CS:5007 Deep Learning

Feb - May 2021

Programming Assignment - 4

Implementing Backpropagation

Due Date: 9^{th} Apr 2021 11:59 PM

Follow the instructions given below carefully:

1. You are allowed to use ONLY the following python libraries: NumPy, Pandas, Matplotlib, random, math, sys
2. You are free to take help from the backpropagation code given in 'Demo 6 - Backpropagation'.
3. You must submit your code in a single python notebook with naming format as follows: Firstname_Lastname_assignment4.ipynb
4. Your code must be properly commented explaining each step clearly.
5. Penalty will be there if any of the above instructions are not followed.
6. Zero marks will be provided for plagiarised code.
7. NOTE: This assignment is **GRADED** and the total points for this assignment is **5**. An extra BONUS of 2 points will also be given to the students whose performance is in top 5.

Questions

1. (5 points) Implement your own fully connected feed-forward neural network with backpropagation from scratch for the task of classification on the given dataset.

   - **Dataset**
     This dataset is about the confirmed COVID-19 cases in Ontario. The original data comes from the following file "*Confirmed positive cases of COVID19 in Ontario*" at: `https://data.ontario.ca/dataset/confirmed-positive-cases-of-covid-19-in-ontario`. The original dataset is resampled and divided into training and test data, namely, data_train.csv and data_test.csv respectively, for this assignment. The target label is the column 'Outcome1' which is present only in data_train.csv. The labels are strings and three different labels are there as follows: Fatal, Resolved, Not Resolved. You should convert categorical features to numerical or one-hot encoded feature, as appropriate.
   - **Task**
     Classify the confirmed COVID-19 cases among the three different classes: Fatal, Resolved, Not Resolved. Setup and train your own fully connected feed-forward neural network model to perform this task.
   - **Evaluation Criteria**

     (a) (2 points) We will evaluate **Mean F1-Score** based on the predictions made by your model on data_test.csv. The F1 score, commonly used in information retrieval, measures accuracy using the statistics precision p and recall r. Precision is the ratio of true positives (tp) to all predicted positives ($tp + fp$). Recall is the ratio of true positives to all actual positives ($tp + fn$). The F1 score is given by:

     $$F1 = 2\frac{p \cdot r}{p + r} \text{ where } p = \frac{tp}{tp + fp},\ r = \frac{tp}{tp + fn}$$

     The F1 metric weights recall and precision equally, and a good retrieval algorithm will maximize both precision and recall simultaneously. Thus, moderately good performance on both will be favoured over

Row_ID	Outcome1
1	Fatal
2	Resolved
3	Not Resolved

extremely good performance on one and poor performance on the other.
You must submit your predictions in a .csv file with the name of the file as 'Predictions.csv'. The file should contain two columns: Row_ID and Outcome1. The file should contain a header and have the following format:

(b) (3 points) Write a report including the following:

i. Give a detailed explanation of your approach which must include: Data preprocessing, model architecture, loss function, optimizer used and relevant hyperparameters.

ii. Mention how you evaluated your model's generalization performance. Plot the train accuracy and validation accuracy per iteration curve and write your observations. **NOTE**: You must submit the report in **PDF format** with the name of the file as 'Report.pdf'.