Assignment-3 Report

Samreet Singh Dhami
2018CSB1120

Fundamentals of Data Science(FDS)
IIT Ropar

Abstract

The Report contains three sections in accordance with the three questions in the assignment.The assignment is based on important concepts of data science i.e. PCA,LDA,tSNE,Linear and Non-linear SVMs. The three sections are-

1. Principal Component Analysis and Eigenfaces for Face Recognition
2. Dimensionality Reduction and Visualization with PCA, LDA and tSNE
3. Data Classification with Linear and Non-linear SVMs

1 PCA and Eigenfaces for Face Recognition

Eigenfaces (https://en.wikipedia.org/wiki/Eigenface) was proposed by Turk and Pentland in the 1980s for face recognition, and represent one of the early breakthroughs in the field of Computer Vision. The „trick in this approach involves significantly reducing the dimensionality of a high-dimensional vector (a 320 x 240 pixel image can be viewed as a 76800-dimensional vector) to far fewer dimensions employing Principal Component Analysis, and being able to reconstruct each training image as a linear combination of "eigenfaces", which resemble ghost faces.
PCA is used in this section,The main goal of PCA is dimensionality reduction. It has many applications in visualisation, feature extraction, data compression, etc. The idea behind it is to linearly project original data onto a lower dimensional subspace offering the principal components (eigenvectors) maximum variance of the projected data and/or minimum distortion error from the projection. and the major advantage is that by projecting our data into a smaller space, we're reducing the dimensionality of our feature space... but because we've transformed our data in these different "directions," we've made sure to keep all original variables in our model.

1.1 tSNE Plot

Each face image is transformed into to a 100-D vector. After projecting the data into 2 dimensions via tSNE, **Fig:1** is the plot of the points corresponding to the train+test faces for the identities-George W Bush, Gerhard Schroeder and Tony Blair.As we know that TSNE will form cluster only when there are sufficient points in a population distribution(meaning when data is not sparse).Therefore from the graph, it is clear that data is very sparse.

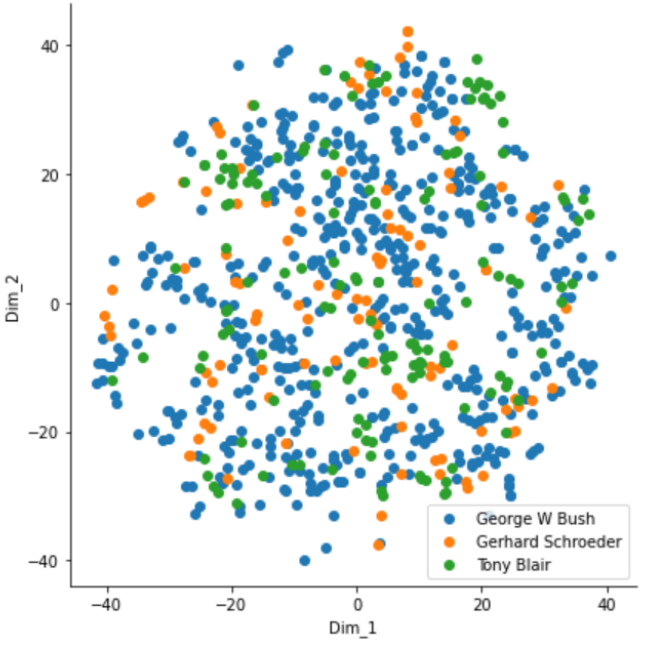


Figure 1: Scatter Plot

Accuracy: 0.6374269005847953

	precision	recall	f1-score	support
Colin Powell	0.58	0.73	0.64	78
Donald Rumsfeld	0.70	0.50	0.58	38
George W Bush	0.69	0.81	0.74	159
Gerhard Schroeder	0.22	0.07	0.10	30
Tony Blair	0.57	0.32	0.41	37
accuracy			0.64	342
macro avg	0.55	0.49	0.50	342
weighted avg	0.61	0.64	0.61	342

Figure 2: Classification Report(n=100)

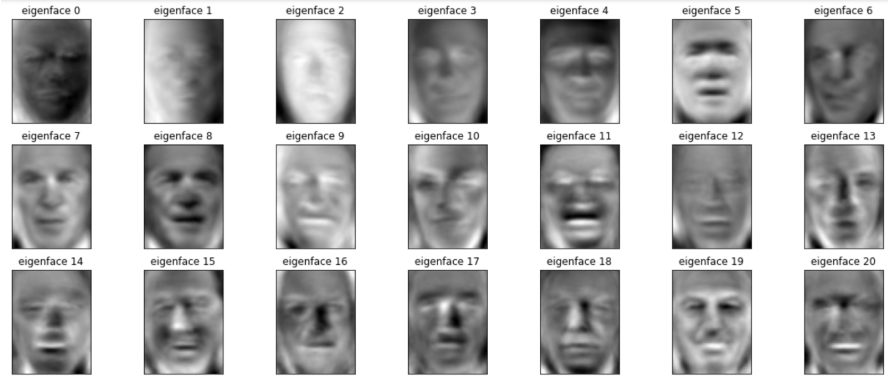


Figure 3: Eigenfaces

Accuracy: 0.6140350877192983

	precision	recall	f1-score	support
Colin Powell	0.52	0.72	0.61	78
Donald Rumsfeld	0.59	0.45	0.51	38
George W Bush	0.70	0.77	0.73	159
Gerhard Schroeder	0.27	0.10	0.15	30
Tony Blair	0.58	0.30	0.39	37
accuracy			0.61	342
macro avg	0.53	0.47	0.48	342
weighted avg	0.60	0.61	0.59	342

Figure 4: Classification Report(n=30)

1.2 Nearest Neighbor classifier on the dataset

KNN classifier used on training dataset after splitting the data into 70:30 ratio, The Classification Report is **Fig:2**. The accuracy comes out to be 63.74% which could have been better if we have used other advanced classifiers.

1.3 Variance Retained for x eigenfaces

The value of x asked in the question comes out to be 30. which means atleast 30 eigenfaces are required to retain 80% variance of the training dataset. The classification report for n=30 is shown in **Fig:4**. The accuracy is decreased by some amount. It is observed that-
1. More the components, more the variance retained.
2. If we increase the value of components in PCA the accuracy generally increases for KNN classifier.

2 Dimensionality Reduction and Visualization with PCA, LDA and tSNE

In this section, We have employed the various dimensionality reduction techniques to visualize the Fisher Iris dataset. Fisher Iris comprises 150 4-dimensional data samples arising from three Iris flower species. The four features measured for each data sample are: sepal length and width, petal length and width.
LDA is a dimensionality reduction technique which reduce the number of dimensions(i.e. variables) in a dataset while retaining as much information as possible.

2.1 Data Exploration using PCA

PCA is used to reduce the dimensionality of the Fisher Iris Dataset to 2. Upon projecting each of the 150 data samples onto this 2-dimensional feature space, data is plotted (**Fig:5**) and following observations are made-

1. High values along eigen-direction 1 correspond to high values of sepal length,petal width and petal length but not sepal width.
2. Graph of petal length and petal width with eigendirection-1 are approximately linear which indicates high postitive correlation.
3. Graph for sepal width with eigen-direction2 is also kind of linear showing high positive correlation.
4. High values along eigen-direction 2 correspond to high values of petal width and petal length.

2.2 Dimensionality reduction using LDA

LDA is employed on combinations of two classes of dataset. Therefore, 3 combinations are created and LDA is employed to get a one-dimensional plot for each of the 3 pairs of classes.Moreover, 3-class Fisher Iris data is projected on the 2D feature space that maximizes inter-class scatter while minimizing intra-class scatter.
Shown in**Fig:6**

In the 2-D graph, the species are clustered and are separated. The irish-setosa species has the cluster away from the other two species.This shows LDA was successful on the data. The 1-D graph(3 colours for 3 pairs) is somewhat symmetric about y-axis and can be linearly separated only to some extent as there is a bit overlapping.

2.3 tSNE plots

4-D Fisher Iris data is projected onto (a) 2D and (b) 3D via tSNE.The two types metric parameter used are euclidean(**Fig:7**) and correlation(**Fig:8**).

As it can be observed from the two graphs, the 2-D graphs are good in both the cases but the 3-D graph is better for euclidean. Basically the metric parameter is the distance function like ‘braycurtis’, ‘canberra’, ‘chebyshev’, ‘cityblock’, ‘correlation’, ‘cosine’, ‘dice’, ‘euclidean’ etc. Therefore, it turns out to be euclidean distance function is better in this case.

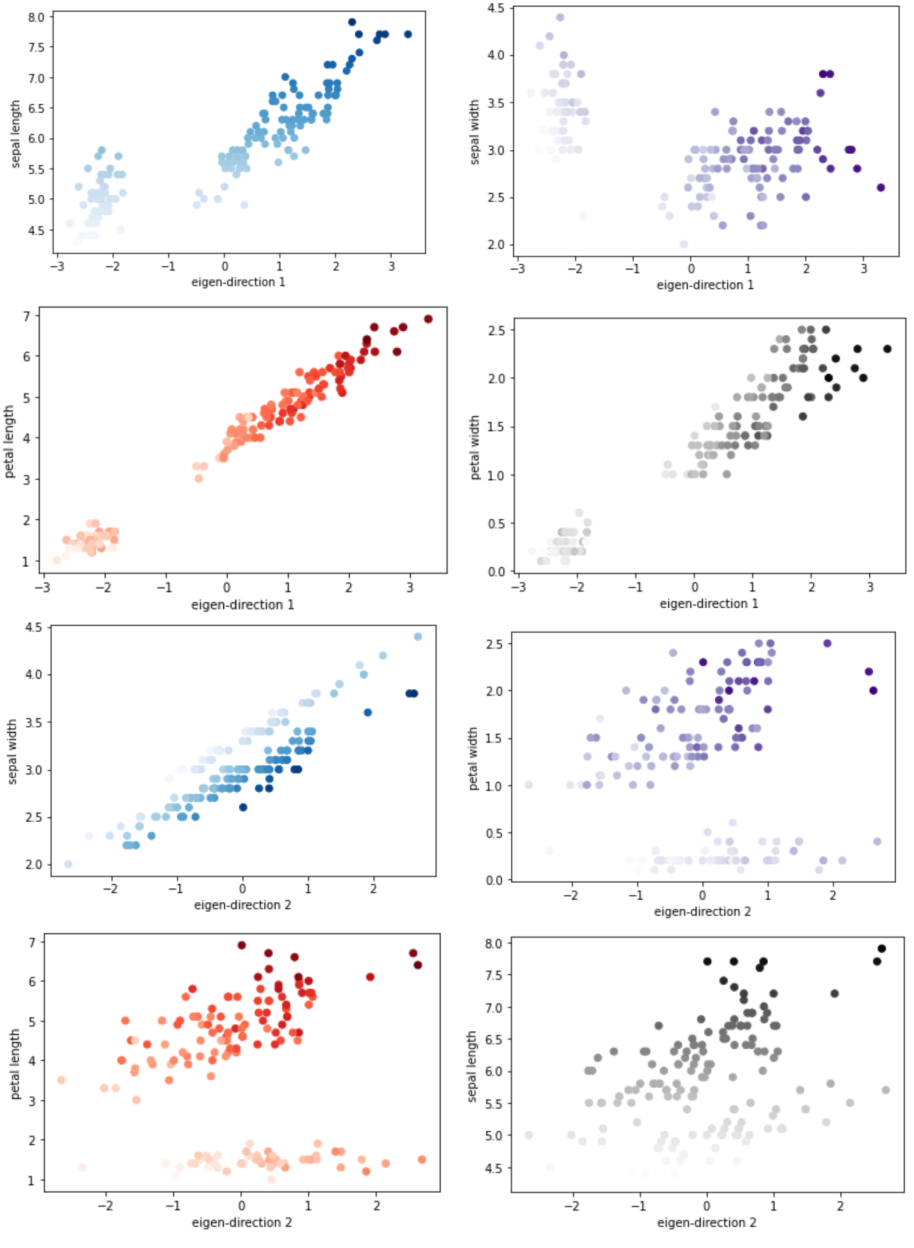


Figure 5: Data Exploration(Darker the colour higher is the value)

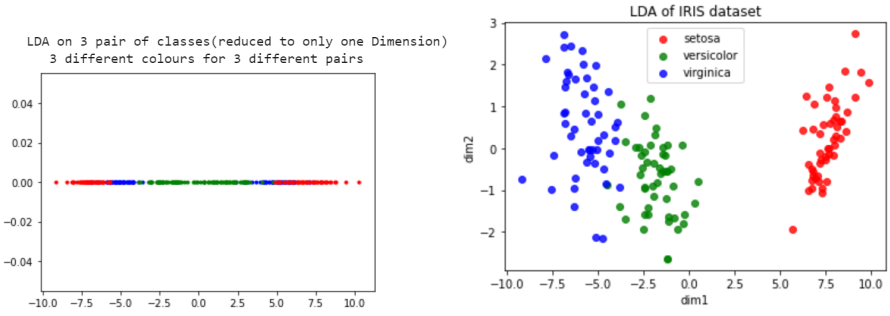


Figure 6: LDA Graphs

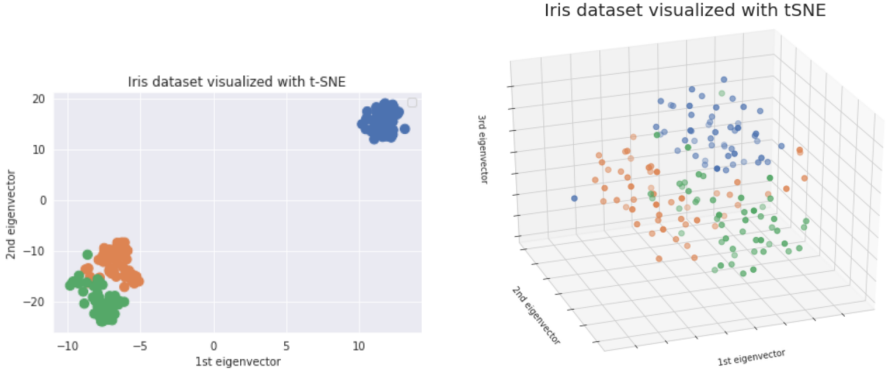


Figure 7: **tSNE Graphs (metric:"euclidean")**->(Blue:Setosa,Green:Versicolor,Orange:Virginica)

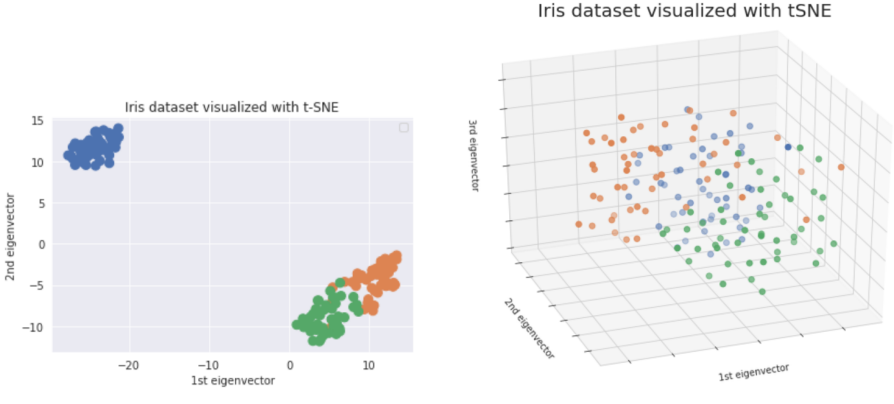


Figure 8: **tSNE Graphs (metric:"correlation")**->(Blue:Setosa,Green:Versicolor,Orange:Virginica

3 Data Classification with Linear and Non-linear SVMs

In this section also, Iris data is used but 4D Iris data is converted to 2D by considering only two features: sepal length and sepal width. The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is.

3.1 SVM model synthesis linear kernel

After splitting the data into train and test dataset, SVM model synthesis is employed. The max-margin hyperplane (i.e, the hyperplane center along with the margin edges) for the three pairs of classes, along with a plot of the two class-data is shown in **Fig:9** along with their classification report.The classification reports of SVM(linear) for C=1,C=0.001 and C=1000 is also shown in the figure.

C is a parameter of the SVC learner and is the penalty for misclassifying a data point. When C is small, the classifier is okay with misclassified data points (high bias, low variance). When C is large, the classifier is heavily penalized for misclassified data and therefore bends over backwards avoid any misclassified data points (low bias, high variance).

C is also called the penalty parameter. Therefore, the low value of C leads to poor accuracy as misclassification is not taken too much into consideration. But after a certain value of C, accuracy is not much affected.

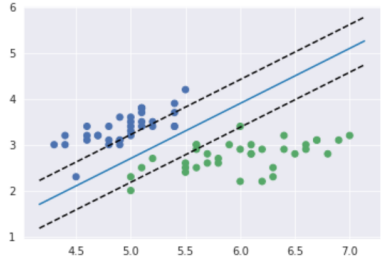
3.2 SVM model synthesis RBF kernel

After splitting the data into train and test dataset, SVM model synthesis is employed using Radial Basis Function(RBF). The centre of the hyperplane for the three pairs of classes, along with a plot of the two class-data is shown in **Fig:10** along with their classification report.The classification reports of SVM(RBF) for C=1,C=0.001 and C=1000 is also shown in the figure.

The importance of C is described previously can also be observed from this classification report. If we compare linear kernel and RBF kernel, for good accuracy RBF requires a higher value of C.
We can also observe the pair of versicolor and virginica in both kernels, for C=1, the classifier is clearly tolerant of misclassified data point. There is much intermixing of the two set of datapoints.

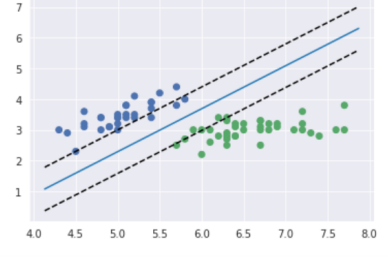
----------Taking 0:setosa and 1:versicolor----------

	precision	recall	f1-score	support
0	1.00	1.00	1.00	16
1	1.00	1.00	1.00	14
accuracy			1.00	30
macro avg	1.00	1.00	1.00	30
weighted avg	1.00	1.00	1.00	30



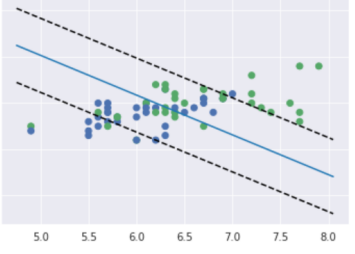
----------Taking 0:setosa and 2:virginica----------

	precision	recall	f1-score	support
0	0.95	1.00	0.97	19
2	1.00	0.91	0.95	11
accuracy			0.97	30
macro avg	0.97	0.95	0.96	30
weighted avg	0.97	0.97	0.97	30



----------Taking 1: versicolor and 2:virginica--------

	precision	recall	f1-score	support
1	0.68	0.81	0.74	16
2	0.73	0.57	0.64	14
accuracy			0.70	30
macro avg	0.71	0.69	0.69	30
weighted avg	0.70	0.70	0.69	30



----------Classification report with C=1----------

	precision	recall	f1-score	support
0	1.00	1.00	1.00	17
1	0.79	0.69	0.73	16
2	0.64	0.75	0.69	12
accuracy			0.82	45
macro avg	0.81	0.81	0.81	45
weighted avg	0.83	0.82	0.82	45

----------Classification report with C=0.001----------

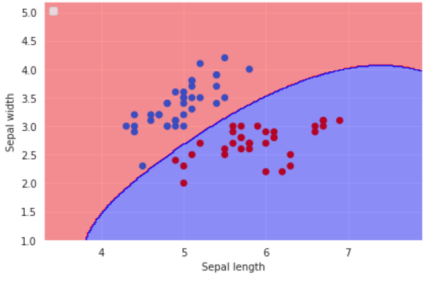
	precision	recall	f1-score	support
0	0.00	0.00	0.00	17
1	0.00	0.00	0.00	16
2	0.27	1.00	0.42	12
accuracy			0.27	45
macro avg	0.09	0.33	0.14	45
weighted avg	0.07	0.27	0.11	45

----------Classification report with C=1000----------

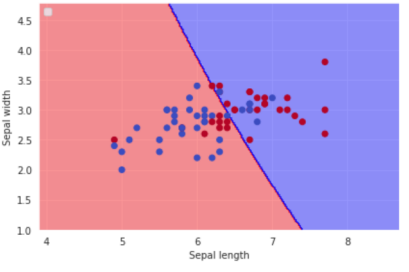
	precision	recall	f1-score	support
0	1.00	1.00	1.00	17
1	0.79	0.69	0.73	16
2	0.64	0.75	0.69	12
accuracy			0.82	45
macro avg	0.81	0.81	0.81	45
weighted avg	0.83	0.82	0.82	45

Figure 9: SVM Graphs and classification report(kernel:linear)

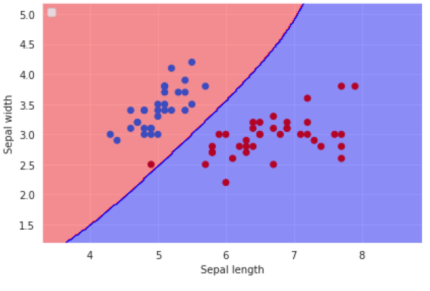
----------Taking 0:setosa and 1:versicolor as a pair-------



----------Taking 1:versicolor and 2:virginica as a pair-------



----------Taking 0:setosa and 2:virginica as a pair-------



----------Classification report with C=0.001--------

	precision	recall	f1-score	support
0	0.00	0.00	0.00	18
2	0.40	1.00	0.57	12
accuracy			0.40	30
macro avg	0.20	0.50	0.29	30
weighted avg	0.16	0.40	0.23	30

----------Classification report with C=1------------

	precision	recall	f1-score	support
0	0.00	0.00	0.00	18
2	0.40	1.00	0.57	12
accuracy			0.40	30
macro avg	0.20	0.50	0.29	30
weighted avg	0.16	0.40	0.23	30

----------Classification report with C=1000---------

	precision	recall	f1-score	support
0	1.00	0.94	0.97	18
2	0.92	1.00	0.96	12
accuracy			0.97	30
macro avg	0.96	0.97	0.97	30
weighted avg	0.97	0.97	0.97	30

Figure 10: SVM Graphs and classification report(kernel:RBF)