Adalines and Widrow Hoff Delta Rule

Sayeda Suaiba Anwar
Lecturer, Department of CSE
East Delta University

Adalines

- Adaline stands for Adaptive Linear Element.
- A neural network that adapts a system to minimize the “error” signal using supervised learning.
- Acts as a filter to sort input data patterns into two categories.
- Used in virtually all high speed modems and telephone switching systems to cancel out the echo of a reflected signal in a transmission line.
- Invented by Bernard Widrow and M.E (Ted) Hoff of Stanford University in the early 1960s.

Adalines

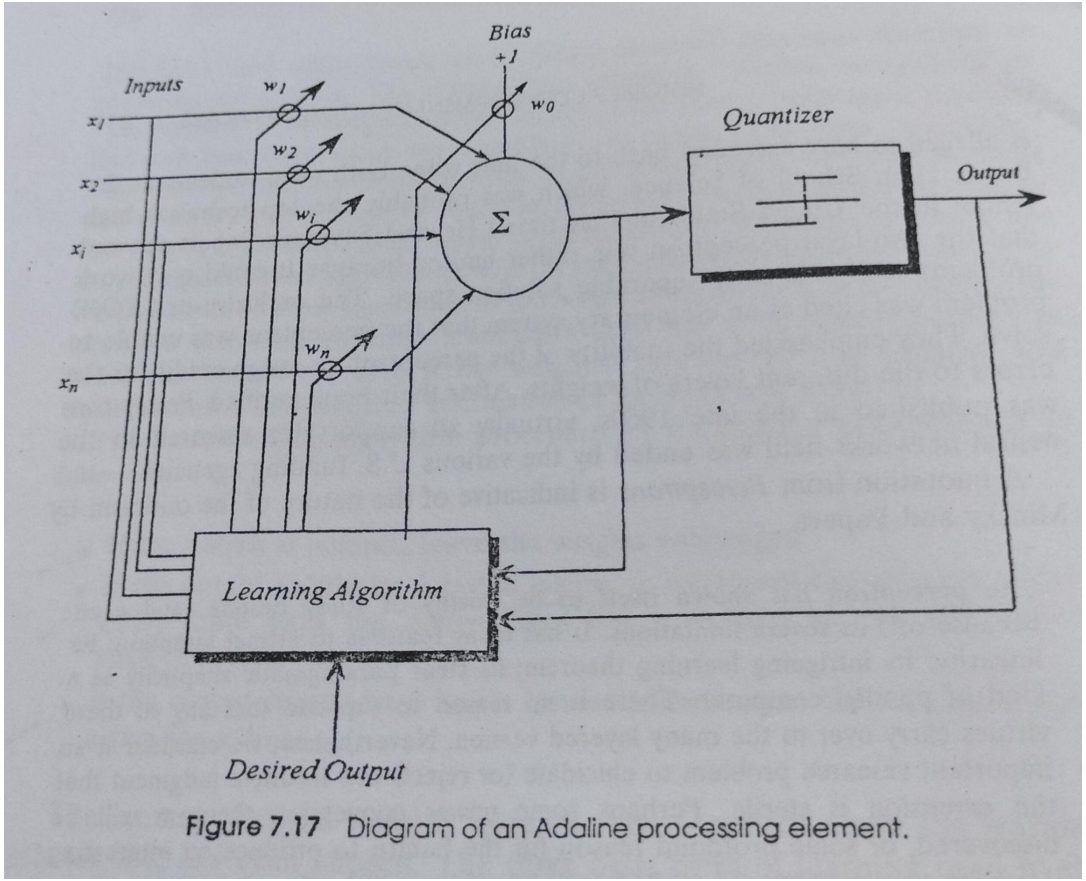


Figure 7.17 Diagram of an Adaline processing element.

Adalines

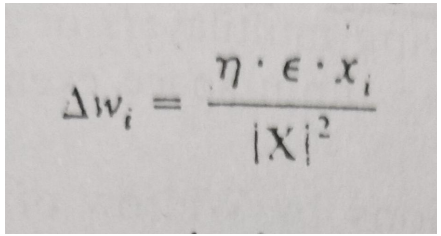
- The quantizer is a threshold-type nonlinear function with +1 and -1 as the limiting values(i.e if the summation of the weighted inputs is positive, the output of the system will be +1; and if it is zero or negative, the output will be -1).
- The learning algorithm uses the difference between the desired output and the output of the summation (not the output of the quantizer) to produce the error Ɛ function used to adjust the weights.
- Prior to the beginning of the training, all weights must be adjusted to random values.

Adalines

With the adaline, an input pattern is presented to the processing elements that filters it for a specific category.

If the input matches the category, the processing element output is +1 and if it does not match the category the processing element output will be -1.

The learning rule is “Delta Rule”, also known as the “Widrow-Hoff Learning Rule”.



$$\Delta w_i = \frac{\eta \cdot \epsilon \cdot x_i}{|X|^2}$$

Here, η is the learning constant, ε is the error, x_i is the i-th input(-1 or +1) and X is the input vector.

Adalines

Each weight is adjusted so that the error is equally distributed among the weights. So the equation becomes -

$$\Delta w_i = \frac{\eta \cdot \epsilon \cdot x_i}{|X|^2} = \frac{\epsilon \cdot x_i}{(N+1)|X|^2}$$

Where (N+1) is the number of inputs plus the bias input and 1/(N+1) replaces the learning constant. This means the error is uniformly distributed to the (N+1) inputs.

Adalines

Most of the time, the convergence of the learning process in the Adaline is very fast.

However, the nature of initial randomization can have a major effect on the speed of convergence.

Adaline is capable of classifying linearly separable patterns only.

Widrow-Hoff Delta Rule

The Widrow–Hoff delta learning rule can be derived by considering the node of Figure 8.4, where T is the target or desired value vector and I is defined by equation (8.1-1) as the dot product of the weight and input vectors and is given by

$$I = \sum_{i=1}^{n} w_i x_i \qquad (8.2\text{-}1)$$

Widrow-Hoff Delta Rule

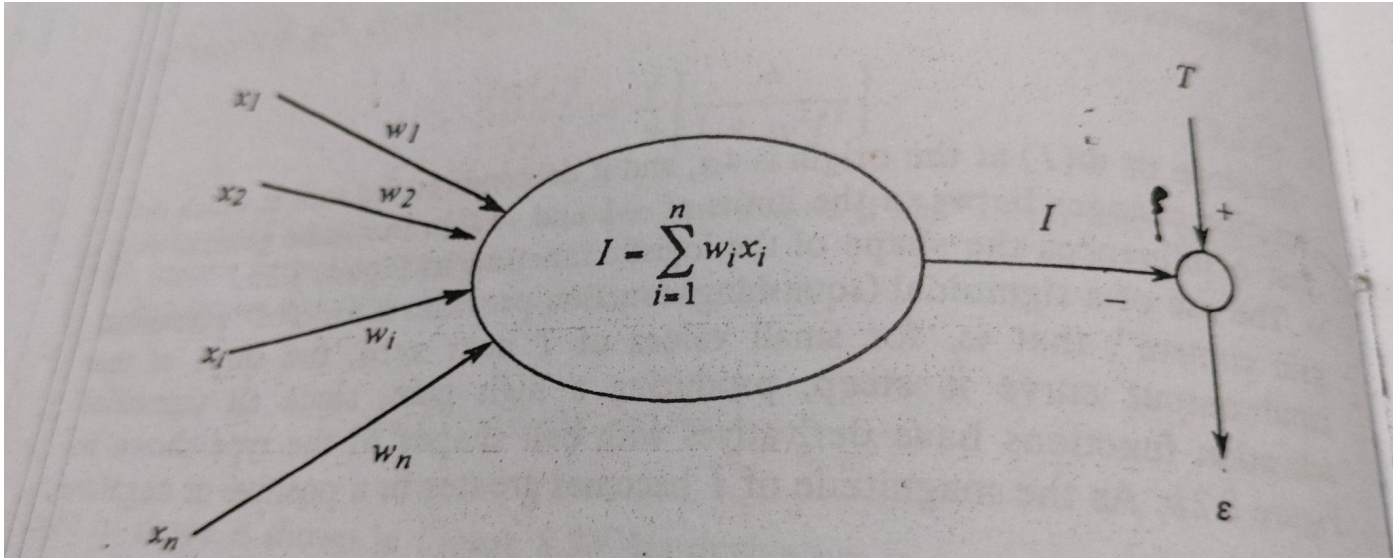


Figure 8.4 Neuron without activation function but with a target value T and an error ε.

Widrow-Hoff Delta Rule

For this derivation, no quantizer or other nonlinear activation function is included, but the result presented here is equally valid when such nonlinear elements are included.

From Figure 8.4, we see the error function ε as a function of all weights w_i, and we see the squared error ε^2 to be

$$\varepsilon = (T - I) \tag{8.2-2}$$

$$\varepsilon^2 = (T - I)^2 \tag{8.2-3}$$

Widrow-Hoff Delta Rule

The gradient of the square error vector is the partial derivatives with respect to each of these i weights:

$$\frac{\partial \varepsilon^2}{\partial w_i} = -2(T - I)\frac{\partial I}{\partial w_i} = -2(T - I)x_i \qquad (8.2\text{-}4)$$

Since this gradient involves only the ith weight component, the summation of equation (8.2-1) disappears.

For demonstration purposes, let us consider a neuron with only two inputs, x_1 and x_2. The square error is now given by

$$\begin{aligned}
\varepsilon^2 &= [T - w_1x_1 - w_2x_2]^2 \\
&= T^2 + w_1^2x_1^2 + w_2^2x_2^2 - 2Tw_1x_1 - 2Tw_2x_2 + 2w_1x_1w_2x_2 \\
&= w_1^2[x_1^2] + w_1[-2x_1(T - w_2x_2)] + [(T - w_2x_2)^2] \\
&= w_2^2[x_2^2] + w_2[-2x_2(T - w_1x_1)] + [(T - w_1x_1)^2] \qquad (8.2\text{-}5)
\end{aligned}$$

Widrow-Hoff Delta Rule

The minimum square error occurs when the partial derivatives of square error with respect to the weights w_1 and w_2 are set equal to zero:

$$\frac{\partial \varepsilon^2}{\partial w_1} = -2[T - w_1 x_1 - w_2 x_2]x_1 = 0 \qquad (8.2\text{-}6)$$

$$\frac{\partial \varepsilon^2}{\partial w_2} = -2[T - w_1 x_1 - w_2 x_2]x_2 = 0 \qquad (8.2\text{-}7)$$

Since x_1 and x_2 cannot be zero, the quantities in the brackets, which are identical for both equations, must be zero. This gives

$$T - w_1 x_1 - w_2 x_2 = 0 \qquad (8.2\text{-}8)$$

from which the location of the minimum in the w_1 and w_2 dimensions are

$$w_1 = \frac{T - w_2 x_2}{x_1} \qquad (8.2\text{-}9)$$

$$w_2 = \frac{T - w_1 x_1}{x_2} \qquad (8.2\text{-}10)$$

Widrow-Hoff Delta Rule

Substitution of either of these values into equation (8.2.5) gives the minimum square error to be zero.

In real world, the minimum square error is never equal to zero because of nonlinearities, noise and imperfect data.

The presence of noise with a sigmoidal activation function will give a minimum square error that is not zero.

Widrow-Hoff Delta Rule

The Widrow Hoff Delta training rule provides that the change in each weight vector component is proportional to the negative of its gradient:

$$\Delta w_i = -K\frac{\partial \varepsilon^2}{\partial w_i} = K \cdot 2(T - I)x_i = 2K\varepsilon x_i$$

Where K is a constant of proportionality. The negative sign is introduced because a minimization process is involved. It is common to normalize the input vector component x_i by adding $|X|^2$. Then the equation becomes

$$\Delta w_i = \left[2K|X|^2\right]\frac{\varepsilon x_i}{|X|^2} = \frac{\eta \varepsilon x_i}{|X|^2}$$

Where

$$\eta = 2K|X|^2$$

Widrow-Hoff Delta Rule

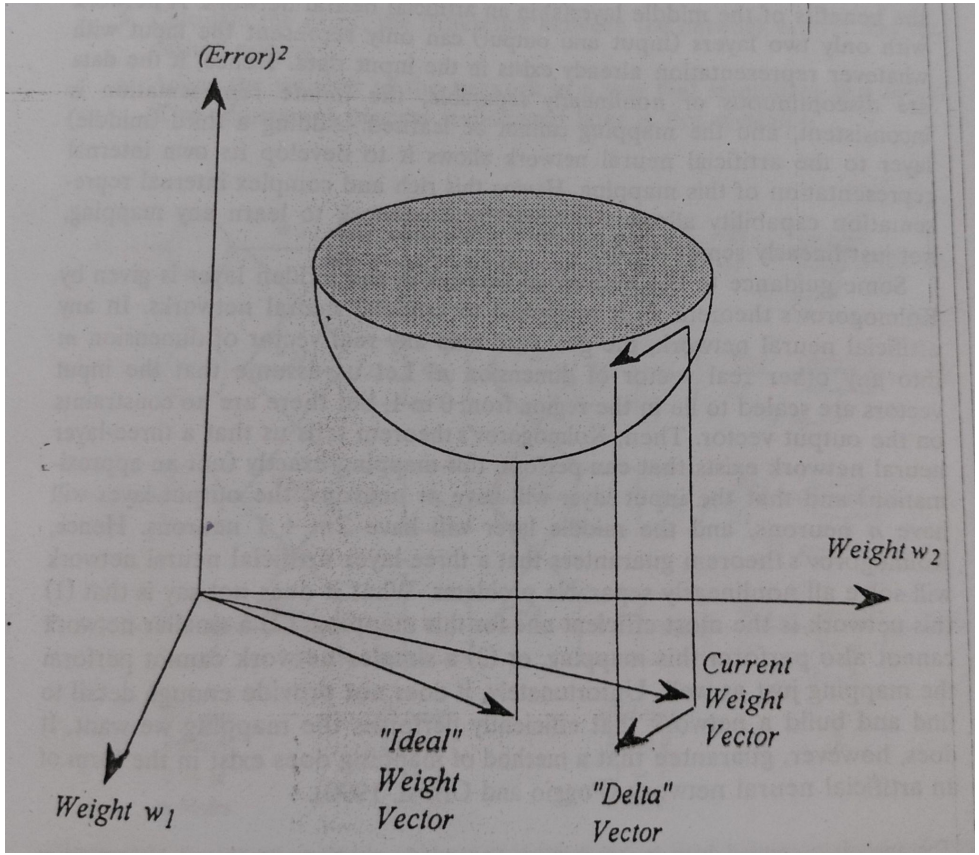
A geometrical interpretation of the delta rule is that it involves a gradient descent algorithm to minimize the square error. When the square error is viewed in three dimensions $(w_1, w_2, \varepsilon^2)$ the square error surface is a paraboloid of revolution with the weight vector descending toward the minimum value along a gradient vector on the surface of the paraboloid. The projection of this gradient vector on the w_1–w_2 plane is the delta vector as shown in Figure 8.6. The delta rule moves the weight vector along the negative gradient of the curved surface toward the ideal weight vector

Widrow-Hoff Delta Rule

position. Because it follows the gradient, it is called a *gradient descent* or *steepest descent* algorithm. Since the gradient is the most efficient path to the bottom of the curved surface, the delta rule is the most efficient way to minimize the square error. There is, however, one caveat that must be added here: This statement is true only if the weight vector is descending toward a global minimum. If there are local minima, which are common with multidimensional problems, other techniques must be used to ensure that a solution (i.e., a weight configuration) is not trapped in one of these local minima.

The Widrow-Hoff delta training

Geometric Interpretation of Delta Rule



Madalines

- Acronym for “Many Adalines”.
- Involves the use of several Adalines as the middle layer of a three layered neural network.
- The input layer is an input buffer to ensure that all inputs go to each of the adalines.
- The output layer is a single unit that combines the output of all adalines.
- Sometimes this output unit gives a +1 when the majority of the inputs are +1 and -1 when they are not (i.e Voting Majority).