

1 Multilayer Perceptrons

It's really a misnomer to call Multilayer Perceptrons *multi-layer*. It's not a perceptron with multiple layers but instead it is a network of multiple "neuron-like" processing units but not every layer in the network is a perceptron.

Recall that the perceptron equation:

$$Z = \sum_{i=0}^N W_i x_i + b$$

What would happen if you applied multiple layers of this equation?

2 History

Dudes named Rumelhart, Hinton, and Williams introduced/discovered this algorithm in the mid-'80s. Rumelhart was a mathematical psychologist working at UC San Diego. He led a research group on neural networks, picking up where Rosenblatt left off after the Minsky book pretty much destroyed public funding for this type of research.

Rumelhart wanted to train a neural network with multiple layers and sigmoidal units instead of threshold units (as in the perceptron) or linear units but they did not know how to do that. Rumelhart came up with an idea called "Backpropagation" but it was Hinton that used it in conjunction with the training model Rumelhart developed for sigmoidal multilayered perceptrons.

This is what is typically referred to as a multilayered perceptron today.

3 The Algorithm

Formally, a multilayered perceptron is one where these conditions are true:

- a linear function that aggregates the input values

- a sigmoid function (the activation function)
- an output function (i.e. a threshold function for classification)
- a loss function that computes the overall error of the network
- a learning procedure to adjust the weights of the network, i.e., the “backpropagation” algorithm

4 The Math

4.1 Linear Function

The linear function is very similar to the single layer perceptron. It is a weight sum of the inputs plus a bias. The difference is that the weights are matrix and no longer a vector.

$$z_m = \sum_{i=0}^N W_{m,n} x_m + b$$

4.2 Sigmoid Function

Every element of the Z_m becomes an input to the sigmoid function.

$$a_m = \sigma(z_m) = \frac{1}{1 + e^{-z_m}}$$

The output of the sigmoid function is a vector of size m where every element is a unit in the hidden layer.

a stands for activation! Which is typically how the output of the hidden layer is described. Using the sigmoid function is somewhat arbitrary, you can use other non-linear functions here like \tanh or $ReLU$.

4.3 Output Function

Similar to single layer but adjusted for sigmoid.

$$f(a) = \begin{cases} 1, & a \geq 0.5 \\ 0, & \text{else} \end{cases}$$

4.4 Cost function

We need a measure of how well the network is performing. We did this with the single layer too. This is used in the training step. This often goes by different names in the literature such *objective function*, *loss function*, or *error function*.

Typically this cost function is tuned for your problem at hand. The one that Rumelhart came up with was

$$E = \frac{1}{2} \sum_k (a_k - y_k)^2$$

This is commonly known as the sum of squared errors function. It's pretty basic but often gets the job done.

4.5 Training: Forward & Back Propagation

Multilayer Perceptron training is composed of two steps. Forward propagation where information flows forward in the network to compute predictions and the error. And then the back propagation stage where the error derivatives is calculated and used to update the weight matrix.

The forward propagation stage is just chaining the linear, sigmoid and output functions (just like the single layer!).

The back propagation part is more complicated but it's basically asking "how does the error change if we tweak the weights"? This question has additional questions that need answers first. How does the error change when we change the activation a_m a bit? How does the activation change if we fiddle with linear function a bit?

It takes a fair amount of calculus to derive all of this. I'll leave it as an exercise to the reader. The outcome of that derivation can be solved for using the gradient descent algorithm.