Dimensionality Reduction and State Vector Machines

Shivank Sharma
2018EEB1180

Indian Institute of Technology Ropar
Rupnagar, Punjab

Abstract

This document summarizes the results of dimensionality reduction and visualization performed on various datasets. It also includes the variation of results by tuning the parameters while using the SVMs for the classification on Iris Dataset. The datasets used in this assignment are **Labelled Faces in the Wild** and **Fisher Iris dataset**. The main techniques used in this assignment include *Principal Component Analysis, tSNE and LDA*. It also includes the results of *SVM classification* with varying parameters.

1 Datasets

1.1 Labelled Faces in the Wild

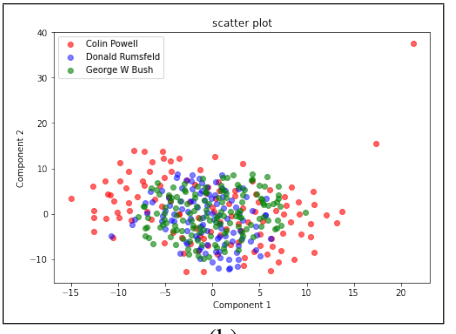
The data set contains more than 13,000 images of faces collected from the web. Each face has been labelled with the name of the person pictured. 1680 of the people pictured have two or more distinct photos in the data set. The only constraint on these faces is that they were detected by the Viola-Jones face detector. The dataset that was downloaded by the given command contains 1140 examples from 5 personality faces.

1.2 Fisher Iris dataset

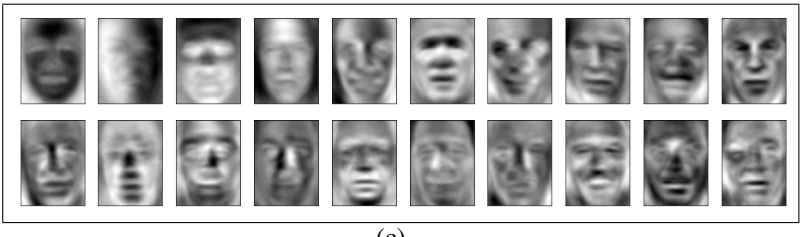
The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphological variation of Iris flowers of three related species. Two of the three species were collected in the Gaspé Peninsula "all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus".

The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.

(a)



(b)



(c)

Figure 1: (a)Some examples from dataset 'Labelled Faces in the Wild'. (b)Three personalities projected to 2D via tSNE. (c)First 20 eigenfaces.

2 Principal Component Analysis and Eigenfaces for Face Recognition

Downloaded the dataset by the given command and split it into 70:30 for training and test set. The first step performed was PCA. Originally the dataset contained 2914 dimensions which were converted to 100 dimensions by Principal Component Analysis. Then three personalities were chosen to perform tSNE (included **class1: Donald Rumsfeld, class3: Gerhard Schroeder and class4: Tony Blair**). Then I performed tSNE to convert each chosen class to 2-dimensional space. The class-wise plot is given in figure 1(b).

Observation:

The figure shows that all the data points for the three chosen personalities lie together. Perhaps this is because all have very similar characteristics like all are the faces of men, beardless and of similar age. However the biggest factor they share is that all are faces and have eyes, nose and mouth. And this may be the reason that all the points are together. However there are one or two outliers may be because of very different angle or expression on the face. Overall they are all similar.

Then fit a K-Nearest Neighbour Classifier on the training set by taking the number of neighbours to be 10. The overall accuracy of the model came out to be 63.1%. The classification report for the model is given below in table 1.

The next step is to plot the first 20 eigenfaces. the plot is given in figure 1(c).

Later I found out the number of eigenfaces required to retain 80% variance of the original training set. The 80% variance is retained by the first 31 eigenfaces. Then I performed the K-Nearest Neighbour classification again on the training dataset but this time taking only the

Classification Report - 100 Components				
	precision	recall	f1-score	support
Colin Powell	0.64	0.76	0.69	78
Donald Rumsfeld	0.57	0.21	0.31	38
George W Bush	0.64	0.87	0.74	159
Gerhard Schroeder	0.50	0.07	0.12	30
Tony Blair	0.53	0.22	0.31	37
accuracy			0.63	342
macro avg	0.58	0.42	0.43	342
weighted avg	0.61	0.63	0.58	342

Table 1: K Nearest Neighbour Classification Report - 100 Components

Classification Report - 31 Components(80%variance)				
	precision	recall	f1-score	support
Colin Powell	0.49	0.51	0.50	78
Donald Rumsfeld	0.44	0.21	0.29	38
George W Bush	0.58	0.77	0.66	159
Gerhard Schroeder	0.39	0.23	0.29	30
Tony Blair	0.31	0.11	0.16	37
accuracy			0.53	342
macro avg	0.44	0.37	0.38	342
weighted avg	0.50	0.53	0.50	342

Table 2: K Nearest Neighbour Classification Report - 31 Components(80%variance)

first 31 eigenfaces. However this time the accuracy declined to 53.2%.

The classification report for the model is given below in table 2.

3 Dimensionality Reduction and Visualization with PCA, LDA and tSNE

I converted the 4 dimensional data to 2 dimensional by using Principal Component Analysis. Then I plotted the data and observed it on the basis of initial parameters (Petal Length, Petal Width, Sepal Length, Sepal Width) one by one.

Observations:

The higher values of petal length and petal width lie across the higher values of component 1 or the eigen-direction 1. However sepal length also slightly followed this relation but it also varied along the component 2 or the eigen-direction 2. But for sepal width the trend was somewhat different. The lower values for sepal width corresponded with the higher values of component 1 and lower values of component 2. Whereas for higher values of sepal width the trend was vice-versa ie. higher values corresponded with higher values of component 2 and lower values of component 1.

All these patterns can be observed in figure 2.

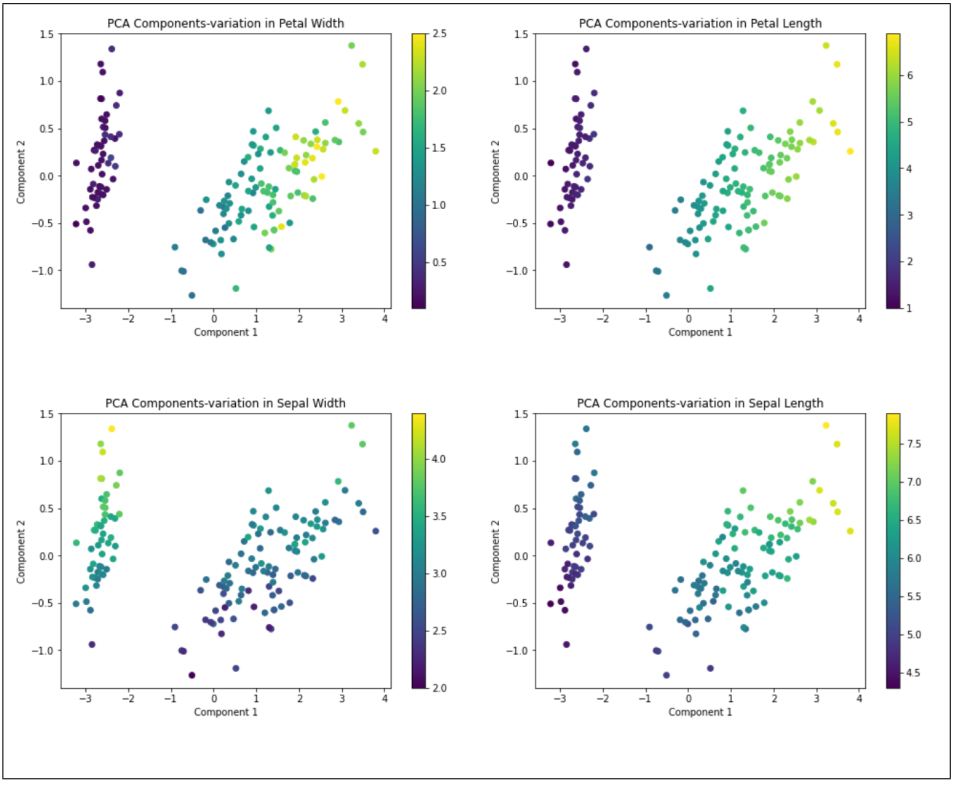


Figure 2: Iris Dataset plotted in 2D via PCA and distributed on the basis of the values of each feature.

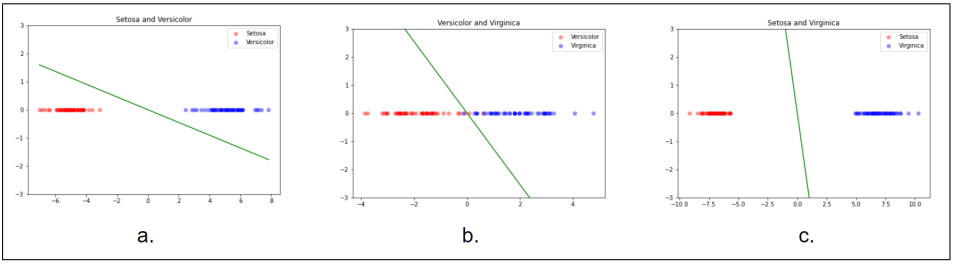


Figure 3: Classwise plot of Iris dataset projected to 1D via LDA along with the lines of separation.

3.1 Performing LDA on the dataset

Later I performed LDA by considering two classes at a time to plot the data in 1 dimension. The lines of separation are also plotted along with the data. The results are given in figure 3. Then I plotted the 3 class data on a 2 dimensional plane by performing LDA.The result is given in figure 4(a).

3.2 Applying tSNE on data with different metrics

The next task was to project the 4 dimensional Fisher iris data on 2 dimensional and 3 dimensional space by using tSNE. This was to be done by considering different metric parameters for each case. I chose **"Euclidean", "Chebyshev"** and **"Mahalanobis"** as my metric parameters.

Observations:

From the 2 dimensional plots it can be clearly observed that using different distance metrics in tSNE provide different separation among the classes. The choice of metric depends on the type of data and varies accordingly. However for the given dataset **Euclidean** and **Chebyshev** metrics provide better separation than the **"Mahalanobis"**. It can be seen that "Mahalanobis" is of no use for this dataset for separation.

However for 3 dimensional plots, the observation was a bit difficult as all three chosen metrics performed nearly the same. But if observed carefully we can see that here also the **"Mahalanobis"** metric performed worse than the **Euclidean** and **Chebyshev**. As in the case of Mahalanobis the data is completely mixed up.

The plots are given in figure 4(b)(c)(d) and figure 5.

4 Data Classification with Linear and Non-linear SVMs

I converted the 4 dimensional data to 2 dimensional by considering only **petal length** and **petal width**. By choosing two classes at a time,I divided the data in a 70:30 training and test set. Then I have fit three SVM models on the data. The **kernel** taken into account was '**Linear**' and the value of **C** was **1**. The related figures including the class data, max-margin hyperplane and support vectors are given in the figure 6. The classification report for the three models are given below in table 3, 4 and 5.

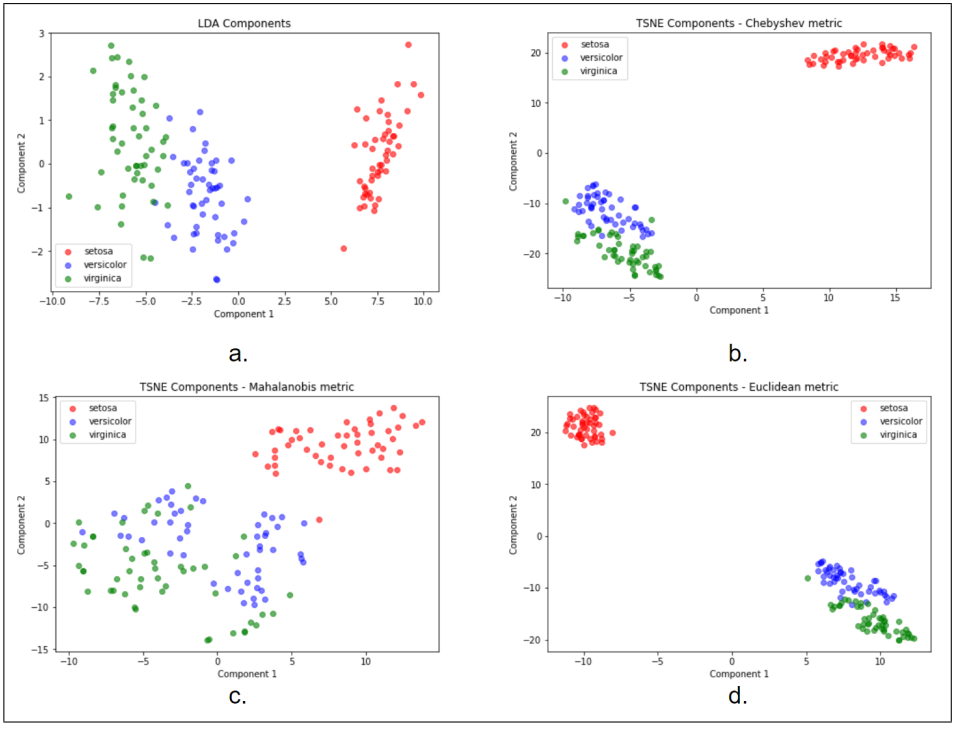


Figure 4: (a)LDA plot of Iris dataset in 2d. (b)tSNE plot of Iris dataset in 2D considering Chebyshev metric. (c)tSNE plot of Iris dataset in 2D considering Mahalanobis metric. (d)tSNE plot of Iris dataset in 2D considering Euclidean metric.

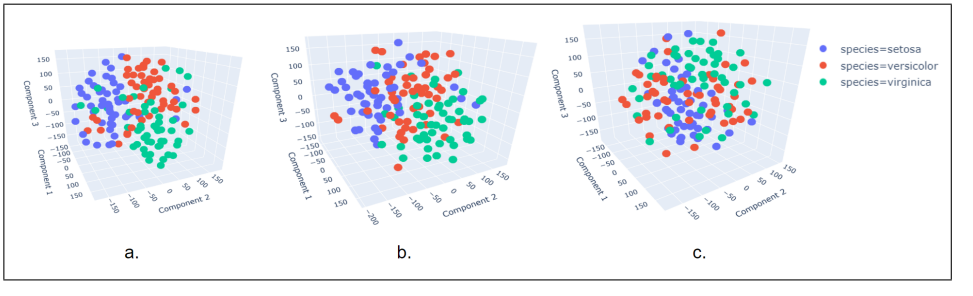


Figure 5: (a)tSNE plot of Iris dataset in 3D considering Euclidean metric. (b)tSNE plot of Iris dataset in 3D considering Chebyshev metric. (c)tSNE plot of Iris dataset in 3D considering Mahalanobis metric.

Classification Report - Setosa-Versicolor, linear kernel, C=1				
	precision	recall	f1-score	support
Setosa	1.00	1.00	1.00	13
Versicolor	1.00	1.00	1.00	12
accuracy			1.00	25
macro avg	1.00	1.00	1.00	25
weighted avg	1.00	1.00	1.00	25

Table 3: Classification Report - Setosa-Versicolor, linear kernel, C=1

Classification Report - Versicolor-Virginica, linear kernel, C=1				
	precision	recall	f1-score	support
Versicolor	0.92	0.92	0.92	12
Virginica	0.92	0.92	0.92	13
accuracy			0.92	25
macro avg	0.92	0.92	0.92	25
weighted avg	0.92	0.92	0.92	25

Table 4: Classification Report - Versicolor-Virginica, linear kernel, C=1

Classification Report - Setosa-Virginica, linear kernel, C=1				
	precision	recall	f1-score	support
Setosa	1.00	1.00	1.00	10
Virginica	1.00	1.00	1.00	15
accuracy			1.00	25
macro avg	1.00	1.00	1.00	25
weighted avg	1.00	1.00	1.00	25

Table 5: Classification Report - Setosa-Virginica, linear kernel, C=1

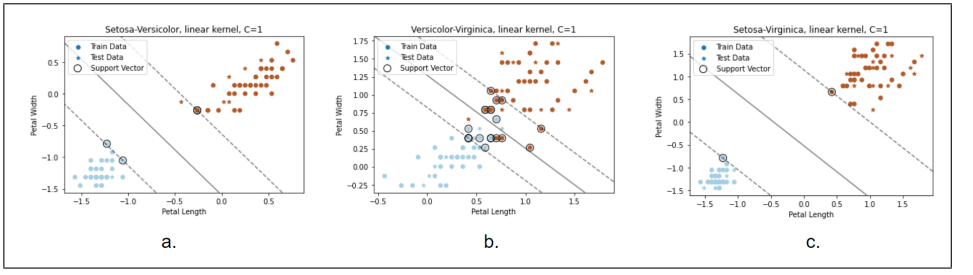


Figure 6: Classwise SVM classification of Iris dataset with linear kernel and C=1 along with the max-margin hyperplane and support vectors.

Classification Report - Setosa-Versicolor, linear kernel, C=0.001				
	precision	recall	f1-score	support
Setosa	0.00	0.00	0.00	13
Versicolor	0.48	1.00	0.65	12
accuracy			0.48	25
macro avg	0.24	0.50	0.32	25
weighted avg	0.23	0.48	0.31	25

Table 6: Classification Report - Setosa-Versicolor, linear kernel, C=0.001

4.1 Fitting SVM - Linear kernel, different C

The next task was to fit the SVM on the same data with the kernel to be "Linear" but for different values of C. The values of C were taken to be 0.001 and 1000.

Observation:

The **smaller value of C** results in a model looking for a larger-margin separating plane and in this process it misclassifies more points. This results in decrease in the accuracy and model performs bad even in cases where classes are easily separable. Smaller value of C lets the weights increase, leading to a large training error. It can be compared to underfitting of the model.

The larger value of C lets the model choose small-margin hyperplanes if that hyperplane lets the model to classify all the training points correctly. The large values of C decrease the weights too much and in doing so it will lead to loss in generalization properties of our model ie. it will be more prone to outliers and leads to overfitting on the training set.

The classification reports for all three model and C value of 0.001 and 1000 are given in table 6, 7, 8, 9, 10 and 11.

4.2 Fitting SVM - RBF kernel, different C

Next I have fit the SVM on the three pairs of data again but now by taking the kernel to be "**rbf**" (**Radial Basis Function**).

Observation:

RBF kernel is used when we want our boundaries to be curved shapes. However the C parameter defines the regularization in the model. When the value of **C = 0.001** then the model under fitted the training set leading to very low accuracies even on easily separable

Classification Report - Versicolor-Virginica, linear kernel, C=0.001				
	precision	recall	f1-score	support
Versicolor	0.48	1.00	0.65	12
Virginica	0.00	0.00	0.00	13
accuracy			0.48	25
macro avg	0.24	0.50	0.32	25
weighted avg	0.23	0.48	0.31	25

Table 7: Classification Report - Versicolor-Virginica, linear kernel, C=0.001

Classification Report - Setosa-Virginica, linear kernel, C=0.001				
	precision	recall	f1-score	support
Setosa	0.40	1.00	0.57	10
Virginica	0.00	0.00	0.00	15
accuracy			0.40	25
macro avg	0.20	0.50	0.29	25
weighted avg	0.16	0.40	0.23	25

Table 8: Classification Report - Setosa-Virginica, linear kernel, C=0.001

Classification Report - Setosa-Versicolor, linear kernel, C=1000				
	precision	recall	f1-score	support
Setosa	1.00	1.00	1.00	13
Versicolor	1.00	1.00	1.00	12
accuracy			1.00	25
macro avg	1.00	1.00	1.00	25
weighted avg	1.00	1.00	1.00	25

Table 9: Classification Report - Setosa-Versicolor, linear kernel, C=1000

Classification Report - Versicolor-Virginica, linear kernel, C=1000				
	precision	recall	f1-score	support
Versicolor	0.85	0.92	0.88	12
Virginica	0.92	0.85	0.88	13
accuracy			0.88	25
macro avg	0.88	0.88	0.88	25
weighted avg	0.88	0.88	0.88	25

Table 10: Classification Report - Versicolor-Virginica, linear kernel, C=1000

Classification Report - Setosa-Virginica, linear kernel, C=1000				
	precision	recall	f1-score	support
Setosa	1.00	1.00	1.00	10
Virginica	1.00	1.00	1.00	15
accuracy			1.00	25
macro avg	1.00	1.00	1.00	25
weighted avg	1.00	1.00	1.00	25

Table 11: Classification Report - Setosa-Virginica, linear kernel, C=1000

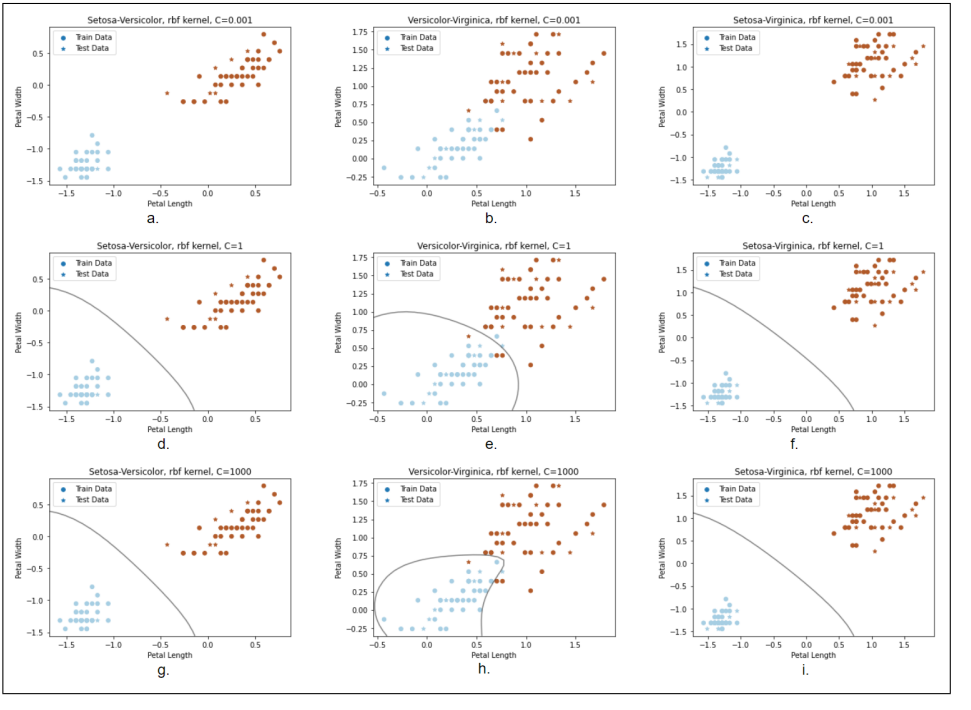


Figure 7: (a)(b)(c)Classwise SVM classification of Iris dataset with rbf kernel and C=0.001 along with center hyperplane. (d)(e)(f)Classwise SVM classification of Iris dataset with rbf kernel and C=1 along with center hyperplane. (g)(h)(i)Classwise SVM classification of Iris dataset with rbf kernel and C=1000 along with center hyperplane.

Classification Report - Setosa-Versicolor, RBF kernel, C=0.001				
	precision	recall	f1-score	support
Setosa	0.00	0.00	0.00	13
Versicolor	0.48	1.00	0.65	12
accuracy			0.48	25
macro avg	0.24	0.50	0.32	25
weighted avg	0.23	0.48	0.31	25

Table 12: Classification Report - Setosa-Versicolor, RBF kernel, C=0.001

Classification Report - Versicolor-Virginica, RBF kernel, C=0.001				
	precision	recall	f1-score	support
Versicolor	0.48	1.00	0.65	12
Virginica	0.00	0.00	0.00	13
accuracy			0.48	25
macro avg	0.24	0.50	0.32	25
weighted avg	0.23	0.48	0.31	25

Table 13: Classification Report - Versicolor-Virginica, RBF kernel, C=0.001

data. This leads to misclassification of more points and hence a poor performance. It performs so bad that its decision boundary is completely away from the data. So while test the accuracy it just classifies one of the class completely and misclassifies the full other class. It can be see from the figure 7(a)(b)(c).

For the value of **C = 1**, the model performed very well. It got an accuracy of 100% on easily separable data in contrast to the small value of C. Also for class data that was mixed in case of **Versicolor-Virginica**, it performed well and got an accuracy of 96% which was more than in the case with "Linear" kernel.

For the value of **C = 1000**, the model performed well in case of easily separable data and got an accuracy of 100% but it also performed well on mixed case of **Versicolor-Virginica** and got the accuracy of 100%.

The plot for all three cases are shown in figure 7. The classification reports for all the models in this case are given below in table 12-20.

Classification Report - Setosa-Virginica, RBF kernel, C=0.001				
	precision	recall	f1-score	support
Setosa	0.40	1.00	0.57	10
Virginica	0.00	0.00	0.00	15
accuracy			0.40	25
macro avg	0.20	0.50	0.29	25
weighted avg	0.16	0.40	0.23	25

Table 14: Classification Report - Setosa-Virginica, RBF kernel, C=0.001

Classification Report - Setosa-Versicolor, RBF kernel, C=1				
	precision	recall	f1-score	support
Setosa	1.00	1.00	1.00	13
Versicolor	1.00	1.00	1.00	12
accuracy			1.00	25
macro avg	1.00	1.00	1.00	25
weighted avg	1.00	1.00	1.00	25

Table 15: Classification Report - Setosa-Versicolor, RBF kernel, C=1

Classification Report - Versicolor-Virginica, RBF kernel, C=1				
	precision	recall	f1-score	support
Versicolor	0.92	0.92	0.92	12
Virginica	0.92	0.92	0.92	13
accuracy			0.92	25
macro avg	0.92	0.92	0.92	25
weighted avg	0.92	0.92	0.92	25

Table 16: Classification Report - Versicolor-Virginica, RBF kernel, C=1

Classification Report - Setosa-Virginica, RBF kernel, C=1				
	precision	recall	f1-score	support
Setosa	1.00	1.00	1.00	10
Virginica	1.00	1.00	1.00	15
accuracy			1.00	25
macro avg	1.00	1.00	1.00	25
weighted avg	1.00	1.00	1.00	25

Table 17: Classification Report - Setosa-Virginica, RBF kernel, C=1

Classification Report - Setosa-Versicolor, RBF kernel, C=1000				
	precision	recall	f1-score	support
Setosa	1.00	1.00	1.00	13
Versicolor	1.00	1.00	1.00	12
accuracy			1.00	25
macro avg	1.00	1.00	1.00	25
weighted avg	1.00	1.00	1.00	25

Table 18: Classification Report - Setosa-Versicolor, RBF kernel, C=1000

Classification Report - Versicolor-Virginica, RBF kernel, C=1000				
	precision	recall	f1-score	support
Versicolor	0.92	0.92	0.92	12
Virginica	0.92	0.92	0.92	13
accuracy			0.92	25
macro avg	0.92	0.92	0.92	25
weighted avg	0.92	0.92	0.92	25

Table 19: Classification Report - Versicolor-Virginica, RBF kernel, C=1000

Classification Report - Setosa-Virginica, RBF kernel, C=1000				
	precision	recall	f1-score	support
Setosa	1.00	1.00	1.00	10
Virginica	1.00	1.00	1.00	15
accuracy			1.00	25
macro avg	1.00	1.00	1.00	25
weighted avg	1.00	1.00	1.00	25

Table 20: Classification Report - Setosa-Virginica, RBF kernel, C=1000