Outline

- Backpropagation algorithm
- Hidden layer representations

Threshold Units: Summary

Perceptron training rule guaranteed to succeed if

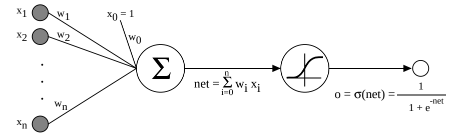
- Training examples are linearly separable
- Sufficiently small learning rate η

Linear unit training rule uses gradient descent

- Converges to hypothesis with minimum squared error
- given sufficiently small learning rate η
- even when training data not describable in H
- (this is our checkers learning algorithm)

Sigmoid unit allows gradient descent for threshold unit

Sigmoid Unit



$\sigma(x)$ is the sigmoid function

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

Nice property: $\frac{d\sigma(x)}{dx} = \sigma(x)(1-\sigma(x))$

Gradient of error E with respect to weight w_i

$$\frac{\partial E}{\partial w_i} = -\sum_{d \in D}(t_d - o_d)o_d(1-o_d)x_{i,d}$$

Backpropagation Algorithm

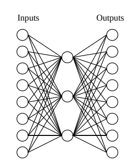
Initialize all weights to small random numbers.
Until satisfied, Do

- For each training example, Do
  1. Input the training example to the network and compute the network outputs
  2. For each output unit k
  $$\delta_k \leftarrow o_k(1-o_k)(t_k - o_k) \quad (1)$$
  3. For each hidden unit h
  $$\delta_h \leftarrow o_h(1-o_h)\sum_{k \in outputs} w_{h,k}\delta_k \quad (2)$$
  4. Update each network weight $w_{i,j}$
  $$w_{i,j} \leftarrow w_{i,j} + \Delta w_{i,j}$$
  where
  $$\Delta w_{i,j} = \eta \delta_j x_{i,j} \quad (3)$$

More on Backpropagation

- Gradient descent over entire *network* weight vector
- Easily generalized to arbitrary directed graphs
- Will find a local, not necessarily global error minimum
  - In practice, often works well (can run multiple times)
- Often include weight *momentum* α
  $$\Delta w_{i,j}(n) = \eta \delta_j x_{i,j} + \alpha \Delta w_{i,j}(n-1)$$
- Minimizes error over *training* examples
  - Will it generalize well to subsequent examples?
- Training can take thousands of iterations $\rightarrow$ slow!
- Using network after training is very fast

Learning Hidden Layer Representations



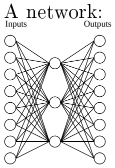
A target function:

Input		Output
10000000	$\rightarrow$	10000000
01000000	$\rightarrow$	01000000
00100000	$\rightarrow$	00100000
00010000	$\rightarrow$	00010000
00001000	$\rightarrow$	00001000
00000100	$\rightarrow$	00000100
00000010	$\rightarrow$	00000010
00000001	$\rightarrow$	00000001

Can this be learned??

Learning Hidden Layer Representations

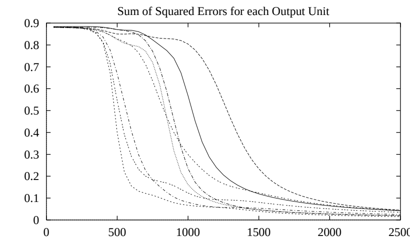
A network:



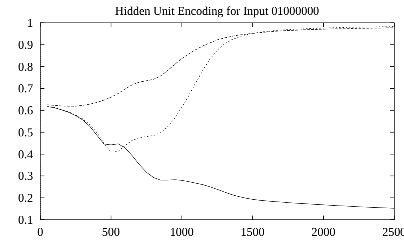
Learned hidden layer representation:

Input		Hidden Values			Output
10000000	→	.89 .04 .08	→		10000000
01000000	→	.01 .11 .88	→		01000000
00100000	→	.01 .97 .27	→		00100000
00010000	→	.99 .97 .71	→		00010000
00001000	→	.03 .05 .02	→		00001000
00000100	→	.22 .99 .99	→		00000100
00000010	→	.80 .01 .98	→		00000010
00000001	→	.60 .94 .01	→		00000001

Training



Training



Training