Machine Learning Assignment 3

Dr Burcu Can

February, 2021

1 The Dataset

You are given an IMDB dataset of 50K Movie Reviews that involves binary sentiment information (positive or negative) for each review. The dataset consists of 25K movie reviews for training and 25K movie reviews for testing.

The dataset is given as a CSV file. Each row has got the review and its associated sentiment column (positive or negative).

2 Task 1

Perform preprocessing on the given dataset such as case removal, tokenization, stopword removal, space removal etc. You can apply any other preprocessing operation of your choice.

3 Task 2

Build a 1-layer multi-layer perceptron (MLP) to predict the sentiment of each review in the test set. Describe your model along with the regularization, activation functions, input/output dimensions. Give the evaluation scores.

4 Task 3

Build a 2-layer MLP to predict the sentiment of each review in the test set. Describe your model along with the regularization, activation functions, input/output dimensions. Give the evaluation scores. Compare it with the 1-layer MLP that you implemented in the previous task.

5 Task 4

Finally, compare your neural models with at least one traditional classification algorithm.

4 Report

Write a report on your findings. Enclose your source code.

https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

5 Submission Information

- Length: Maximum 1,500 words in total
- The source code also needs to be submitted along with the report.
- Format of the files to be submitted: a zip folder including your report in .pdf format, and your source codes.
- Deadline: 10 June 2021, 12pm (GMT)
- Submitted via Canvas

6 Marking

- Task 1: 20%
- Task 2: 20%
- Task 3: 20%
- Task 4: 20%
- Overall report: 20%