Introduction to Machine Learning (CSCI-UA.473): Homework 3

Instructor: Lerrel Pinto

October 4, 2022

Submission Instructions

You must typeset the answers using LaTeX and submit them as a single PDF file. Name the pdf file as ⟨Your-NetID⟩_hw3.pdf. Submit the files through the following Google Form - https://forms.gle/ud89h4dESqEDVg8S9 The due date is **October 18, 2022, 11:59 PM**. You may discuss the questions with each other but each student must provide their own answer to each question.

Problem Statement

Spam email classification using Support Vector Machine (SVM)

In this assignment you will use a SVM to classify emails into spam or non-spam categories. And report the classification accuracy for various SVM parameters and kernel functions. You have to submit the report file in pdf format. **No programs need to be submitted**.

Data Set Description

An email is represented by various features like frequency of occurrences of certain keywords, length of capitalized words etc. A data set containing about 4601 instances are available in this link (data folder):
https://archive.ics.uci.edu/ml/datasets/Spambase

The data format is also described in the above link. You have to randomly pick 70% of the data set as training data and the remaining as test data.

Assignment Tasks

In this assignment, you can use any SVM package to classify the above data set. You should use one of the following languages: C/C++/Java/Python. You have to study the performance of the SVM algorithms and submit a report in pdf format. The report should contain the following sections:

- **Methodology:**

  - Mention the libraries used in your solution.

  - Explain the details of the methodology used to solve the homework - how did you read the data, scaler/normalizer used for the data, the SVM package used, data split, kernel functions, etc.

- **Experimental Results:**

  - You have to use each of the following three kernel functions (a) Linear, (b) Quadratic, (c) RBF.

  - 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 that 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. The following format can be used for the table -

Kernel	**C**	10^{-2}	10^{-1}	10^{-0}	10^{1}	10^{2}	10^{3}	10^{4}
Linear	Train Accuracy							
	Test Accuracy							
Quadratic	Train Accuracy							
	Test Accuracy							
RBF	Train Accuracy							
	Test Accuracy							

  - Provide an intuition for the results observed for different kernels and different values of C.