Figure 29-5. *Backpropagation*

The backpropagation algorithm works by computing the cost function at the output layer by comparing the predicted output of the neural network with the actual outputs from the dataset. It then employs gradient descent (earlier discussed in Chapter 16) to calculate the gradient of the cost function using the weights of the neurons at each successive layer and update the weights propagating back through the network.
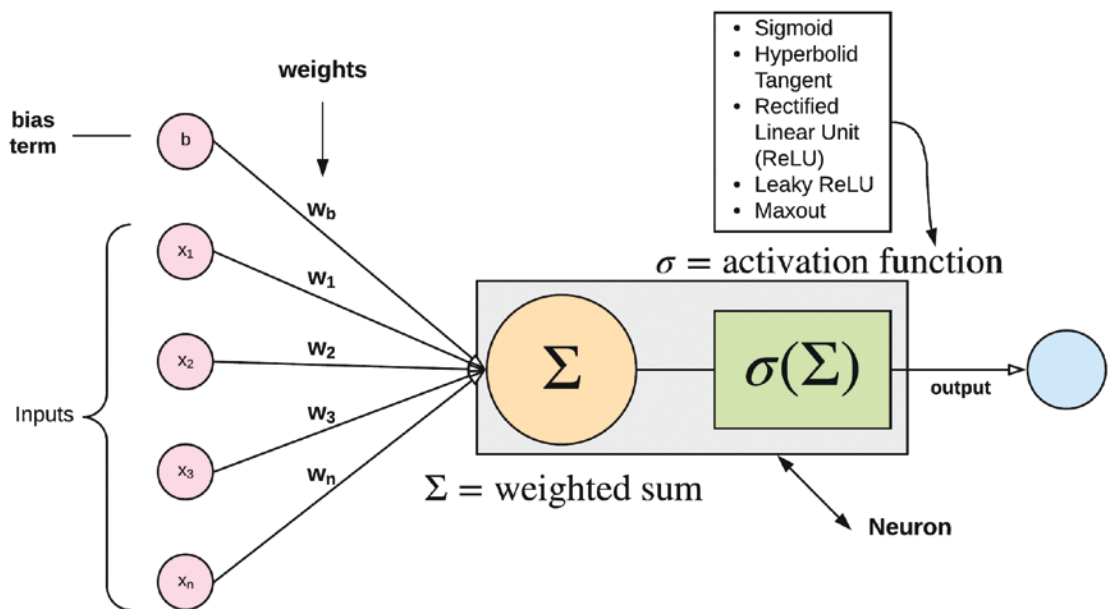
Activation Functions

Up till now, we have mentioned activation functions. Now let's go a bit deeper into what activation functions are and why do we have them.

Activation functions act on the weighted sum in the neuron (which is nothing more than the weighted sum of weights and their added bias) by passing it through a non-linear function to decide if that neuron should fire (propagate) its information or not to the succeeding neural layers.

In other words, the activation function determines if a particular neuron has the information to result in a correct prediction at the output layer for an observation in the training dataset. Activation functions are analogous to how neurons communicate and transfer information in the brain, by firing when the activation goes above a particular threshold value.

These activation functions are also called non-linearities because they inject non-linear capabilities to our network and can learn a mapping from inputs to output for a dataset whose fundamental structure is non-linear. An illustration of passing the weighted sum of weights and biases through an activation function is shown in Figure 29-6.

Figure 29-6. *Activation function*

The following are examples of activation functions used in a neural network:
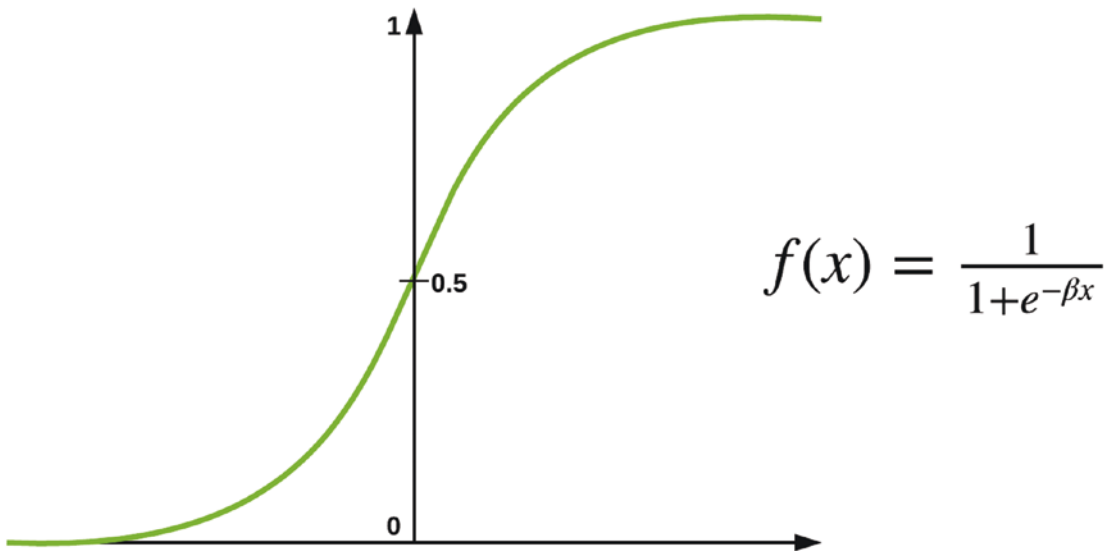
- Sigmoid

- Hyperbolic tangent (tanh)

- Rectified linear unit (ReLU)

- Leaky ReLU

- Maxout

Let's briefly examine them.

Sigmoid

The sigmoid function illustrated in Figure 29-7 is a non-linear function that brings (or squashes) the activations to fall within a range of 0 and 1. This brings large negative and positive numbers to 0 and 1, respectively. The neurons typically begin firing when the function output is above a threshold of 0.5.

Sigmoid Activation Function

$$f(x) = \frac{1}{1+e^{-\beta x}}$$

Figure 29-7. *Sigmoid activation function*

However, a significant drawback of the sigmoid function is its susceptibility to a phenomenon called exploding and vanishing gradients. In the process of optimizing the weights of the network during backpropagation, the gradients can become disproportionately small or large with their activations concentrated at either 0 or 1.