Homework 4

COM SCI X 450.4 - Machine Learning

Thursday 30th July, 2020

1 Introduction

This homework will be divided into 2 parts.

In the first part, you will work on image data. You will train a model to predict if an image contains a dog or a cat. You will then use Deep Learning models to extract features from images like we learned in class. For the second part, you will perform predictions on textual data. You will be working on the sentiment analysis task for 2 different social media datasets.

2 Part I - Image Classification (48 points + 10 bonus)

For this section, you will download the dataset of images containing cats and dogs from Canvas (it is a subset from the original Kaggle competition). You will follow the Notebook and complete the necessary steps in the code.

2.1 Exercise 1 - Making predictions using pixels

- What algorithm did you decide to use?
- What is the size of your feature vector?
- What hyper-parameters were used?
- Report your results using the classification report. Is your model better at predicting dogs or cats?

2.2 Exercise 2 - Using image features extracted from DL models

- Is this algorithm faster or slower than the previous one? Why?
- What is the size of your feature vector?
- What hyper-parameters were used?
- Report your results using the classification report. Did your model improve?

- Please explain the steps taken and why there are differences in the results.

Extra Points: Test at least 3 Deep Learning models for feature extraction and report the results. Which one performs better? Why do you think it performs better(you might have to do some research on what those models are doing)?**(10 extra points)**

3 Part 2 - Text Classification (48 points +10 bonus)

For this section, you will download the dataset of text labeled for sentiment analysis, containing data from Twitter and Reddit, from Canvas (it has been taken from the Twitter and Reddit Sentimental analysis Dataset on Kaggle. You will follow the Notebook and complete the necessary steps in the code.

3.1 Exercise 3 - Using count-based methods to represent text

- What algorithm did you decide to use?
- What is the size of your feature vector?
- What hyper-parameters were used?
- What is the difference between the two models? Which one is better and why do you think this happens.

3.2 Exercise 4 - Train a sentiment analysis model on Twitter data

- How many epochs did it take to converge?
- Add your history plot (by using the *plot_history* function) to your report. Did your model overfit?
- (Bonus) Modify the hyper-parameters. What is the best results you can get? What happens if you add more layers?

3.3 Exercise 5 - Compare the models

- Which algorithms took longer to train?
- Which algorithms took longer to make predictions?
- Which algorithms got the best performance? Why?

4 How long did it take? (4 Points)

Please add to your written document how many hours you took to finish this assignment.

5 Deliverable

You will submit your assignment before Sunday, July 19th, before 11:59pm. You should submit a zipped file with:

- The Python Notebook with the code you wrote
- A separate PDF document with the answers to the questions.

Start early, especially if you are not familiar with Python! If you require help, please contact instructor. Office hours can be scheduled if necessary as long as the instructor is contacted with 48 hour notice.