AI - LAB
Assignment-6 Report

Support Vector Machine

Arvind Kumar M - 200020008

Tarun Saini - 200010051

Contents

1 Introduction

Given a data set of email spam or not spam, we have to train it using SVM and classify and make prediction. Furthermore we have to analyse the accuracy of train and test set for different kernel i.e. rbf, quadratic and linear and different values of generalisation parameter c.

2 Libraries Used

- For pre-processing and training data sets we are using **sklearn** - python library.
- For reading the data set and making a dataframe we are using **pandas** - python library

3 Methodology

3.1 Pandas

- read_csv method of pandas will read the data set and store it in the form of a dataframe which will be easier to access and modify.

3.2 Sklearn - StandardScaler

- **Preprocessing** module in sklearn provides **StandardScaler** method which normalises all the features given.
- This will give improvement in running time for training data using SVM especially for linear kernel.
- We have also learnt that it is good to preprocess the data before training.

3.3 Sklearn - train_test_split

- **model_selection** module in sklearn provides **train_test_split** method which splits the given data randomly into train and test set given a ratio
- As per the problem statement, we are using 70% randomly chosen data for training and remaining 30% for testing our model.

3.4 Sklearn - SVC

- **svm** module in sklearn provides **SVC** method which can be used to train data set using specific kernel and generalisation paramter c.
- For this assignment, we are training the model with rbf (default in the SVC), Quadratic, linear and analysing them for different value of c and finding the optimal value for the same.

4 Training using SVC

Intially after getting the dataframe using pandas and preprocessing split the data into train and test set.

Syntax : X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)

Then we have to create a model using SVC with required parameters and kernel values as

- kernel = "rbf" for rbf
- kernel = "poly" and degree = 2 for Quadratic
- kernel = "linear" for Linear

Syntax : model = SVC(kernel="linear" ,C = 20)

Then we have to fit the model for the train data set which will train the model.

Syntax : model.fit(X_train, y_train)

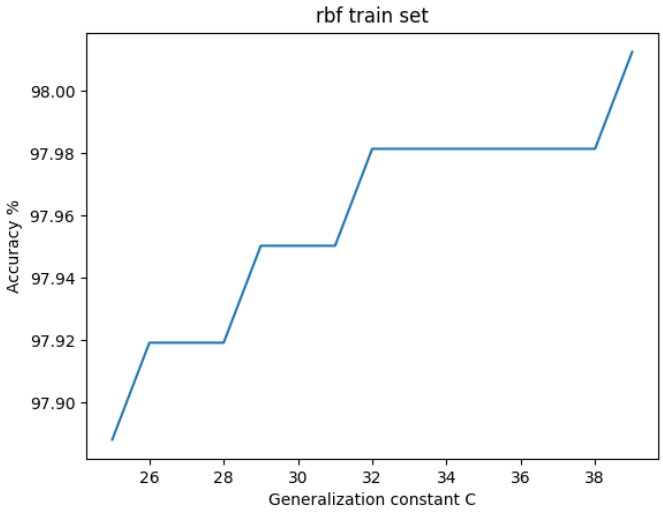
Accuracy of the svm model can be find using score method.

Syntax : model.score(X_test, y_test)

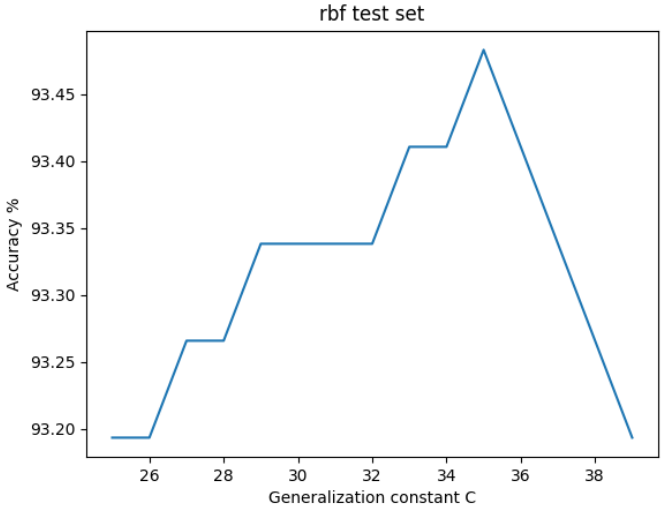
5 Results

5.1 RBF kernel

Training set accuracy for different values of c

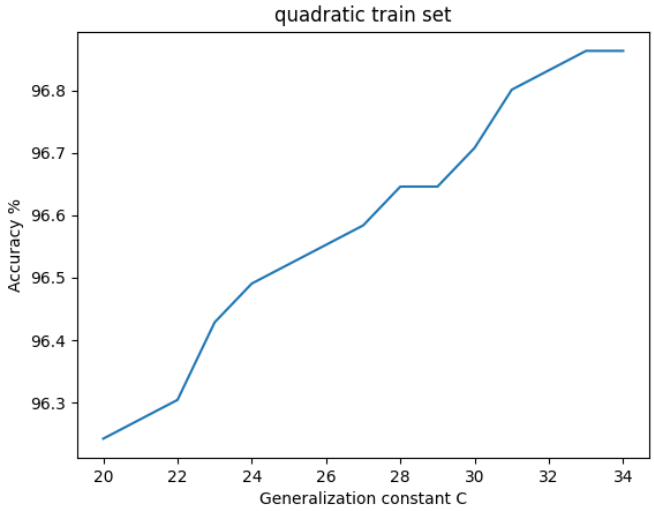


Test set accuracy for different values of c

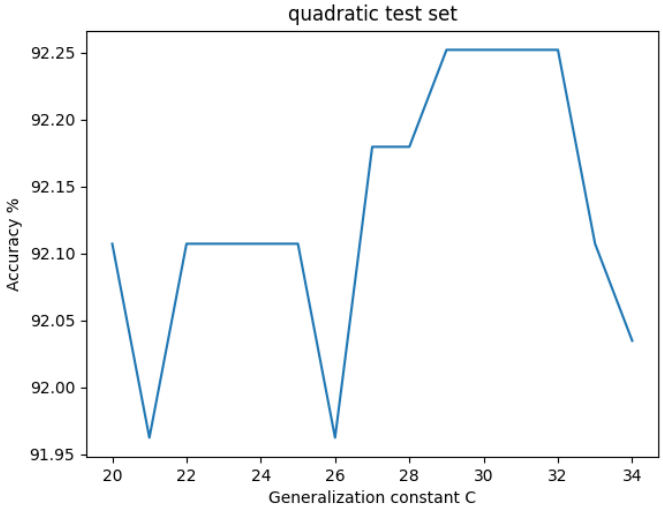


5.2 Quadratic kernel

Training set accuracy for different values of c

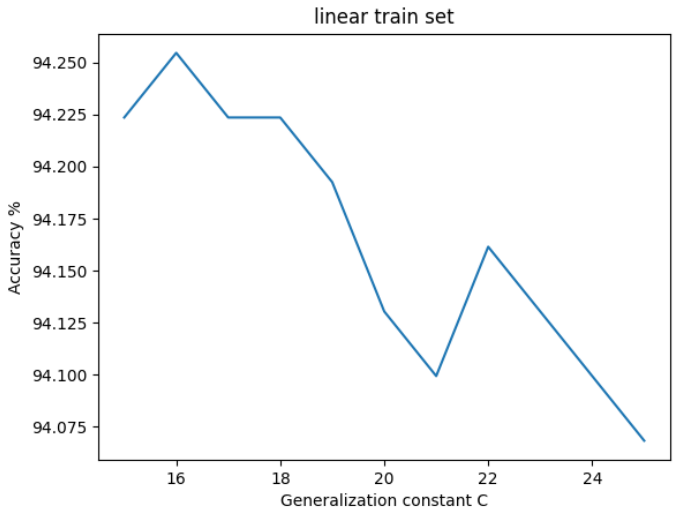


Test set accuracy for different values of c

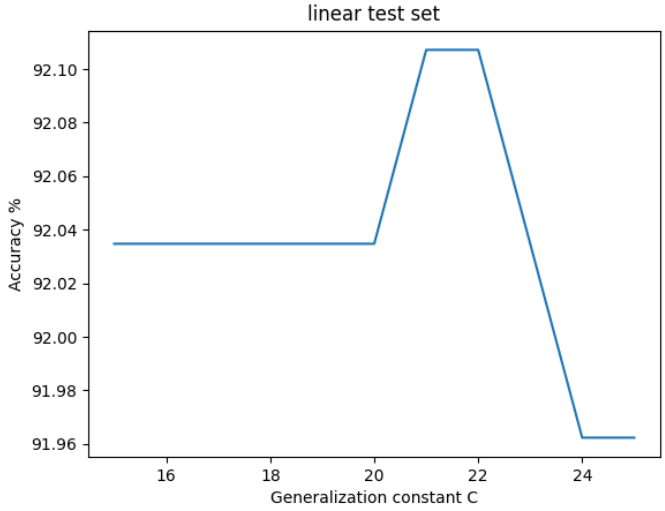


5.3 Linear kernel

Training set accuracy for different values of c



Test set accuracy for different values of c



6 Optimal c value for each kernel

- Rbf - **35**

  with train set accuracy as **97.98%** and test set accuracy as **93.48%**

- Quadratic - **30**

  with train set accuracy as **96.70%** and test set accuracy as **92.25%**

- Linear - **21**

  with train set accuracy as **94.10%** and test set accuracy as **92.10%**

The above values are got from experimenting different values which can be infered from the above graphs for test set accuracy since test accuracy is more important than train set accuracy.