Amirali Molaei

Iran University of Science and Tech.
www.iust.ac.ir

Assignment 3 Problems

Advanced Data Mining: Spring 1401: Dr. Minaei
Due Friday, Tir 3th, 1401

Contents

Figures

Problem 1

In this section, you're going to implement algorithms to perform classification and clustering tasks on the MNIST dataset. For each task, you should consider the corresponding items specified in the following sections:

Classification

1. Utilize a Deep Neural Network model to learn this task
2. Include the following function in your code:

```
class DNN(nn.Module):
    # Deep Neural Network class here.
    # Hints: 1- Define the architecture of the class in the init method.
    #        2- Define a method that takes training data as input
    #           and outputs softmaxed logits.
    def __init__(self, feature_size, output_size, seed):
        super(DNN, self).__init__()

    def forward(self, state):

        return # Softmax of the logits
```

1) Deep Neural Network class Structure

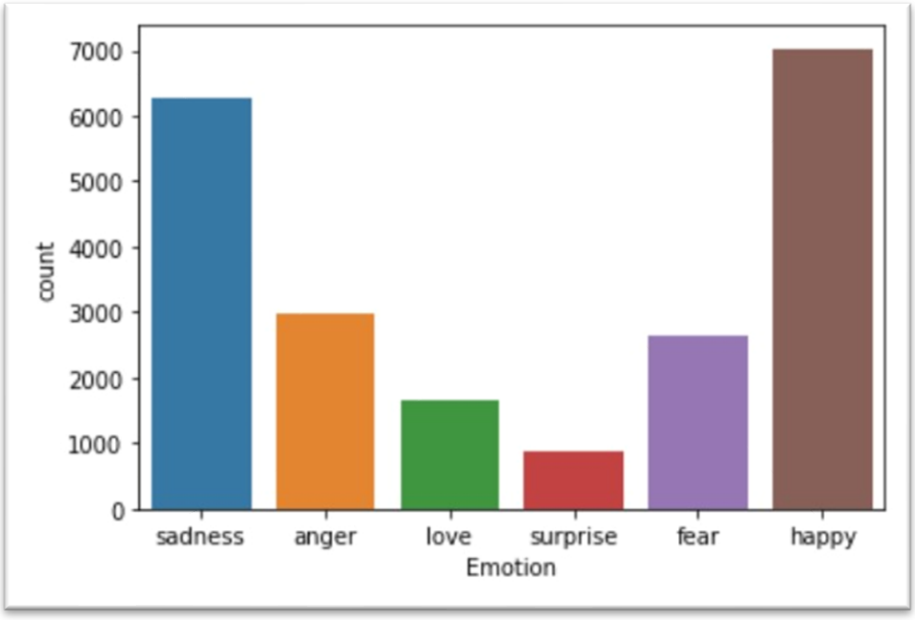
3. Define a predict function that takes a test sample as input and outputs a label.

Clustering

1. Utilize a clustering algorithm that you've learned from the course for cluster analysis.
2. Compare the performance with the classification section with proper metrics.

Problem 2

In this problem, you're going to utilize huggingface transformer models for Multiclass classification on the twitter emotion dataset. The dataset is available on **Kaggle**. The task is defined as predicting the emotion from a user's twit that is written on the Twitter platform. The label distribution is as follows:

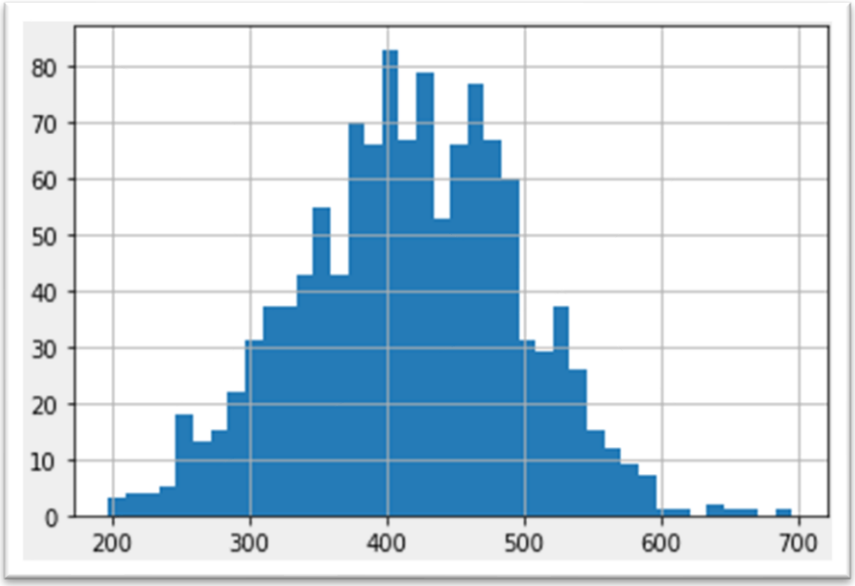


2) Label Distribution of the twitter emotion dataset

Consider the following items in your implementation:

Considerations

1. You should only use the transformers library for model implementation (not any other libraries, e.g., simpletransformers, etc).
2. You have to use the Pytorch framework for your implementation
3. Show sentence length distribution like *3*) Example of sentence Length distribution representation by histogram *3*, then use this information for encoding your data. (note that this is Not the distribution for your dataset!)
4. Utilize four different transformer models: Bert, Roberta, XLnet, and Distilbert. then provide a table like *4*) The format of *the* evaluation table *4* for evaluation and comparison.
5. Visualize the performance of the models with a confusion matrix



3) Example of sentence Length distribution representation by histogram

model	epoch	Learning rate	Training accuracy	Training loss	Precision	Recall	F-Score(weighted)
Bert-large-uncased							
Xlnet-large-uncased							
Roberta-large							
Distillbert-base-uncased							

4) The format of the evaluation table

Questions

1. As you can see in *2*) Label Distribution of *the* twitter emotion dataset 2, the dataset is imbalanced. How can this affect the learning process?

2. Why did you use the Transformers for this task?

Bonus Part

A simple way to tackle the imbalanced dataset problems is data augmentation. Use this technique to balance the label distribution, then repeat the process of the previous section for every model.

Notes

- You should write a report on how you implemented your code. **No score** will be given to a code without a **report**!
- If you have any questions, feel free to ask. You can ask your questions in the Telegram group.
- Please upload your assignments as a zipped folder with all necessary components. Upload your file in *HW3_ADM_YourStudentID_YourName.zip* format.