3. Its derivative is always nonzero, so Gradient Descent can always roll down the slope. When the activation function is a step function, Gradient Descent cannot move, as there is no slope at all.

5.

a. X: $m \times 10$

b. Wh: 10×50, bh: 50×1

c. Y: $m \times 3$

d. $Y = \text{relu}(X \times Wh + bh) * Wa + b0$

6. Spam or ham: only one neuron and using logistic function

MNIST Classification: 10 neurons and using SoftMax function

7. Backpropagation is a technique used to train artificial neural networks. It first computes the gradient of the cost function with regards to every model parameter, and then it performs a *Gradient Descent* step using these gradients.

8. Number of hidden layers

Number of neurons per layer

Activation function.