MIT Department of Brain and Cognitive Sciences
9.641J, Spring 2005 - Introduction to Neural Networks
Instructor: Professor Sebastian Seung

The delta rule

Learn from your mistakes

If it ain’t broke, don’t fix it.

Outline

- Supervised learning problem
- Delta rule
- Delta rule as gradient descent
- Hebb rule

Supervised learning

- Given examples

$$R^N \to \{0,1\}$$

$$x_1 \to y_1$$

$$x_2 \to y_2$$

$$x_3 \to y_3$$

$$\vdots$$

- Find perceptron such that

$$y_a = H\left(w^T x_a\right)$$

Example: handwritten digits

- Find a perceptron that detects “two”s.

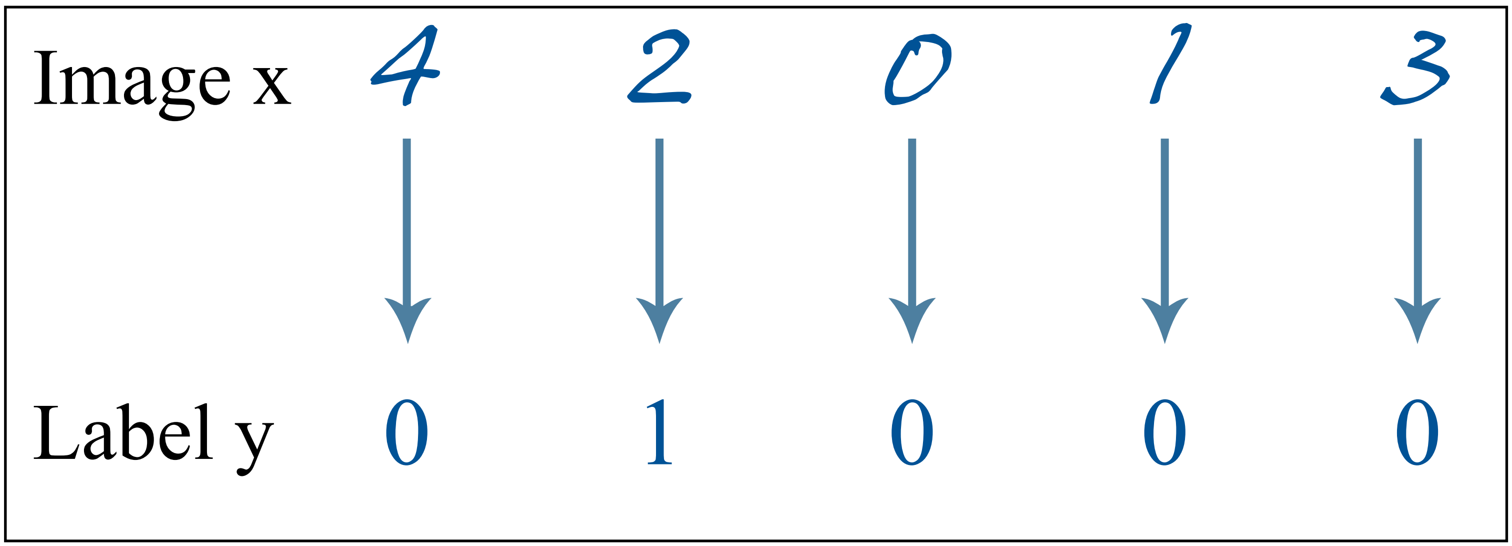

Figure by MIT OCW.

Delta rule

$$\Delta w = \eta\left[y - H\left(w^T x\right)\right]x$$

- Learning from mistakes.
- “delta”: difference between desired and actual output.
- Also called “perceptron learning rule”

Two types of mistakes

- False positive $y = 0, \; H\left(w^T x\right) = 1$
  - Make *w* less like *x*.

$$\Delta w = -\eta x$$

- False negative $y = 1, \; H\left(w^T x\right) = 0$
  - Make *w* more like *x*.

$$\Delta w = \eta x$$

- The update is always proportional to *x.*

Objective function

- Gradient update

$$\Delta w = -\eta \frac{\partial e}{\partial w} \qquad e(w,x,y) = \left| y - H\left(w^T x\right) \right| \left| w^T x \right|$$

- Stochastic gradient descent on

$$E(w) = \left\langle e(w,x,y) \right\rangle$$

- *E*=0 means no mistakes.

Perceptron convergence theorem

- Cycle through a set of examples.
- Suppose a solution with zero error exists.
- The perceptron learning rule finds a solution in finite time.

If examples are nonseparable

- The delta rule does not converge.
- Objective function is not equal to the number of mistakes.
- No reason to believe that the delta rule minimizes the number of mistakes.

Memorization & generalization

- Prescription: minimize error on the training set of examples
- What is the error on a test set of examples?
- Vapnik-Chervonenkis theory
  - assumption: examples are drawn from a probability distribution
  - conditions for generalization

contrast with Hebb rule

$$\Delta w = \eta y x \qquad \qquad \Delta w = \eta \left(y - \langle y \rangle \right) x$$

- Assume that the teacher can drive the perceptron to produce the desired output.
- What are the objective functions?

Is the delta rule biological?

- Actual output: anti-Hebbian

$$\Delta w = -\eta H\left(w^T x\right)x$$

- Desired output: Hebbian

$$\Delta w = \eta y x$$

- Contrastive

Objective function

- Hebb rule
  - distance from inputs
- Delta rule
  - error in reproducing the output

Supervised vs. unsupervised

- Classification vs. generation
- I shall not today attempt further to define the kinds of material [pornography] … but I know it when I see it.
  - – Justice Potter Stewart

Smooth activation function

- same except for slope of f

$$\Delta w = \eta f'\left(w^T x\right)\left[y - f\left(w^T x\right)\right]x$$

- update is small when the argument of *f*□ has large magnitude.

Objective function

- Gradient update

$$\Delta w = -\eta \frac{\partial e}{\partial w} \qquad e(w,x,y) = \frac{1}{2}\left[y - f\left(w^T x\right)\right]^2$$

- Stochastic gradient descent on

$$E(w) = \left\langle e(w,x,y) \right\rangle$$

- E=0 means zero error.

Smooth activation functions are important for generalizing the delta rule to multilayer perceptrons.