



Artificial Intelligence Programming

More Neural Networks

Chris Brooks

Department of Computer Science

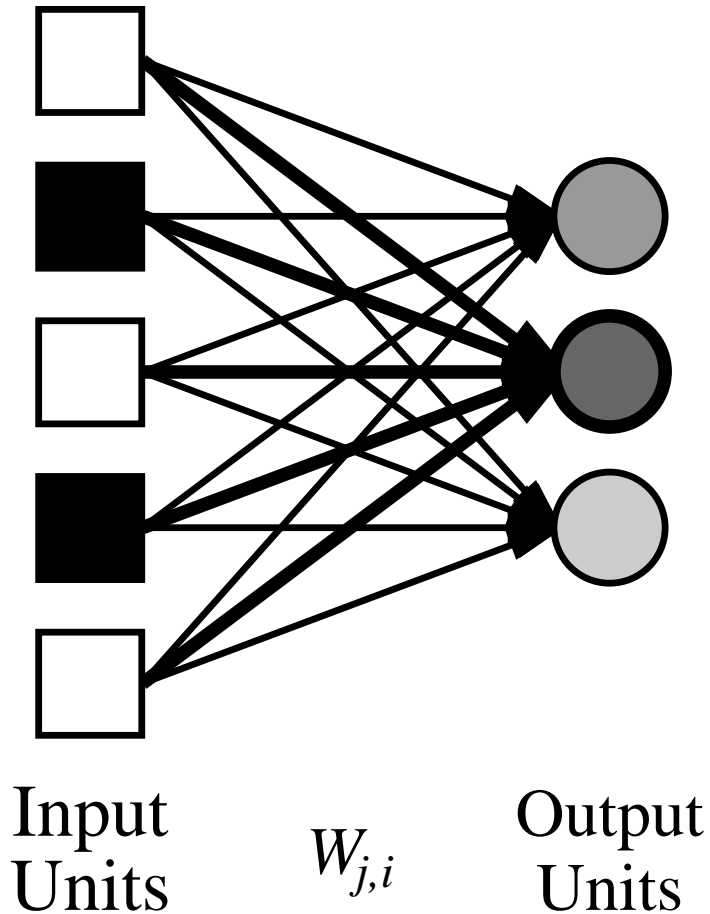
University of San Francisco



Neural networks - refresher

- A network is composed of layers of nodes
 - input, hidden, output layers in feedforward nets
- An input is applied to the input units
- The resulting output is the value of the function the net computes.

Perceptrons - refresher



- Single-layer networks (perceptrons) are the simplest form of NN.
- Easy to understand, but computationally limited.
- Each input unit is directly connected to one or more output units.

Delta rule - refresher

- The appeal of perceptrons is the ability to automatically learn their weights in a supervised fashion.
- The weight updating rule is known as the *delta rule*.
- $\Delta w_i = \alpha \sum_{d \in D} (t_d - o_d) x_{id}$
- Where D is the training set, t_d is expected output and o_d is actual output.

Multilayer Networks

- While perceptrons have the advantage of a simple learning algorithm, their computational limitations are a problem.
- What if we add another “hidden” layer?
- Computational power increases
 - With one hidden layer, can represent any continuous function
 - With two hidden layers, can represent any function
- Problem: How to find the correct weights for hidden nodes?

Multilayer Network Example

Output units

a_i

$W_{j,i}$

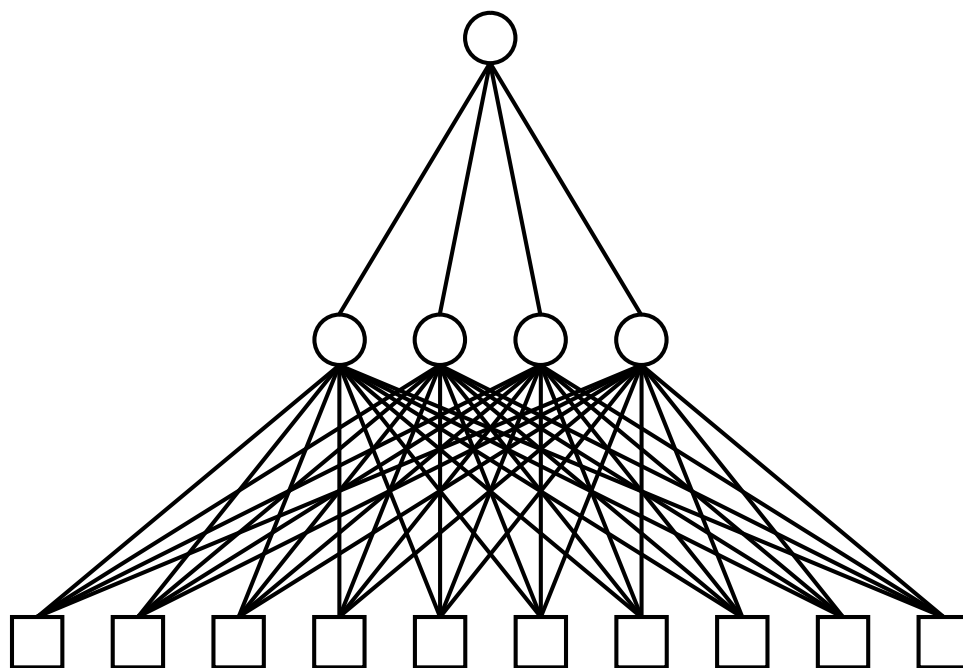
Hidden units

a_j

$W_{k,j}$

Input units

a_k



Backpropagation

- Backpropagation is an extension of the perceptron learning algorithm to deal with multiple layers of nodes.
- Nodes use sigmoid activation function, rather than the step function
 - $g(input_i) = \frac{1}{1+e^{-input_i}}$.
 - $g'(input_i) = g(input_i)(1 - g(input_i))$ (good news here - to compute g' , we just need g)
- We will still “follow the gradient”, where g' gives us the gradient.

Backpropagation

- Updating input-hidden weights:
- Idea: each hidden node is responsible for a fraction of the error in δ_i .
- Divide δ_i according to the strength of the connection between the hidden and output node.
- For each hidden node j
- $\delta_j = g(input)(1 - g(input)) \sum_{i \in outputs} W_{j,i} \delta_i$
- Update rule for input-hidden weights:
- $W_{k,j} = W_{k,j} + \alpha * input_k * \delta_j$

Backpropagation Algorithm

- The whole algorithm can be summed up as:

While not done:

for d in training set

Apply inputs of d, propagate forward.

for node i in output layer

$$\delta_i = output * (1 - output) * (t_{exp} - output)$$

for each hidden node j

$$\delta_j = output * (1 - output) * \sum W_{j,i} \delta_i$$

Adjust each weight

$$W_{j,i} = W_{j,i} + \alpha * \delta_i * input_j$$

Theory vs Practice

- In the definition of backpropagation, a single update for all weights is computed for all data points at once.
 - Find the update that minimizes total sum of squared error.
- Guaranteed to converge in this case.
- Problem: This is often computationally space-intensive.
 - Requires creating a matrix with one row for each data point and inverting it.
- In practice, updates are done incrementally instead.

Stopping conditions

- Unfortunately, incremental updating is not *guaranteed* to converge.
- Also, convergence can take a long time.
- When to stop training?
 - Fixed number of iterations
 - Total error below a set threshold
 - Convergence - no change in weights

Comments on Backpropagation

- Also works for multiple hidden layers
- Backpropagation is only guaranteed to converge to a local minimum
 - May not find the absolute best set of weights
- Low initial weights can help with this
 - Makes the network act more linearly - fewer minima
- Can also use random restart - train multiple times with different initial weights.

Momentum

- Since backpropagation is a hillclimbing algorithm, it is susceptible to getting stuck in plateaus
 - Areas where local weight changes don't produce an improvement in the error function.
- A common extension to backpropagation is the addition of a momentum term.
 - Carries the algorithm through minima and plateaus.
- Idea: remember the “direction” you were going in, and by default keep going that way.
- Mathematically, this means using the second derivative.

Momentum

- Implementing momentum typically means remembering what update was done in the previous iteration.
- Our update rule becomes:
- $\Delta w_{ji}(n) = \alpha \Delta_j x_{ji} + \beta \Delta \mathbf{w}_{ji}(n-1)$
- To consider the effect, imagine that our new delta is zero (we haven't made any improvement)
- Momentum will keep the weights “moving” in the same direction.
- Also gradually increases step size in areas where gradient is unchanging.
 - This speeds up convergence, helps escape plateaus and local minima.

Design issues

- As with GAs, one difficulty with neural nets is determining how to *encode* your problem
 - Inputs must be 1 and 0, or else real-valued numbers.
 - Same for outputs
- Symbolic variables can be given binary encodings
- More complex concepts may require care to represent correctly.

Design issues

- Like some of the other algorithms we've studied, neural nets have a number of parameters that must be tuned to get good performance.
 - Number of layers
 - Number of hidden units
 - Learning rate
 - Initial weights
 - Momentum term
 - Training regimen
- These may require trial and error to determine
- Alternatively, you could use a GA or simulated annealing to figure them out.

Radial Basis Function networks

- One problem with backpropagation is that every node contributes to the output of a solution
- This means that all weights must be tuned in order to minimize global error.
- Noise in one portion of the data can have an impact on the entire output of the network.
- Also, training times are long.
- Radial Basis Function networks provide a solution to this.

Radial Basis Function networks

- Intuition: Each node in the network will represent a portion of the input space.
- Responsible for classifying examples that fall “near” it.
 - Much like k-NN
- Vanilla approach: For each training point $(x_i, f(x_i))$, create a node whose “center” is x_i .
- The output of this node for a new input x will be $W * \phi(|x - x_i|)$
- Where W is the weight, and $\phi = \exp(-\frac{x^2}{2\sigma^2})$
- ϕ is a *basis function*.

Radial Basis Function networks

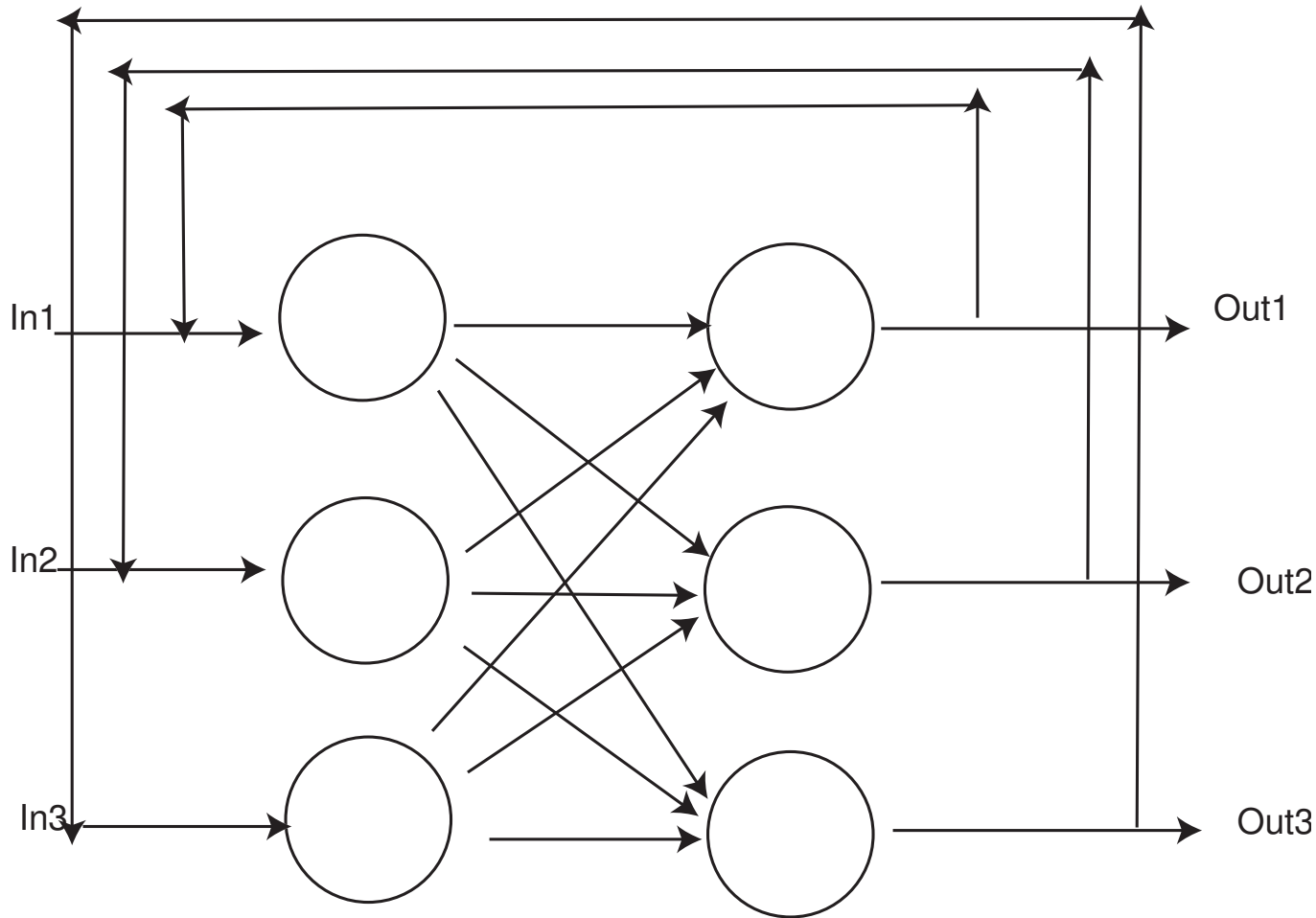
- Each node has a “zone of influence” where it can classify nearby examples.
- Training due to misclassification will only affect nodes that are near the misclassified example.
- Also, network is single-layer.
- Weights can be trained by writing a matrix equation:
 - $\Phi \mathbf{W} = \mathbf{t}$
 - $\mathbf{W} = \Phi^{-1} \mathbf{t}$
- Inverting a matrix is a much faster operation than training with backpropagation.

Recurrent NNs

- So far, we've talked only about feedforward networks.
 - Signals propagate in one direction
 - Output is immediately available
 - Well-understood training algorithms
- There has also been a great deal of work done on recurrent neural networks.
 - At least some of the outputs are connected back to the inputs.

Recurrent NNs

- This is a single-layer recurrent neural network



- Notice that it looks a bit like an S-R latch.

Hopfield networks

- A Hopfield network has no special input or output nodes.
- Every node receives an input and produces an output
- Every node connected to every other node.
- Typically, threshold functions are used.
- Network does not immediately produce an output.
 - Instead, it oscillates
- Under some easy-to-achieve conditions, the network will eventually stabilize.
- Weights are found using simulated annealing.

Hopfield networks

- Hopfield networks can be used to build an *associative memory*
- A portion of a pattern is presented to the network, and the net “recalls” the entire pattern.
- Useful for letter recognition
- Also for optimization problems
- Often used to model brain activity

Neural nets - summary

- Key idea: simple computational units are connected together using weights.
- Globally complex behavior emerges from their interaction.
- No direct symbol manipulation
- Straightforward training methods
- Useful when a machine that approximates a function is needed
 - No need to understand the learned hypothesis