PRNN - 24, Assignment 3

Prof Prathosh A P

April 1, 2024

General Instructions:

- This assignment involves two datasets one on vision and one on text.
- **Animal Image Dataset:-** This dataset contains colour images of animals from 10 different categories. The link for the dataset can be found here - Every folder contains images from one class. The link for the dataset is here https://www.kaggle.com/datasets/alessiocorrado99/animals10.
- **News Text data:** - This dataset contains around 210k news headlines from 2012 to 2022 from HuffPost. Each record in the dataset consists of the following attributes: **Category:** category in which the article was published. **Headline:** the headline of the news article. **Authors:** list of authors who contributed to the article. **Link:** link to the original news article. **Short description:** Abstract of the news article. **Date:** publication date of the article. Dataset be downloaded here - https://www.kaggle.com/datasets/rmisra/news-category-dataset.
- You **ARE NOT ALLOWED** to use any off-the-shelf Python-based ML libraries such as SK-learn, Pytorch, and others. However, you can use the functions and classes that you built in the previous assignments.
- We will run a plagiarism check on both your report and the codes. Any suspicion of copying would lead to a harsh penalty from negative marks in the assignment to a failing grade in the course, depending upon the severity. Therefore, kindly refrain from copying others' codes and/or reports.

1 General Tasks

The following implementations have to be made on both datasets.

1. Implement a self-attention block from scratch. Make the token length and number of attention layers as the hyper-parameters.
2. Implement PCA as a class with projected dimensionality as a hyper-parameter.
3. Implement K-means clustering with different distance metrics as hyper-parameters.
4. Implement a Decision tree classifier with multiple impurity functions as hyper-parameters.
5. Implement gradient boosting that can take multiple classifiers as inputs and perform assembling.

2 Vision Dataset

Here the goal is a 10-class classification problem in the supervised case and clustering in the unsupervised case. The metrics for the classification case are accuracy and F1 score while for the clustering it is Normalized Mutual Information (NMI). The dataset has to be divided into train-validation-test splits in a 60:10:30 ratio and the results have to be reported via cross-validation.

1. Solve the 10-class classification problem with a CNN with images as inputs.

2. Use PCA to reduce the feature dimensions and apply MLP on dimensionality-reduced features. Compare the results with the CNN.
3. Use a transformer model with self attention on PCA'ed features.
4. Run K-means with different values on both raw pixes and PCA'ed data features.
5. Use the Ensemble of CNN/MLP/Decision tress in an ada-boost framework and compare the results with non-Ensemble models.
6. Extract the features from a pre-trained Imagenet, append and train an MLP on top of it.

3 Text Tasks

Here use TF-IDF as embeddings (https://en.wikipedia.org/wiki/Tf%E2%80%93idf. You may use SkLearn only for this. Define a 12-class classification problem using the top 12 categories. The input features are taken from headlines.

1. Solve the 12-class classification problem using an MLP. Append all the data with zeros to convert them into the same size.
2. Repeat the above experiment with a transformer with self-attention.
3. Implement a Random forest and a Gradient boosted tree and compare the results.