CMPE 462 Assignment 1

Department of Computer Engineering
Boğaziçi University

Spring 2024

In this assignment, you will implement the following classifiers. You can use libraries such as numpy, scipy, matplotlib in your experiments. However, you are expected to implement the learning algorithms regarding the following classifiers from scratch. Therefore, you are not allowed to use scikit-learn functions of the following classifiers. If training and test splits are not provided in the datasets, please randomly split your data into training and test. Please submit a PDF report containing the link to our code, your answers, and references. Please cite all the resources used in the assignment. If you ever use an AI tool such as ChatGPT, please acknowledge. Each group member should be able to answer questions regarding any of the sections below.

1 Perceptron Learning Algorithm (PLA) (25 pts)

Please implement the PLA algorithm from scratch. You are given two toy datasets, `data_large` and `data_small` under `PLA_data` folder, of 2-D inputs and two categories. Please train your perceptron on the datasets and address the following.

1. Compare the number of iterations required to converge for the large and small datasets.

2. Plot the datasets using different colors and markers to indicate the classes. Next, plot the decision boundaries. Please provide two separate plots, one for large and one for small datasets.

3. Please repeat the training multiple times with different initial points and compare the learned weights. You may use the small dataset. Is PLA sensitive to initialization? If so, what do you think is the reason? Please explain.

2 Logistic Regression (40 pts)

Implement full batch (GD) and stochastic gradient descent (SGD) algorithms for logistic regression and regularized logistic regression from scratch. Please use the dataset at `https://archive.ics.uci.edu/dataset/545/rice+cammeo+and+osmancik` to train and evaluate your models. Please address the following.

1. Familiarize yourself with the dataset. Explore the normalization techniques and apply one if necessary. Explain what you did to prepare your data for training.
2. Use 5-fold cross-validation to determine the value of the regularization parameter.
3. Please report the training and test performance. Create a table including the training and test classification accuracies for the following models:
   - logistic regression trained with GD
   - logistic regression trained with SGD
   - logistic regression regularized by the square of the weight vector's ℓ_2 norm trained with GD
   - logistic regression regularized by the square of the weight vector's ℓ_2 norm trained with SGD
4. Compare the training times of GD and SGD for converging to a similar model (models with similar test accuracies). Compare the changes in their loss values at each iteration for GD and each epoch for SGD on the same plot.
5. Investigate the effect of the step size in SGD on convergence. For this purpose, train models using an initial step size larger and smaller than the one you used to report your best performance. Remember that the step size is not fixed in SGD. Compare their convergences by plotting the loss values with respect to epochs.

3 Naive Bayes (35 pts)

In this section, you will implement a Naive Bayes classifier for the dataset at `https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic`. In this dataset, input samples have continuous values. You may assume that each attribute from each category follows the normal distribution with different means and variances.

1. Please report the training and test accuracy of the Naive Bayes classifier you trained.

2. Compare the number of parameters if you did not consider the conditional independence assumption.

3. Train your logistic regression classifier on this dataset and compare its performance with Naive Bayes. Outline the fundamental differences between the two classifiers. If one performs better than the other, discuss the underlying reasons.