Shanghai Jiao Tong University

SE3332: Machine Learning

Fall 2023

Lab 1

Due: Nov, 7th, 2023 23:59

1 Objectives

The objective of this project is two-fold:

1. To acquire a better understanding of classification methods by implementing Logistic Regression and Supporting Vector Machine with gradient descent.
2. Learn how to implement neural networks using PyTorch.
3. To compare the performance of your implementations by conducting an empirical comparative study on real-world data sets.

2 Major Tasks

The project consists of the following tasks:

1. Implement a logistic regression model using gradient descent method for classification.
2. Implement an SVM model using gradient descent method for binary classification.
3. Implement the MLP model using PyTorch for classification.
4. Conduct a comparative study of the above three methods.
5. To write up a project report.

Each of these tasks will be elaborated in the following subsections.

2.1 Logistic Regression with gradient descent method

In class we have elaborated how to implement logistic regression using gradient descent and provide a pseudo code in the slides. You are expected to implement this algorithm using Python. The gradient-descent algorithm requires that the step size parameter η be specified. Try out a few values (<1) and choose one that can lead to stable convergence. You may terminate the learning procedure if the improvement between iterations is not larger than a small threshold or if the number of iterations has reached a prespecified maximum number. Since the solution found may depend on the initial weight values chosen randomly, you may repeat each setting multiple times and report the average classification accuracy.

2.2 Support Vector Machine with gradient descent method (primal)

The Support Vector Machine algorithm discussed in class uses gradient descent to minimize the loss function. You are expected to implement this algorithm using Python. Again you need to specify the learning parameters such as η.

2.2 MLP using PyTorch

In class we have demonstrated how to implement MLP using PyTorch. You are expected to practice this tutorial and realize an MLP by yourself.
Firstly, you should learn how to use the popular deep learning framework PyTorch from online materials (e.g., https://pytorch.org/tutorials/). For your convenience, we also provide a tutorial about PyTorch in the slides.
Like logistic regression and SVM, you are also required to tune parameters of your implementation and show the optimal results.

2.3 A Comparitive Study

You can download a dataset from canvas. Every row in the "X" files stores features of one example while the "Y" files stores the labels in corresponding row. As is always the case, **the test sets should not be used for classifier training but only for measuring the classification accuracy.** You can do some processing before using the data.

For the data set, the following methods will be compared with respect to the classification accuracy on the training and the test sets, respectively:

- Logistic Regressiong using gradient descent
- SVM using gradient descent
- MLP using PyTorch

You are expected to also report the time required by each of the methods to complete the task, excluding the time needed for loading the data files. This may be done using the *time* function.

2.4 Report Writing

In your report, you need to present the **parameter settings** and the **experimental results**. Besides reporting the classification accuracy (for both training and test data) in numbers, graphical aids should also be used to compare the performance of different methods visually. For the CPU time information, you may just report it in numbers.

3 Tips

You are expected to do the implementation all by yourself so you can gain a better understanding of the algorithm. Python is the preferred language choice which can

allow you to do fast prototyping possibly at the expense of run-time efficiency. You may also use some other programming languages such as C++ and Java if you insist, but this is not recommended because you then cannot take advantage of the powerful and convenient matrix manipulation capabilities and built-in functions provided by Python.

Your programs should be written in such a way that the TA can run them easily to verify the results reported by you.

3 Project Submission

Project submission should be done electronically using the canvas.

There should be two main files in your submission:

1. Project report (with filename report): preferably in PDF format.
2. Source code and a README file: The code is preferably in .ipynb format. Turn in all necessary code for running your program as well as a brief user guide for the TA to run the programs easily to verify your results, all compressed into a single ZIP file. **The data should not be submitted to keep the file size small unless you do other preprocessing operations.**

We hope you can create your work according to the directory structure shown in the figure below. **Again, do not submit data.**
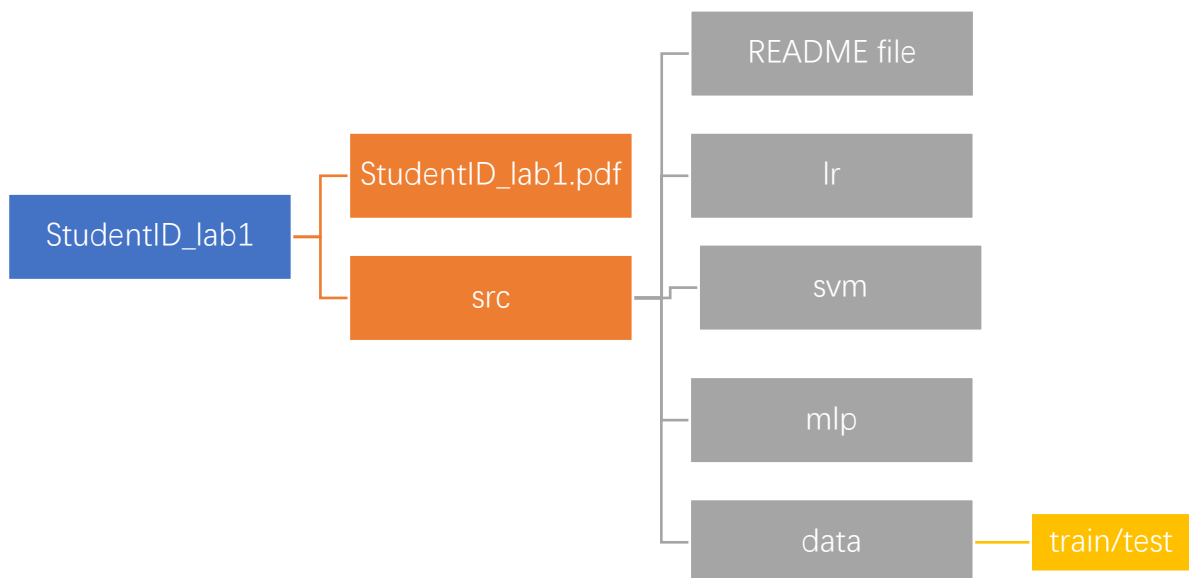
4 Grading Scheme

This project will be counted towards [20pt/3] 6.67% of your final course grade. The weights for different tasks are as follows:

• Implement MLP in the PyTorch framework [8pt]

    Complete the code [4pt]

    The accuracy on the test set reach 80% [2pt]

    The accuracy on the training set reach 85% [2pt]

• Implementation of SVM using gradient descent [4pt]

    Complete the code [2pt]

    The accuracy on the test set reach 80% [2pt]

• Implementation of Logistic regression using gradient descent [4pt]

    Complete the code [2pt]

    The accuracy on the test set reach 80% [2pt]

• Project report [4pt in total]

    Completeness (implementatios, hyperparameters, evaluation metrics, changes of loss, change of accuracy during training). [2pt]
    Compare the three methods and analyze the results. [2pt]

6 Academic Integrity

While you may discuss with your fellow classmates on general ideas about the project, your submission should be based on your own independent effort. In case you seek help from any person or reference (from the Web or other sources), you should state it very clearly in your submission. Failure to do so is considered plagiarism which will be handled with appropriate disciplinary actions.

Please contact Hu Chao (TA) for any questions.