DASE7140 Final Project

Mysterious Time Series

1 Description

A customer offered us a few samples of their measurements, whose background of application was kept a secret from us. The training `dataset` consisted of 90 files with a file name format `XXXXX-Y.csv` where `XXXXX` is a randomly generated string with no actual meaning, and `Y` (0 or 1) is the label of the corresponding data. As we can infer from the labels, this is a binary classification task. Opening each file, we see a floating point number series with a length of 8,192.

The customer also told us how they classified the data. They calculated the spectral density of the time series of 8,192 points, observed the spectrum, and then assigned a label to the data. However, the criteria of the observation were not unveiled. Whether to utilize the information depends on you.

The `test` dataset has a similar format with 110 files. The only difference is that the filenames include no label.

2 Your tasks

T1| You are required to design three different models (either neural networks or traditional machine learning algorithms) to solve the classification problem. You need to provide three training Python scripts (`train1.py`, `train2.py`, and `train3.py`) to save the models respectively, and three inference Python scripts (`inference1.py`, `inference2.py`, and `inference3.py`) to load the corresponding models and save the predicted results as plain TXT files `results1.txt`, `results2.txt`, and `results3.txt` for the test dataset.

Note that, each line of `results?.txt` should be the corresponding label for a data file in the following order. In case some students overlook the sorted order, the output should be in two columns, the first is the filenames without extension, the second is the classification results, and in-between is a white space.

```
>>> sorted(os.listdir())
['00D9c.csv', '06ade.csv', '06d3E.csv', '0F13B.csv', '0F5F3.csv',
'11A9E.csv', '13FA2.csv', '1B5bF.csv', '1BA7e.csv', '1D24d.csv',
'1Da11.csv', '1bD0D.csv', '21A6f.csv', '23394.csv', '260BF.csv',
'26ACa.csv', '29a4d.csv', '2B4ba.csv', '2DBE6.csv', '3AF85.csv',
'3ab7A.csv', '44Bc3.csv', '44eb7.csv', '49e5c.csv', '4BaB7.csv',
'4CaF0.csv', '4db91.csv', '53a5E.csv', '57b9D.csv', '57d5D.csv',
'5D8f3.csv', '5EE36.csv', '5ab7f.csv', '5fa18.csv', '63EbE.csv',
'64aaa.csv', '67130.csv', '67CfD.csv', '6Cb39.csv', '6cb19.csv',
'81afa.csv', '87484.csv', '8E0ce.csv', '921ba.csv', '9533a.csv',
'99a8A.csv', '9E7ef.csv', 'A8cdd.csv', 'AeABd.csv', 'B194b.csv',
'B1dDf.csv', 'B8601.csv', 'B8eCc.csv', 'B9D78.csv', 'B9c0b.csv',
'BDbd5.csv', 'BEfAe.csv', 'BccDB.csv', 'Bea5e.csv', 'BfCB5.csv',
'C4923.csv', 'C4DBC.csv', 'C6F13.csv', 'CA6D5.csv', 'CbC5e.csv',
'D4f14.csv', 'D9eA6.csv', 'De21A.csv', 'De504.csv', 'E24D8.csv',
'EC270.csv', 'ED28a.csv', 'EaA2d.csv', 'EcBAa.csv', 'Ece1b.csv',
'F3273.csv', 'F454e.csv', 'F45d4.csv', 'FCEfe.csv', 'FDa2a.csv',
'FEABB.csv', 'Fd6E7.csv', 'a2Ce0.csv', 'a367b.csv', 'aAc90.csv',
'aDd28.csv', 'ad07C.csv', 'b2C56.csv', 'b73Ba.csv', 'bAE4b.csv',
'bC2F6.csv', 'bDCeb.csv', 'bF8BB.csv', 'bFDd4.csv', 'bc9Ed.csv',
'c631e.csv', 'c6C09.csv', 'cbF72.csv', 'd2bD9.csv', 'd37ec.csv',
'd5A4f.csv', 'd5d48.csv', 'd652C.csv', 'dAf32.csv', 'e8Cee.csv',
'eBa3d.csv', 'eDBd8.csv', 'f3E64.csv', 'fD0C9.csv', 'fa08e.csv']
```

For example, the **first 10 lines** of your `results?.txt` may look like

```
00D9c 1
06ade 0
06d3E 0
0F13B 0
0F5F3 0
11A9E 1
13FA2 1
1B5bF 0
1BA7e 1
1D24d 0
```

Note that the 0–1 series here is generated by `np.random.randint(2, size=110)`, so it doesn't serve as a hint about the task.

T2| You are required to write a report that contains

- Description of the three models;
- Comparison between the performances of the models;
- Analyses of the results.

Your deliverables should include

1. The training scripts: `train1.py`, `train2.py`, and `train3.py`;
2. The inference scripts: `inference1.py`, `inference2.py`, and `inference3.py`;
3. The inference results: `results1.txt`, `results2.txt`, and `results3.txt`
4. A report: `report.pdf`

3 Important hints

The teaching assistants will evaluate your submission as follows:

1. Whether all the codes are runnable.
2. The completeness of the report. Extra insights will gain bonus points.

3. The results `results1.txt`, `results2.txt`, and `results3.txt` will be evaluated, and the one with the highest accuracy will be used for ranking. The marking criteria for rankings may be modified according to the average performance of all the students' submissions. To provide intuitive understanding, assuming the full mark for this part to be 12 points, here is an example:

   - Top 20% will get 12 points.
   - Top 20%–60% will get 8 points (If there's a tie at the top 60% place, all students at this place get 8 points).
   - For the remaining: ACC$\geq$0.70, 4 points; 0.60$\leq$ACC$<$0.70, 2 points; ACC$<$0.6, 0 points.