CS-564 ML: Assignment 4

Date:- 17-Nov-2021 **Deadline:- 22-Nov-2021**

Instructions:

1. All the assignments should be completed and uploaded before the deadline.
2. Markings will be based on the correctness and soundness of the outputs. Marks will be deducted in case of plagiarism.
3. Proper indentation and appropriate comments are mandatory.
4. You should zip all the required files and name the zip file as roll_no.zip, eg. 1921cs28.zip. In case of groups, merge them as 1921cs28_1921cs29.zip
5. Upload your assignment (the zip file) in the following link: https://www.dropbox.com/request/EfMGNCp6OxhYIOnTtKmd
6. Make necessary assumptions if required. For further clarification, you can write an email: ratneshkr.joshi@gmail.com
7. Reporting parts of the questions must be done in a separate file.

Dataset: bbc news dataset
https://www.dropbox.com/s/tp3l54tnatvbldf/bbc.csv?dl=0

Dataset Info: The assignment targets document classification of the BBC news dataset. The dataset/files has already been converted to a single csv file for easier processing. The dataset has 2 features, first is the 'Article' which is the input for the task while the second 'Class' features classifies the article into one of 5 classes : business, entertainment, politics, sport and tech . You can find more info at http://mlg.ucd.ie/datasets/bbc.htm

What you have to do:

1. Implement the Feed Forward Neural Network for the sequence classification task. Implement the model with a minimum 2 hidden layers. Run the model for 5 epochs. Use the seed value of '1' for consistent results. Evaluate the model on the basis of performance metrics (accuracy, precision, recall) and number of parameters. Write your results in a separate file when submitting. Steps are
    a. Preprocess/clean the data(remove stopwords etc.)
    b. The input to Neural Network will be the articles, with tokens represented with a 100 dimension vector each(ie. Embedding size = 100).
    c. Make sure to pad the input to the maximum article length or fix a max length and truncate the longer sequences if you are facing memory issues.
    d. Use 70:10:20 split for training, validation and testing
    e. Use of dataloader is optional.

2. The general implementation for Neural networks for classification tasks often uses ReLU for hidden layers and sigmoid/softmax for final output. Run your model(created above) with different activation layers (Relu, softmax/sigmoid, tanh) at their appropriate positions and calculate the performance change for each combination. Specify which works best in case of your model. Do the same for various optimization algorithms available (SGD, Adam etc). State the results in the file when submitting.

3. Evaluation:
    a. Perform 3-fold cross-validation
    b. Compute the overall accuracy for each of the 3-folds (for all the classes together)
    c. Compute the class-wise accuracy for each of the 3-folds