Report – Assignment 6

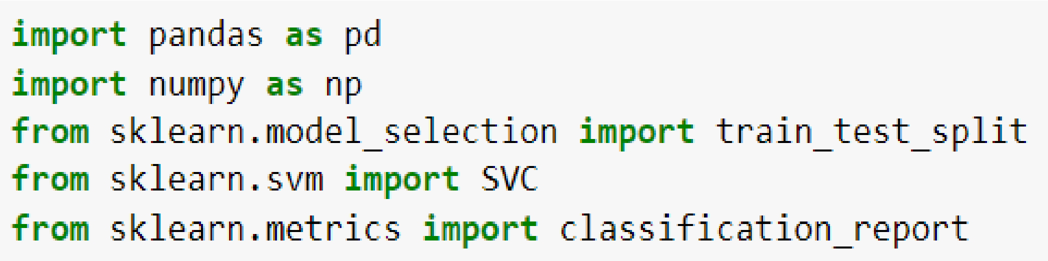
Bhavesh Borse (200010005), Eshita Pagare (200010016)

March 16, 2022

1 Introduction

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.

2 Libraries and Packages



```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.metrics import classification_report
```

2.1 Scikit-Learn Package

Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

3 Kernels

3.1 Linear Kernel

Linear Kernel is used when the data is Linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are a large number of Features in a particular Data Set.

3.2 Polynomial Kernel

In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.

For this dataset we are using Quadratic kernel for which degree is 2.

3.3 RBF Kernel

In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification.

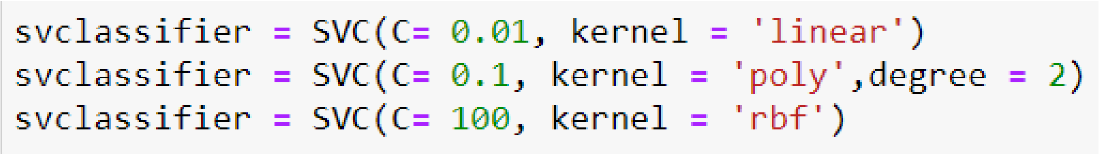
4 Methodology

In sklearn library different number of functions are used to process the data. For SVM we are using function SVC(C= c , kernel = ‘linear’)

For regularization use C argument and for kernel use kernel argument to pass the decision function.

We can vary different parameter by changing the argument in the SVC function. Different C values are used for each of the three kernels linear, polynomial and RBF.

As we have to use Quadratic kernel for this data we specify degree = 2 for polynomial.



```
svclassifier = SVC(C= 0.01, kernel = 'linear')
svclassifier = SVC(C= 0.1, kernel = 'poly',degree = 2)
svclassifier = SVC(C= 100, kernel = 'rbf')
```

5 Experimental Results

Kernels	C value	Accuracy
Linear	0.01	94.74%
	0.1	92.54%
	0.5	92.54%
	1	90.11%
Quadratic	0.5	66.40%
	50	70.22%
	5000	82.88%
	50000	89.84%
RBF	0.5	70.14%
	50	80.84%
	5000	92.54%
	50000	95.55%

As we increase the C accuracy decreases for linear kernel, smaller value of C working well with highest value of 94.74% with C = 0.01 as it looks for larger margin separating hyperplane. For Quadratic kernel it performs poorly as the C value decreases, accuracy is highest with value of 89.84% with C = 50000. RBF perform poorly for smaller C but with larger C it performs as equal to linear kernel with highest value of 95.55% with C = 50000.

6 Conclusion

The training data is linear separable because for kernel other than linear model perform poorly with low value of C. For very tiny values of C, we get false predictions, often even if our training data is linearly separable. For higher values of C, RBF kernel is preferred over quadratic as it takes lesser time than linear kernel and is as equally accurate as linear kernel.