867002105_HW9

Charles Dotson

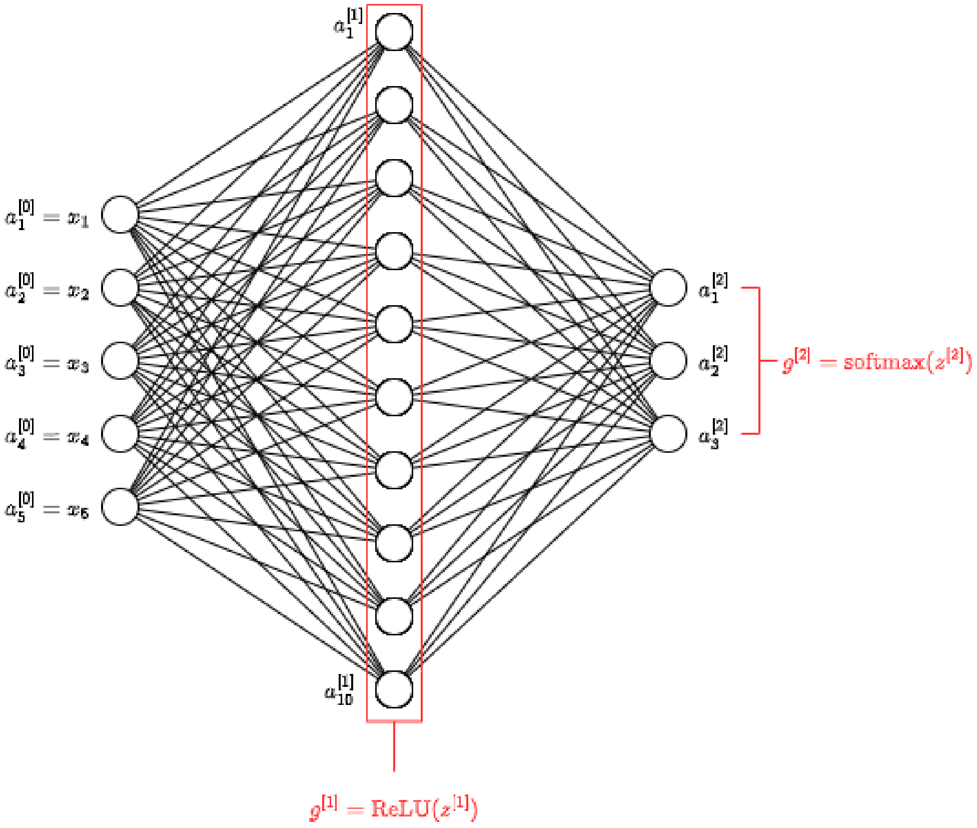
November 11, 2022

1 * Neural network architecture: given input feature vectors $x^{(i)} \in \mathbb{R}^5$, we want to predict one out of three classes for each input. Design a neural network with 1 input layer, 1 hidden layer, and 1 output layer. Specify the following architecture parameters:

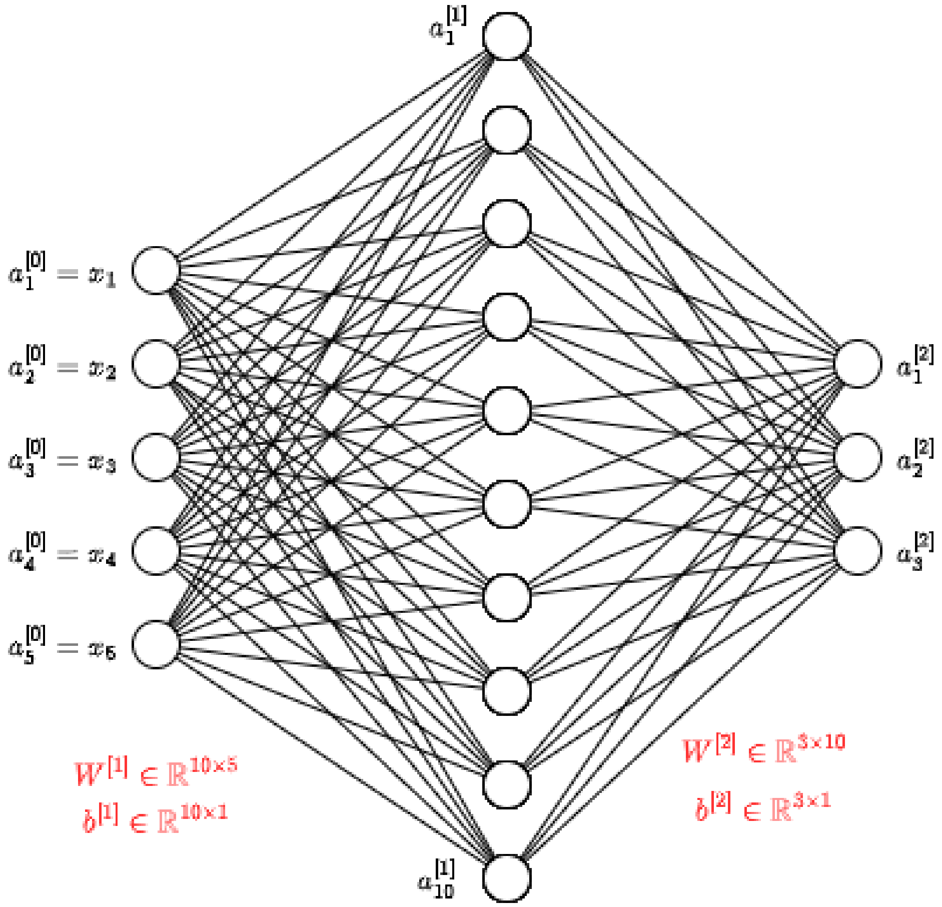
- number of neurons of each layer (you have to select one for the hidden layer as well);
- activation function $g^{[l]}$ for each layer $ l = 1, 2$. Make sure to think about which activation function is suitable for predicting one out of more than two classes. You may need to refer back to the logistic regression note.

I created a diagram for this. I think it easier to understand and more concise. I will list the neurons as this is not in the diagram but I thought may be obvious. Just to be safe $n^{[0]} = 5$, $n^{[1]} = 10$, $n^{[0]} = 3$. The input and hidden layer are not counting the bias term. I made $n^{[1]}$ as just $2(n^{[0]})$ or 2(number of features). Output layer must be three as we have three classes and are identifying which trainning example is part of which class.

This leads to $g^{[l]}$. Since we are classifying multiclass we cannot use sigmoid becasue it is not binary. For the hidden layer I chose the trusty ReLU(z).



2 For the above network, list the weight and bias parameters $((W^{[l]},\ b^{[l]})$ for $l = 1,\ 2)$ necessary to define the neural network. Make sure you specify the number of rows and columns of each such parameter.



$W^{[1]} \in \mathbb{R}^{10\times5}$

$b^{[1]} \in \mathbb{R}^{10\times1}$

$W^{[2]} \in \mathbb{R}^{3\times10}$

$b^{[2]} \in \mathbb{R}^{3\times1}$

3 * Given B training examples, feeding them through an MLP will generate B column activation vectors at the output layer, collected in the matrix $A \in \mathbb{R}^{n[L] \times B}$. Let $A^{[L-1]}$ be the matrix B activation vectors at the penultimate (second last) layer. Let $Y \in \{0,1\}^{n[L] \times B}$ be the ground truth label matrix, with the k-th column $y^{(k)}$ being the one-hot label vector for the k-th training example. Prove that

$$\frac{\partial J}{\partial W^{[L]}} = \frac{1}{B}(A^{[L]} - Y)(A^{[L-1]})^T \in \mathbb{R}^{n[L] \times n[L-1]}$$

is indeed the average of the gradients of the loss function J with respect to $W^{[L]}$ over the B training examples. Your answer should contain the gradients for each of the B training examples.

For SGD method we can say the following.

$$J(Y, A^{[L]}) = -Y \log A^{[L]} - (1 - Y)\log(1 - A^{[L]})$$

$$= \sum_{k=1}^{B} -y^{(k)} \log a^{[L](k)} - (1 - y^{(k)})\log(1 - a^{[L](k)})$$

$$z^{[L]} = W^{[L]} a^{[L-1]} + b^{[L]}$$

$$\frac{\partial J}{\partial W^{[L]}} = \frac{\partial J(Y, A^{[L]})}{\partial z^{[L]}} \times \frac{\partial z^{[L]}}{\partial W^{[L]}}$$

$$= \sum_{k=1}^{B} (a^{[L](k)} - y^{(k)})(a^{[L-1](k)})^T$$

$$= \frac{1}{B}\sum_{k=1}^{B} (a^{[L](k)} - y^{(k)})(a^{[L-1](k)})^T \quad \text{(due to this being an SGD problem)}$$

$$= \frac{1}{B}(A^{[L]} - Y)(A^{[L-1]})^T$$

4 MLP can also be used for regression on multiple target values. Given training examples $\{x^{(i)}; y^{(i)}\}_{i=1}^{m}$ where $y^{(i)} \in \mathbb{R}^k$ and $x^{(i)}\mathbb{R}^n$. MSE loss will be appropriate for training an MLP to predict $\hat{y}^{(i)}$, so that $\hat{y}^{(i)}$ is as close as possible to $y^{(i)}$ for the input $x^{(i)}$. The MSE loss function will then be

$$J = \frac{1}{2m}\sum_{i=1}^{m}\|\hat{y}^{(i)} - y^{(i)}\|_2^2$$

where $\|w\|_2$ is the 2-norm of the vector w. Find the partial derivative of J w.r.t $\hat{y}^{(i)}$. IS the softmax activation function appropriate for regression MLP? Justify your answer.

5 * (Graduate only) This question is related to Project 3. In the function explain of the NN class, you will be asked to find $\frac{\partial z_c^{[L]}}{\partial a^{[0]}}$, the gradient of $z_c^{[L]}$ w.r.t the input data $a^{[0]}$, where $z_c^{[L]} = W_c^{[L]} a^{[L-1]}$ and $W_c^{[L]}$ is the c-th row of the parameter matrix $W^{[L]}$, and $a^{[0]} \in \mathbb{R}^n$ is the input vector. Use the following hints to find $\frac{\partial z_c^{[L]}}{\partial a^{[0]}}$.