[Project Code: SESVM]
Spam Email Classification using Support Vector Machine

Project Duration: 26-Feb-2023 ~~ 18-Mar-2023
Submission Information: (via) CSE-Moodle

Objective:
In this assignment, you will use SVM to classify emails into spam or non-spam categories. And report the classification accuracy for various SVM parameters and kernel functions. An email is represented by various features like frequency of occurrences of certain keywords, length of capitalized words etc. The data set **SpamBase** contains about 4601 instances to make up a SPAM E-mail Database.

Your Tasks:

1. *Building a SVM Classifier*
   a. **Pre-processing the data**:
      i. Randomly pick 70% of the data as a training set and the rest as a test set.
      ii. Normalize each feature of the dataset to have zero mean and unit variance. Note that while normalizing the features, their mean and variance should be computed over the train split only. Once, the mean and variance are computed using only the train split, you normalize the test split using the mean and variance computed over the train split.
   b. **Training the model**:
      i. Note that training requires solving the dual optimization problem. To solve the dual optimization problem you can use any python packages like: CVXOPT or Scipy.optimize.minimize
      ii. Implement the following three kernels: (a) linear, (b) quadratic and (c) radial basis function
   c. **Making predictions:** Write a function that takes new datapoint as input and predicts the class
   d. **Evaluation:** Finally, you should generate results on the given data and compare its results with the sklearn module (sklearn.svm)

2. *Hyper-parameter Tuning*
   a. For each of the kernels, you have to report training and test set classification accuracy for the best value of generalization constant C. The best C value is the one which provides the best test set accuracy that you have found out by trial of different values of C. Report accuracies in the form of a comparison table, along with the values of C.

3. *Visualization*
   a. Consider the model (with best hyper-parameters) and plot the decision boundary and the support vectors, on both train & test set. (You may use suitable python packages for this task)

4. *Report*
   a. Prepare a concise report or manual, spanning 2 to 3 pages, that details the results of your work, along with your observations and explanations. Be sure to include a clear and thorough overview of the methods used, the results obtained, and your interpretation of the findings.

Data Filename: **spambase.data**
(Please note that the dataset may contain missing values. To handle these missing values, you should use appropriate techniques and clearly explain your methods in the report)

Dataset description: **spambase.DOCUMENTATION**

Submission Details: (to be submitted in CSE-Moodle, **by one representative of the group**)

1. ZIPPED folder containing code (with comments) and the dataset files
2. Report (in pdf format)

Submission Guidelines:

1. You may use one of the following languages: C / C++ / Java / Python.
2. Your program should run on a Linux Environment.
3. Your program should be standalone and should not use any special purpose library for Machine Learning. (Apart from numpy, pandas and the one's mentioned in tasks). And, you can use libraries for other purposes, such as formatting and visualization of data.
4. You should submit the program file and a README file with instructions to run the code.
5. You should name your file as <GroupNo_ProjectCode.extension>.
   (e.g., *Group99_SESVM.zip* for code-distribution and *Group99_SESVM.pdf* for report)
6. The submitted program file *should* have the following header comments:
   # Group Number
   # Roll Numbers : Names of members (listed line wise)
   # Project Number
   # Project Title
7. Submit through CSE-MOODLE only.
   Link to our Course page: https://moodlecse.iitkgp.ac.in/moodle/course/view.php?id=508

You should not use any code available on the Web. Submissions found to be plagiarized or having used ML libraries (except for parts where specifically allowed) will be awarded zero marks.

For any questions about the assignment, contact the following TA:
Rijoy Mukherjee (Email: rijoy.mukherjee@gmail.com)