Homework 4:
Support Vector Machine

Machine Learning and Data Mining (Fall 2024)

Student Name: Student ID:

Lectured by: Shangsong Liang
Sun Yat-sen University

Your assignment should be submitted to the email that will be provided by the TA
Deadline of your submission is: 23:59PM, November 30, 2024
Do NOT Distributed This Document and the Associated Datasets

Problem: Implementing SVM to complete a classification task

The zip includes `train.txt` and `test.txt` as the training and test datasets, respectively. `train.txt` contains 43,957 labeled data, while `test.txt` contains 4,885 unlabeled data. Each line in the .txt files represents a data entry, with labels as either "less than 50K" or "greater than 50K". Detailed descriptions of the data feature can be found at `https://archive.ics.uci.edu/ml/datasets/Adult`.

Write a Python program to train a SVM on the training dataset using stochastic gradient descent (SGD) [1], test it on the test dataset, and report the test accuracy. Please do not use pre-built packages; instead, implement the SVM yourself.

Hint:
(1) Gradients of SVM:

$$\nabla_a = \begin{cases} \lambda a & \text{if } y_k(a^T x_k + b) \geq 1 \\ \lambda a - y_k x_k & \text{otherwise} \end{cases}$$

$$\nabla_b = \begin{cases} 0 & \text{if } y_k(a^T x_k + b) \geq 1 \\ -y_k & \text{otherwise} \end{cases}$$

(2) The first column in the data represents the ID, not a feature. Data preprocessing (normalization) on the different features is necessary.

(3) Adjust the learning rate accordingly.

Submission:

Submit a document (Formats such as PDF or docx are acceptable) including at least the following:

1. Key SVM concepts and an outline of your implementation approach.
2. Screenshot showing the printed accuracy on the test dataset.
3. Train the SVM with a regularization term and test at least the following values for the regularization constant: $[1e-3, 1e-2, 1e-1, 1]$. List the accuracy for each value, and explain how and why the accuracy trends with changes in the regularization constant.
4. Screenshots of the code.
5. Complete code attached at the end of the document or using separate files.

Besides the PDF, submit the source code files as well.

[1] Please go to internets to search materials on how to use SGD to train SVM.