IE343: Statistical Machine Learning

Assignment 4

Professor: Seyoung Yun

(TA : Mingyu Kim)

1 Instruction

This assignment is to implement a "Support Vector Machine". The task is binary classification $x \in \mathbb{R}^2, y \in \{0, 1\}$. You should write codes for implementing the `__call__` function in the `(./App/rbf.py)` and `fit` function in the `(./App/support_vector_classifier.py)` and run the algorithm for solving two tasks. In the first task, the data set is same as what we used in assignment 2 and 3. The data set of the second task is the new generated synthetic data, in which all classes cannot be clearly classified. Both tasks can be executed by running `(./SVM.py)`. Some of details of the source codes are as follows :

* There are "App", "Data", "Data2" and "Result" directories and `(./SVM.py)`. Students have to run this '.py' files to make sure that the SVM algorithm has no errors. After running this file, you can get a accuracy and its standard deviation.
* When it comes to "App" directory, there are "Pre_processing" directory and several source codes such as data_import and evaluation. It is important that you should complete the "__call__" methods in `(./App/rbf.py)` and "fit"method in `(./App/support_vector_classifier.py)` to properly run the SVM algorithm and improve performance.
* In this assignment, The SVM can be implemented to use numerical approximation by using the gradient descent algorithm.
* In other aspects, an appropriate kernel function is the most important factor to determine the performance of the SVM algorithm. The students have 3 types of kernel functions such as "Polynomial", "Euclidean" and "Radial Basis Function". The "Polynomial" and "Euclidean" kernels are given for students to understand how the kernel function works. Instead, students must implement the "Radial basis function (RBF)" kernel based on the given two kernels. In this assignment, you should look for the appropriate hyper parameters and kernel functions by trial and error approach. Prior to running `./SVM.py`, you must manually determine these values.
* To help your implementation, there is a boolean variable to choose a data set between the previous data set used in the assignment 2 and 3 and new synthetic data at (`./SVM.py`)

In this assignment, we must describe the model that best fits the individual data set. Like Assignment 2 and 3, the average accuracy and its standard errors for the entire data set are also provided. Furthermore, you need to visualize the training and testing data with the decision boundary. `(./Results/svm_result_data set_(i).png)`
Therefore, you must answer the questions in English and submit a report as a PDF file with your completed source code. (No limit of pages) The report should includes theoretical background for the SVM and your opinion about this model. (Please write a report concisely). In addition, you have to complete this source codes in the attached python files.

2 Data

As we mentioned before, there are two data sets. First, the synthetic data is used for the assignment 2 and 3. `(.App/Data)`. Second, it comes from a synthetic data set where all classes cannot be clearly classified. `(.App/Data2)`. Regarding to the data set of the first task, we expect that students can study how sensitive the model is to outliers. On the other hand, students can also learn how SVM works for the second data set where all classes cannot be clearly classified due to overlapped regions.
In both problems, there are 10 data folds. The classifiers should be trained by only one of the data sets. After training your model, you should compare the binary 'y' values with fitted values in terms of "accuracy"

Previous Synthetic Data

* The first 5 data sets have 50 training data and 20 test data, but the last 5 data sets have 60 training data and 24 test data with outliers.
* You should use the same numbered data sets while training and testing.

New Synthetic Data

* All data sets have 100 training data points and 100 test data points.
* There are overlapped regions where all classes are not able to be clearly classified.
* You should use the same numbered data sets while training and testing.

3 Implementation

This code should be written using the 'Numpy' package. Other packages are not allowed. (Scipy, Tensorflow, pytorch and etc.) You must write down appropriate methods `__call__` function in the `(./App/rbf.py)` and `fit` function in the `(./App/support_vector_classifier.py)`. Except for the these methods, you don't need to modify any source code. Furthermore, you should submit all source codes to check out your codes working.

4 Question

A. Mathematically, write down derivation processes of the SVM algorithm. You need to define "margin (soft and hard)" and primal and dual representation of optimization formulations.

B. Write down the gradient of the lagrangian loss function w.r.t dual variables and its constraints.

C. After applying the SVM, write down the comparable average accuracy and its standard error and make a plot with train, test data and the decision boundary in any data fold in the **data set used in assignment 2 and 3**. Remember that you need to attach two kinds of results and plots, such as, with outliers as well as without outliers. Especially, You need to elaborate the selection process of an appropriate kernel function to describe this data set. Also explain your opinion about the SVM algorithm in terms of sensitivity.

D. After applying the SVM, write down the comparable average accuracy and its standard error and make a plot with train, test data and the decision boundary in any data fold in the **new data set (data 2)**. Especially, You need to elaborate the selection process of an appropriate kernel function to describe this data set. Also explain your opinion about the SVM algorithm in terms of cluttered data points.
(Your score will be graded by proportional assessment against the student's highest accuracy. The total score for this problem will be doubled than other questions.)

E. Compare the results from logistic regression, LDA and decision tree with the results from the SVM. In particular, you should attach plots from each method and explain what difference between these method have.