The use of Neural Networks in Text Categorization

Abhishek Sethi

What is a Neural Network

- An attempt at creating a learning system that mimics the brain
  - Brain consists of billions of neurons, interconnected, firing based on external stimuli
- A group of nodes connected to create a highly parallel system
  - The nodes are the “neurons” per se [4]

Typical "Neuron"

- Multiple inputs with an associated weight for each
  - Can be (+) or (-) weights
- Fire if the sum of the product of the weights and their associated input is greater than a threshold value

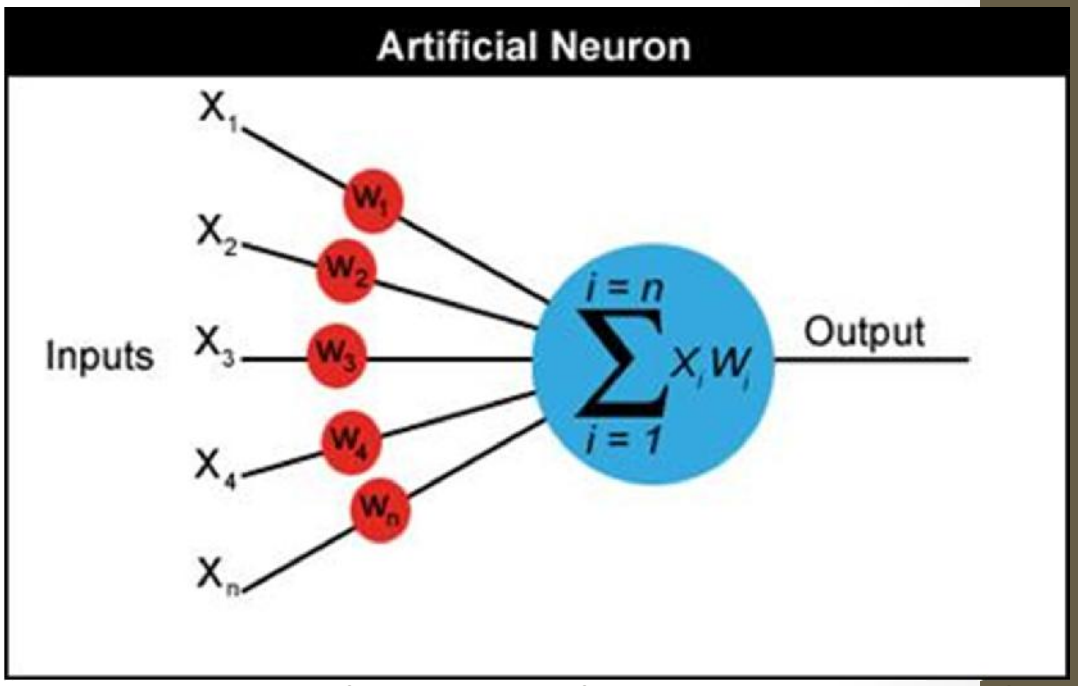

Figure 1: A typical neuron with respective input and output values [4]

Firing Rule

- Rules stating what calculation would be used to determine what neuron fires for given input [3]
- Example
  - For pattern recognition
  - Use Hamming distance for calculating difference between two strings
  - If input closer to the set that fires a 1, the input induces a 1. If closer to 0, fire a 0. If equidistant, then output is undefined

Training a Neural Network

- Use a training data set
- Multiple methods for training
- Idea: Adjust weights of input until the network begins to produce the desired output
  - Examples in pattern recognition, text recognition, genomic interpretation, etc

Types of Networks

- Feedforward network
- Feedback network

Feedforward Network

- 3 layers
  - Input, Hidden, Output
- Inputs start at input layer and move up
- All neurons at each layer connected to all neurons at next layer
- Any number of neurons allowed

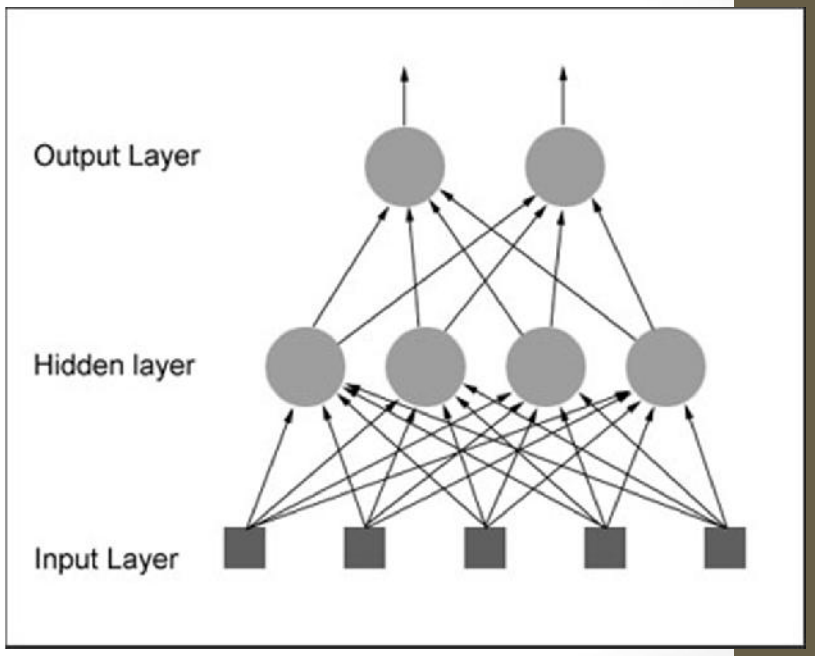

Figure 2: Example of a typical feedforward network [4]

Feedback Network

- More complex
- Signals travels in both directions
- Still 3 layers
- Dynamically changing based on input

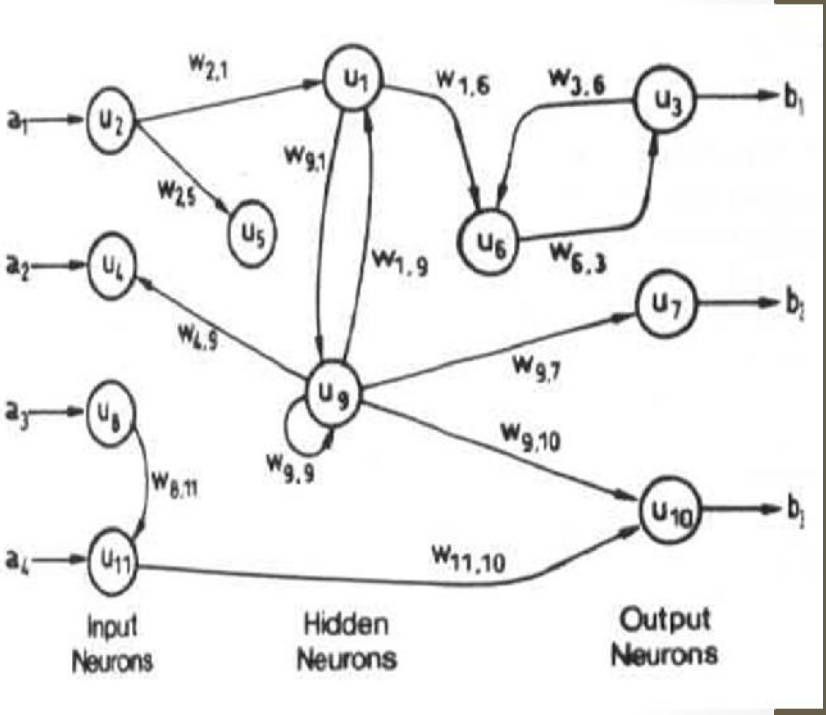

Figure 3: Example of a feedback network [3]

Back Propagation Algorithm

- Training algorithm
  - Adjust input weights until the desired output is obtained [3]
- Calculate the Error derivative (EW)
  - How error responds to changes in input weight
- 1st, calculate EA (rate at which error changes as neuron activity changes)
  - At output layer: Difference between actual and desired output
  - At hidden layer: ID all weights between hidden and output layer, multiply weights by output neuron EA's, sum together the products for EA of neuron
- Work from output layer, and go towards inputs
- EW = product of EA & activity through incoming connection

Applications

- Sales forecasting
- Data validation
- Medical imaging
- Text categorization
- Functional Genomics

Why am I interested?

- Model the brain
- Can have applications in bioinformatics

What has been done (a very small example)

1. Text categorization with improved Back Propagation Neural Network [1]
2. Hierarchical Neural Networks [5]
3. Multilabel Neural Networks for text categorization and application to genomics [2]

What still can be done

- Optimize Back Propagation Algorithm
  - Slow convergence, bad for large networks
- Improve Neural Network learning methods
- Improve performance
- More generalized so weights aren't needed
- Extend uses in IR outside of document categorization/classification

References

1. Li, C H, and S C. Park. "A Novel Algorithm for Text Categorization Using Improved Back-Propagation Neural Network." *Lecture Notes in Computer Science*. (2006): 452-460. Print.

2. Min-Ling, Zhang S, and Zhou S. Zhi-Hua. "Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization." *Ieee Transactions on Knowledge and Data Engineering*. 18.10 (n.d.): 1338-1351. Print.

3. Neural Networks. http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

4. Neural Network Tutorial. http://www.ai-junkie.com/ann/evolved/nnt1.html

5. Ruiz, Miguel E, and Padmini Srinivasan. "Hierarchical Text Categorization Using Neural Networks." *Information Retrieval*. 5.1 (2002): 87-118. Print.

Any Questions?