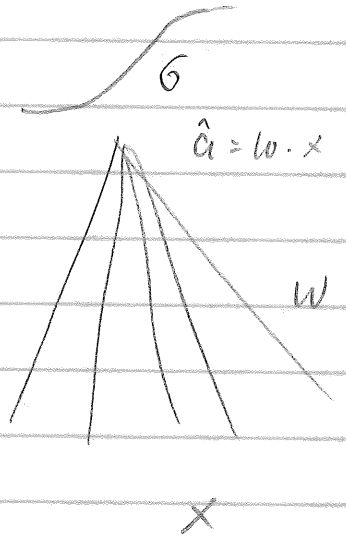
8 BACKPROP

SO FAR $\vec{y} = \sigma(w \cdot x)$

chain rule



matching loss for sigmoid

$$L_y(w \cdot x) = \frac{1}{2}(\sigma(w \cdot x) - y)^2$$

$$L_y(w \cdot x) = \ln(1 + e^{w \cdot x}) - w \cdot x\, y$$

$$\frac{\partial L_y(\hat{a})}{\partial \hat{a}} = \underbrace{(\sigma(\hat{a}) - y)\,\sigma'(\hat{a})}_{:= \delta(\hat{a})}$$

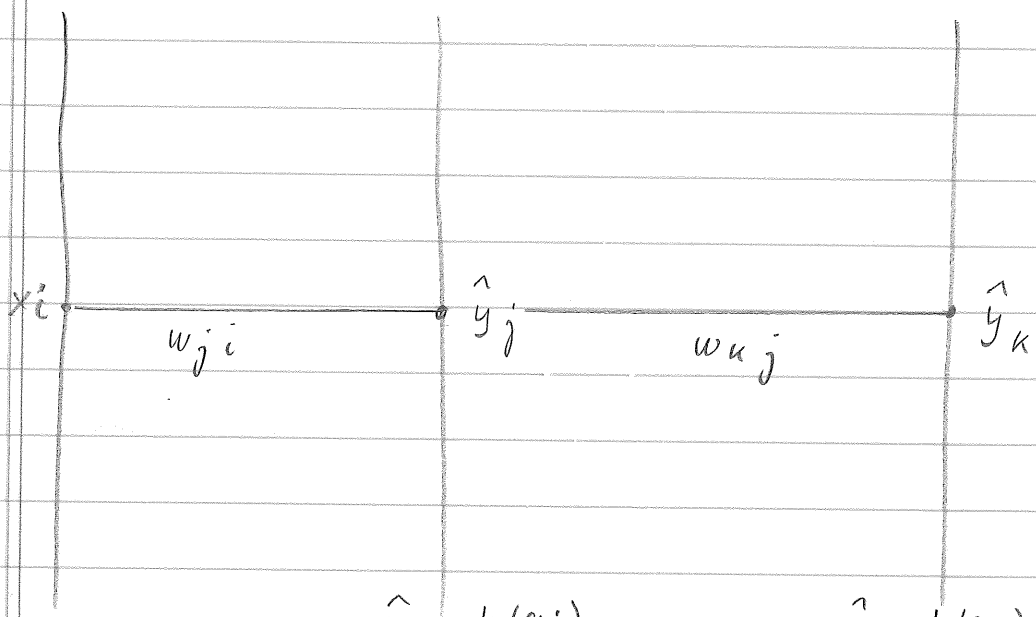
$$\delta(\hat{a}) = (\sigma(\hat{a}) - y)$$

$$\frac{\partial L_y(\hat{a})}{\partial w_i} = \delta(\hat{a})\, x_i$$

Chain rule for computing derivs

$f(x) = \sin(x^2)$ $\quad f'(x) = \cos(x^2)\, 2x$

$f(x) = \sin e^{-x^2}$ $\quad f'(x) = (\cos e^{-x^2})(e^{-x^2})(-2x)$



$\hat{y}_j = h(a_j)$ $\quad$ $\hat{y}_k = h(a_k)$ $\quad$ INITIALLY IDENTITY OUTPUT

$a_j = \sum_i w_{ji} x_i$ $\quad$ $a_k = \sum_j w_{kj} \hat{y}_j$

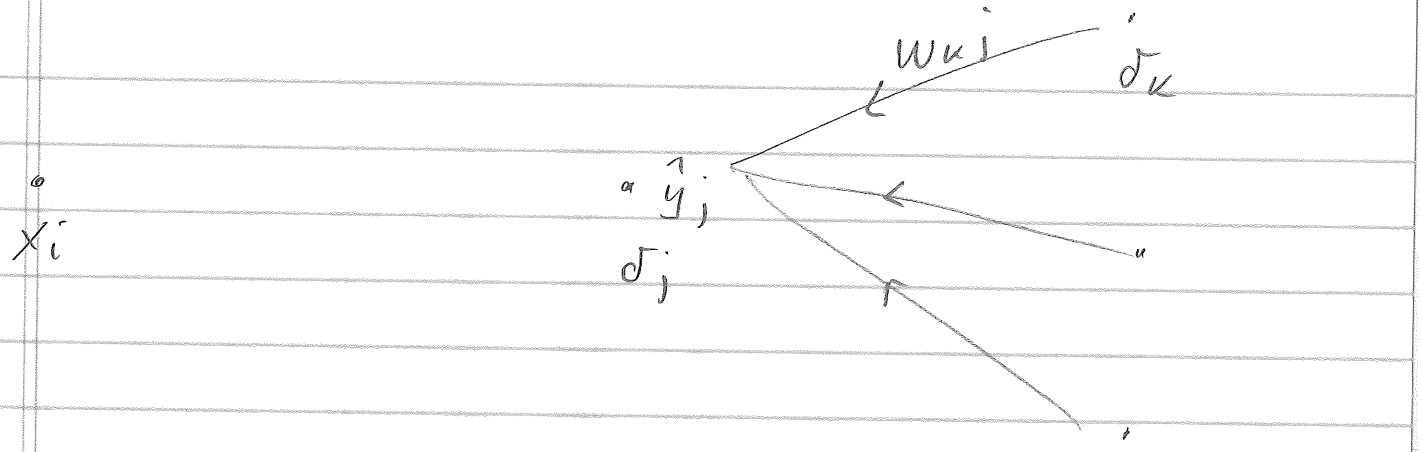
$$\frac{\partial L}{\partial w_{ji}} = \underbrace{\frac{\partial L}{\partial \hat{a}_j}}_{:= \delta_j} \underbrace{\frac{\partial \hat{a}_j}{\partial w_{ji}}}_{x_i}$$

FOR BIAS VARS INPUTS ARE 1: $\frac{\partial L}{\partial b_j} = \delta_j$

$$\frac{\partial L(\dots \hat{a}_k \dots)}{\partial \hat{a}_j} = \sum_k \frac{\partial L}{\partial \hat{a}_k} \frac{\partial \hat{a}_k}{\partial \hat{a}_j}$$

$$= \sum_k \delta_k \frac{\partial \sum_{j'} w_{kj'} \hat{y}_{j'}}{\partial \hat{a}_j}$$

$$= \sum_k \delta_k \, w_{kj} \, h'(a_j) = h'(a_j) \sum_k \delta_k w_{kj}$$



Basic Backprop templet stays the same
But

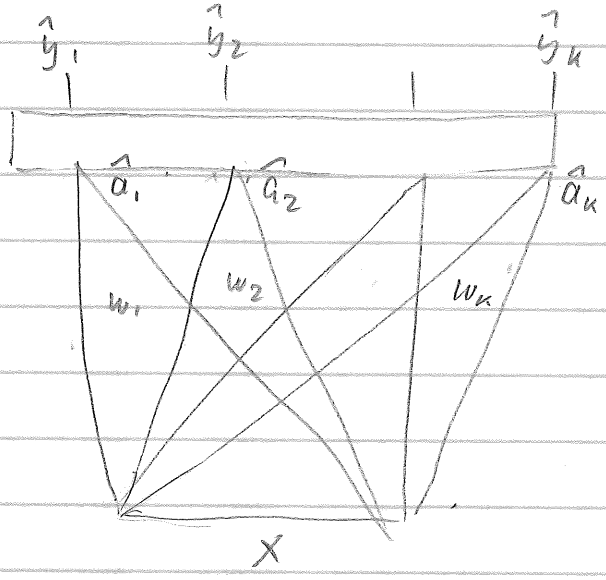
- different transfer function
  provide different local derivatives

- different losses
  change update at output layer

QUESTIONS:

WHAT ARE THE BEST
TRANSFER FUNCTIONS &
LOSS FUNCTIONS
FOR NEURAL NETS

MULTICLASS SIGMOID (SOFTMAX)



$$\hat{a}_i = w_i \cdot x$$

$$\hat{y}_i = \frac{e^{\hat{a}_i}}{\sum_j e^{\hat{a}_j}}$$

MULTICLASS LOGISTIC LOSS

$$RE([y_1, \dots, y_n], [\hat{y}_1, \dots, \hat{y}_n])$$

$$= \sum y_i \ln \frac{y_i}{\hat{y}_i}$$

c

$$RE([0, 0, 1, 900], [\hat{y}_1, \dots, \hat{y}_k])$$

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

$k = 2$

$$\hat{y} = \left[\frac{e^{\hat{a}_1}}{e^{\hat{a}_1} + e^{\hat{a}_2}} , \frac{e^{\hat{a}_2}}{e^{\hat{a}_1} + e^{\hat{a}_2}} \right]$$

$$\hat{a}_1 = w_1 \cdot x, \quad \hat{a}_2 = w_2 \cdot x$$

$k-1$ WEIGHT VECTORS SUFFICE

$$\hat{y} = \left[\frac{1}{z}, \frac{e^{\hat{a}_2 - \hat{a}_1}}{z}, \ldots, \frac{e^{\hat{a}_k - \hat{a}_1}}{z} \right]$$

WHERE $z = 1 + \sum_{i=2}^{k} e^{\hat{a}_i - \hat{a}_1}$

$$\hat{a}_i - \hat{a}_1 = \underbrace{(w_i - w_1)}_{\tilde{w}_i} \cdot x \qquad 2 \le i \le k$$

$$\frac{\partial RE(y, \hat{y})}{\partial w_i} = (\hat{y}_i - y_i)\, x$$

$RE(y, \hat{y})$ IS MATCHING LOSS FOR MULTICLASS SIGMOID