Introduction to Neural Networks..

Feed forward Network Functions.

The linear models for regression and classification are based on linear combinations of fixed non linear basis functions $\phi_j(x)$ and take the form

$$y(x, w) = f\left(\sum_{j=1}^{M} w_j \cdot \phi_j(x)\right)$$

where $f(\cdot)$ is a non linear activation function in the case of classification and is the identity in the case of regression.

"Our goal is to make rather extend this model by making the basis functions $\phi_j(x)$ depend on parameters and then to allow these parameter to be adjusted, along with the coefficients $\{w_j\}$, during training."

This leads to the basic neural network model, which can be described as a series of functional transformation First we construct M linear combinations of the input variables $x_1, \ldots, x_D$ in the form

$$a_j = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}$$

where $j = 1 \cdots M$. and superscript (1) indicates that the corresponding params are of first layer.

These activation functions are then transformed using a differentiable, nonlinear activation function $h(\cdot)$ to give

$$z_j = h(a_j).$$

These quantities correspond to output of basis functions, in context of neural networks, known as hidden units.

Some $h(\cdot)$ functions are –

- tanh
- Sigmoid
- ReLU
- Maxout
- Leaky ReLU etc.

These values are again linearly combined to give output unit activations.

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)}$$

where $k = 1, \ldots, K$ and K is the total number of outputs.

The choice of activation function is determined by the nature of the data and the assumed distribution of target variables.

Thus, for standard regression problems, the activation is the identity so that

$$y_k = a_k.$$

For classification problems

$$y_k = \sigma(a_k).$$

where

$$\sigma = \frac{1}{1 + \exp(-a)}$$

complete Picture –

$$y_k(x, w) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right) \quad \text{– (1)}$$

Thus, the neural network model is simply a non-linear function from a set of input variables $\{x_i\}$ to set output variable $\{y_k\}$ controlled by a vector w of adjustable parameter.

The process of evaluating (1) can be interpreted as a forward propagation of information through the network.

We will absorb the bias parameter into weight parameters by defining an additional input var x_0 whose value is clamped at $x_0 = 1$

Hence, $a_j = \sum_{i=0}^{D} w_{ji}^{(1)} x_i$

We can similarly absorb the second-layer biases into second-layer weights

Hence

$$y_k(x, w) = \sigma\left(\sum_{j=0}^{M} w_{kj}^{(2)} h\left(\sum_{i=0}^{D} w_{ji}^{(1)} x_i\right)\right)$$