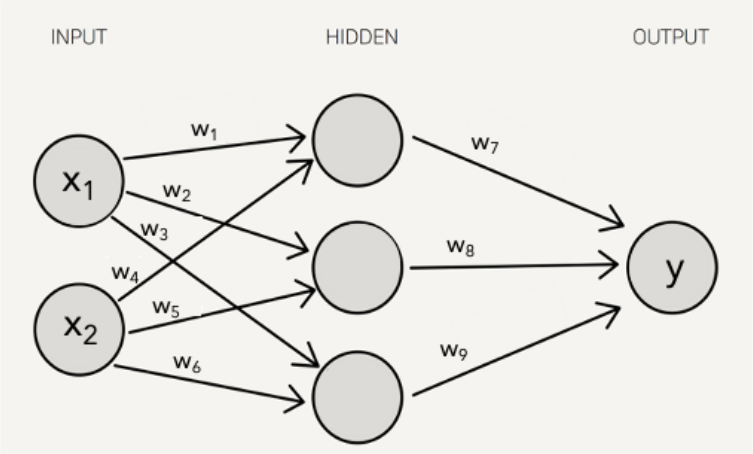
Homework 4. Neural Networks and Interpretability
COSC425, Fall 2020

Due: November 23, 11:59pm

1 Understanding Neural Networks [30 pts]

For the following questions, consider the following neural network stucture:



Using this structure, you should now undergo the process for training the neural network. As done in our lectures, you are learning the appropriate weights for one set of inputs for the XOR operation. You should use the sigmoid activation function where necessary. Your training procedure should begin with the inputs and outputs as:

$$x_1 = 1, x_2 = 0, y = 1$$

Your weights are randomly initialized to the following values:

$$w_1 = 0.1, w_2 = 0.1, w_3 = 0.5, w_4 = 0.9, w_5 = 0.1, w_6 = 0.9, w_7 = 0.2, w_8 = 0.3, w_9 = 0.4$$

For each of the following questions, you will be asked to show the outcome of various parts of the training process. You are being provided with a blank image of the neural network structure image above. **It is recommended that you draw on top of (or annotate) the blank image to communicate the outcomes of each training step procedure.** The blank image is available on Canvas and can be pasted into this PDF.

1. Complete the first Feed Forward iteration of the training process. Show all parts of your work for both the Hidden and Output Layer. Report this iteration's calculated Y. [10pts]

2. Complete the first Backpropagation iteration of the training process. Show all parts of your work for updating each weight. Report the updated set of weights, w_1 - w_9. [10pts]

3. Complete a second Feed Forward iteration. Show all parts of your work for both the Hidden and Output Layer. Report this iteration's calculated Y and specifically state how your prediction's accuracy has changed, if at all. [10pts]

2 Programming Neural Networks [40pts]

For the following questions, you will implement and evaluate Keras classifiers for the Iris dataset. In "class", we ran through two Python scripts that introduced us to programming neural networks with Keras and Tensorflow using the Iris dataset. You can make use of these scripts however you like. You can find them both on Canvas (Module: Neural Networks II; `nn.py` and `nn_tuning.py`).

1. Using the Iris dataset, implement a Keras classifier that uses a neural network model with one input layer, four hidden layers, and an output layer. The input layer has four neurons, each hidden layer has four neurons, and the output layer has three neurons. Each layer should use the sigmoid activation function. Evaluate the classifier with KFold validation across five splits with shuffling. Run the evaluation 10 times, and report the performance mean and standard deviation for each run. Alongside the performance metrics, take a screenshot of your model's implementation in your source code. [15pts]

2. Following your implementation in the prior question, you decide that it makes more sense to simplify the model with two layers. However, you still need to identify the appropriate set of hyperparameters for your neural network. Implement a neural network model that allows you to configure the number of neurons and the activation functions in the network's hidden layers. Use Scikit-Learn's `GridSearchCV` function to identify the optimal hyperparameters. [15pts]

You should explore the following hyperparameter possibilities across both hidden layers:

- **Activation Functions:** linear, sigmoid, tanh, and relu.
- **Number of Neurons:** 1, 2, 5, 10, 15, 20, 25, 30.

Note: Tensorflow will throw warnings. It is safe to ignore them.

(a) Below, include a screenshot of your model's implementation that clearly shows your `Sequential()` Keras model and its ability to configure the number of neurons and the activation functions in both hidden layers. [5pts]

(b) State the optimal number of neurons and choice of activation function for both layers as observed via GridSearchCV. [5pts]

(c) Re-run the GridSearchCV technique on your neural network again, and you will find that your output may suggest a different set of hyperparameters perform best. With this variability in mind, what steps could you take with GridSearchCV to know that you've truly reached the optimal set of hyperparameters? [5pts]

3 Interpreting Neural Networks [40 pts]

It's your first week at Google as a machine learning engineer in the newly-minted Interpretability Assessment and Assurance team. You've been tasked with evaluating the interpretability of ***Inception***, a pre-trained model that recognizes objects in images. You know that Google plans to use the model to build a new e-Commerce competitor to Amazon, so you've decided to see how well the model's recognition works on gold-standard data: Amazon's product images.

To evaluate the model's interpretability, you have set out to identify the strengths and weaknesses of the Inception model using LIME[1]. A Python script coontaining the LIME tutorial is available on Canvas (Module: Interpretable Machine Learning; `main.py`). Your task is as follows:

1. Run `pip install lime` on your machine;
2. Read over the LIME tutorial online; And
3. On your local machine, run `main.py`, which implements the tutorial in a Python script.

Afterwards, you should address the following questions:

1. Browse Amazon.com and collect an image for ten different types of products. Try and find a diverse set of images that include different people, objects, etc. Paste each image below. Apply LIME to each image and report its Top-5 predictions and their associated probabilities. [20pts]

There is additional room on the following page.

[1] https://github.com/marcotcr/lime/

.

2. Using Matplotlib, plot each image with its boundaries that "explain" its classification. Alongside each plot, provide a single sentence that states whether you believe the classification is well-explained by the bounded area. [10pts]

Note: There is an additional question on the following page.

3. Based on your observations in Question 1 and 2, is the Inception model interpretable? Are there types of images that seem to fail or succeed? Rationalize your answer with Miller's definition of "Good Explanations". [10pts]