ML805 - Advanced Machine Learning

Assignment for the module of learning with noise

Due: 15 March 2025, 23:59PM

Introduction

The objective of this assignment is to estimate a transition matrix estimator and two classification algorithms that are robust to label noise.

Three input datasets are given. For each dataset, the training and validation data contains class-conditional random label noise, whereas the test data is clean. You need to build **at least two different classifiers** trained and validated on the noisy data, which can have a good classification accuracy on the clean test data. You are required to compare the robustness of the two algorithms to label noise.

For the first two datasets, the transition matrices are provided. You can directly use the given transition matrices for designing classifiers that are robust to label noise.

For the last dataset, the transition matrix is not provided. You are required to build **a transition matrix estimator** to estimate the transition matrix. Then, employ your estimated transition matrix for classification. Your estimated transition matrix must be included in your final report. Note that to validate the effectiveness of your transition matrix estimator, you could use your estimator on the first two datasets and compare your estimation to the given transition matrices.

Data prepossessing is allowed, but please remember to clarify and justify it in the report carefully.

1 A Guide to Using the Datasets

Three image datasets with .npz format are provided.

1.1 Attributes Contained in a Dataset

The following code is used to load a dataset and check the shape of its attributes.

```
import numpy as np
# Remember to replace the $FILE_PATH
dataset = np.load($FILE_PATH)
Xtr = dataset['Xtr']
Str = dataset['Str']
Xts = dataset['Xts']
Yts = dataset['Yts']
print(Xtr.shape)
print(Str.shape)
print(Xts.shape)
print(Yts.shape)
```

1.1.1 Training and validation data

The variable **Xtr** contains the **features** of the training and validation data. The shape is $(n, image\_shape)$ where n represents the total number of the instances.

The variable **Str** contains the **noisy labels** of the n instances. The shape is $(n,)$. For all datasets, the class set of the noisy labels is $\{0, 1, 2, 3\}$.

Note that do not use all the n examples to train your models. You are required to independently and randomly sample **80%** of the n examples to train a model and use the rest **20%** examples to validate the model.

1.1.2 Test data

The variable **Xts** contains **features** of the test data. The shape is $(m, image\_shape)$, where m represents the total number of the test instances.

The variable **Yts** contains the **clean labels** of the m instances. The class set of the clean labels is also $\{0, 1, 2, 3\}$.

1.2 Dateset Description

1.2.1 FashionMINIST0.3.npz

Number of the training and validation examples $n = 24000$.

Number of the test examples $m = 4000$.

The shape of each example $image\_shape = (28 \times 28)$.

The transition matrix $T = \begin{bmatrix} 0.7 & 0.3 & 0 & 0 \\ 0 & 0.7 & 0.3 & 0 \\ 0 & 0 & 0.7 & 0.3 \\ 0.3 & 0 & 0 & 0.7 \end{bmatrix}$.

1.2.2 FashionMINIST0.6.npz

Number of the training and validation examples $n = 24000$.

Number of the test examples $m = 4000$.

The shape of each example $image\_shape = (28 \times 28)$.

The transition matrix $T = \begin{bmatrix} 0.4 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.4 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.4 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.4 \end{bmatrix}$.

1.2.3 CIFAR.npz

Number of the training and validation examples $n = 20000$.

Number of the test examples $m = 4000$.

The shape of each example $image\_shape = (32 \times 32 \times 3)$.

The transition matrix T is unknown.

2 Performance Evaluation

The performance of each classifier will be evaluated with the top-1 accuracy metric, that is,

$$\text{top-1 accuracy} = \frac{\text{number of correctly classified examples}}{\text{total number of test examples}} * 100\%.$$

To have a rigorous performance evaluation, you need to **train each classifier at least 3 times with the different training and validation sets generated by random sampling**. Then report both the **mean** and the **standard deviation** of the test accuracy.

3 Tasks

You need to implement at least **two label noise robustness classifiers** with at least one not taught in this course and test their performance on the three datasets. You need to implement **an estimator to estimate the transition matrix**. You are allowed to use external libraries for optimization and linear algebraic calculation.

3.1 Image Classification with Known Flip Rates

For the first two datasets, the transition matrices are provided. You can directly use the given transition matrices for designing classifiers that are robust to label noise. As mentioned in the **section 2**, for each classifier, you should report the mean and the standard derivation of the test accuracy.

3.2 Image Classification with Unknown Flip Rates

For the last dataset, Since the transition matrix is not provided, you need to implement an estimator to **estimate the transition matrix**. Then use the estimated transition matrix to build a noise robust classifier. Note that you can use the provided transition matrices of the first two datasets to validate the effectiveness of your transition matrix estimator. You need to include your estimated transition matrix in the final report. You also need to report the mean and the standard derivation of the test accuracy for each of your designed noise robustness classifiers. Both estimation accuracy of the transition matrix and the test accuracy on the last dataset contribute to the final mark.

3.3 Report

The report should be organized similar to research papers, and should contain the following sections:

- In **abstract**, you should briefly introduce the topic of this assignment, your methods, and describe the organization of your report.
- In **introduction**, you should first introduce the problem of learning with label noise, and then its significance and applications. You should give an overview of the methods you want to use.
- In **related work**, you are expected to review the main idea of related label noise methods (including their advantages and disadvantages).
- In **methods**, you should describe the details of your classification models, including the formulation of the objective functions, the theoretical foundations or views (if any) of the objective functions, and the optimization methods. You should describe the details of the transition matrix estimation methods, theoretical foundations (if any), and optimization algorithms.
- In **experiments**, you should introduce your experimental setup (e.g., datasets, algorithms, evaluation metric, etc.). Then, you should show the experimental results, compare, and analyze your results. If possible, give your personal reflection or thoughts on these results.
- In **conclusion**, you should summarize your methods, results, and your insights for future work.
- In **references**, you should list all references cited in your report and formatted all references in a consistent way.
- In **appendix**, you should provide instructions on how to run your code.

The layout of the report:

- Font: Times New Roman; Title: font size 14; Body: font size 12
- Length: maximum 10 pages (reference not included).

Note: Submissions **must** be typeset in LaTex using the provided template.

4 Submissions

Detailed instructions are as follows:

1. The submission contains two parts: **report** and **source code**.
   (a) report (a pdf file): the report should include each member's details (student id and name).
   (b) code (a compressed folder)
      i. algorithm (a sub-folder): your code could be multiple files.
      ii. data (an empty sub-folder): although two datasets should be inside the data folder, please **do not** include them in the zip file. We will copy those datasets to the data folder when we test the code.
2. The report (file type: pdf) and the codes (file type: zip) must be named as student's name and ID. For example, "Tongliang Liu AI40059.pdf".
3. Your submission should include the report and the code. A plagiarism checker will be used.
4. You need to clearly provide instructions on how to run your code in the appendix of the report.
5. A penalty of minus 5% marks per each day after due (email late submissions to TA and confirm late submission dates with TA). The maximum delay is 5 days, after that assignments will not be accepted.

5 Marking scheme

Category	Criterion	Marks	Comments
Report [80]	Abstract [3] •problem, methods, and organization Introduction [6] •the problem you intend to solve •the importance of the problem Previous work [8] •previous relevant methods used in literature •their advantages and disadvantages Label noise methods with known flip rates [23] •pre-processing (if any) •label noise methods' formulation •cross-validation method for model selection or avoiding overfitting (if any) •experiments •discussions Noise rate estimation method [12] •noise rate estimation method's formulation •experiments •discussions Label noise methods with unknown flip rates [10] •pre-processing (if any) •label noise methods' formulation (if different from above) •cross-validation method for model selection or avoiding overfitting (if any) •experiments •discussions Conclusions and future work [3] •meaningful conclusions based on the results •meaningful future work suggested		

	Presentation [8] •academic style, grammatical sentences, no spelling mistakes •good structure and layout, consistent formatting •appropriate citation and referencing •use graphs and tables to summarize data Other [7] •at the discretion of the assessor: illustrate outstanding comprehensive theoretical analysis, demonstrate the insightful and comprehensive assessment of the significance of their results, provide descriptions and explanations that have depth but clarity, and are concisely worded		
Code [20]	•reasonable code running time •well organized, commented and documented		

Note: Marks for each category is indicated in square brackets. The minimum mark for the assignment will be 0 (zero).