Assignment 2: Learning from a Data Set

[Advanced Machine Learning Course]

Overview

Your job includes,

1) Read the description of the task and download the data set

2) Implement an algorithm and output the prediction

3) Write a report

4) Submit your work

TA's job is to evaluate your work.

Here are some baselines for your consideration.

CAUTION: DO YOUR JOB ALONE!

Task Decription and dataset download

The Story: We create a weakly supervised image classification task using the Fashion-MNIST dataset. Each image is 28x28 pixels, represented as a flattened array of 784 pixel values (ranging from 0 to 255). Each label is an integer value ranging from 0 to 9. In the training set, 20% of the image labels are retained, while the remaining 80% are discarded to serve as unlabeled training data. We ensure the balance of sampling, the specific details are as follows:

- **train_image_labeled.csv**: The labeled images consist of a total of 12,000 samples, encompassing 10 classes, with 1,200 images per class.
- **train_label.csv**: The training image labels consist of a total of 12,000.
- **train_image_unlabeled.csv**: The shuffled unlabeled images data consist of a total of 48,000 samples, with 4,800 unlabeled images per class. You need to fully utilize these unlabeled data in your algorithm.
- **test_image.csv**: The test data consists of a total of 10,000 test images.

Download:

Original data file

Task: Predict "label" for each record in the **test_image.csv**

Implementation and Output

Implementation: It is up to you to implement any learning algorithm with any programming language. It is encouraged that you can make some detail analysis about the difficulty of this task, and this will be good to you to find out appropriate learning algorithm. Problem analysis and innovative thoughts will help you get higher score.

Output: The output of your learning algorithm should be a txt file "yourId.txt" which contains 1000 lines without header, each line is prediction for the corresponding example. Please do not make confusion about the order of test example, otherwise you may get a very low accuracy.

How to write and what to write

Your report should includes:

1) Your understand and analysis of the problem;

2) The motivation of your algorithm and introduction of the background of your algorithm;

3) Full technical details of your algorithm, especially including pseudo code of your algorithm;

4) Description or analysis of the performance you got;

5) Conclusion and (optional) discussion

CAUTION: NOT PLAGIARIZE! OTHERWISE, YOU WILL GET PUNISHMENT!

Please use MSWord template or LaTeX template to write your report in chinese with english abstract. Attention, please transform your source file to PDF file for submission.

Name your PDF file with "report.pdf".

How to submit

Your submission includes:

1) 'yourId.txt' file : containing 10000 lines of predictions;

2) 'report.pdf' file : your report;

3) source file of your algorithm (Do not submit the whole data set or the trained model to FTP)

Please carefully check out your submission.

Note that "yourId" should be replaced with your ID and the name of the files should not be other names.

Pack all your files into a single compressed file (compress in ZIP format).

Name the compressed file using your student ID and version number, e.g., "AB12345678_v1.zip". We will take your file with highest version number as your final homework, e.g., "AB12345678_v2.zip"
Please delete the .bak, i.e., the backup files from your final zip files.

Upload your compressed file to FTP:

sftp://www.lamda.nju.edu.cn

switch path to: /D:/courses/AML25/assignment2

username: aml25

password: course01234!@#$

Evaluate your work

Evaluate of your prediction: According to your "yourId.txt", we will use macro F1 score to evaluate your prediction. As for macro F1 score, you may refer to https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html.

Evaluate of your report: Novel idea, sound techniques, and beautiful writing gain you high scores. See also The evaluation of your report in Assignment 1.

Evaluate of your source code: Fake and plagiarized source codes receives low scores.

Baselines

Some classic algorithms such as Random Forest, Logistic Regression, XGBoost GBDT, SVM, MLP, GBDT etc., serve as baselines. The performance of these algorithms can be referenced, and we look forward to you providing better solutions.

Download:

Baselines

About the DEADLINE and Score.

Deadline: 2025.05.09

Scores: 25%

- 下载

- 投屏

高速下载