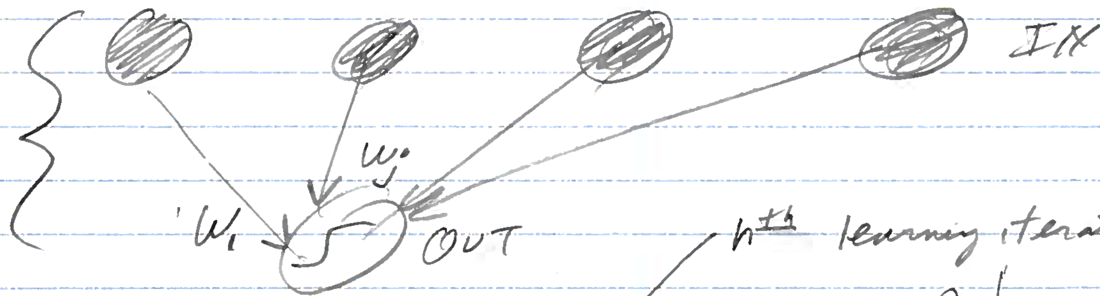
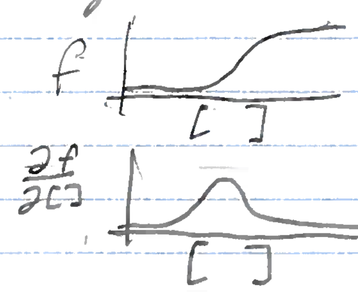
David Kleinfeld
10 March 2016
Preliminary notes on two layer network

Perceptron "Warm Up"



n^{th} learning iteration



output $\qquad \hat{y}(n) = f\left[\sum_j w_j(n)\, x_j(n)\right]$

error $\qquad \mathcal{E}(n) = \overset{\text{Desired}}{y(n)} - \hat{y}(n)$

Square error on n^{th} iteration $\qquad E(n) = \frac{1}{2}\left[y(n) - \hat{y}(n)\right]^2$

Change in error w.r.t. change in weight

$$\frac{\partial E(n)}{\partial w_j} = -\left[y(n)-\hat{y}(n)\right]\frac{\partial \hat{y}(n)}{\partial w_j(n)}$$

$$\text{"} \quad = -\left[y(n)-\hat{y}(n)\right]\frac{\partial \hat{y}(n)}{\partial [\text{arg.}]}\,\frac{\partial [\text{arg.}]}{\partial w_j}$$

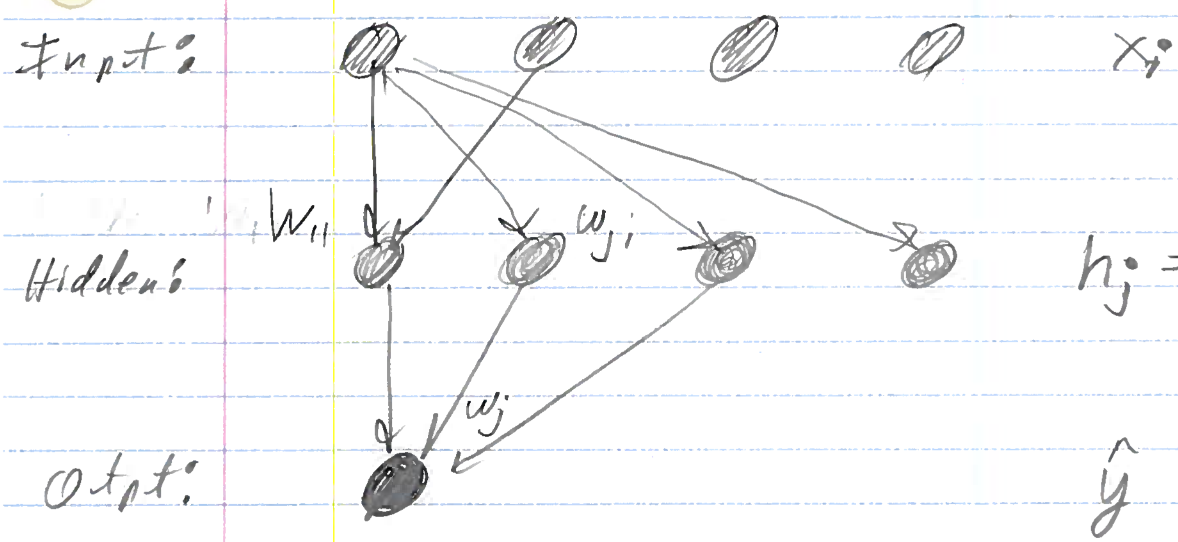
$$\text{"} \quad = -\left[y(n)-\hat{y}(n)\right]\hat{y}(n)\left[1-\hat{y}(n)\right] x_j(n)$$

Learning Rule $\qquad \Delta w_{ij}(n) = -\eta \dfrac{\partial E(n)}{\partial w_j(n)}$

$$\text{"} \quad = +\underbrace{\eta\, \hat{y}(n)\left[1-\hat{y}(n)\right]}_{\text{weight sensitivity}}\ \underbrace{\left[y(n)-\hat{y}(n)\right]}_{\text{error}}\ \underbrace{x_j(n)}_{\text{input}}$$

Repeat Δw_{ij} over a set of patterns $(\vec{x}, y)$

Two layer Network (Single output)



$$h_j = f\left[\sum_i w_{ji} x_i\right]$$

$$\hat{y} = f\left[\sum_j w_j h_j\right]$$

Error @ output: $\epsilon(n) = \underset{\text{desired}}{y(n)} - \underset{\text{actual}}{\hat{y}(n)}$

$$\hat{y}(n) = \sum_j w_j(n)\, h_j(n)$$

$$'' \quad = f\left[\sum_j w_j\, f\left(\sum_i w_{ji} x_i\right)\right] \qquad \text{Dropping "n"}$$

$$E(n) = \tfrac{1}{2}\epsilon^2(n) = \tfrac{1}{2}\left[y(n) - \hat{y}(n)\right]$$

Keep in mind $\dfrac{\partial f[\cdot]}{\partial[\cdot]} = f[\cdot]\left(1 - f[\cdot]\right)$

Note:

Linear Hidden layer $\Longrightarrow$ Same as Perceptron

linear hidden piece

$$f\left[\qquad\qquad\right]$$

$$f\left[\sum_i x_i \sum_j w_j w_{ji}\right]$$

$$f\left[\sum x_i\, w_i'\right]$$

Output layer

$$\frac{\partial E(n)}{\partial w_j} = \frac{\partial E(n)}{\partial \epsilon(n)} \frac{\partial \epsilon(n)}{\partial w_j}$$

$$= -[y(n) - \hat{y}(n)] \frac{\partial \hat{y}(n)}{\partial w_j}$$

$$[\cdot] = \sum_j w_j h_j(n)$$

$$= -[y(n) - \hat{y}(n)] \frac{\partial \hat{y}(n)}{\partial [\cdot]} \frac{\partial [\cdot]}{\partial w_j}$$

$$= \underbrace{-[y(n) - \hat{y}(n)]\, \hat{y}(n)\, [1 - \hat{y}(n)]}_{\text{output term} \,\equiv\, \delta(n)} \underbrace{h_j(n)}_{\substack{\text{input from} \\ \text{hidden layer}}}$$

$$\frac{\partial E(n)}{\partial w_{ij}} = \frac{\partial E(n)}{\partial \epsilon(n)} \frac{\partial \epsilon(n)}{\partial w_{ji}}$$

$$= -[y(n) - \hat{y}(n)] \frac{\partial \hat{y}(n)}{\partial w_{ji}}$$

$$\frac{\partial \left[\sum_j w_j \overbrace{f\left(\sum_i w_{ji} x_i\right)}^{h_j(n)}\right]}{\partial w_{ji}} \qquad w_j \frac{\partial h_j(n)}{\partial [\cdot]} \frac{\partial [\cdot]}{\partial w_{ji}}$$

$$= -[y(n) - \hat{y}(n)] \underbrace{\frac{\partial \hat{y}(n)}{\partial [\cdot]}}_{\hat{y}(n)[1 - \hat{y}(n)]} \underbrace{\frac{\partial [\cdot]}{\partial w_{ji}}}_{w_j h_j(n)(1 - h_j(n)) x_i}$$

$$= \underbrace{-[y(n) - \hat{y}(n)]\, \hat{y}(n)\, [1 - \hat{y}(n)]}_{\delta(n),\ \text{a non-local term from output}}\, w_j(n)\, \underbrace{h_j(n)\, [1 - h_j(n)]}_{\substack{\text{local} \\ \text{output}}}\, \underbrace{x_i(n)}_{\substack{\text{local} \\ \text{input}}}$$

$$\therefore \Delta w_j(n) = -\eta \frac{\partial E(n)}{\partial w_j} \quad \text{and} \quad \Delta w_{ij}(n) = -\eta \frac{\partial E(n)}{\partial w_{ij}}$$