Chain rule

- Single variable.

If $z = f(y)$

& $y = g(x)$.

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}.$$

- Multiple variables

$z = f(y_1 \cdots y_n)$.

$y_1 = g_1(x)$

$\vdots$

$y_n = g_n(x)$.

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y_1}\frac{\partial y_1}{\partial x} + \frac{\partial z}{\partial y_2}\frac{\partial y_2}{\partial x} + \cdots \frac{\partial z}{\partial y_n}\frac{\partial y_n}{\partial x}.$$

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

$$E = \frac{1}{2}\sum_{j=1}^{M} (y_j - t_j)^2$$

Each

$$y_j = \sigma(u_j').$$

$$u_j' = \sum_{i=1}^{N} w_{ij}' h_i$$

$$h_i = \sigma(u_i).$$

$$u_i = \sum_{k=1}^{K} w_{ki} x_k$$

7

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial u_j}\frac{\partial u_j'}{\partial w_{ij}}$$ (Because other y variables are not involved.)

$$\frac{\partial E}{\partial y_j} = (y_j - t_j)$$

$$\frac{\partial y_j}{\partial u_j'} = y_j(1-y_j).$$

$$\frac{\partial u_j'}{\partial w_{ij}} = h_i$$

$$w_{ij}'^{(new)} = w_{ij}'^{(old)} - \eta \frac{\partial E}{\partial w_{ij}}$$

$$\boxed{w_{ij}'^{(new)} = w_{ij}'^{(old)} - \eta\, EI_j'\, h_i}$$ 80

Also $$\frac{\partial E}{\partial u_j'} = \underbrace{(y_j - t_j)(y_j)(1-y_j)}_{EI_j'}$$

$$\frac{\partial E}{\partial w_{ki}} = \frac{\partial E}{\partial h_i} \frac{\partial h_i}{\partial u_i} \frac{\partial u_i}{\partial w_{ki}}$$

$$= \frac{\partial E}{\partial h_i} h_i(1-h_i) x_k$$

$$\frac{\partial E}{\partial h_i} = \sum_{j=1}^{M} \frac{\partial E}{\partial u'_j} \frac{\partial u'_j}{\partial h_i}$$

$$= \sum_{j=1}^{M} EI'_j \, w'_{ij}$$

So $$\frac{\partial E}{\partial w_{ki}} = \underbrace{\left(\sum_{j=1}^{M} EI'_j w'_{ij}\right)(h_i)(1-h_i)}_{EI_i} x_k$$

$$w_{ki}^{(new)} = w_{ki}^{(old)} - \eta \frac{\partial E}{\partial w_{ki}}$$

$$\Rightarrow \boxed{w_{ki}^{(new)} = w_{ki}^{(old)} - \eta \, EI_i \, x_k}$$

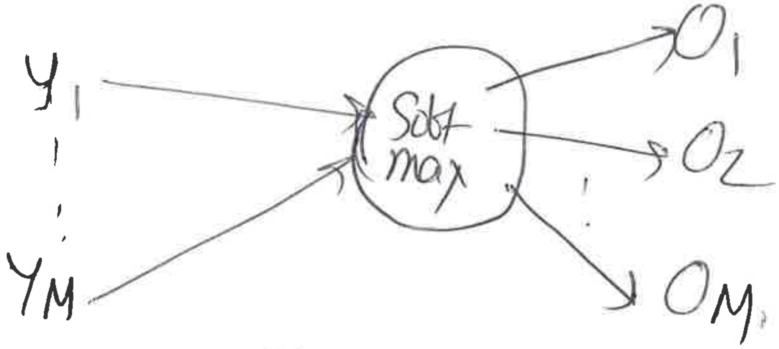
84

Softmax

— If we want a classifier we may compute the probability of various possible output values.

— Sigmoid fn doesn't give us that. (Its input-output).

— To make the output a probability we use a softmax module.



$$O_i = \frac{e^{y_i}}{\sum_{j=1 \cdots M} e^{y_j}}$$ — normalization.

- $y_1 \cdots y_M$ – may have -ve values (if we don't have sigmoid before)
  ~~– may not be~~ or between 0..1
  – may not add up to 1.

Softmax guarantee each O_i is betwee 0 & 1 & their sum add up to 1.