PRML: Assignment 10: Support Vector Machine Classifier

Problem Statement:

Spam email classification using Support Vector Machine: 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 Python. You have to study performance of the SVM algorithms. *You have to submit a report in pdf format. The report should contain the following sections:*

1. Methodology: Details of the SVM package used.

2. Experimental Results:

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

ii. 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.

Submission Guidelines: