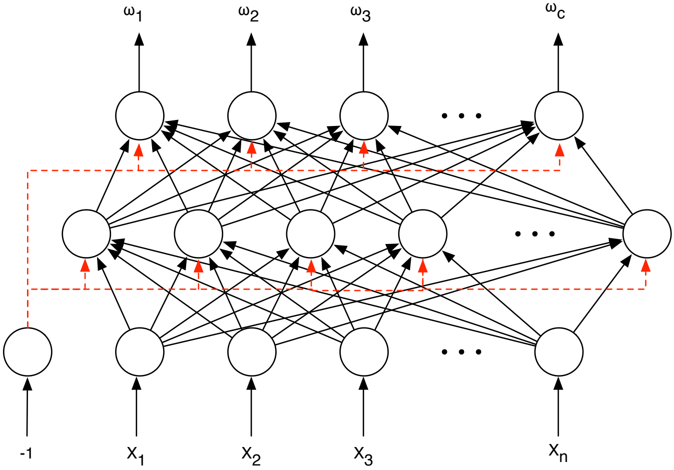
Pragmatics for Backpropagation in Neural Networks

Lei Hu, Paul Romanczyk, & Erin Wirch
lxh9868@rit.edu par4249@rit.edu aew8948@rit.edu

September 30, 2010

Outline

Neural Networks



Backpropagation

- Learning for neural networks
- Supervised learning
- Used in feed-forward networks

Pragmatics for Backpropogation

- Activation Function Properties
- Scaling Input
- Number of Hidden Units
- Learning Rate
- Momentum
- Adding Noise
- Hints
- Stopped Training

Activation Function Properties

- Required Properties
  - $f(\cdot)$ and $f'(\cdot)$ are continuous.
  - $f(\cdot)$ is non-linear.
- Desired Properties
  - $f(\cdot)$ saturates.
  - $f(\cdot)$ is monotonic.
- The sigmoid has all of these properties.
  - $f(x) = \frac{1}{1+e^{-x}}$

Scaling Input

- **Problem:** With more than one metric, you have more than one type of measurement. These measurements need to interact such that no metric is more heavily weighted than another
- **Solution:** Scale the inputs so that the neural network treats each feature with equal weight

Scaling Input—Standardizing

- Shift the numbers so that the average over the training set data is zero
- Scale each feature so that the variance is the same in each metric

Number of Hidden Units

- The number of hidden units determines the number of weights
- Weights are degrees of freedom

Choosing the Number of Hidden Units

- Too few hidden units means a poor fit to the training data, too many means overfitting
- The goal is to find a happy medium with low test error
- Rule of thumb: n/10

Learning Rate

- $w_{new} = w_{old} - \eta\delta\mu$
- η is the learning rate.
- Learning rate can affect the quality of the final network

Optimal Learning Rate

- The optimal learning rate could lead to the local minimum in one learning step.
- $\eta_{opt} = \left(\frac{\partial^2 J}{\partial w^2}\right)^{-1}$
- $\eta \approx 0.1$

Optimal Learning Rate

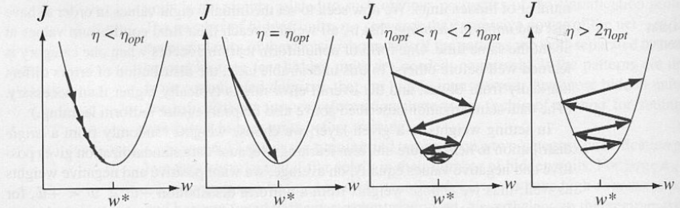


FIGURE 6.16. Gradient descent in a one-dimensional quadratic criterion with different learning rates. If $\eta < \eta_{opt}$, convergence is assured, but training can be needlessly slow. If $\eta = \eta_{opt}$, a single learning step suffices to find the error minimum. If $\eta_{opt} < \eta < 2\eta_{opt}$, the system will oscillate but nevertheless converge, but training is needlessly slow. If $\eta > 2\eta_{opt}$, the system diverges.

Figure 6.16 from [Duda et al., 2001]

Optimal Learning Rate

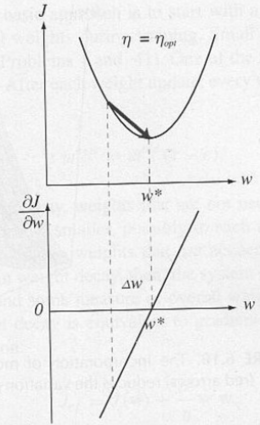


FIGURE 6.17. If the criterion function is quadratic (above), its derivative is linear (below). The optimal learning rate η_{opt} ensures that the weight value yielding minimum error, w^*, is found in a single learning step.

Figure 6.17 from [Duda et al., 2001]

Hu, Romanczyk, & Wirch

Momentum

- Momentum changes the learning rule and could pull the network out of plateaus.
- $w(m+1) = w(m) + (1-a)\Delta w_{bp}(m) + a\Delta w(m-1)$
- $\Delta w_{bp}(m) = -\eta\delta\mu$
- $\Delta w(m) = w(m) - w(m-1)$
- $a \approx 0.9$

Momentum

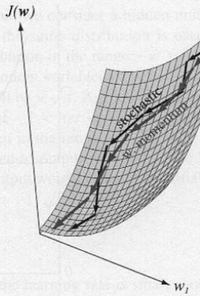


FIGURE 6.18. The incorporation of momentum into stochastic gradient descent by Eq. 37 (red arrows) reduces the variation in overall gradient directions and speeds learning.

Figure 6.18 from [Duda et al., 2001]

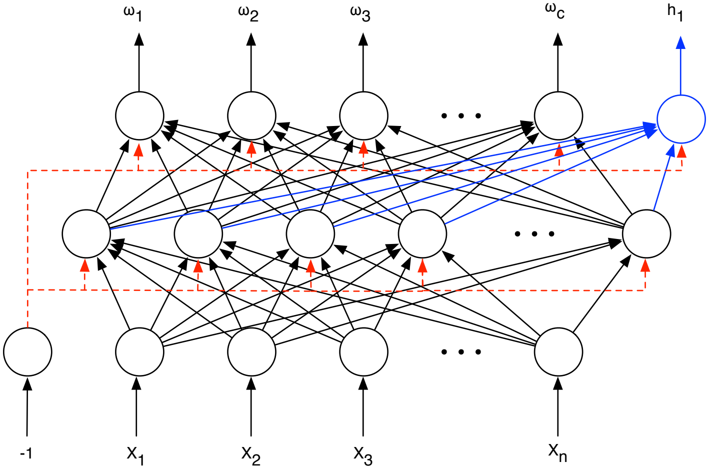
Adding Noise

- Add a different random noise to the data on each training run.
- Noise has the effect of blurring the neural network.
- The trained network will be more general than one trained without noise.
- Noise addition can be used to generate more data.

Hints

- Add extra outputs during learning to try to control the evolution of the NN.
- Hints are not calculated during classification.
- Hints provide additional, but related information to help the classification.

Hints



Stopped Training

- Training on one data set may lead to over fitting
  - The NN is being fit to the sample and not the population
  - Loss of generallity
- Having a second test set can be used to reduce over fitting.
  - Stop training when testing error begins to increase.
- Other stopping criteria:
  - Stop training when training error is below a predetermined threshold.
  - The average training error stops decreasing.
  - The change in average training error is small.

References

Duda, R., Hart, P., and Stork, D. (2001).
Pattern Classification.
John Wiley and Sons, 2^{nd} edition.

Kuncheva, L. (2004).
Combining pattern classifiers: methods and algorithms.
Wiley-Interscience.

Swingler, K. (1996).
Applying neural networks: a practical guide.
Morgan Kaufmann.

Questions