IE643 Deep Learning - Theory and Practice **July-Dec 2023**

Assignment 2: Due On 9^{th} October 2023 (11:59 PM IST)

1 Instructions

Answer all questions. Write your answers clearly. You can score a maximum of 45 marks in this assignment.

Make sure that your answers and plots are visible in the python notebook (.ipynb) file. Upload in moodle, the .ipynb files correspoding to the questions as a single zip file named as "IE643_rollno_assignment2.zip". All your files within the zip file should follow similar naming convention. There will be no extensions to the submission deadline.

The links for all related files used in Assignment 2 are provided in moodle.

2 Assignment Questions

1. **[Use only Python]** Use the code template for multi-layer perceptron (or feed-forward network) coded from scratch, posted in moodle. Answer the following:
   (a) The linear activation function is given by $\text{linear}(z) = z$. Implement the python functions to compute $\text{linear}(z)$ and its gradient. [**2 marks**]
   (b) Recall that you might have implemented the ReLU activation function given by $\text{ReLU}(z) = \max\{z, 0\}$ and its gradient. Now, implement the python functions to compute the following activation function called **SRELU** given by $\text{SRELU}(z) = z\sigma_{\tanh}(\log(\exp(z) + 1))$ and its gradient, where $\sigma_{\tanh}(q) = \frac{e^q - e^{-q}}{e^q + e^{-q}}$. [**2 marks**]
   (c) Consider another activation function called **SSIG** given by $\text{SSIG}(z) = \frac{\alpha z}{1+|\alpha z|}$ with $\alpha = 5$. Write python functions to compute the SSIG activation function and its (sub-)gradient. [**2 marks**]
   (d) Consider an appropriate neural network architecture where each hidden layer has only SRELU activation functions and the output layer has a logistic sigmoid activation function. Illustrate the exploding gradient and vanishing gradient problems in this network. Justify the architecture you used, indicate how you checked the exploding gradient and vanishing gradient problems, and explain your observations. [**3 marks**]
   (e) Consider an appropriate neural network architecture where each hidden layer has only SSIG activation functions and the output layer has a logistic sigmoid activation function. Illustrate the exploding gradient and vanishing gradient problems in this network. Justify the architecture you used, indicate how you checked the exploding gradient and vanishing gradient problems, and explain your observations. [**3 marks**]
   (f) Among the networks considered in questions 1d, 1e, describe which network was more prone to the vanishing gradient issue and which network was more prone to the exploding gradient issue. Use appropriate justifications for your observations involving the quantities used during backpropagation. [**5 marks**]

(g) Suppose you had a linear activation function at the output layer in the networks you considered in questions 1d, 1e. Compare and contrast the exploding gradient and vanishing gradient issues in the networks with linear activations at the output layer with that of networks with logistic sigmoid activation function at the output layer. Use appropriate justifications for your observations involving the quantities used during backpropagation. Using these observations, can you comment on the behavior if you had ReLU activation at the output layer (without actually running the experiments)? [**5 marks**]

2. **[Use only Python]** Use the code template for multi-layer perceptron (or feed-forward network) coded from scratch, posted in moodle. Consider the CharRecgn data set posted in moodle. In the file CharRecgn.csv, the first 784 columns denote the pixel values while the last column contains the class label. Note that there are 6 classes in the data set.

   (a) Write code to read the data into suitable `numpy` arrays for features and labels. [**1 mark**]

   (b) Write the required code to shuffle and split the data set into three sets S_1, S_2 and S_3 such that S_1 contains 70% of the data, S_2 contains 15% of the data and S_3 contains 15% of the data. Make sure that the splits contain similar label distributions. [**2 marks**]

   (c) Design a suitable feed forward neural network with 2 or 3 hidden layers with appropriate activation functions and a corresponding loss function to perform training on the S_1 split of CharRecgn data set. Justify the design choice of your neural network and loss function and implement the loss function in the code. [**3 marks**]

   (d) Illustrate how you will carry out backpropagation for the loss gradients in the layers. Include its implementation in code. [**3 marks**]

   (e) Use **accuracy**, **precision** and **recall** as performance metrics to assess the performance of your neural network. Recall that the following computations are used to compute accuracy, precision and recall:
   accuracy $= \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$ and **precision** $= \frac{\text{TP}}{\text{TP+FP}}$ and **recall** $= \frac{\text{TP}}{\text{TP+FN}}$, where TP, TN, FP, FN denote respectively true positives, true negatives, false positives and false negatives. Implement in code, the computation of **accuracy**, **precision** and **recall** computed over the data set samples. [**3 marks**]

   (f) For the chosen loss function, choose the learning rates from the set $\{0.01, 0.001, 10^{-4}\}$ and mini-batch sizes from $\{10, 20, 30, 40\}$. For each (learning rate, mini-batch size) pair, run the mini-batch stochastic gradient descent algorithm on S_1, with 200 epochs (use more epchs if necessary). For every 5 epochs, record the loss and accuracy, precision and recall achieved on the sets S_1 and S_2. Now plot the loss for every 5 epochs for each (learning rate, mini-batch size) pair on S_2 (use a single plot and different colors for different pairs). Plot the accuracy for every 5 epochs for each (learning rate, mini-batch size) pair on S_2 (use a single plot and different colors for different pairs). Also plot the precision for every 5 epochs for each (learning rate, mini-batch size) pair on S_2 (use a single plot and different colors for different pairs). Similarly plot also the recall for every 5 epochs for each (learning rate, mini-batch size) pair on S_2 (use a single plot and different colors for different pairs). Can you come up with a suitable selection procedure for the best (learning rate, mini-batch size) pair using the experiments conducted? Explain your selection procedure and justify. [**6 marks**]

   (g) Using the best (learning rate, mini-batch size) pair identified above, conduct training using mini-batch SGD on the set $S_1 \bigcup S_2$ with max epochs set to 500. For every 5 epochs, record the loss and accuracy, precision and recall achieved on the sets $S_1 \bigcup S_2$ and S_3. Include a stopping condition such that you can stop the training when the accuracy on the set $S_1 \bigcup S_2$ does not increase significantly for p epochs with a suitable choice for p. Plot the loss on $S_1 \bigcup S_2$ and S_3 in a single plot and comment on the observations. Plot the accuracy on $S_1 \bigcup S_2$ and S_3 in a single plot and comment on the observations. Similarly plot precision on $S_1 \bigcup S_2$ and S_3 in a single

plot and comment on the observations. Also plot recall on $S_1 \bigcup S_2$ and S_3 in a single plot and comment on the observations. [**5 marks**]
