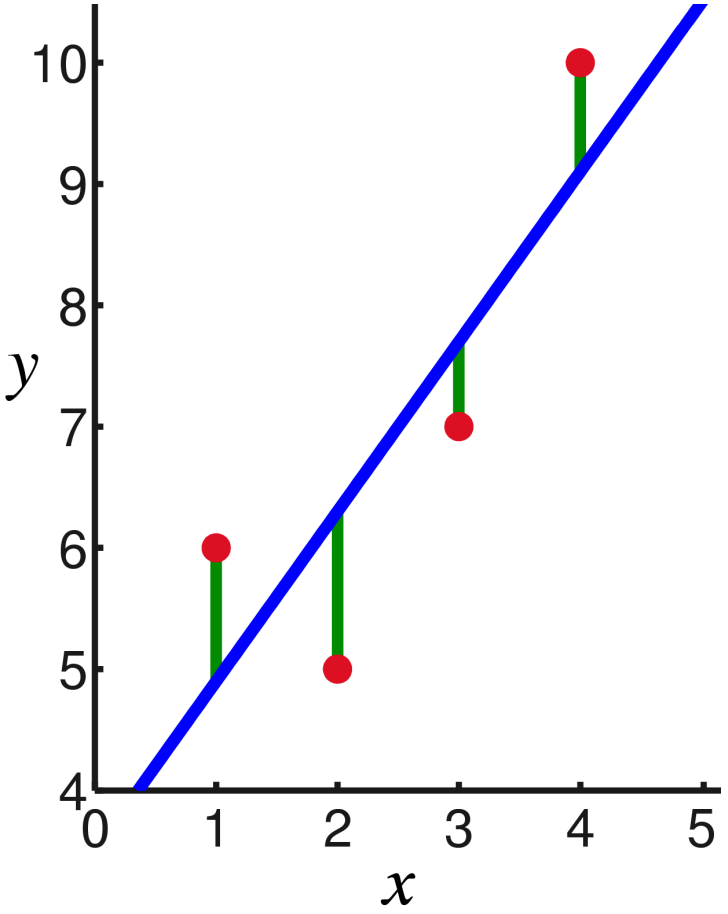
MSDS 7333

Unit 11: Where I decide that its more important to teach advanced methods

Stochastic Gradient Descent

- What is gradient Descent (a/k/a Linear Regression)
- How is this related to linear regression?
  - Linear regression Updates all the slopes (β) at once.
  - Stochastic Gradient Descent only updates a few at a time
    - Why would you do this?
    - Saves Memory!
    - Super fast

Gradient Descent vs Stochastic Gradient Descent



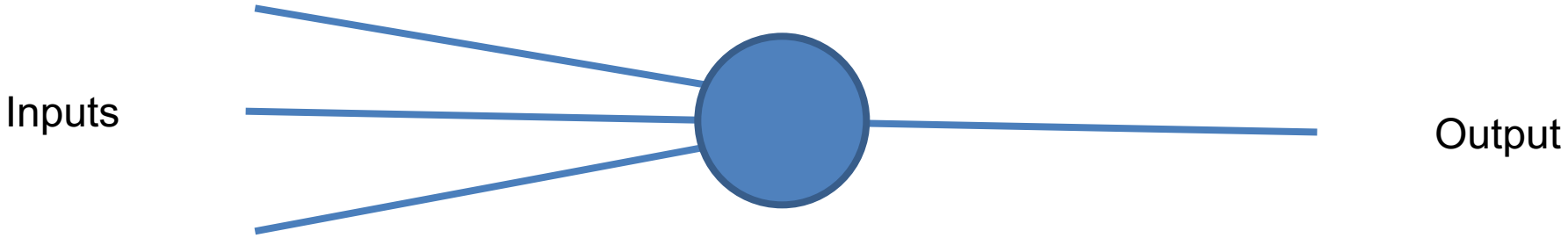
- Gradient descent: Calculate the error for all four points, calculate slope all 4 data points
- Stochastic Gradient Descent: Calculate the error of a single point (at random). Calculate new slope. Pick a second point and repeat *ad nasuem.*
- SGD: Smaller memory required. Faster Calculation. NO NEED TO LOAD ALL DATA INTO MEMORY!

Vowpal Wabbit

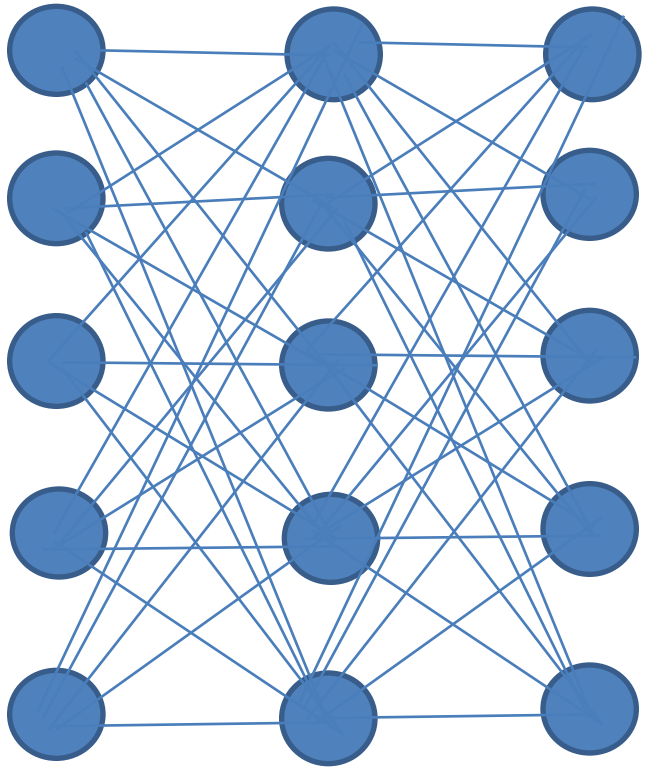
- Super fast implementation of Stochastic Gradient Descent
- Can handle millions of data points
- Can do feature interactions
- Support regularization
- 10,000 lines of code only
- You want to learn to use this
  - Compiled from source
  - Steep learning curve
  - Worth it
  - Out performed a 100 node Hadoop cluster (Macbook)

Neural net

- Modeled on a simple version of a neuron
- A neuron fires if its input reaches a certain potential



From neuron to neural net



Inputs

Hidden Layer

Output

- Each neuron computes an activation function:
  σ(mx +b)
- Example: $1/\sigma(z) \equiv 1/(1 + e^{-z})$
  Aka sigmoid function

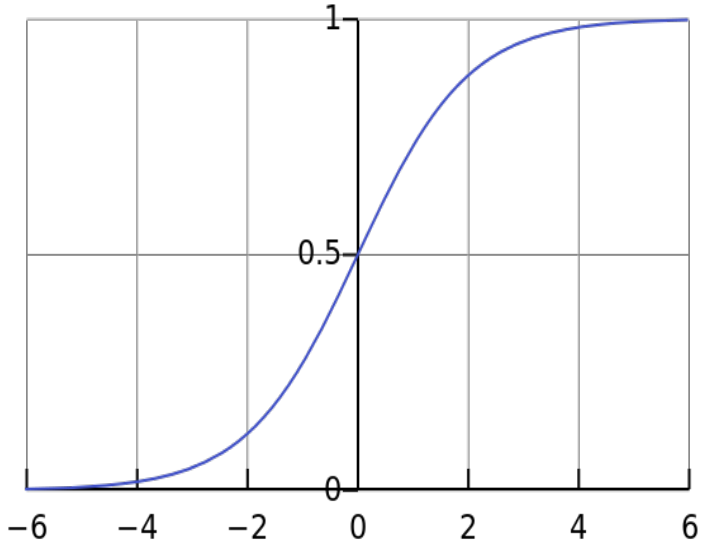
Each neuron is in effect a linear regressor

- Neural net is an ensemble of linear regressors
- Each neuron also has a bias
- Proven that 1 hidden layers can approximate any function
- Math/Computer Intensive

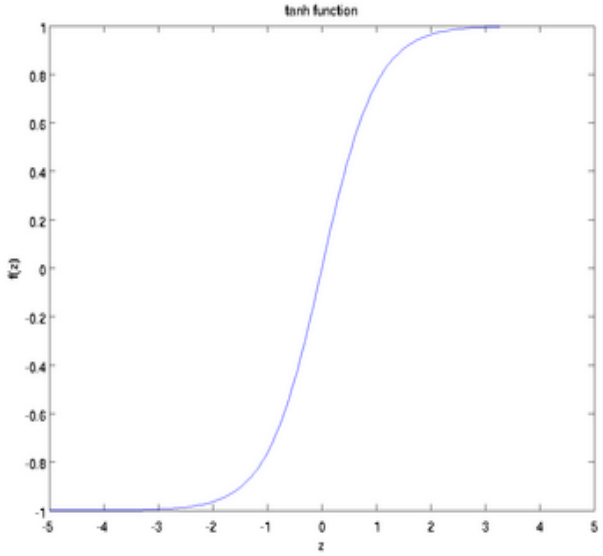
Neural Net architecture is a hyper parameter

- Learning rate
- Dropout
- Number of layers
- Number of neurons

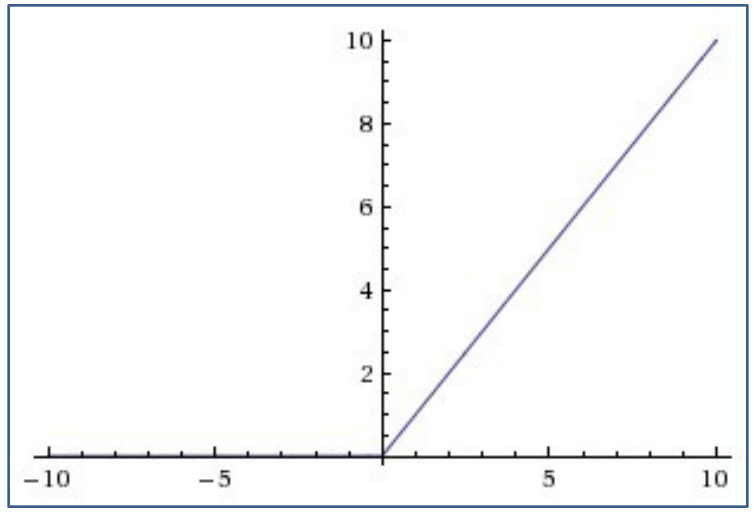
Activation functions



Sigmoid



Tanh



ReLU (Rectified Linear Unit)

Time to play

- https://playground.tensorflow.org/
- http://neuralnetworksanddeeplearning.com/chap4.html