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

~~$\frac{\partial E}{\partial w_{jk}}$~~

$$\frac{\partial E}{\partial w_{jk}} = \underbrace{\frac{\partial E}{\partial y_k} g'(z_k)}_{\delta_k} y_j$$

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_j} \frac{\partial y_j}{\partial w_{ij}} = \underbrace{\frac{\partial E}{\partial y_k} g'(z_k) \overset{\sum_j}{w_{jk}}}_{\delta_k} g'(z_j) y_i$$

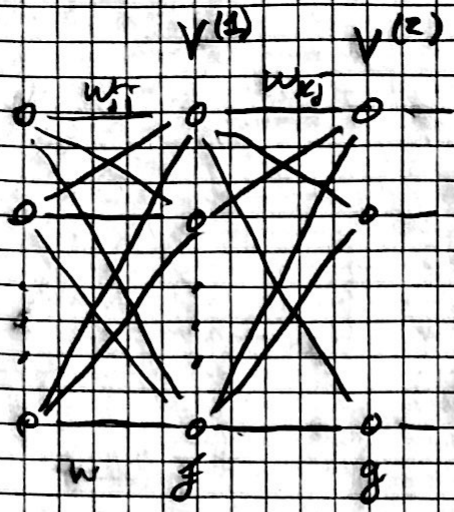
$$= \underbrace{\delta_k w_{jk} g'(z_j)}_{\delta_j} y_i = \delta_j y_i$$

$$\frac{\partial E}{\partial w_{jk}} = \frac{\partial E_{classification}}{\partial w_{jk}} + \frac{\alpha}{2} \frac{\partial \|w\|^2}{\partial w_{jk}}$$

$$E = E_{class} + \alpha \frac{1}{2} \|w\|^2$$

$$\frac{\alpha}{2} \frac{\partial \|w\|^2}{\partial w_{jk}} = \alpha \|w_{jk}\|$$

$$\|w\| = \sum \left(\sqrt{w_i} \right)^2$$

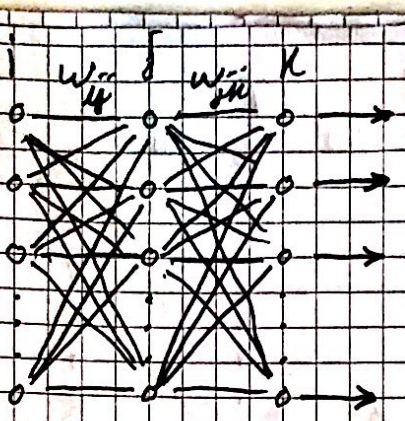


$V^{(2)} \equiv$ len hid j

$V^{(2)} \equiv$ len out k

$$V^{(1)} = f(h^{(1)}) \qquad V^{(2)} = g(h^{(2)})$$

$$\frac{\partial E}{\partial w_{kj}} = C \cdot V^{(2)}(1 - V^{(2)})\, V^{(1)}$$



w_{ij} = input_to_hid 7×256 M

w_{jk} = hid_to_class 10×7 M

err = class_prob − class_target

y_j = hid_output

y_k = class_prob

Wd_coefficient = ~~learning_rate~~ weight decay

model_to_chem $\begin{bmatrix} 11 & 12 & \cdots \\ 21 & \vdots & \vdots \\ 31 & \cdots & \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \end{bmatrix}$

grad should return

$\delta_i^m V_j^{m-1}$ in model form

$[\Sigma_k^{\mu} - O_k^{\mu}] \underbrace{g'(h_k^{\mu})}_{\text{class\_prod}} V_j^{\mu}$

model.hid_to_class

δ_k = class_grad_dk_temp

$\delta_j^{\mu} = g'(h_j^{\mu}) \sum_k w_{jk}^{ij} \delta_k$

$f(x)(1-f(x))$

$[\Sigma^{n} - O^{n}]$ err

$g($

err = (class_prob - dum.target)

cost = transpose (err) * model.fitd_to_class

2.4

a)

- Validation data classification cost
  0,4302

- b) Validation data classification cost
  0,3345

c)

10	22,6128
1	2,3026
0,0001	0,3483
0,001	0,2879
5	2,3026

d)

10	0,3708	0,4217
30	0,2868	0,3171
100	0,3021	0,3686
130	0,3169	0,3876
200	0,3345	0,4302

c)

18	0,3061
37	0,2652
83	0,3112
113	0,3137
236	0,3438

f)

n_hid = 37
learning rate = 0,35
momentum = 0,9
early stop = true
mini_batch = 100

Wd = 0

test cost 0,2825

Wd = 0,001

test cost 0,2571

From over derivation order of

cross entropy

$$\frac{\partial H}{\partial o_i} = -\sum_k y_k \frac{\partial \log(p_k)}{\partial o_i}$$

$$= -\sum_k y_k \frac{\partial \log(p_k)}{\partial p_k} * \frac{\partial p_k}{\partial o_i} =$$

$$= -\sum_k y_k \frac{1}{p_k} * \frac{\partial p_k}{\partial o_i}$$

now serving p_k as the soft max function we get

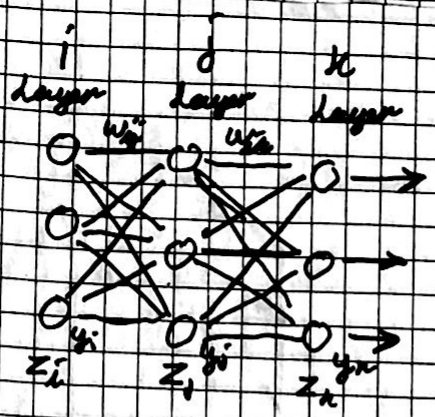
$$-y_i(1-p_i) - \sum_{k \neq i} y_k \frac{1}{p_k}(-p_k p_i)$$

$$= -y_i + y_i p_i - \sum_{k \neq i} y_k p_i$$

$$= p_i \sum_k y_k - y_i$$

$$\sum_k y_k = 1 \Rightarrow$$

$$p_i - y_i$$



y_i, z_k

$y_i = g_i(z_i)$

$z_j = \sum_i w_{ij} y_i$

$y_j = g(z_j)$

$z_k = \sum_j w_{jk} y_i$

$y_k = g(z_k)$

$g(z)$ is some differentiable function

incoming error derivate

$\frac{\partial E}{\partial y_k}$

$\frac{\partial E}{\partial z_k} = \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial z_k} \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial z_k}$

$\frac{\partial E}{\partial w_{jk}} = \frac{\partial E}{\partial z_k} \sum_j \frac{\partial z_k}{\partial w_{jk}} y_j$

(a) $\frac{\partial E}{\partial z_k} =$

b) $\frac{\partial E}{\partial z_j}$

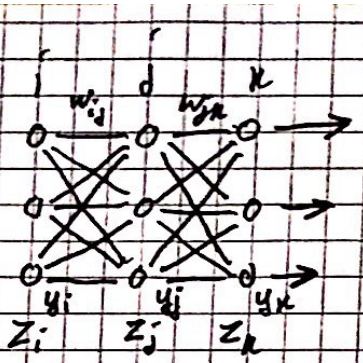
c) $\frac{\partial E}{\partial w_{jk}}$

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

$z_k = W_{jk}^T \cdot y_j$ $\quad \frac{\partial E}{W_{jk}^T} = \frac{\partial E}{\partial z_k} \frac{\partial z_k}{\partial W_{jk}^T}$

$\frac{\partial z_k}{\partial W_{jk}} = \frac{\partial (W_{jk} y_j)}{\partial W_{jk}} = y_j$

$\frac{\partial E}{W_{jk}^T} = y_j \frac{\partial E}{\partial z_k}$ $\quad \frac{\partial E}{W_{jk}^T} = y_j \frac{\partial E_k}{\partial z_k}$



$y_i = g(z_i)$

$z_j = \sum_i w_{ij} y_i$

$y_j = g(z_j)$

$z_k = \sum_j w_{jk} y_j$

$y_k = g(z_k)$

$y_i = g(\sum_j$

$z_j = \sum_i w_{ij} y_i = \sum_i w_{ij} g(z_i)$

$y_j = g(z_j) = g(\sum_i w_{ij} g(z_i))$

$z_k = \sum_j w_{jk} y_j = \sum_j w_{jk} g(\sum_i w_{ij} g(z_i))$

$y_k = g(\sum_j w_{jk} g(\sum_i w_{ij} g(z_i))$

$\frac{\partial E}{\partial y_k} = C$

$\frac{\partial E}{\partial z_k} = \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial z_k}$

$\frac{\partial E}{\partial w_{jk}} = \sum_\mu$

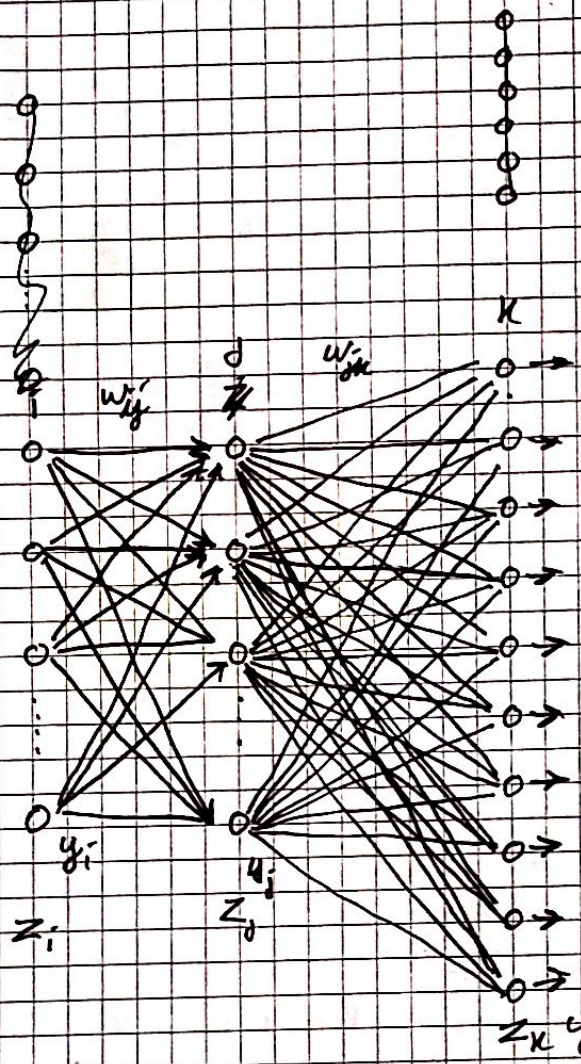
$\frac{\partial E}{\partial z_j} = \frac{\partial E}{\partial y_k} \frac{\partial z_k}{\partial y_j} \frac{\partial y_j}{\partial z_j}$

$\frac{\partial E}{\partial z_j} = C\, g'(z_k) \sum_j w_{jk}\, g'(z_j)$

$\frac{\partial E}{\partial w_{jk}} = \frac{\partial E}{\partial y_k} g'(z_k) \cdot y_j$

$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_j} \frac{\partial y_j}{\partial w_{jk}}$

$\frac{\partial E}{\partial y_j} = \frac{\partial E}{\partial z_j}$



$$y_i = z_i$$

$$z_j = \sum_i w_{ij} z_i$$

$$y_j = \frac{1}{1+e^{-kz_j}}$$

$$z_k = \sum w_{jk} y_j$$

$$y_k = z_k$$

$$E = E_{classification} + \alpha E_{weight\ decay}$$

$$\frac{\partial E_{classification}}{\partial z}$$

$$\frac{\partial E}{\partial z} = \frac{\partial E_{classification}}{\partial z} + \alpha \left\| \frac{\partial E_{classification}}{\partial z} \right\|$$

$$y_j = \frac{1}{1+e^{-kz_j}}$$

$$y_j' = \frac{ke^{kz_j}}{(1+e^{kz_j})^2}$$

$$\delta_i^M = g'(h_i^M)(\xi_i^M - V_i^M)$$

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

$$\frac{\partial E}{\partial y_k} = g'(z_k)[\xi_k^M - V_k]$$

$$\frac{\partial E}{\partial y_j} = g'(z_i)\sum w_{jk}\frac{\partial E}{\partial y_k}$$