad x_{am}\ ]^T$

$x_{window} = x \in \mathbb{R}^{5d}$, col vector!

$x\ [\ x_1 \ x_2 \ \ldots \ x_{10}\ ]$ (each of length d)

$= 10d$

→ dimensionality

→ variables

→ der. w.r.t f_c when $c = y$

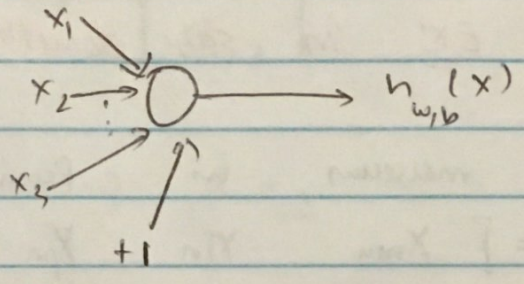
f_c when $c \neq y$

→ Matrix notations

$$x_{window} = [x_{mu} \quad x_{in} \quad x_{pa} \quad x_{aa} \quad x_{am}]$$

$$\nabla_{window} = \begin{bmatrix} \nabla_{x_{mu}} \\ \nabla_{x_{in}} \\ \vdots \\ \nabla_{x_{an}} \end{bmatrix} \in \mathbb{R}^{5d}$$

Neural Nets ⇒

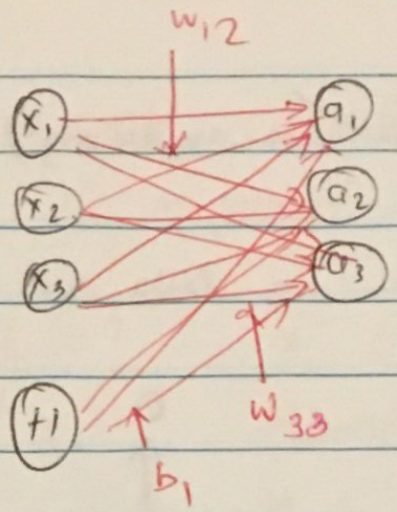


$$h_{w,b}(x) = f(w^T x + b).$$

$$f(z) = \frac{1}{1+e^{-z}}$$

$W \in \mathbb{R}^{a*x}$;

a : no of activations
x : no of dimensions



$$a_1 = f(w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + b_1)$$

$$a_2 = f(w_{21}x_1 + w_{22}x_2 + w_{23}x_3 + b_2)$$

etc.

$$\underset{W}{\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix}} \quad \underset{X}{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}} \quad \underset{B}{\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}}$$

$3 \times 3 \qquad 3 \times 1 \qquad 3 \times 1$

$a = 3 \times 1$

$$z = Wx + B.$$

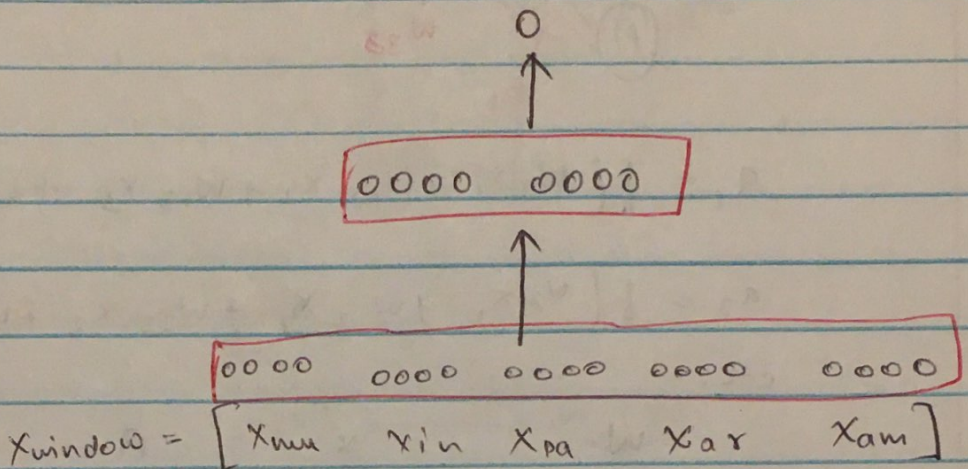
$$z = Wx + b. \qquad z = [z_1 \; z_2 \; z_3]$$

$$a = f(z)$$

$$f(z) = f([z_1 \; z_2 \; z_3]) = [f(z_1), f(z_2), f(z_3)]$$

s = score("Paris is amazing")

s = score(museums in Paris are amazing)



$s = U^T a$. $1 \times 1 = 1 \times 8 \; 8 \times 1$

$a = f(z)$

$z = Wx + b$.

$\therefore \; s = U^T f(Wx + b)$.

$x \in \mathbb{R}^{20 \times 1}$

$W \in \mathbb{R}^{8 \times 20}$

$U \in \mathbb{R}^{8 \times 1}$

$\frac{\partial}{\partial (vec)}$ (scalar) = vector

Task: ~~Complete~~ Complete a few questions in 1st assignment.

$\frac{\partial J}{\partial x}$ = ? σ @ hidden

softmax output.

y == 1-hot

cross entropy.

$h = \text{sigmoid}(xW_1 + b_1)$ $\hat{y} = \text{softmax}(hW_2 + b_2)$

→ what's objective function? Reg / not reg?

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} -\log\left(\frac{e^{f_{y_i}}}{\sum_{c=1}^{C} e^{f_c}}\right) + \lambda \sum_{k=1}^{d} \theta_k^2$$

$\frac{\partial J(\theta)}{\partial x}$ = ? $f_{y_i} = W_{y_i} \cdot x = \sum_{j=1}^{d} (W_{y_i})_j x_j$

$f_c = W_c \cdot x = \sum_{j=1}^{d} (W_c)_j x_j$

$$\therefore J(\theta) = \frac{1}{N} \sum_{i=1}^{N} -\log\left(\frac{\exp\left(\sum_{j=1}^{d} (W_{y_i})_j \cdot x_j\right)}{\sum_{c=1}^{C} \exp\left(\sum_{j=1}^{d} (W_c)_j \cdot x_j\right)}\right) + \lambda \sum_{k=1}^{d} \theta_k^2$$

$$\frac{d}{dx} x^T a = \frac{d}{dx} \sum_i x_i a_i = \frac{d}{dx_1} \sum$$

σ @ hidden $\quad \frac{d}{dx_1} \sum_i x_i a_i$

S @ output. $\quad \sum_i \frac{d}{dx_1} x_i a_i$

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} \lambda \sum_{k=1}^{d} \theta_k^2 - \log \left(\frac{\exp\left(\sum_{j=1}^{d} (w_{y_i})_j \cdot x_j \right)}{\sum_{c=1}^{C} \exp\left(\sum_{z=1}^{d} (w_c)_z \cdot x_z \right)} \right)$$

$$\frac{\partial J(\theta)}{\partial x} = \left[\frac{\partial}{\partial x_1} J(\theta) \quad \frac{\partial}{\partial x_2} J(\theta) \quad \cdots \quad \frac{\partial}{\partial x_d} J(\theta) \right]$$

$$\frac{\partial}{\partial x_1} J(\theta) = \frac{1}{N} \sum_{i=1}^{N} - \frac{\partial}{\partial x_1} \cdot \left(\log \left(\frac{\exp\left(\sum_{j=1}^{d} (w_{y_i})_j \cdot x_j \right)}{\sum_{c=1}^{C} \exp\left(\sum_{z=1}^{d} (w_c)_z \cdot x_z \right)} \right) \right) \quad ②$$

$\log\left(\frac{a}{b}\right) = \log a - \log b.$

$$= \sum_{j=1}^{d} (w_{y_i})_j \, x_j - \sum_{z=1}^{d} (w_c)_z \cdot x_z$$

$$= \frac{\partial}{\partial x_1} \sum_{j=1}^{d} (w_{y_i})_j \cdot x_j - \frac{\partial}{\partial x_1} \sum_{z=1}^{d} (w_c)_z \cdot x_z .$$

$$= (w_{y_i})_1 - (w_c)_1$$

$\frac{\partial}{\partial x} x^2$

$$\frac{d}{dx} e^{-x} = -e^{-x}. \qquad \frac{d}{dz} e^{z} \quad ; \quad \frac{d}{dx} -x$$

$$\log\left(\frac{\exp\left(\sum_{j=1}^{d} (w_{y_i})_j \cdot x_j \right)}{\sum_{c=1}^{C} \exp\left(\sum_{z=1}^{d} (w_c)_z \cdot x_z \right)} \right)$$

$$a = \sum_{j=1}^{d} (w_{y_i})_j \cdot x_j - \log \sum_{c=1}^{C} \exp\left(\sum_{z=1}^{d} (w_c)_z \cdot x_z \right)$$

$$\frac{da}{dx_1} = (w_{y_i})_1 - \frac{1}{\sum_{c=1}^{C} \exp\left(\sum_{z=1}^{d} (w_c)_z \cdot x_z \right)} \cdot \frac{d}{dx_1} \left(\sum_{c=1}^{C} \exp\left(\sum_{z=1}^{d} (w_c)_z \cdot x_z \right) \right)$$

$$= (w_{y_i})_1 - \frac{1}{\sum_{c=1}^{C} \exp\left(\sum_{z=1}^{d} (w_c)_z \cdot x_z \right)} \cdot \sum_{c=1}^{C} \frac{d}{dx_1} \exp\left(\sum_{z=1}^{d} (w_c)_z \cdot x_z \right)$$

$$= (w_{y_i})_1 - \frac{\sum_{c=1}^{C} \exp\left(\sum_{z=1}^{d} (w_c)_z \cdot x_z \right) \cdot (w_c)_1}{\sum_{c=1}^{C} \exp\left(\sum_{z=1}^{d} (w_c)_z \cdot x_z \right)}$$

$$= w_{y_i} - \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{c=1}^{C} \exp\left(\sum_{z=1}^{d} (w_c)_z \cdot x_z \right) \cdot (w_c)_1}{\sum_{c=1}^{C} \exp\left(\sum_{z=1}^{d} (w_c)_z \cdot x_z \right)}$$

$f(g(x))$

$$J(\theta) = CE(y, \hat{y}) = -\sum_{i=1}^{N} y_i \cdot \log(\hat{y}_i)$$

$$\frac{dJ(\theta)}{dx} = ?$$

$$\left[f\left(\underbrace{g(x)}_{z}\right)\right]' = \frac{d}{dz} f(z) \cdot \frac{d}{dx} g(x) \cdot$$

$h = \text{sigmoid}(x W_1 + b_1)$ $\quad \hat{y} = \text{softmax}(h W_2 + b_2)$

$$J(\theta) \rightarrow \theta \rightarrow h \rightarrow x.$$

$$\therefore \frac{dJ(\theta)}{dx} = \frac{dJ(\theta)}{d\theta} \cdot \frac{d}{dh} \hat{y} \cdot \frac{d}{dx} h.$$

$$= \frac{dJ}{d\theta} \cdot \frac{d\theta}{dh} \cdot \frac{dh}{dx}$$

$$J = -\sum_{i=1}^{N} y_i \cdot \log(\text{softmax}(hw_2 + b_2))$$

$$\frac{dJ}{d\theta}$$

$$J = -\sum_{i=1}^{N} y_i \log(\text{softmax}(\theta))$$

$\theta = h w_2 + b_2$ $\quad \frac{d\theta}{dh}$

$h = \sigma(x w_1 + b_1)$ $\quad \frac{dh}{d(\cdot)}$

Let $J = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$

$= -\sum_{i=1}^{N} y_i \log(\text{Softmax}(\theta_i))$

$\therefore J(\theta)$.

$\theta = h w_2 + b_2$

$\therefore \theta(h)$.

$h = \text{sigmoid}(\underbrace{x w_1 + b_1}_{z})$ $\quad h = \sigma(z)$.

$\therefore h(z)$.

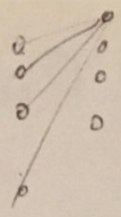
$\therefore \frac{\partial J}{\partial x} = \frac{\partial J(\theta)}{\partial \theta} \cdot \frac{\partial \theta(h)}{\partial h} \cdot \frac{\partial h(x)}{\partial x}$.

$= (\hat{Y} - Y) \cdot w_2 \cdot \frac{d}{dx} \cdot \sigma(x w_1 + b_1)$

$= (\hat{Y} - Y) \cdot w_2 \cdot \frac{d}{dx} \cdot \left[1 + e^{-(x w_1 + b_1)}\right]^{-1}$

$= (\hat{Y} - Y) \cdot w_2 \cdot -1 \left(1 + e^{-(x w_1 + b_1)}\right)^{-2} \cdot \frac{d}{dx} e^{-(x w_1 + b_1)}$

$= \frac{-1}{1 + e^{-(x w_1 + b_1)^2}} \cdot (-1) \cdot e^{-(x w_1 + b_1)}$.



row = # h
col = x dim

$D_h \times D_x$ = $h \times x$

$x \times h$

$1 \times h$

$z = x W_1 + b_1$,

$\therefore \frac{\partial z}{\partial x}$.

$$\frac{\partial \Sigma}{\partial x} = \frac{\partial \Sigma(\theta)}{\partial \theta} \cdot \frac{\partial \theta(h)}{\partial h} \cdot \frac{\partial h(z)}{\partial z} \cdot \frac{\partial z(x)}{\partial x} \cdot$$

$$= \left[\hat{y} - y\right] \cdot * W_2 \odot \sigma(x W_1 + b_1) \odot (1 - \sigma(x W_1 + b_1)) * W_1$$

$\Rightarrow$ $1 * C * C * h$.

$C * h \quad * \quad h *$

$$(\hat{y} - y) \quad \frac{d}{dz} sig(\overbrace{x W_1^T + b_1}^{z})$$

$D_h \times D_x$

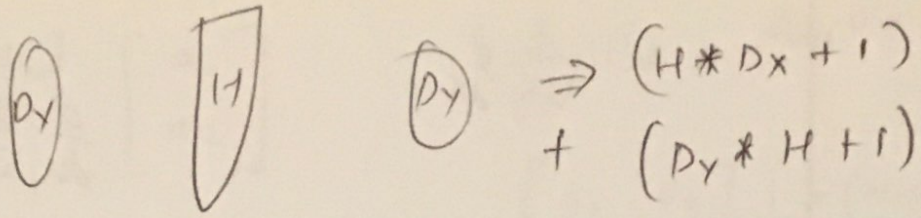
$\downarrow$

$1 * D_x \cdot D_x * D_h$

$1 * D_h$.

$1 * D_h$.

$$\frac{d}{dt} \sigma(z) = \sigma(z) * 1 - \sigma(z).$$



$$(D_x + 1) * H + (H + 1) * D_y$$

$$\sigma(x) \quad \frac{1}{1 + \exp(-x)}.$$

$$\sigma(x) \qquad \sigma'(x) = \sigma(x) \cdot 1 - \sigma(x).$$

$$f = \sigma(x).$$

$f \cdot 1 - f$ $\quad$ 20 × 1

$$z = Wx + b \qquad x \in \mathbb{R}^{20\times 1}$$

$$a = f(z) \qquad W \in \mathbb{R}^{8\times 20}$$

$$\text{score} = U^T a \qquad U \in \mathbb{R}^{1\times 8}$$

$$s = U^T f(Wx + b)$$ → Max-Margin

→ Minimize J

→ $J = \max(0, 1 - s + s_c)$

$\frac{\partial}{\partial x}$

$\frac{\partial}{\partial u_i} \sum v_i x_i$

$x_1 \; x \ldots x_n$

$s = v^T f(wx + b)$

$s_c = v^T \cdot f(w x_c + b)$.

$J = \max(0, 1 - s + s_c)$.

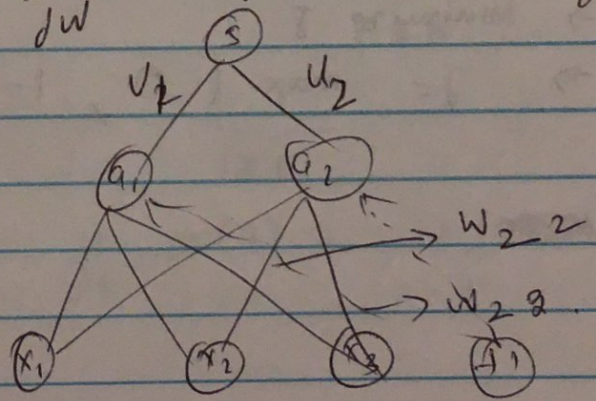
Training with Back-prop:

Compute derivatives of s & s_c w.r.t.

v, w, x, b.

(1) $\frac{ds}{dv} = \frac{\partial}{\partial v} v^T \cdot f(wx + b) = f(wx+b) = a$

(2) $\frac{ds}{dw} = \frac{\partial}{\partial w} \cdot v^T \cdot f(wx+b) = \frac{\partial}{\partial w} v^T \cdot a$.



a_i depends ~~only~~ on w_i

&

$a_i \rightarrow w_i$

$u_i \rightarrow a_i \rightarrow w_i$

$[w_1, v_2] \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$

$\frac{d}{dw} v^T a = \delta w_{11} v^T a$

$[v_1 \quad v_2] \begin{bmatrix} \sigma(w_{11}x_1 + w_{12}x_2 + w_{13}x_3) \\ \sigma(w_{21}x_1) + w_{12}x_2 + w_{13}x_3 \end{bmatrix}$

$\frac{d}{dw_{ij}} (v^T \cdot a) \longrightarrow$

$\frac{d}{dw} v^T a = \begin{bmatrix} \delta_{w_{11}}(u^T a) & \delta_{w_{12}}(v^T a) & \delta_{w_{13}}(u^T a) \\ \vdots & & \\ \delta_{w_{21}}(u^T a) & \delta_{w_{22}}(u^T a) & \delta_{33}(v^T a) \end{bmatrix}$

$\delta_{w_{11}}(u^T a) \rightarrow \delta_{w_{11}} u_1 a_1$

$\delta_{w_{ij}}(v^T a) \rightarrow \delta_{w_{ij}} v_i a_i$

$\therefore \frac{d}{dw_{ij}} v^T a \rightarrow \frac{d}{dw_{ij}} v_i a_i .$

$= v_i \frac{d}{dw_{ij}} a_i = v_i \cdot \frac{d}{dw_{ij}} \sigma(z)$

$z_i = \sum_j w_{ij} \cdot x_j + b$

$\therefore z(w_{ij}) .$

$\delta = v_i \cdot \frac{d}{dz_i} f(z_i) \cdot \frac{d}{dw_{ij}} f(w_{ij})$

$\frac{d}{dw_{ij}} (v^T a) = v_i \sigma'(z_i) x_j$

$$\frac{\partial}{\partial w_{ij}} s = \delta_i \cdot x_j$$

$$\begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \qquad \begin{bmatrix} \delta_1 x_1 & \delta_1 x_2 & \delta_1 x_3 \\ \delta_2 x_1 & \delta_2 x_2 & \delta_2 x_3 \end{bmatrix}$$

$$\therefore \frac{\partial}{\partial w} s = \delta x^T \quad ; \quad \frac{\partial}{\partial w} s_c = \delta_c x_c^T$$

(error msg)

$$\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}$$

$1 \times 3 \times 2 \times 1 \rightarrow 3 \times 2$ $\qquad 3 \times 1 \times 1 \times 2 = 3 \times 2$.

$\Rightarrow$ Outer-product.

$$J = \max(0, 1 - s + s_c)$$

$$\therefore \frac{\partial J}{\partial w} = \frac{\partial J(s)}{\partial w}$$

$$\therefore \frac{\partial}{\partial w} J = -\delta x^T + \delta_c x_c^T$$

①

$$\frac{d}{db} s = \frac{d}{db} v^T a = \frac{d}{db} v^T \cdot f(Wx+b)$$

$$\frac{d}{db_i} \quad \frac{d}{db} s = \begin{bmatrix} \frac{d}{db_1} s & \frac{d}{db_2} s \end{bmatrix}$$

$$\therefore \frac{d}{db_1} s = v^T \cdot f(Wx+b)$$

$$= \sum_i v_i \cdot f(w_i x + b_i)$$

$$\therefore s(a_i) \quad a(z) \cdot \quad z(b)$$

$$\frac{d}{db} \quad \therefore \frac{d}{db_1} c = \frac{d}{dba_i} s(a_i) \cdot \frac{d}{dbz} a(z) \cdot \frac{d}{db} z(b) \cdot$$

$$= \frac{d}{da} v_i^T a \cdot \frac{d}{db_z} f(z) \cdot \frac{d}{db_1} w_i x + b_i$$

$$= v_i \sigma'(z_i) \cdot (1)$$

$$= \delta_i$$

$$\frac{ds}{dx} = \begin{bmatrix} \frac{ds}{dx_1} & \frac{ds}{dx_2} & \frac{ds}{dx_3} \end{bmatrix}$$

$$[u_1 \; u_2] \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}.$$

$$s = v^T \cdot f(Wx + b).$$

$$= v^T a. \quad ; \quad a = f(z) \quad ; \quad z = Wx + b$$

$$s(a) \quad ; \quad a(z) \quad ; \quad z(x).$$

$$\therefore \frac{ds}{dx_1} = \frac{ds(a)}{da} \cdot \frac{da(z)}{dz} \cdot \frac{dz(x)}{dx_1}$$

$$= \quad s = \sum_i v_i a_i \quad ; \quad a_i = f(z_i) \; ; \; z_i = w_i x + b_i$$

$$s(a_i) \qquad a_i(z_i) \qquad z_i(x).$$

$$\begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} (x_1, x_2)$$

$$= v_i \cdot \sigma'(z_i) \cdot \begin{bmatrix} \frac{d z}{dx_1} \end{bmatrix}$$

$$\frac{d}{dx_1} \qquad \frac{d}{dx} s = \begin{bmatrix} \frac{d}{dx_1} s & \frac{d}{dx_2} s \end{bmatrix}$$

$$\frac{d}{dx_1} v^T a$$

~~$a = f(W)$~~

$$\frac{d}{dx_1} v^T f \qquad a = [a_1 \; a_2]$$

$$= \begin{bmatrix} f(w_1 x + b_1) & f(w_2 x + b_2) \end{bmatrix}$$

$$= \begin{bmatrix} f(z_1) & f(z_2) \end{bmatrix}$$

$\frac{\partial}{\partial x} f(x) =$

$z_i = \sum_j w_{ij} x_j$ $\quad \therefore [w_1^T s \; w_2^T s]$ $[w_1^T \; w_2^T] s.$

$z_1 = w_1 x + b_1$ $\qquad z_2 = w_2 x + b_2$. $\qquad w^T s.$

$\therefore$ ~~$s(a)$~~; $s(a_1, a_2)$; $a_1(z_1)$; $a_2(z_2)$

$z_1(x)$; $z_2(x)$.

$\therefore \frac{\partial}{\partial x_1} s(a_1, a_2) = \frac{\partial}{\partial x_1} s(a_1) + \frac{\partial}{\partial x_1} s(a_2)$.

$\Rightarrow$ If $\frac{\partial}{\partial x_1} s(a_1) = \frac{\partial}{\partial a_i} s(a_i) \cdot \frac{\partial}{\partial z_1} z_1(x)$

$= \sigma'(z_1)$.

$s(a) = \sum_i v_i a_i$

$v^T \cdot a$

$v_i \cdot a_i$ $\quad \frac{\partial}{\partial x_1} s = \frac{\partial}{\partial a_1} s(a_1) \cdot \frac{\partial}{\partial z_1} a_1(z_1) \cdot \frac{\partial}{\partial x} z_1(x)$.

$= v_1 \sigma'(z_1) \cdot w_1$

$\Rightarrow \frac{\partial}{\partial x_1} s = \delta_1 w_1$; $\frac{\partial}{\partial x_2} s = \delta_2 w_2$.

$= \delta_1 w_1 + \delta_2 w_2$.

$= \sum_{i=1}^{2} \delta_i w_i$

$$J = \max(0, 1 - s + s_c)$$

↳ Minimize J.

$$\frac{d}{d(x,y)} f(x,y) = \frac{d}{dx} f(x,y) + \frac{d}{dy} f(x,y)$$

$$\frac{dJ}{d(s,s_c)} = \frac{dJ(s,s_c)}{ds} + \frac{dJ(s,s_c)}{ds_c}$$

$$s = a. \quad = \frac{d}{ds}(1 - s + s_c) + \frac{d}{ds_c}(1 - s + s_c)$$

$$=$$

$$J(u, w, x, b) = \max(0, 1 - s + s_c)$$

$$\frac{dJ}{d(u,w,x,b)} = \frac{d(1 - s + s_c)}{d(u,w,x,b)}$$

$$J = 1 - s + s_c$$

$$\therefore J(s, s_c)$$

$$s = u^T a$$

$$\therefore s(a)$$

$$a = f(wx + b) = f(z)$$

$$\therefore a(z).$$

$$z = wx + b$$

$$\therefore z(x).$$

$$\left[f(g(x))\right]' = \frac{\partial f(z)}{\partial z} \cdot \frac{\partial g(x)}{\partial x}$$

$s \to z \to x$; $s \to x$

$s_c \to z \to x$; $s_c \to x$

$$\frac{\partial (1 - s + s_c)}{\partial (x)} \cdot \qquad \frac{\partial}{\partial x} J(s, s_c) \cdot$$

$$= \frac{\partial}{\partial x} J(s, s_c) + \frac{\partial}{\partial x} J(s, s_c) \cdot$$

$$= \frac{\partial}{\partial x} J(s, s_c)$$

$$= \qquad J = \max(0, \underbrace{1 - s + s_c}_{z})$$

$$\therefore J(z).$$

$$z = 1 - s + s_c.$$

$$\therefore z(s, s_c)$$

$s(x)$

$$\frac{\partial J(z)}{\partial z} \cdot \left[\frac{\partial z(s)}{\partial s} \cdot \frac{\partial s(x)}{\partial x} + \frac{\partial z(s_c)}{\partial s_c}\right]$$

$$= 1 \cdot \left[\frac{\partial z(s)}{\partial s} \cdot \frac{\partial s(x)}{\partial x} + \frac{\partial z(s_c)}{\partial s_c} + \frac{\partial s_c(x)}{\partial x}\right]$$

$$\frac{d}{dx} f(x) + g(x) .$$

$$\frac{d}{dx} f(x) + \frac{d}{dx} g(x) .$$

$J = \max(0, 1 - s + s_c)$

Minimize J.

$\nabla J(U, W, X, b) = ?$

① $$\frac{d}{d(x,y)} f(x,y) = \frac{d}{dx} f(x,y) + \frac{d}{dy} f(x,y) .$$

② $$\frac{d}{dx} f(\underbrace{g(x)}_{z}) = \frac{d}{dz} f(z) \cdot \frac{d}{dx} g(x) .$$

$$\frac{dJ}{dX} = \frac{d}{dX} (1 - s + s_c) .$$

$$= 0 - \frac{d}{dx} s + \frac{d}{dx} s_c .$$

$s = U^T a \quad \therefore s(a)$

$a = f(z) \; ; z = W^T x + b \quad \therefore a(z)$

$z = W^T x + b \quad \therefore z(x) .$

$$\therefore \frac{d}{dx} s = \frac{d}{da} s(a) \cdot \frac{d}{dz} a(z) \cdot \frac{d}{dx} z(x) .$$

$$= U \cdot \sigma'(z) \cdot W^T .$$

$$\boxed{\frac{d}{dx} J = 0 - \frac{d}{d(U,W,X,b)} s + \frac{d}{d(U,W,X,b)} s_c}$$

$$\frac{\partial}{\partial(U,W,X,b)} s = \frac{\partial}{\partial U} s + \frac{\partial}{\partial W} s + \frac{\partial}{\partial X} s + \frac{\partial}{\partial b} s$$

$$\frac{\partial}{\partial(U,W,X,b)} s_c = \frac{\partial}{\partial U} s_c + \frac{\partial}{\partial W} s_c + \frac{\partial}{\partial X} s_c + \frac{\partial}{\partial b} s_c$$

$$\frac{\partial}{\partial(U,W,X,b)} J(U,W,X,b) = -\frac{\partial}{\partial(U,W,X,b)} s + \frac{\partial}{\partial(U,W,X,b)} s_c$$

$$s = U^T a = U^T f(z) = U^T f(Wx+b).$$

$$\frac{\partial}{\partial U} s = \frac{\partial}{\partial U} U^T a = \frac{\partial}{\partial U} \sum_i U_i a_i = [a_1, \dots a_n] = a$$

$$\therefore \boxed{\frac{\partial}{\partial U} s = a} \rightarrow ①$$

$2\times3 \times 3\times1$

2×1

$$\frac{\partial}{\partial W} s = \frac{\partial}{\partial a} s(a) \cdot \frac{\partial}{\partial z} f(z) \cdot \frac{\partial}{\partial W} (Wx+b)$$

2×2.

2×2.

$$= U \cdot f'(z) \cdot x$$

2×1 . 2×1 . 1×3

2×1 . 1×2

$$\frac{\partial}{\partial w} s = \frac{\partial}{\partial w} U^T a = \frac{\partial}{\partial w} U^T \cdot f(z) = \frac{\partial}{\partial w} U^T \cdot f(Wx+b)$$

derivative w.r.t. W_{ij}

$$\frac{\partial}{\partial W_{ij}} U^T \cdot f(Wx+b)$$

$1\times 2 \times 2\times 1$

$$= \frac{\partial}{\partial W_{ij}} U_i \cdot f(z)$$

$$= \frac{\partial}{\partial W_{ij}} \sum_i U_i \cdot f(z_i)$$

$z = (Wx+b)$

$z_i = (W_i x + b)$

$$= \sum_i U_i \cdot \frac{\partial}{\partial W_{ij}} f(z_i)$$

$x = 1, 2$

$$= \sum_x U_x \cdot \frac{\partial}{\partial W_{ij}} f(z_x)$$

$$= \sum_z U_z \cdot \frac{\partial}{\partial W_{ij}} f(W_z x + b)$$

$W \in \mathbb{R}^{2\times 3}$; $x \in \mathbb{R}^{3\times 1}$

$Wx \in \mathbb{R}^{2\times 1}$

$$\frac{\partial}{\partial w_{ij}} v^T \cdot f(Wx+b).$$

$$= \frac{\partial}{\partial w_{ij}} \sum_p v_p^T \cdot f(z_p).$$

$i = 2$

$= v_1 f(z_1) + v_2 f(z_2)$

$= v_1 f(w_1 x + b_1) + v_2 f(w_2 x + b_2)$

$= 0 +$

$= \frac{\partial}{\partial w_{ij}}$ when $p \neq i$;

$$\frac{\partial}{\partial w_{ij}} s = \frac{\partial}{\partial w_{ij}} v_i f(z_i).$$

$$\frac{\partial}{\partial w_{ij}} v_i \cdot f(z_i). = \frac{\partial}{\partial w_{ij}} v_i.$$

$$= v_i \cdot \frac{\partial}{\partial w_{ij}} f(z_i) \quad ; \; z_i = w_i x + b$$

$$= v_i \cdot \frac{\partial}{\partial z_i} f(z_i) \cdot \frac{\partial}{\partial w_{ij}} (w_i x + b)$$

$$= \underbrace{v_i \cdot f'(z_i)}_{\delta_i} \cdot x_j$$

$$\boxed{\frac{\partial}{\partial w_{ij}} s = \delta_i x_j}$$

from $\frac{\partial}{\partial W_{ij}} s$ to $\frac{\partial}{\partial W} s$

$$\frac{\partial}{\partial W_{ij}} s = \delta_i \cdot x_j \quad \rightarrow (2)$$

$$= \underbrace{\delta x^T}_{\text{outer product}}$$

$[\delta_1 \ \delta_2] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$

$\delta \in \mathbb{R}^{2\times 1}$: responsibility/ error msg from each activation 'a'.

~~$\delta_1 x_1 \ \delta_2 x_2$~~

$\delta_1 x_1 \ \delta_1 x_2 \ \delta_1 x_3$.

$$\frac{\partial}{\partial b} s = \frac{\partial}{\partial b} v^T a = \frac{\partial}{\partial b} v^T \cdot f(z) = \frac{\partial}{\partial b} v^T f(Wx+b)$$

$$= \frac{\partial}{\partial b} \sum_i v_i \cdot f(z_i)$$

$$\Rightarrow \frac{\partial}{\partial b_i} s = \frac{\partial}{\partial b_i} v_i \cdot f(z_i)$$

$$= v_i \cdot \frac{\partial}{\partial b_i} f(z_i)$$

$$= v_i \cdot \frac{\partial}{\partial z_i} f(z_i) \cdot \frac{\partial}{\partial b_i} (w_i x + b_i)$$

$1\times 2 \times 2\times 1$
$= 1\times 1$

$\delta_1 w_1 + \delta_2 w_2$

$1\times 1 \times$

$= v_i \cdot f'(z_i) \cdot 1$

$= \delta_i$

$$\boxed{\frac{d}{db} s = \delta} \rightarrow (3)$$

$$\frac{d}{dx} s = \frac{d}{dx} \sum_i v_i \cdot f(z_i)$$

$[\delta_1\ \delta_2] \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}$

$(\delta_1\ \delta_2)$ 1×2 $\begin{pmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \\ w_{13} & w_{23} \end{pmatrix}$ 3×2

$$= \sum_i v_i \cdot \frac{d}{dx} f(z_i) \quad ; z_i = w_i x + b_i$$

$$= \sum_i v_i \cdot \frac{d}{dz_i} f(z_i) \cdot \frac{d}{dx} z_i(x)$$

$\begin{pmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \\ w_{13} & w_{23} \end{pmatrix} \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}$ $3\times 2 \quad 2\times 1$

$$= \sum_i v_i \cdot f'(z_i) \cdot \frac{d}{dx} w_i x + b_i$$

$\delta_1 w_1 + \delta_2 w_2$

δ_1

$$= \sum_i \delta_i \cdot w_i = w_{11}\delta_1 + w_{21}\delta_2 = \delta_1$$
$$w_{12}\delta_1 + w_{22}\delta_2$$
$$w_{13}\delta_1 + w_{23}\delta_2$$

$$\begin{bmatrix} \delta_1 w_{11} & \delta_1 w_{12} & \delta_1 w_{13} \\ \delta_2 w_{21} & \delta_2 w_{22} & \delta_2 w_{23} \end{bmatrix}$$

$= w^T \cdot \delta$

$$\therefore \boxed{\frac{d}{dx} s = w^T \delta}$$

$\sum_i u_i f_i(x)$

$\frac{d}{dx} v^T a$

$= \frac{d}{dx} \sum_i u_i f(z_i)$

$= \frac{d}{dx} \sum_i v_i \cdot f(w_i x + b_i)$

$\frac{d}{dx_j} \sum_i v_i \cdot f(w_i x + b_i)$

$= \sum_i v_i \cdot \frac{d}{dx_j} f(z_i)$

$= \sum_i v_i \cdot \frac{d}{dz_i} f(z_i) \cdot \frac{d}{dx_j} (w_i x + b_i)$

$= \sum_i v_i \cdot f'(z_i) \cdot \frac{d}{dx_j} \sum_p w_{ip} x_p$

$= \sum_i v_i \cdot f'(z_i) \cdot w_{ij}$

$\frac{d}{dx_j} \Sigma$ $= \sum_i \delta_i \cdot w_{ij}$ $\quad \delta_1 w_{12} + \delta_2 w_{22}$ $\quad x_2$

$= \delta \cdot W_{\cdot j}^T$

$$\frac{d}{dx} s = \begin{bmatrix} \delta \cdot w_{\cdot 1}^T & \delta w_{\cdot 2}^T & \delta w_{\cdot 3}^T \end{bmatrix}$$

$$\boxed{\frac{d}{dx} s = \delta W^T} \rightarrow (4)$$

$$\frac{d}{dU} s = a$$

$$\frac{d}{dW} s = \delta x^T$$

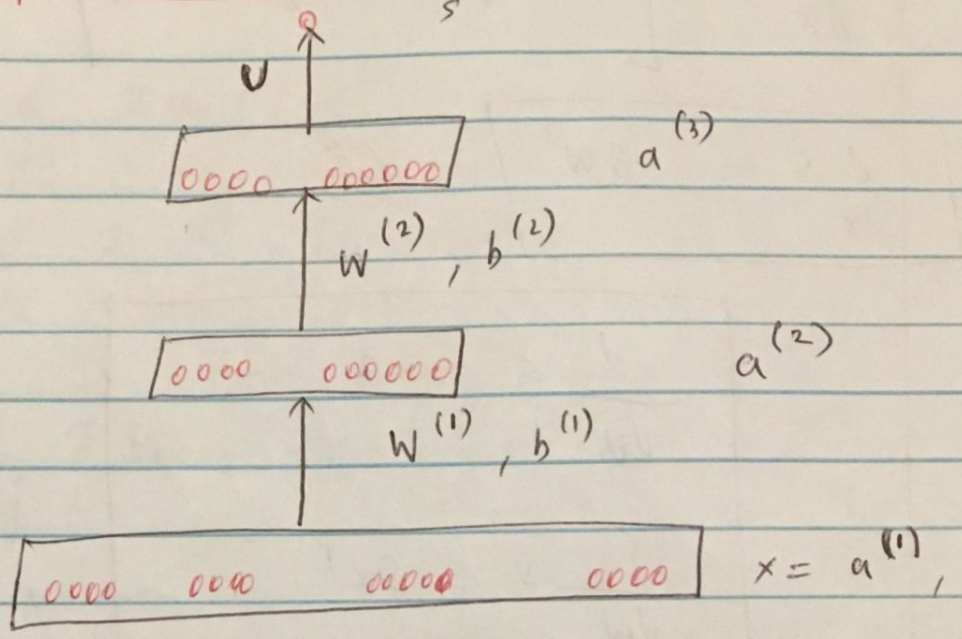
$$\frac{d}{db} s = \delta$$

$$\frac{d}{dx} s = \delta W^T$$

; $\delta = [\delta_1 \quad \delta_2]$

$= [U_1 f'(z_1) \quad U_2 f'(z_2)$

2- layer neural net:



$$x = z^{(1)} = a^{(1)}$$

$$z^{(2)} = W^{(1)} x + b^{(1)}$$

$$a^{(2)} = f(z^{(2)})$$

$$z^{(3)} = W^{(2)} \cdot a^{(2)} + b^{(2)}$$

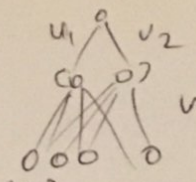
$$a^{(3)} = f(z^{(3)})$$

$$s = U^T a^{(3)}$$

$s = v^T \cdot a$

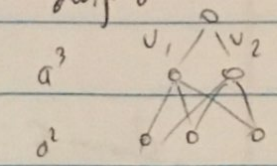
$= v^T \cdot b(z_i)$

$\frac{\partial s}{\partial w_{ij}} = 2 v_i f(z_i) \times \frac{\partial}{\partial w_{ij}}$ $v_i f(z_i)$



$W \begin{bmatrix} \cdot \end{bmatrix}$

$v_i \frac{\partial}{\partial w_{ij}} b(z_i)$



$s = v^T a^{(3)}$

$= v^T f(z^{(3)})$

$= v^T \cdot f(w^{(2)} a^{(2)} + b^{(2)})$

$= v^T \cdot f(w^{(2)} f(z^{(2)}) + b^{(2)})$

$= v^T \cdot f(w^2 \cdot f(w^{(1)} a^{(1)} + b^{(1)}) + b^{(2)})$

$= v^T \cdot f(w^{(2)} \cdot f(w^{(1)} \cdot x + b^{(1)}) + b^{(2)})$

$\frac{\partial s}{\partial w_{ij}} = v_i \cdot f'(z_i) \cdot x_j$

$= \delta_i \cdot x_j$

$\begin{bmatrix} \delta_1 x_1 & \delta_1 x_2 & \delta_1 x_3 \\ \delta_2 x_1 & \delta_2 x_2 & \delta_2 x_3 \end{bmatrix}$

$\frac{\partial s}{\partial w_{ij}^{(2)}} = \delta_i^{(3)} \cdot a_j^{(2)} = \delta_i^{(3)} \cdot a_j^{(2)}$

$\begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$

$\therefore \frac{\partial s}{\partial w^{(2)}} = \underbrace{\delta^{(3)} \cdot a^{(2)^T}}_{\text{outer product}}$

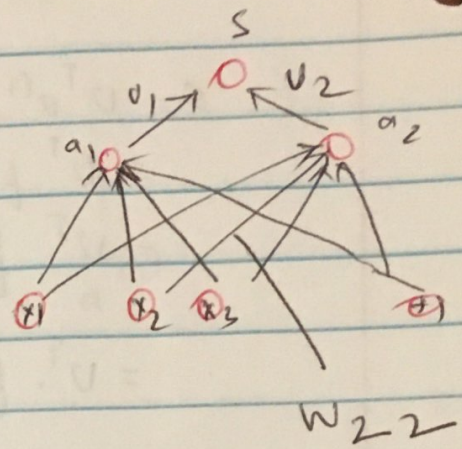
$\delta_1^{(3)}$ $\delta_2^{(3)}$

$f'(z_1^{(3)})$

$\delta^{(3)} = v \circ f'(z^{(3)})$

$\frac{\partial s}{\partial w^{(1)}} = \underbrace{\delta^{(2)} \cdot a^{(1)^T}}_{\text{outer prod.}}$

$\delta^{(2)} = \circ f'(z^{(2)})$



$$\frac{\partial}{\partial w_{ij}} s = \frac{\partial}{\partial w_{ij}} v^T a$$

$$= \frac{\partial}{\partial w_{ij}} \sum_y v_y a_y$$

$$= \frac{\partial}{\partial w_{ij}} \sum_y v_y f(W_y x + b)$$

w_{ij} appears only in w_i,

i.e, when $i = y$.

$$\therefore \frac{\partial}{\partial w_{ij}} s = \frac{\partial}{\partial w_{ij}} v_i \cdot f(W_i x + b)$$

$$= v_i \cdot \frac{\partial}{\partial w_{ij}} f(\underbrace{W_i x + b}_{z_i})$$

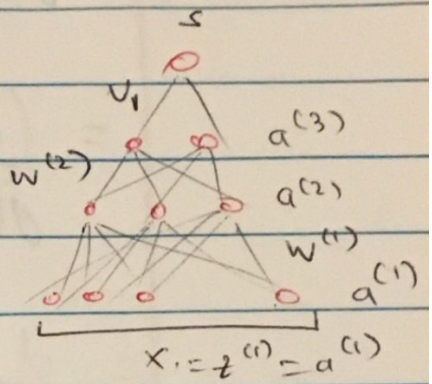
$$= v_i \cdot f'(z_i) \cdot \frac{\partial}{\partial w_{ij}} (W_i x + b)$$

$$= v_i \cdot f'(z_i) \cdot x_j$$

$$\boxed{\frac{\partial}{\partial w_{ij}} s = \delta_i x_j}$$

$$\therefore \frac{\partial}{\partial w_{ij}^{(1)}} s = \delta_i \cdot x_j$$

For 2 layer Neural Net:



$$\frac{\partial}{\partial w_{ij}^{(2)}} s = \delta_i^{(3)} \cdot a_j^{(2)}$$

$$\frac{\partial}{\partial w} s = \underbrace{\delta^{(3)} a^{(2)}}_{\text{outer product.}}$$

$$\frac{\partial}{\partial w_{ij}^{(1)}} s = \delta_i^{(2)} \cdot a_j^{(1)}$$

$$\delta_i^{(3)} = v_i \cdot f'(z_i^{(3)})$$

$$\delta_i^{(3)} = v_i \cdot f'(z_i^{(3)})$$

$$\delta^{(3)} = \underbrace{v \circ f'(z^{(3)})}_{\text{Hamard product}}$$

$\delta_i^{(2)} = ?$

∂

X_{np}^{m} $\quad$ $X_{j}^{(i)}$

i → layer no

X → node name
m → layer no
n → row
p → col

$$\frac{d}{dw_{ij}^{(1)}} \cdot s = \frac{d}{dw_{ij}^{(1)}} \cdot v^T a$$

$$= \frac{d}{dw_{ij}^{(1)}} \cdot \sum_p v_p \cdot a_p^{(3)}$$

$$= \frac{d}{dw_{ij}^{(1)}} \sum_p v_p \cdot \frac{d}{dw_{ij}^{(1)}} \cdot a_p^{(3)}$$

$$= v_1 \cdot \frac{d}{dw_{ij}^{(1)}} a_1^{(3)} + v_2 \cdot \frac{d}{dw_{ij}^{(1)}} a_2^{(3)}$$

$$= v_1 \cdot \frac{d}{dw_{ij}^{(1)}} f(z_1^{(3)}) + v_2 \frac{d}{dw_{ij}^{(2)}} f(z_2^{(3)})$$

$z_1^{(3)}$

$$= v_1 \frac{d}{dw_{ij}^{(1)}} f(w_1^{(2)} a^{(2)} + b_1^{(2)})$$

$z_2^{(3)}$

$$+ v_2 \frac{d}{dw_{ij}^{(1)}} f\left(w_2^{(2)} a^{(2)} + b_2^{(2)}\right)$$

$\begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{pmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} w_{11}x_1 + w_{12}x_2 + w_{13}x_3 \\ w_{21}x_1 + w_{22}x_2 + w_{23}x_3 \end{bmatrix} \begin{matrix} z_1 \\ z_2 \end{matrix}$

$2 \times 3 \times 3 \times 1$

2×1

$\frac{\partial}{\partial w_{ij}}$

$$= v_1 f'(z_1^{(3)}) \frac{d}{dw_{ij}^{(1)}} \left(w_1^{(2)} a^{(2)} + b_1^{(2)} \right)$$

$$+ v_2 f'(z_2^{(3)}) \frac{d}{dw_{ij}^{(1)}} \left(w_2^{(2)} a^{(2)} + b_2^{(2)} \right)$$

$$= v_1 f'(z_1^{(3)}) \cdot \frac{d}{dw_{ij}^{(1)}} w_1^{(2)} a^{(2)}$$

$$+ v_2 f'(z_2^{(3)}) \frac{d}{dw_{ij}^{(1)}} \cdot w_2^{(2)} a^{(2)}$$

$$= v_1 f'(z_1^{(3)}) \cdot w_1^{(2)} \frac{d}{dw_{ij}^{(1)}} a^2 + v_2 f'(z_1^{(3)}) \frac{d}{dw_{ij}^{(1)}} w_2^{(2)} a^{(2)}$$

$w_2^{(2)}$

$$a^2 = f(z^{(2)}) = f(\underbrace{w^{(1)}x + b^{(1)}}_{z^{(2)}})$$

$$\frac{d}{dw_{ij}^{(1)}} a^{(2)} = \frac{d}{dw_{ij}} f(w^{(1)}x + b^{(1)})$$

$\begin{bmatrix} 0 & x_j \\ x_j & 0 \end{bmatrix}$

$$= f'(z^{(2)}) \cdot \frac{d}{dw_{ij}} (w^{(1)}x + b^{(1)})$$

$$= f'(z^{(2)}) \begin{bmatrix} x_j & 0 \end{bmatrix}$$

$$= f'(z^{(2)}) x_j$$

$f(W_p^{(2)} a^{(2)} + b)$

$$\sum_p v_p \cdot \frac{\partial}{\partial w_{ij}^{(1)}} a_p^{(3)}$$

$$= \sum_p v_p \cdot \frac{\partial}{\partial w_{ij}^{(1)}} f(z_p^{(3)})$$

$$= \sum_p v_p \cdot \frac{\partial}{\partial w_{ij}^{(1)}} f(W_p^{(2)} a^{(2)} + b_p^{(2)})$$

$$= \sum_p v_p \, f'(z_p^{(3)}) \cdot \frac{\partial}{\partial w_{ij}^{(1)}} (W_p^{(2)} a^{(2)} + b_p^{(2)})$$

$$= \sum_p v_p \, f'(z_p^{(3)}) \, W_p^{(2)} \cdot \frac{\partial}{\partial w_{ij}^{(1)}} a^{(2)}$$

$$= \sum_p v_p \, f'(z_p^{(3)}) \, W_p^{(2)} \cdot \frac{\partial}{\partial w_{ij}^{(1)}} f(z^{(2)})$$

$$= \sum_p v_p \, f'(z_p^{(3)}) \, W_p^{(2)} \cdot f'(z^{(2)}) \frac{\partial}{\partial w_{ij}^{(1)}} [w^{(1)} x + b^{(1)}]$$

$$= \sum_{p=1}^{2} v_p \, f'(z_p^{(3)}) \, W_p^{(2)} \cdot f'(z^{(2)}) \, x_j$$

$$= \sum_p \underbrace{\delta_p^{(3)} \cdot W_p^{(2)} \cdot f'(z^{(2)})}_{\delta_p^{(2)}} \underbrace{x_j}_{a_j^{(1)}}$$

$2\times3 \times 3\times1 \quad 2\times1$

$$\frac{d}{dw_{ij}^{(1)}} \begin{bmatrix} w_{11}^{1}x_1 + w_{12}^{1}x_2 + w_{13}^{1}x_3 \\ w_{21}^{1}x_1 + w_{22}^{1}x_2 + w_{23}^{1}x_3 \end{bmatrix} \qquad \begin{bmatrix} 0 \\ x_j \end{bmatrix} \text{ or } \begin{bmatrix} x_j \\ 0 \end{bmatrix}$$

$$\therefore \frac{d}{dw_{ij}^{(1)}} s = \sum_p \delta_p^{(3)} \cdot w_p^{(2)} \cdot f'(z^{(2)})\, x_j$$

$\delta_p^{(3)}$: 1×1, $w_p^{(2)}$: 1×1

$$= f'(z^{(2)}) \cdot x_j \left[\sum_p \delta_p^{(3)} \cdot w_p^{(2)} \right]$$

$$= f'(z^{(2)})\, x_j \left[\delta^{(3)} w^{(2)} \right]$$

$$= \underbrace{w^{(2)T} \delta^{(3)} \circ f'(z^{(2)})}_{\delta_i^{(2)}} \cdot \underbrace{x_j}_{a_j^{(1)}}$$

$2\times3 \cdot 2\times1$

$3\times1 \cdot 2\times1$

3×1

$f'(z^{(2)})\, x_j$ $\qquad [z_1^{(2)}, z_2^{(1)}]$

$$\left[w^{(1)T} \cdot \delta^{(3)} \right] \circ f'(z^{(1)})\, x_j$$

3×1 $\qquad 1\times1$

$w^{(1)T} \cdot \delta^{(3)} \delta^{(2)}$

element-wise hamard.

3×1

$$\frac{\partial}{dw_{ij}^{(1)}} s = \delta^{(2)} x_j.$$

$$\delta^{(2)} \cdot x^T$$

2×3

$w^{(2)}$

$\delta_1^3 \quad \delta_2^3$

$w_{11}^2 \quad w_{12}^2$

$\delta_1^{(3)} \quad \delta_2^{(3)}$

$\left[w_{11}^2 \; w_{12}^2 \; w_{13}^2 \right] \left[w_{21}^2 \; w_{22}^2 \; w_{23}^2 \right]$

$\left[\delta_1^{(3)} \; \delta_2^{(3)} \right]$

$\begin{bmatrix} w_{11}^{(2)} & w_{12}^{(2)} & w_{13}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} & w_{23}^{(2)} \end{bmatrix}$

$1\times2 \quad \cdot \quad 2\times3$

1×3

$x_1 x_2 x_3 \quad \cdot \quad \delta^{(2)} x_1 \quad \delta^{(2)} x_2 \quad \delta^{(2)} x_3$

$\delta^{(2)} x_1 \quad \delta^{(2)} x_2 \quad \delta^{(2)} x_3$

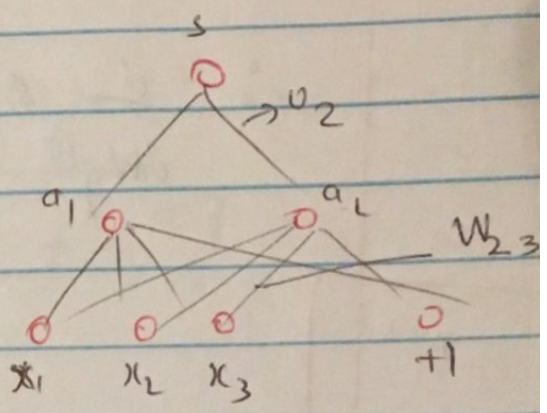
$$\delta^{(l)} = \left((W^{(l)})^T \delta^{(l+1)}\right) \circ f'(z^{(l)})$$

$J = \max(0, 1 - s + s_c)$

$s = v^T a$

$= v^T f(z)$

$= v^T \cdot f(Wx + b)$



Min J.

$\therefore \nabla J_{(v, W, x, b)} = ?$

$\nabla J = -\nabla s + \nabla s_c$

$\nabla J = \nabla s \Rightarrow$

$$\frac{ds}{dv} = \frac{d}{dv} v^T a = \frac{d}{dv} \sum_i v_i a_i = a \quad \rightarrow (1)$$

$$\frac{ds}{dw} \Rightarrow \frac{d s}{dw_{ij}} = \frac{d}{dw_{ij}} \sum_p v_p f(w_p x + b)$$

$$= \sum_p v_p \cdot \frac{d}{dw_{ij}} f(w_p x + b)$$

$$= \sum_p v_p \cdot f'(z_p) \frac{d}{dw_{ij}} (w_p x + b)$$

$$= \sum_p \delta_p \cdot \frac{\partial}{\partial w_{ij}} (w_p x + b)$$

→ when $i \neq p$ $\quad \frac{\partial}{\partial w_{ij}} (w_p x + b) = 0$

$i == p$ $\quad \frac{\partial}{\partial w_{ij}} (w_i x + b) = x_j$

$$\therefore \frac{\partial}{\partial w_{ij}} s = 0 + \delta_i x_j = \delta_i x_j$$

$\delta_1 \; \delta_2$ $\quad x_1 \; x_2 \; x_3$

$$\boxed{\frac{\partial}{\partial W} s = \delta x^T} \quad \text{outer prod} \quad \rightarrow (2)$$

$$\frac{\partial}{\partial b_j} s = \frac{\partial}{\partial b_j} \sum_i v_i \cdot f(w_i x + b_i) \qquad \partial b_i$$

$$= \sum_i v_i \cdot \frac{\partial}{\partial b_j} f(z_i)$$

$$= \sum_i v_i \cdot f'(z_i) \frac{\partial}{\partial b_j} (w_i x + b_i)$$

$= \sum_i \delta_i$ $\quad$ when $j \neq i$, $\frac{\partial}{\partial b_j} (z_i) = 0$

else 1

$$= v_i \cdot f'(z_i)$$

$$\frac{\partial}{\partial b_i} s = \delta_i \qquad \therefore \frac{\partial}{\partial b} s = \delta$$

$$\frac{\partial}{\partial b} s = \delta \quad \rightarrow (3)$$

$$\frac{\partial}{\partial x_j} s = \frac{\partial}{\partial x_j} v^T a = \frac{\partial}{\partial x_j} \sum_i v_i a_i$$

$$= \sum_i v_i \cdot \frac{\partial}{\partial x_j} f(z_i)$$

$$= \sum_i v_i \cdot f'(z_i) \frac{\partial}{\partial x_j} (w_i x + b_i)$$

$$= \sum_i v_i \cdot f'(z_i) \cdot w_{ij}$$

$$= \sum_i \delta_i w_{ij}$$

$$= \delta w_{\cdot j}^T = w_{\cdot j} \cdot \delta^T$$

$$\therefore \frac{\partial}{\partial x} s = \begin{bmatrix} w_{\cdot 1} \delta^T & w_{\cdot 2} \delta^T & w_{\cdot 3} \delta^T \end{bmatrix}$$

$$= \delta \cdot W^T$$

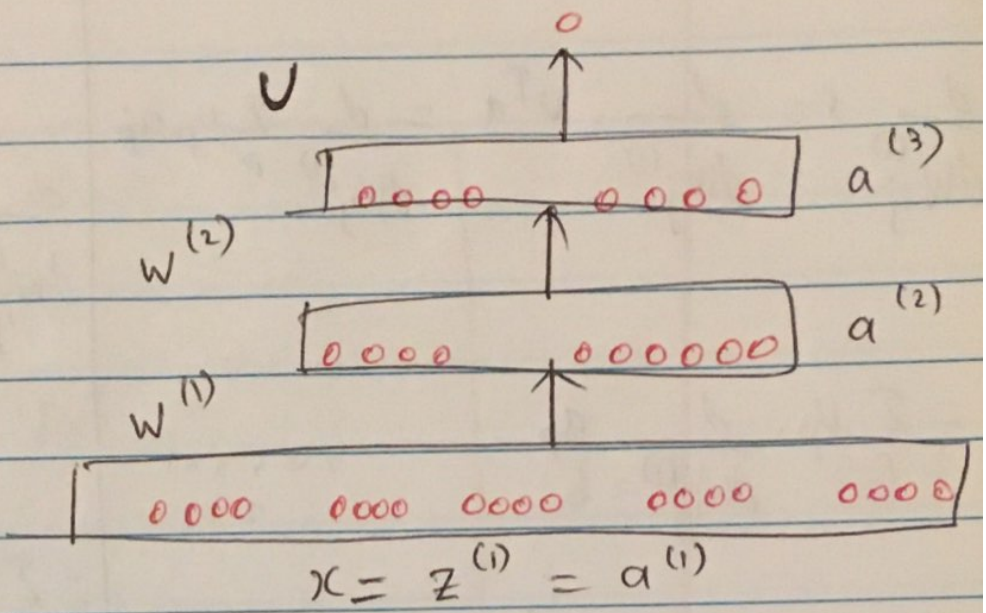
$$\boxed{\frac{\partial}{\partial x} s = W^T \delta} \quad \rightarrow (4)$$

w_1 w_{12} v_{13}
w_{11} w_{21} w_{23}.

$\delta_1 w_{1j}$
$+ \delta_2 w_{2j}$

$[\delta_1 \ \delta_2] \begin{bmatrix} w_{1j} \\ w_{2j} \end{bmatrix}$

$$x = z^{(1)} = a^{(1)}$$
$$z^{(2)} = W^{(1)} x + b^{(1)}$$
$$a^{(2)} = f(z^{(2)})$$
$$z^{(3)} = W^{(2)} a^{(2)} + b^{(2)}$$
$$a^{(3)} = f(z^{(3)})$$
$$s = U^T a^{(3)}$$



$$\frac{\partial}{\partial W^{(2)}} s = \delta^{(3)} a^{(2)T}$$

$$\delta^{(3)} = U \circ f'(z^{(3)})$$

$$\frac{\partial}{\partial W^{(1)}} s = \delta^{(2)} a^{(1)T}$$

$$\delta^{(2)} = (\quad) \circ f'(z^{(2)})$$

$$\frac{\partial}{\partial w_{ij}^{(1)}} s = \frac{\partial}{\partial w_{ij}^{(1)}} \cdot v^T a = \frac{\partial}{\partial w_{ij}^{(1)}} \sum_p v_p \cdot a_p$$

$$= \sum_p v_p \cdot \frac{\partial}{\partial w_{ij}^{(1)}} \cdot a_p$$

$$= \sum_p v_p \cdot \frac{\partial}{\partial w_{ij}^{(1)}} \cdot f(z_p)$$

$$= \sum_p v_p \cdot f'(z_p) \cdot \frac{\partial}{\partial w_{ij}^{(1)}} \cdot z_p$$

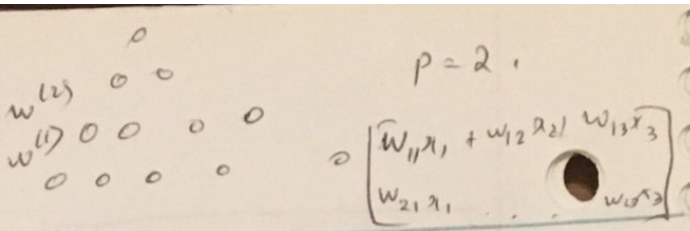
$$= \sum_p v_p \cdot f'(z_p) \frac{\partial}{\partial w_{ij}^{(1)}} \cdot \left(w^{(2)} a^{(2)} + b^{(2)}\right)$$

$$= \sum_p v_p \cdot f'(z_p) \cdot w^{(2)} \cdot \frac{\partial}{\partial w_{ij}^{(1)}} a^{(2)}$$

$$= \sum_p v_p \cdot f'(z_p) \cdot w^{(2)} \cdot \frac{\partial}{\partial w_{ij}^{(1)}} \cdot \left(w^{(1)} x + b^{(1)}\right)$$

$2 \times 3 \quad 3 \times 1$

2×1



$$\frac{\partial}{\partial w_{ij}^{(1)}} w^{(1)} x = \begin{bmatrix} x_j & 0 \end{bmatrix}$$

or

$$\begin{bmatrix} 0 & x_j \end{bmatrix}$$

$$= x_j$$

$$= \sum_p v_p \cdot f'(z_p) \cdot w^{(2)} x_j$$

$1 \times 1 \quad (1) \quad 2 \times 3 \quad 1 \times 1$

$$= \sum_p \delta_p^{(3)} \cdot w^{(2)} \cdot x_j$$

$$\delta^{(3)} x_j$$

$$\delta_1^{(3)} w^{(1)} x_j + \delta_2^{(3)} w^{(2)} x_j$$

$$= x_j \cdot \delta^{(3)} \cdot w^{(2)}$$

$$= \left(a_j^{(1)} \circ \delta^{(3)}\right) \cdot w^{(2)}$$

$$\frac{\partial}{\partial v} s = a$$

$$\frac{\partial}{\partial w} s = \delta x^T$$

$$\frac{\partial}{\partial b} s = \delta$$

$$\frac{\partial}{\partial x} s = w^T \delta$$

$$\delta = \begin{bmatrix} \delta_1 & \delta_2 \end{bmatrix}$$

$$\begin{bmatrix} w_{11}x_1 + w_{12}x_2 + w_{13}x_3 \\ w_{21}x_1 + w_{22}x_2 + w_{23}x_3 \end{bmatrix}$$

$2\times3 \quad 3\times1$
2×1

$$\frac{\partial}{\partial w_{ij}^{(1)}} s = \frac{\partial}{\partial w_{ij}^{(1)}} v^T f(z^{(3)})$$

$$= \frac{\partial}{\partial w_{ij}^{(1)}} \sum_{p=1}^{2} v_p \cdot f(z_p^{(3)})$$

$$= \sum_{p=1}^{2} v_p \cdot \frac{\partial}{\partial w_{ij}^{(1)}} f(z_p^{(3)})$$

$$= \sum_{p=1}^{2} v_p \cdot f'(z_p^{(3)}) \cdot \frac{\partial}{\partial w_{ij}^{(1)}} z_p^{(3)}$$

Ep

$$= \sum_{p=1}^{2} \delta_p^{(3)} \frac{\partial}{\partial w_{ij}^{(1)}} w^{(2)} a^{(2)} + b^{(2)}$$

$$= \sum_{p=1}^{2} \delta_p^{(3)} w^{(2)} \frac{\partial}{\partial w_{ij}^{(1)}} w^{(1)} x + b^{(1)}$$

$$= \sum_{p=1}^{2} \delta_p^{(3)} w^{(2)} \frac{\partial}{\partial w_{ij}^{(1)}} f(z^{(2)})$$

$$= \sum_{p=1}^{2} \delta_p^{(3)} w^{(2)} f'(z^{(2)}) \frac{\partial}{\partial w_{ij}^{(1)}} w^{(1)} x + b^{(1)}$$

$$= \sum_{p=1}^{2} \delta_p^{(3)} w^{(2)} f'(z^{(2)}) x_j$$

$\delta^{(2)}$

$$= \sum_{p=1}^{2} \delta_p^{(3)} \; w^{(2)} \; f'(z^{(2)}) \; a_j^{(1)}$$

$1\times1 \quad 2\times3 \quad 2\times1 \quad 1\times1$

$$\delta^{(2)} = U \circ f'(z^{(2)})$$

$$\frac{d}{dW_1^{(1)}} s = \delta^{(2)} a^{(1)T}$$

$=$

$\delta^{(2)} =$

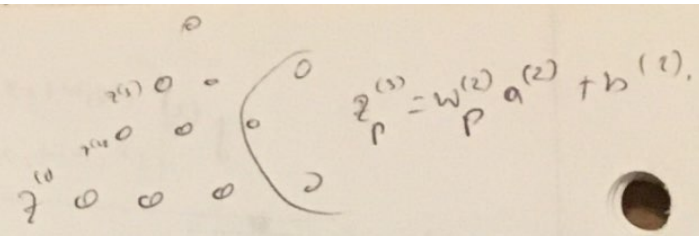
$$= f'(z^{(2)}) \; a_j^{(1)} \left[\sum_{p=1}^{2} \delta_p^{(3)} \; w^{(2)} \right]$$

$$= f'(z^{(2)}) \; a_j^{(1)} \circ \delta^{(3)} \; w^{(2)T}$$

$$= \underbrace{\left(w^{(2)T} \delta^{(3)}\right) \circ f'(z^{(2)})}_{\delta^{(2)}} \; a_j^{(1)}$$

$$\delta^{(2)} = \left(w^{(2)T} \delta^{(3)}\right) \circ f'(z^{(1)})$$

$\frac{\partial}{\partial w_{ij}^{(1)}} f = ?$



$$\frac{\partial}{\partial w_{ij}^{(1)}} v^T a = \frac{\partial}{\partial w_{ij}^{(1)}} \sum_p v_p a_i$$

$$= \sum_p v_p \cdot \frac{\partial}{\partial w_{ij}^{(1)}} a_p^{(3)}$$

$$= \sum_p v_p \cdot \frac{\partial}{\partial w_{ij}^{(1)}} f(z_p^{(3)})$$

$$= \sum_p v_p \cdot f'(z_p^{(3)}) \cdot \frac{\partial}{\partial w_{ij}^{(1)}} z_p^{(3)}$$

$$= \sum_p v_p \cdot f'(z_p^{(3)}) \cdot \left[w_p^{(2)} \cdot \frac{\partial}{\partial w_{ij}^{(1)}} \cdot f(w_i^{(1)} a^{(1)} + b^{(1)}) \right] \quad z^{(2)}$$

$$= \sum_p v_p \cdot f'(z_p^{(3)}) \cdot \left[w^{(2)} \cdot \left(f'(z^{(2)}) \cdot \frac{\partial}{\partial w_{ij}} z^{(2)} \right) \right]$$

$$= \sum_p s_p^{(3)} \cdot \left[w^{(2)} \cdot \left(f'(z^{(2)}) \cdot \frac{\partial}{\partial w_{ij}} (w^{(1)} x_i + b^{(1)}) \right) \right]$$

$$= \sum_p s_p^{(3)} \cdot \left[w^{(2)} \cdot \left(f'(z^{(2)}) \cdot x_j \right) \right]$$

$$= \sum_p s_p^{(3)} \cdot \left[w_p^{(2)} \cdot \left(f'(z^{(2)}) \cdot a_j^{(1)} \right) \right] \quad \delta^{(3)} \cdot \alpha$$

1×1 1×2 1×2 2×1 1×2

$\delta^{(2)}$ $a^{(1)}$

$\delta^{(2)} = U \circ f'(z^{(2)})$ | $\times 2$ $\quad 2 \times 1$

$\delta^{(1)} = \delta^{(3)} W^{(2)T} \circ f'(z^{(2)}) \; a_j^{(1)}$

$\frac{d}{dW^{(1)}} q = \underbrace{\left(W^{(2)T} \delta^{(3)}\right) \circ f'(z^{(2)})}_{\delta^{(2)}} a^{(1)}$

$= \delta^{(2)} a^{(1)}$.

$2 \times 2 \cdot 2 \times 1 \quad T$

$(2 \times 1) \; 3 \times 1$

$2 \times 1 \times 1 \times 5$

2×5

(2) $\delta^{(3)} = U \circ f'(z^{(3)})$

$\delta^{(2)} = \left(W^{(2)} \delta^{(3)}\right) \circ f'(z^{(2)})$

$\therefore \boxed{\delta^{(l)} = \left(W^{(l)} \delta^{(l+1)}\right) \circ f'(z^{(l)})}$

$\therefore$ For last hidden layer, $\delta^{n} = U \circ f'(z^{(n)})$

$\delta^{(l)} = \left(W^{(l)T} \delta^{(l+1)}\right) \circ f'(z^{(l)})$

$\frac{d}{dW^{(2)}} S = \delta^{(3)} a^{(2)T}$

$\frac{d}{dW^{(1)}} S = \delta^{(2)} . a^{(1)T}$

$$\sum_p \delta_p^{(3)} \circ \left[W_p^{(2)} \cdot \left(f'(z^{(2)}) \circ a_j^{(1)} \right) \right]$$

$$= \sum_p \delta_p^{(3)} \circ \left[W_p^{(2)T} \cdot f'(z^{(2)}) \right] \circ a_j^{(1)}$$

$$= \delta^{(3)} \alpha$$

$$\delta^{(3)} = \begin{bmatrix} \delta_1^{(3)} & \delta_2^{(3)} \end{bmatrix}$$

$$a^{(1)} = \begin{bmatrix} a_1^{(1)} & a_2^{(1)} & a_3^{(1)} \end{bmatrix}$$

$$\left(f'(z^{(2)}) \circ a_j^{(1)} \right) \circ \left[\sum_p \delta_p^{(3)} \; W_{p}^{(2)T} \right]$$

$$\left(f'(z^{(2)}) \circ a_j^{(1)} \right) \circ \left[\delta^{(3)} \cdot W^{(2)T} \right]$$

$$\left[\delta^{(3)} \cdot W^{(2)T} \right] \circ f'(z^{(2)}) \circ a_j^{(1)}$$

$$\left[\delta^{(3)} \cdot W^{(2)T} \right]$$

$$\left[\delta^{(3)} \cdot W^{(2)T}\right] \cdot f'(z^{(2)}) \quad a_j^{(1)}$$

$1\times 2 \qquad 1\times 2 \qquad 1\times 1$

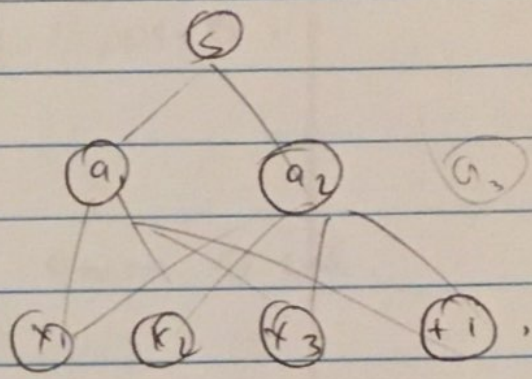
$\delta^{(3)} \; W^{(2)T} \; f'(z^{(2)}) \; a_j^{(1)}$

o o o o o D_y

o o o o H. (+1).

o o o (+1)

x



$(D_x * H) + (H * D_y) + \cancel{2}$

$\cancel{4}\,x$

$\cancel{12}$

$D_H \times \cancel{(x+1)}$.

$\left[(D_x+1) * H\right]$

$+ \left[(H+1) * D_y\right]$

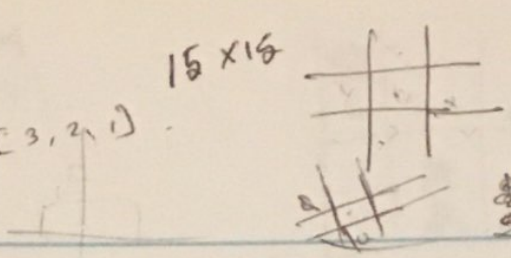
7 4 8x8 boards

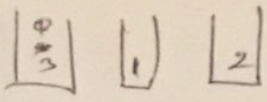
def hanoi(

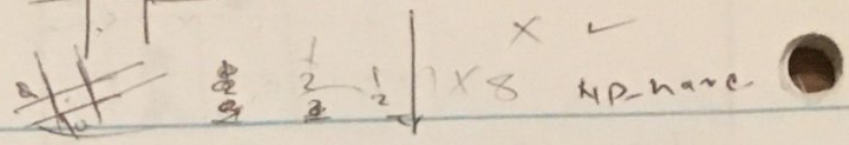
1
2
3

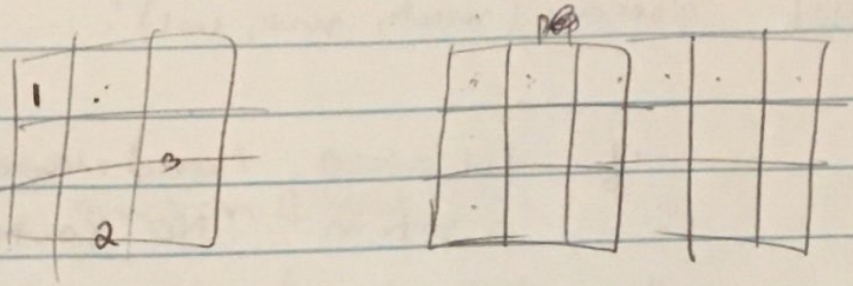
15 x 15

[3, 2, 1]

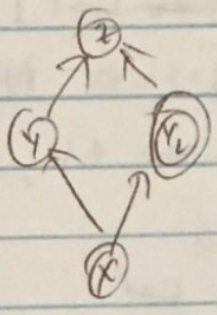








→ Compute ∇ Loss example wise.



$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y_1}\frac{\partial y_1}{\partial x} + \frac{\partial z}{\partial y_2}\cdot\frac{\partial y_2}{\partial x}$$

$$\frac{\partial z}{\partial x} = \sum_{i=1}^{n} \frac{\partial z}{\partial y_i}\frac{\partial y_i}{\partial x}$$

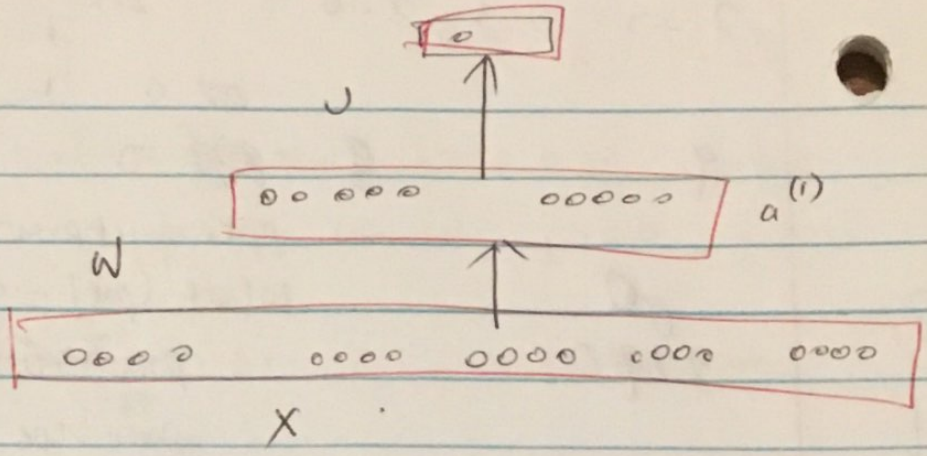
canPlace

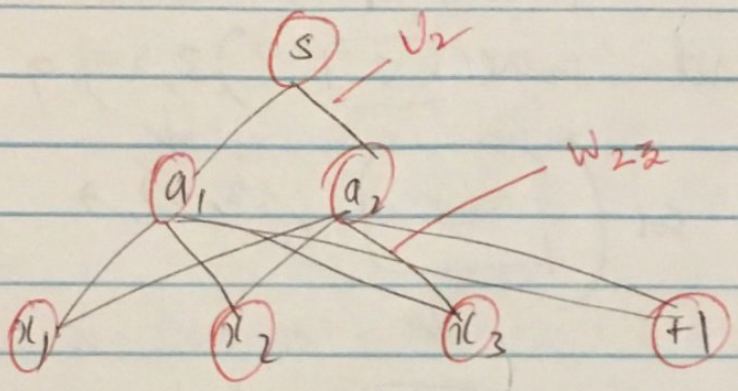
→ ~~check~~ (board, pos, num)

// 'Y' if ok

'N' if not ok.

→ Complete getPos





$s = U^T a = U^T \cdot f(z) = U^T \cdot f(W^T x + b)$

$s \in \mathbb{R}^{1*1}$; $U^T \in \mathbb{R}^{1*2}$; $z \in \mathbb{R}^{2*1}$

$W \in \mathbb{R}^{2*3}$; $b \in \mathbb{R}^{2*1}$

$1*2$ $\quad$ $2*3 \cdot 3*1$ $\quad$ $1*1$

$1*2$ $\quad$ $2*1$ $\quad$ $2*1$

$1*2 \cdot 2*1$

$1*1$

$J = \max(0, 1 - s + s_c)$

$s > s_c \Rightarrow 0$

$s < s_c \Rightarrow p$

$s = = s_c \Rightarrow 1$

Minimize J.

$\therefore \nabla J_{(v, w, x, b)} = ?$

$$\frac{dJ}{d\alpha} = \frac{-ds}{d\alpha} + \frac{\partial s_c}{d\alpha}$$

$$\therefore \frac{ds}{dv} = \frac{d}{dv} v^T \alpha = \frac{d}{dv} \sum_i v_i a_i = a$$

$$\therefore \boxed{\frac{ds}{dv} = a} \rightarrow (1)$$

$$\frac{ds}{dw} = \frac{d}{dw} v^T a$$

Try for w_{ij}; $\frac{d}{dw_{ij}} v^T a = \frac{d}{dw_{ij}} \sum_i v_i a_i$

$= v_i \frac{d}{dw_{ij}} a_i = v_i \frac{d}{dw_{ij}} f(z_i)$

$= v_i \frac{d}{dw_{ij}} f(\overbrace{w_i x + b_i}^{z_i}) = v_i \cdot \frac{d}{dz_i} f(z_i) \cdot \frac{d}{dw_{ij}} (w_i x + b_i)$

$= v_i \cdot f'(z_i) \cdot x_j$

$= v_i \cdot v'$

$v_1 f(z_1).x_1 \quad v_1 f(z_1)x_2 \quad v_1 f(z_1)x_3$

$v_2 f'(z_2)x_1 \quad \ldots \quad v_2 f(z_2)x_3$

$$\therefore \frac{\partial}{\partial w_{ij}} s = \underbrace{v_i \cdot f'(z_i)}_{\delta_i} \cdot x_j = \delta_i x_j$$

From $w_{ij} \longrightarrow w$:

$$\frac{\partial}{\partial w} s = ?$$

Let $v = [v_1 \quad v_2]$

$f'(z) = [f'(z_1) \quad f'(z_2)]$

$x = [x_1 \quad x_2 \quad x_3]$.

$\delta = [\delta_1 \quad \delta_2 \quad \delta_3]$.

$$\boxed{\frac{\partial}{\partial w} s = \underbrace{\delta x^T}_{\text{outer product}}} \longrightarrow (2)$$

$$\frac{\partial}{\partial x_j} s = \frac{\partial}{\partial x_j} v^T a = \frac{\partial}{\partial x_j} \sum_i v_i . a_i$$

$$= \sum_i v_i . \frac{\partial}{\partial x_j} a_i$$

$$= \sum_i v_i . \frac{\partial}{\partial x_j} f(z_i) .$$

$$= \sum_i v_i . \frac{\partial}{\partial x_j} f(w_i x + b_i)$$

$$= \sum_i v_i \frac{d}{dz_i} f(z_i) \cdot \frac{d}{dx_j} (w_i x + b_i)$$

$$= \sum_i v_i \cdot f'(z_i) \cdot w_{ij}$$

$$= \sum_i \delta_i \cdot w_{ij} \qquad = \delta \cdot W_{\cdot j}^T$$

$$\therefore \boxed{\frac{d}{dx_j} s = \delta W_{\cdot j}^T}$$

From $x_j \longrightarrow x$

$$\boxed{\frac{d}{dx} s = \delta W^T} \longrightarrow (3)$$

$$\frac{d}{db} s = ?$$

$$\frac{d}{db_i} v^T a = \frac{d}{db_i} \sum_i v_i \cdot a_i$$

$$= v_i \cdot \frac{d}{db_i} a_i \qquad = v_i \cdot \frac{d}{db_i} f(\underbrace{w_i x + b_i}_{z_i})$$

$$= v_i \cdot \frac{d f(z_i)}{d z_i} \frac{d}{db_i} (w_i x + b_i)$$

$$= v_i \cdot f'(z_i) \cdot 1$$

$$= \delta_i$$

$$\therefore \frac{\partial}{\partial b_i} s = \delta_i$$

$$\boxed{\frac{\partial}{\partial b} s = \delta} \rightarrow (4)$$

1, 2, 3, 4 $\Rightarrow$

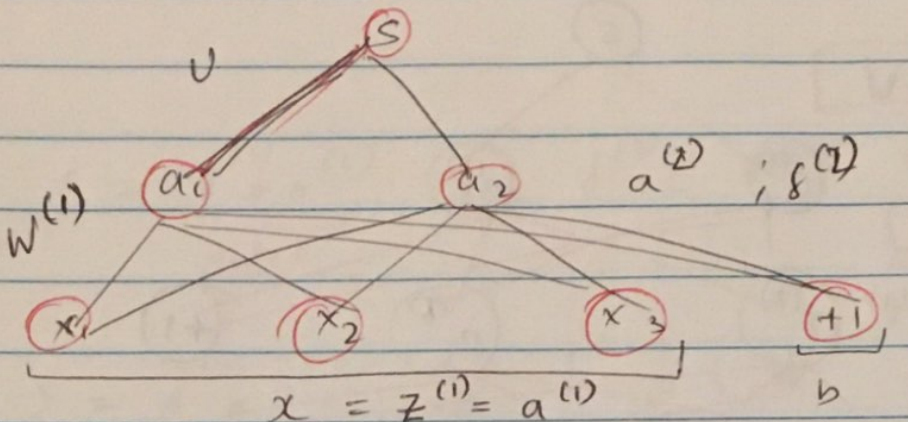
$$\frac{\partial}{\partial U} s = a$$

$$\frac{\partial}{\partial W} s = \underbrace{\delta x^T}_{\text{Howard}} \quad ; \quad s = [\delta_1, \delta_2]$$

$$= [U_1 \sigma(z_1) \quad U_2 \sigma(z_2)]$$

$$\frac{\partial}{\partial x} s = \delta W^T$$

$$\frac{\partial}{\partial b} s = \delta$$



$$\frac{\partial}{\partial U} s = a \; ; \quad \frac{\partial s}{\partial W} = \delta x^T \; ; \quad \delta = \begin{bmatrix} \delta_1 & \delta_2 \end{bmatrix}$$

$$= \begin{bmatrix} U_1 \sigma' f(z_1) & U_2 \sigma f(z_2) \end{bmatrix}$$

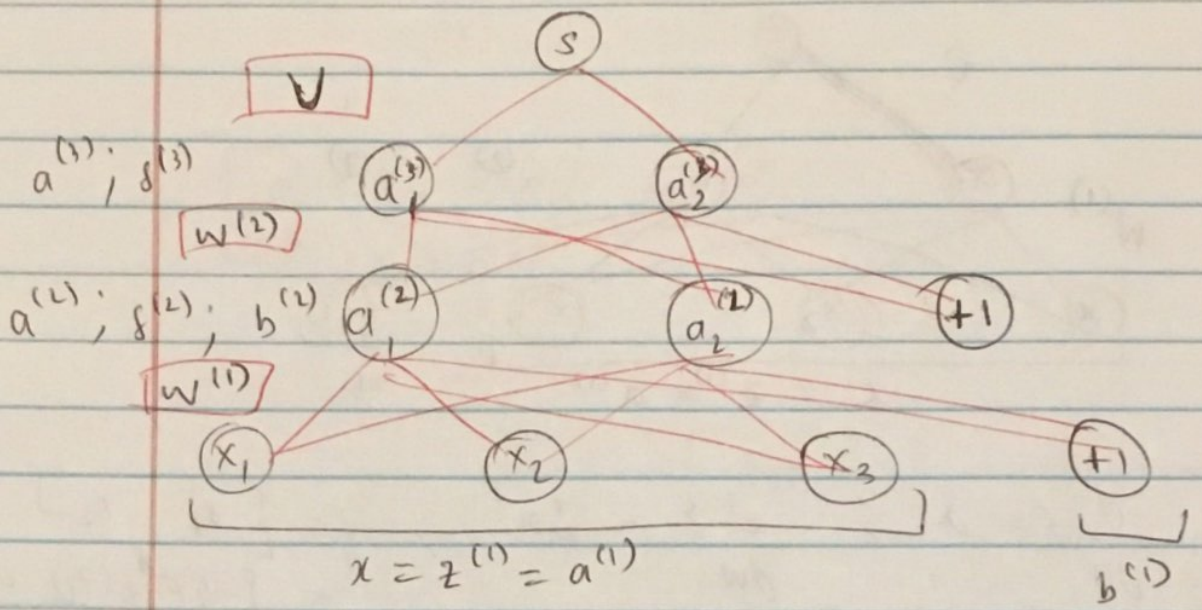
$$\frac{\partial}{\partial x} s = \delta W^T \; ; \quad \frac{\partial}{\partial b} s = \delta$$

Renaming:

$$\frac{\partial}{\partial U} s = a^{(2)} \; ; \quad \frac{\partial}{\partial W^{(1)}} s = \delta^{(2)} \cdot a^{(2)} \; ; \quad \frac{\partial}{\partial a^{(1)}} s = \delta^{(2)} \cdot a^{(1)}$$

$$\frac{\partial}{\partial b^{(1)}} = \delta^{(1)}$$

2 layer Neural Network:



To find:

$\frac{d}{dV} S$

$\frac{d}{dW^{(2)}} S$

$\frac{d}{dW^{(1)}} S$

$\frac{d}{da^{(1)}} S$

$\frac{d}{db^{(1)}} S$

$\frac{d}{db^{(2)}} S$

$\frac{d}{da^{(2)}} S$

$\circ$ ⇒ element
$\bullet$ ⇒ dot
⇒ outer

250 → 25%

x → 100

$\frac{\partial}{\partial v} s = a^{(3)}$ 25

$\frac{\partial}{\partial W^{(2)}} s = \underbrace{\delta^{(3)} \bullet a^{(2)}}_{\text{outer}}$; $\delta^{(3)} = \underbrace{V \circ b'(z^{(3)})}_{\text{element wise}}$

$\frac{\partial}{\partial W^{(1)}} s = \underbrace{\delta^{(2)} \bullet a^{(1)}}_{\text{outer}}$; $\delta^{(1)} = (W^{(2)^T} \delta^{(3)}) \circ f'(z^{(1)})$

2×1

2×1 ~~3×1~~

$\frac{\partial}{\partial x} s = \frac{\partial}{\partial a^{(1)}} s = \delta^{(1)} a^{(1)}$

$\therefore \boxed{\delta^{(1)} = \left(W^{(1)^T} \cdot \delta^{(1+1)}\right) \circ f'(z^{(1)})}$

(1) cr limit; acc. our bal, open dche.

30

30 × 25 $\frac{30 \times}{4}$

$J(\theta)$

$$\nabla J(\theta) \approx \lim_{\epsilon \to 0} \frac{J(\theta+\epsilon) - J(\theta+\epsilon)}{2\epsilon}$$

chose $\epsilon \approx 10^{-9}$

→ If FP is coorect

→ use Gradient check to chell BP.

$J(U, V, W, X, b)$

$\frac{1}{2}($

$J(\theta) = \frac{1}{2}\theta - \frac{3}{5}$

$\nabla J(\theta) = \frac{1}{2}$.

$$g(\theta) = \frac{J(\theta+\epsilon) - J(\theta-\epsilon)}{2\epsilon}$$

~~$= J($~~

$J(\theta+\epsilon) = \frac{1}{2}\theta + \frac{1}{2}\epsilon - \frac{3}{5}$

$= \frac{1}{2}\theta + \frac{10^{-9}}{2} - \frac{3}{5}$

$J(\theta-\epsilon) = -\frac{1}{2}\theta + \frac{10^{-9}}{2} + \frac{3}{5}$

$= \frac{10^{-9}}{\frac{2}{2 \times 10^{-9}}} * = \frac{1}{2} 10^{-9} \times \frac{1}{2 \times 10^{-9}}$

$= 1/4$.

$\theta = [\theta_1 \quad \theta_2]$.

$J(\theta) = \frac{1}{2}\theta - \frac{1}{3}$.

$\nabla J(\theta) =$

$\frac{\partial}{\partial \theta_i} J(\theta) = \frac{J(\theta_i + \epsilon) - J(\theta_i - \epsilon)}{2\epsilon}$

$\frac{\partial}{\partial \theta_i} J(\theta) \quad [\theta_1 \quad \theta_2] \quad [\epsilon \quad 0] = \epsilon_i$

$\epsilon * \epsilon_i$

$\therefore \theta^{(i+)} = \theta + \epsilon \vec{e}_i$

$\frac{\partial}{\partial \theta_i} J(\theta) = \frac{J(\theta^{i+}) - J(\theta^{i-})}{2\epsilon}$

$\theta^{(i-)} = \theta - \epsilon \vec{e}_i$

$\frac{\partial}{\partial \theta_i} J(\theta) = \frac{J(\theta^{i+}) - J(\theta^{i-})}{2\epsilon}$

$J(\theta)$.

$\frac{\partial}{\partial \theta} J(\theta) =$ ~~$J(\theta+$~~

$\frac{\partial}{\partial \theta} J(\theta) = \lim_{\theta \to 0} \frac{J(\theta+\epsilon) - J(\theta-\epsilon)}{2\epsilon}$

$\theta = [\theta_1 \ \theta_2 \ \theta_3 \ \cdots \ \theta_n]$

$\theta \in \mathbb{R}^{n*1}$

$\frac{\partial}{\partial \theta_i} J(\theta) =$

$\theta^{i+} = \theta + e_i * \epsilon$

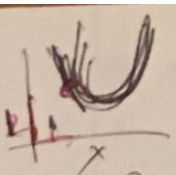
$\theta^{i-} = \theta - e_i * \epsilon$

$\frac{\partial}{\partial \theta_i} J(\theta) = \lim_{\epsilon \to 0} \frac{J(\theta^{i+}) - J(\theta^{i-})}{2\epsilon}$

chore $\epsilon \approx 10^{-4}$

$\frac{\partial}{\partial \theta_i} J(\theta) = \frac{J(\theta^{i+}) - J(\theta^{i-})}{2\epsilon}$

$\nabla_{\theta_i} J(\theta) \approx \frac{\partial}{\partial \theta_i} J(\theta)$



$J(\theta) = x^2$

$y = x^2$

$\nabla J(\theta) = 2x.$ $\frac{dy}{dx} = 2x = 2(1) = 2.$

$J(\theta) = h_\theta(x$

$J(\theta) = \sum_{i=1}^{m} (h_\theta(x^i) - y^i)^2$ $\quad err^i$

$[1 \ 2]$

$J(\theta) = cost = sum([1 \ 4])$

$= 5.$

$\nabla J(\theta) = grad = 2 \ err^i$

$= 2([1 \ 2])$

$= 2 \times 3 = 6$

$= [2 \ 4]$

$(5, [2 \ 4])$

$[1, 2]$

$\downarrow$

$\downarrow$

$\downarrow$

$(5, [2 \ 4])$

$f(\theta) = \theta^2$; $\frac{f(\theta+\epsilon) - f(\theta+\epsilon)}{2\epsilon}$

[1 2]

↓

b

↓

(5 , [1 4])

[1 2]

↓

∇

↓

$2 \cdot 10^{-4}$

123.456

$Y = x^2$; $\frac{\partial Y}{\partial x} = 2x$

$Y = 15,241.383$; $\nabla Y = 246.912$.

~~x = [2.3.c~~

$x = [123.456]$.

~~Cost = f(x)~~ cost, grd = f(x)

f(x)

↓ ↓

Cost = 15,241.383 grad = [246.912]

[1, 2]

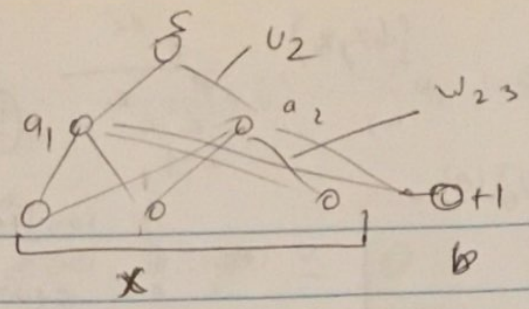
↓

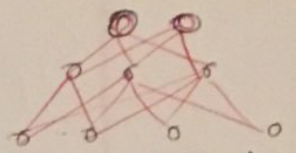
f(x)

cost = 5 grad = [2, 4]

↑ ↑

g(x)





$\frac{ds}{dW}$; $\frac{ds}{dW}$; $\frac{ds}{dx}$; $\frac{ds}{db}$.

$$s = U^T a = U^T f(z) = U^T f(z_1, z_2)$$

$$s = U^T \left(f(z_1)^{th} \quad f(z_2)^{th} \right)$$

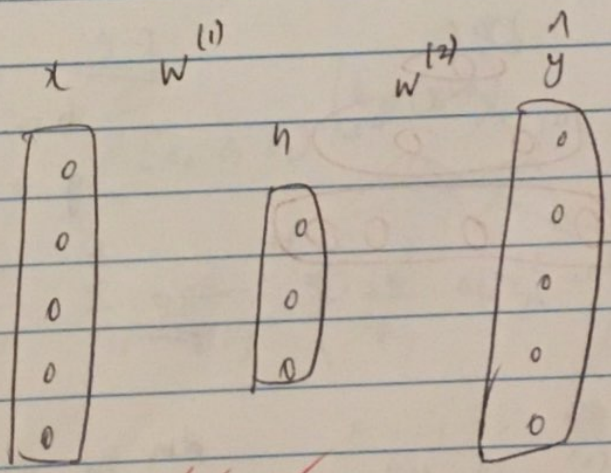
$$s = U^T \left(f(W_1^T x)^{th} \quad f(W_2^T x)^{th} \right)$$

$$s = U^T \left(\sigma(W_1^T x)^{th} \quad \sigma(W_2^T x)^{th} \right)$$

$$\boxed{\sigma(z) = 1/1+e^{-z}}$$

$(D_x+1 * H)$
$+ (D_H+1 * D_y)$

$s = U^T a$



$h = \text{sigmoid}(W^{(1)} x + b^{(1)})$

$D_H * D_x \cdot D_x \times 1$

$D_H * 1 + D_H * 1$

$D_H * 1$

$\hat{y} = \text{softmax}(W^{(2)} h + b^{(2)})$

$D_y * D_H \cdot D_H * 1$

$D_y * 1 + D_y * 1$

$D_y * 1$

$\sigma(z) = \frac{1}{1+e^{-z}}$ $\quad f(x, x_i)$ $\quad \frac{e^{x_i}}{\sum_{i=1}^{d} e^{x_i}}$ (+1, −1).

Repeat {

$\theta = \theta + \eta \frac{\partial}{\partial \theta} J(\theta)$

}

$\theta = \theta + \eta \nabla_\theta J(\theta)$

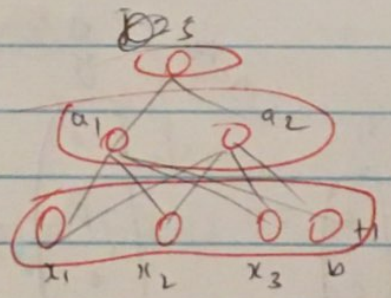
y_i	$\hat{y}_i$	$y_i \log(\hat{y}_i)$
1	0	$\approx +\infty$
1	1	≈ 0
0	0	$\approx$
0	1	$\approx$

$$\frac{\partial}{\partial(x,y)} f(x,y) = \frac{\partial}{\partial(x)} f(x,y) + \frac{\partial}{\partial y} f(x,y)$$

$$\frac{\partial s}{\partial(U, \theta, W, x, b)} = \frac{\partial s}{\partial U} + \frac{\partial s}{\partial W} + \frac{\partial s}{\partial x} + \frac{\partial x}{\partial b}.$$

Forward: σ, s.

Backprop:



FP:

$h = \text{Sigmoid}(W^{(1)} x + b^{(1)})$

$y = \text{Softmax}(W^{(2)} h + b^{(y)})$

BP:

$$J(W^{(1)}, W^{(2)}, b^{(1)}, b^{(2)}) = -\sum_{i=1}^{dm} y^{(i)} \log(\hat{y}^{(i)})$$

$$LE = -\quad CE(y, \hat{y}) = -\sum_{i=1}^{m} y_i \log(\hat{y}_i)$$

$$-\infty \leq \log(x) \leq 0 \qquad x \to 1$$

$$0 \leq x \leq 1 \qquad \log$$

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$$

$$\log(x) \to 0 \quad ; x \to 1$$

$$\log(x) \to -\infty \quad ; x \to 0$$

---

* Needed: $\text{Cat}, \text{grad}(W_1, W_2, b_1, b_2)$.

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i)^2$$

$$\frac{d}{dx} x^2 \qquad 2x \frac{d}{dx}(xy)$$

$$\nabla J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \frac{d}{d\theta} (h_\theta(x^i) - y^i)$$

$$= \frac{1}{m} \sum_{i=1}^{m} \frac{d}{d\theta} (h_\theta(x^i)) - 0$$

$$\theta_0 x_0 \quad \vdots \quad \theta_j x_j^i$$

$$= \frac{1}{m} \sum_{i=1}^{m} \frac{d}{d[\theta_0\ \theta_1\ \theta_2]} \left(h_\theta(x^i)\right)$$

$$= \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{d} \frac{\partial}{\partial \theta_j} \left(\sum_{k=1}^{d} \theta_k x_k^i \right)$$

$$k = j \; ; \; x_j^i$$

$$= \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{d} x_j^i$$

$\frac{d}{d(x,y)} f(x,y)$ $\frac{\partial}{\partial x} f(x,y) + \frac{\partial}{\partial y} f(x,y)$.

Repeat

{

$$\theta = \theta + \alpha \nabla J(\theta)$$

}

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (h_\theta(x^i) - y^i)^2$$

$$\frac{\partial}{\partial \theta} J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \frac{\partial}{\partial \theta} (h_\theta(x^i) - y^i)$$

$$= \sum_{i=1}^{m} \frac{d}{d\theta} (h_\theta(x^i)) \; (h_\theta(x^i) - y^i)$$

$$= \sum_{i=1}^{m} \frac{d}{d\theta} \sum_{j=1}^{d} \theta_j x_j^i$$

$$= \sum_{i=1}^{m} \left[\frac{d}{d\theta_1} \sum_{j=1}^{d} \theta_j x_j^i \quad \frac{d}{d\theta_2} \sum_{j=1}^{d} \theta_j x_j^i \quad \cdots \quad \frac{d}{d\theta_d} \theta_j x_j^i \right]$$

$$= \sum_{i=1}^{m} [x_1^i \; x_2^i \; \cdots \; x_j^i]$$

$$= \sum_{i=1}^{m} x^i \; (h_\theta(x^i) - y^{(i)})$$

$$\therefore \theta = \theta + \alpha \sum_{i=1}^{m} x^i \; (h_\theta(x^i) - y^{(i)})$$

$$\theta_j := \theta_j + \alpha \nabla_j J(\theta)$$

$$\theta_j := \theta_j + \alpha \sum_{i=1}^{m} x_j^i \left(h_\theta(x^i) - y^i\right)$$

$$\theta := \theta + \alpha \sum_{i=1}^{m} x^i \left(h_\theta(x^i) - y^i\right).$$

$$\theta := \theta + \alpha$$

$$\begin{bmatrix} x_1^i & x_2^i \\ 4 & 2 \end{bmatrix} \quad J(W_1, W_2, b_1, b_2) =$$

Repeat.

{

$$[W_1, W_2, b_1, b_2] := [W_1, W_2, b_1, b_2] + \eta \nabla J(W_1, W_2, b_1, b_2)$$

}

$$\nabla J(W_1, W_2, b_1, b_2) = \nabla_{W_1} J(W_1, W_2, b_1, b_2) + \nabla_{W_2} J(\text{params})$$

$$+ \nabla$$

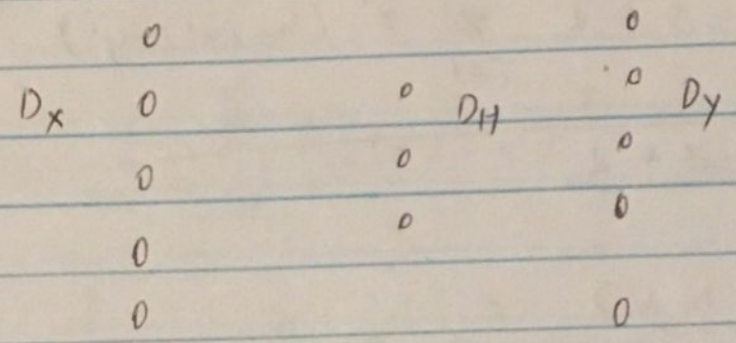
$$\frac{\partial}{\partial W^{(l)}} \delta = \delta^{(l+1)} \cdot a^{(l)^T}$$

$$\delta^{(l)} = \left(W^{(l)^T} \cdot \delta^{(l+1)}\right) \odot f'\left(z^{(l)}\right)$$

$$\frac{\partial}{\partial b^{(l)}} \delta = \delta^{(l)}$$

Needed: $\frac{d}{dW^{(1)}} \quad \frac{d}{dW^{(2)}} \quad \frac{d}{db^{(1)}} \quad \frac{d}{db^{(2)}} = ?$

cost = ?



No of params: $((D_x+1) * D_H) + (D_y * (D_H+1))$

$$\text{cost}(y, \hat{y}) = -\sum_{i=1}^{m} y_i \log(\hat{y}_i)$$

Can:

Forward: σ, Softmax

Backprop:

$\frac{d}{dW^{(1)}} \cdots \frac{d}{db^{(2)}}$.

Cannot:

Cost: $\hat{y}_i$

1, 2, 3. ... n

$x + 2x + 3x = n \quad n \geq 1$

Abstract solution:

Gweights, Gbiases = backprop();

cost = $-\sum_i y_i \cdot \log(\hat{y}_i)$ →

return (cost, Gweights, Gbiases)

$\hat{y}_i$ = Softmax ($W_2^T h + b_2$)

$D_y * D_H \cdot D_H \times 1$ | $D_y \times 1$

$D_y \times 1$

$D_y * 1$

$\hat{y}_i$ = softmax (β)

$D_y \times 1$

$D_y * 1$

$\sigma(W^{(1)} x + b^{(1)})$

$a^{(1)} = \sigma(W^{(1)T} x + b^{(1)})$

$y_i = [0 \quad 1]$

$\hat{y}_i = [0.09 \quad 0.01]$ (ε = small.

$0 + \log(0.01)$

initialize network weights

do

forEach ex: train example;

pred = forward(nn, ex)

actual = label(ex)

error = actual − pred

Compute ΔW_{HO}

Compute ΔW_{IH}

update n/w weights

until params wont change

return nn .

$a^{(3)}$

$W^{(2)}$

$W^{(1)}$ $a^{(2)}$ $y^{(2)}$

$x = a^{(1)}$ $b^{(1)}$

$a^{(1)}, a^{(2)}$

$\frac{d}{dW^{(1)}}$