Shanghai Jiao Tong University

SE332-1: Machine Learning

Fall 2022

Project 1

Due: Nov, 11st, 2022 11:11pm

1 Objectives

The objective of this project is two-fold:

1. To acquire a better understanding of classification methods by implementing Support Vector Machine in two ways: with gradient descent method, and using a public-domain software package.
2. To compare the performance of your two implementations by conducting an empirical comparative study on real-world data sets.

2 Major Tasks

The project consists of the following tasks:

1. To implement an SVM model using gradient descent method for binary classification.
2. To use the SVM model implemented in the sklearn package for binary classification.
3. To conduct empirical study to compare the above two methods.
4. To write up a project report.

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

2.1 Support Vector Machine with gradient descent method (primal)

The Support Vector Machine algorithm discussed in class uses gradient descent to minimize the loss function. Your implementation may be based on this gradient-descent algorithm which 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.

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.

2.2 Support Vector Machine with sklearn

Using the Support Vector Machine model implemented in the sklearn package.

2.3 Empirical 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 set and the test set separately:

- SVM using gradient descent
- SVM using sklearn

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.

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

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 Some Programming Tips

As is always the case, good programming practices should be applied when coding your program. Below are some common ones but they are by no means complete:

• Using functions to structure program clearly
• Using meaningful variable and function names to improve readability

• Using indentation
• Using consistent styles
• Including concise but informative comments

For Python in particular, you are highly recommended to take full advantage of the built-in functions. Also, using loops to index individual elements in matrices and arrays should be avoided as much as possible. Instead, block indexing without explicitly using loops is much more efficient. Proper use of these implementation tricks often leads to speedup by orders of magnitude.

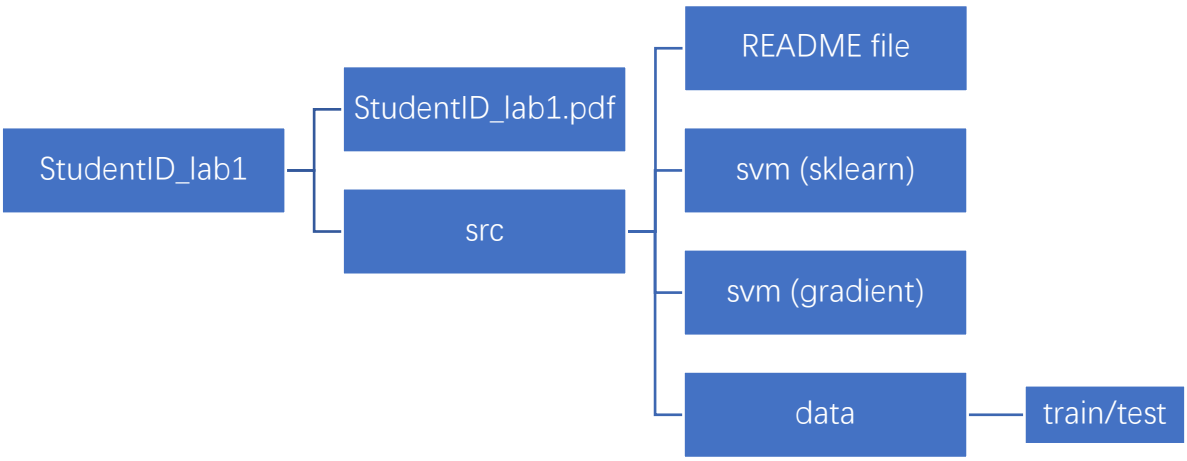
4 Project Submission

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.**



5 Grading Scheme

This project will be counted towards 10% of your final course grade. The weights for different tasks are as follows:

- Using SVM in the sklearn package [2% in total]

  Complete the code [2%]

- Implementation of SVM using gradient descent [4% in total]

  Complete the code [2%]

  The accuracy on the test set reach 80% [1%]

  The accuracy on the test set reach 85% [1%]

- Project report [4% in total]

  For SVM using sklearn, describe the parameters you chose and show the evaluation metrics of the model. [1%]
  For SVM using gradient descent, describe your implementation and show the evaluation metrics of the model. [1%]
  For SVM using gradient descent, describe and show the changes of loss and accuracy during training. [1%]
  Compare the two method and analyze the results. [1%]

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 CHEN Meng (TA) for any questions.