Programming Assignment4

CS669: Pattern Recognition (Feb-Jun 2018)

Sukesh Kumar Das (Roll no. d17025)

Joe Johnson (Roll no. d17024)

Group 2

School of Computing and Electrical Engineering

Indian Institute of Technology Mandi

Email:skd_sentu@rediffmail.com

joe4robotics@gmail.com

---

1 Introduction

In the given assignment, two type of dataset-artificial data and real world data are given. In the first type, linearly separable data and non linearly separable data has been provided. The artificial data is 2D three class data. From the real world three class scene image data, bag of visual words(BOVW) feature is extracted using histogram feature. Principle component analysis(PCA) is employed for dimentionality reduction of the BOVW features. Fisher linear discriminant analysis (FDA) is also applied to data set1 and data set2. Three kinds of classifiers-Gaussian mixture model(GMM), perceptron-based classifier and support vector machine(SVM) are used through out the assignment for the purpose of classification.

1.1 Principle component analysis(PCA)

It is a linear method of direction reduction. Here all classes are considered and hence it is also called unsupervised dimention reduction technique. PCA can be interpreted as those direction in which variance is significant. Directions should mantain the orthonormality principle. Overall it finds the directions of projections that are l in number(i.e.l<<d Where d is the dimention of the original data) in such a way that it minimizes the cost function. Projected data,

$$\bar{y}_n = \sum_{i=1}^{d} a_{in}\bar{q}_i. \tag{1}$$

Where q_i be the orthonormal direction of projections. Cost function for reduced dimention(l) is

$$J = \frac{1}{N}\sum_{n=1}^{N} ||\bar{y}_n - \hat{\bar{y}}_n||^2. \tag{2}$$

where,

$$\hat{\bar{y}}_n = \sum_{i=1}^{l} a_{in}\bar{q}_i. \tag{3}$$

The cost function then can be simplified to lagrangian multiplier form such that $\bar{q}_i^T\bar{q}_i = 1$. And it can be simplified to

$$\Sigma\bar{q}_i = \lambda_i\bar{q}_i. \tag{4}$$

Which is eigen decomposition of covariance matrix. Where λ_is are eigen values and $\bar{q}_i$s are the corresponding eigen vectors. To get the directions on which variances are more,λ_is are sorted in decending order and take leading l eigen values and take directions corresponding to them. Finally data is projected on few directions.

1.2 Linear discriminant analysis(LDA)

Earlier we have seen PCA can reduce dimention of data and preserved orthonormality principle but it does not guarantee that discrimination ability can be preserved. But it preserves common information. LDA is not a classifier rather it is a reduced representation. Here we look for the direction in which seperability of projected data is maximum. Mean of the projected data of $'+ve'$ class,

$$m_+ = \frac{1}{N_+}\sum_{n=1}^{N_+} a_n = \frac{1}{N_+}\sum_{n=1}^{N_+} \bar{w}^T\bar{x}_n = \bar{w}^T\bar{\mu}_+. \tag{5}$$

Where μ_+ is the mean vector of the original examples. Similarly, $m_- = \bar{w}^T\bar{\mu}_-$

Total deviation of examples from the center(μ) is defined as scatter. Scattar matrix(S) is given by

$$S = \sum_{n=1}^{N}(x_n - \mu)(x_n - \mu)^T. \tag{6}$$

Scatter matrix of the original $+ve$ class,

$$S_+ = \sum_{n=1}^{N_+}(\bar{x}_n - \bar{\mu}_+)(\bar{x}_n - \bar{\mu}_+)^T. \tag{7}$$

Scatter matrix of the original $-ve$ class,

$$S_- = \sum_{n=1}^{N_-}(\bar{x}_n - \bar{\mu}_-)(\bar{x}_n - \bar{\mu}_-)^T. \tag{8}$$

Total within class scatter matrix,

$$S_w = S_+ + S_-. \tag{9}$$

Between class scatter matrix(S_B) represents the deviation of mean of $+ve$ class from the mean of $-ve$ class.

$$S_B = (\bar{\mu}_+ - \bar{\mu}_+)(\bar{\mu}_+ - \bar{\mu}_+)^T. \tag{10}$$

Now separability/Fisher discriminant ratio is

$$J(w) = \frac{(m_+ - m_-)^2}{S_+^2 + S_-^2} = \frac{\bar{w}^T S_B \bar{w}}{\bar{w}^T S_w \bar{w}} \tag{11}$$

After maximizing J -

$$S_B \bar{w} = \lambda \bar{w} \tag{12}$$

Where λ is a constant. It gives the magnitude that does not affect the direction. Now the direction($\bar{w}$) corresponding to maximum valued eigen vectors of λ is

$$\bar{w} = \lambda S_w^{-1}(\bar{\mu}_+ - \bar{\mu}_-) \tag{13}$$

After getting the direction($\bar{w}$), we can get more separable univariate data(a_n) by projecting multivariate data($\bar{x}_n$) on the direction obtained.

$$a_n = \bar{w}^T \bar{x}_n \tag{14}$$

1.3 Single sample perceptron:

The goal of the single sample prceptron learning is to find a linear function/hyperplane that can discriminate data of two classes such that data of one class is on one side of the line and another class is on other side. For correct classification

$$when\ y_n = +1,\ \bar{w}^T \bar{x}_n + w_0 = \bar{a}^T \bar{z}_n \geq 0\ and\ when\ y_n = -1,\ \bar{w}^T \bar{x}_n + w_0 = \bar{a}^T \bar{z}_n < 0 \tag{15}$$

Where

$$\bar{a} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} \quad and \ \bar{z} \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} \tag{16}$$

In training period, weight vector $\bar{a}$ gets updated untill unless misclassified matrix gets empty. And at (k+1)th iteration, weight vector is

$$\bar{a}(k+1) = \bar{a}(k) + \eta \sum y_n \bar{z_n}. \tag{17}$$

1.4 Support vector machine(SVM)

Support vector machine are motivated by preprocessing the data to represent patterns in a high dimension-typically much higher than the original features space. With an appropriate non linear mapping to a sufficiently higher dimension, data from two categories can always be separated by a hyperplane. The support vectors are the training samples that define the optimal separating hyperplanes. In formally speaking they are the patterns most informative for the classification task [1].

$$Margin = \frac{g(\bar{x})}{||w||} = \frac{2}{||w||}. \tag{18}$$

Our objective to minimize the margin. or maximize $\frac{1}{2}||w||^2$ w.r.t $\bar{w}$ constrained by the seperation into a unconstrained problem by the method of Lagrange undetermined multipliers. To support nonlinearity, we use kernel method in SVM. Here in the assignment, we use the following three kernels.

1.4.1 Linear kernel

$$K(\bar{x}_m, \bar{x}_n) = \bar{x}_m^T \bar{x}_n. \tag{19}$$

1.4.2 Polynomial kernel

$$K(\bar{x}_m, \bar{x}_n) = (a\bar{x}_m\bar{x}_n + b)^p \tag{20}$$

1.4.3 Gaussian/RBF kernel

$$K(\bar{x}_m, \bar{x}_n) = \exp(\frac{-||\bar{x}_m - \bar{x}_n||^2}{\sigma}) \tag{21}$$

2 Gaussian mixture model(GMM) on the reduced dimensional representations of Dataset-2 obtained using PCA

First 24 dimentional histogram features are extracted form scene image and then 32 dimentional BOVW features are obtained using k-means clustering. Finally 32 dimentional BOVW was reduced by using PCA and first l(i.e.1,2,4,8,16) principle components were taken. GMM was built with different number of mixtures and results are observed.

2.0.4 Performance Analysis of GMM based classifier for scene image data using PCA

Table 1 shows the confusion matrix(for no of Gaussian=1 and no of principle components=4) and the table 2 shows the total performance of the GMM based classifier for different no of Gaussian and different number of principle components and higest performance is highlighted.

Table 1

Confusion matrix when BOVW of scene image data is considered for GMM based classifier for k=1 and l=4.

		Actual			
		Coast	Kennel	V court	Total
Prediction	Coast	37	6	2	45
	Kennel	10	37	7	54
	V court	3	7	41	51
	Total	50	50	50	

The performance measure of the optimal value of k(=1) and l(=4) interms of accuracy, recall, precesion and F-score is obtained by using eq(13-24).

$$Accuracy\ of\ the\ system = \frac{No\ of\ examples\ correctly\ classified}{Total\ no\ of\ test\ samples} \times 100\% == 76.66\% \tag{22}$$

$$Recall\ for\ class\ Coast = \frac{True\ Positive\ for\ Class\ Coast}{Total\ Actual\ Coast} = 0.74. \tag{23}$$

$$Precision\ for\ class\ Coast = \frac{True\ Positive\ for\ Class\ Coast}{Total\ Predicted\ Coast} = 0.82. \tag{24}$$

$$F - measure\ for\ class\ Coast = \frac{2 \times (Precision \times Recall)}{Precision + Recall} = 0.78. \tag{25}$$

$$Recall\ for\ class\ Kennel = \frac{True\ Positive\ for\ Class\ Kennel}{Total\ Actual\ Kennel} = 0.74. \tag{26}$$

$$Precision\ for\ class\ Kennel = \frac{True\ Positive\ for\ Class\ Kennel}{Total\ Predicted\ Kennel} = 0.69. \tag{27}$$

$$F - measure\ for\ class\ Kennel = \frac{2 \times (Precision \times Recall)}{Precision + Recall} = 0.71. \tag{28}$$

$$Recall\ for\ class\ VC = \frac{True\ Positive\ for\ Class\ VC}{Total\ Actual\ VC} = 0.82. \tag{29}$$

$$Precision\ for\ class\ VC = \frac{True\ Positive\ for\ Class\ VC}{Total\ Predicted\ VC} 0.80. \tag{30}$$

$$F - measure\ for\ class\ VC = \frac{2 \times (Precision \times Recall)}{Precision + Recall} = 0.78. \tag{31}$$

$$Mean\ recall\ = \frac{1}{M}\sum_{i=1}^{M} Recall for\ class\ C_i = 0.77. Where\ M = \ No.\ of\ classes. \tag{32}$$

$$Mean\ precision\ = \frac{1}{M}\sum_{i=1}^{M} Precision for\ class\ C_i = 0.77. Where\ M = \ No.\ of\ classes. \tag{33}$$

Table 2

Performance Analysis of GMM based classifier for scene image data with different no of mix(k) and PCs(l):

Sl. No	l	k	Accuracy	Recall(Coast,Kennel,VC,Mean)	Precesion(Coast,Kennel,VC,Mean)	F-measure(Coast,Kennel,VC)
1	1	1	58.67	0.4,0.64,0.72,0.59	0.47,0.52,0.8,0.59	0.43,0.57,0.43
2	1	2	57.67	0.26,0.76,0.68,0.57	0.43,0.49,0.81,0.58	0.33,0.59,0.33
3	1	4	61.33	0.46,0.66,0.72,0.61	0.55,0.54,0.77,0.62	0.5,0.60,0.5
4	2	1	74.7	0.7,0.72,0.82,0.75	0.78,0.65,0.82,0.75	0.74,0.69,0.74
5	2	2	69	0.56,0.68,0.82,0.67	0.67,0.58,0.84,0.70	0.61,0.62,0.60
6	2	4	70	0.60,0.74,0.76,0.70	0.71,0.61,0.81,0.71	0.65,0.67,0.65
7	4	1	77	0.74,0.74,0.82,0.77	0.82,0.69,0.80,0.77	0.78,0.71,0.78
8	4	2	73	0.78,0.66,0.76,0.73	0.74,0.69,0.78,0.73	0.76,0.68,0.76
9	8	1	73	0.72,0.66,0.82,0.73	0.72,0.69,0.79,0.73	0.72,0.67,0.72
10	8	2	63	0.70,0.52,0.80,0.67	0.61,0.67,0.74,0.67	0.65,0.58,0.65
11	12	1	73	0.68,0.70,0.82,0.73	0.76,0.67,0.77,0.73	0.71,0.68,0.71
12	16	1	71	0.66,0.62,0.84,0.71	0.70,0.62,0.79,0.70	0.68,0.62,0.68

2.0.5 Observation

Here we can observe that the classifier yields 70% accuracy when no of Gaussians is 4 and first principle component is taken into consideration. Earlier(2nd assignment) when we used 32 dimensional BOVW feature in building GMM,

we got 38.67% accuracy using k=1 and larger no of k values, GMM was not converging. By projecting original data on priciple orthogonal directions we get the good results.

3 Gaussian mixture model(GMM) on the reduced dimensional representations obtained using LDA

GMM basd classifiers are buid for the three types of data set(artificial linearly separable, artificial non-linearly separable and scene image data set). For scene image data, BOVW is extracted. Multidimentional features are then projected on higest separable direction using LDA discussed above followed by modelling using GMM with different no of Gaussians. Fig1 shows the general block diagram for data classification using FDA.

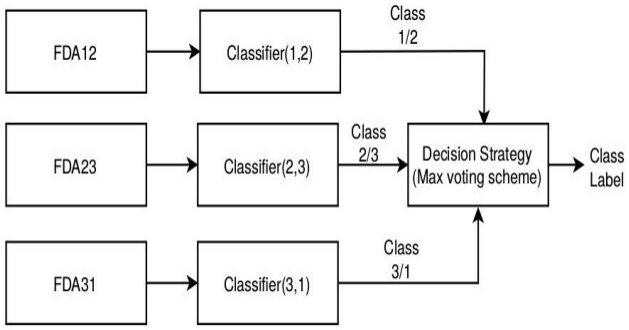

Fig. 1. General block diagram of a classifier using FDA.

3.1 GMM on reduced dimension of artificial linearly separable data obtained by LDA

2D data is projected on the single direction that leads to high separability. Projected data is used for classification using GMM with different numbers of Gaussian.

3.1.1 Performance Analysis of GMM based classifier for artificial linearly separable data

Table 3 shows the confusion matrix for all case and the table 4 shows the total performance of the GMM based classifier for different values of k. Fig2 shows the decision boundary when no of Gaussian is 2.

Table 3

Confusion matrix when artificial linearly separable data is considered for GMM based classifier.

		Actual			
		C1	C2	C3	Total
Prediction	C1	125	0	0	125
	C2	0	125	0	125
	C3	0	0	125	125
	Total	125	125	125	

Table 4

Performance Analysis of GMM based classifier for artificial linearly separable data:

Sl. No	k	Accuracy	Recall(C1,C2,C3,Mean)	Precesion(C1,C2,C3,Mean)	F-measure(C1,C2,C3)
1	1	100	1,1,1,1	1,1,1,1	1,1,1
2	2	100	1,1,1,1	1,1,1,1	1,1,1
3	8	100	1,1,1,1	1,1,1,1	1,1,1

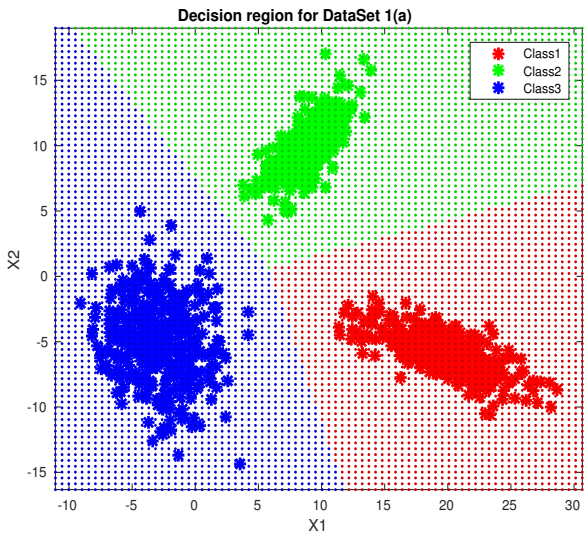

Fig. 2. Decision region plot for all the three classes together with the training data superimposed(k=2).

3.2 GMM on reduced dimension of artificial non-linearly separable data obtained by LDA

2D non-linearly separable data is projected on the single direction that leads to high separability. Projected data is used for classification using GMM with different numbers of Gaussians.

3.2.1 Performance Analysis of GMM based classifier for artificial non-linearly separable data

Table 5 shows the confusion matrix for k=2 and the table 6 shows the total performance of the GMM based classifier for different values of k. Fig3 shows the decision boundary when no of Gaussian is 2.

Table 5

Confusion matrix when artificial non-linearly separable data is considered for GMM(k=2) based classifier.

		Actual			
		C1	C2	C3	Total
Prediction	C1	109	6	14	129
	C2	5	119	0	124
	C3	11	0	111	122
	Total	125	125	125	

Table 6

Performance analysis of GMM based classifier for artificial non-linearly separable data:

Sl. No	k	Accuracy	Recall(C1,C2,C3,Mean)	Precesion(C1,C2,C3,Mean)	F-measure(C1,C2,C3)
1	1	90.40	0.87,0.95,0.89,0.90	0.84,0.96,0.91,0.90	0.86,0.96,0.90
2	2	90.40	0.87,0.95,0.89,0.90	0.84,0.96,0.91,0.90	0.86,0.96,0.90
3	16	88.80	0.90,0.90,0.86,0.89	0.79,0.97,0.93,0.90	0.84,0.93,0.89

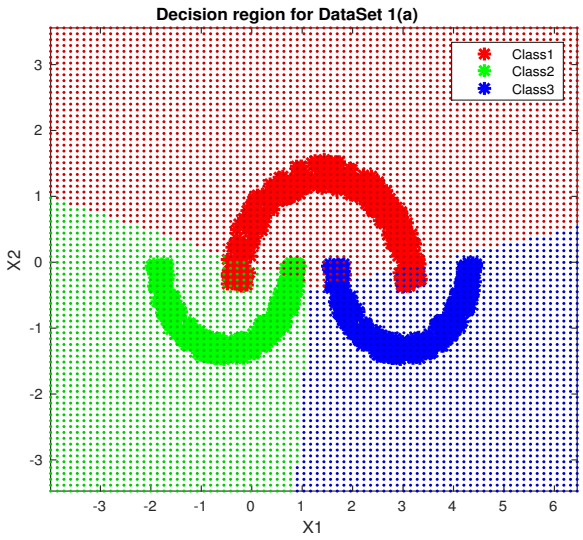

Fig. 3. Decision region plot for all the three classes together with the training data superimposed(k=2).

3.3 GMM on reduced dimension of scene image data obtained by LDA

32 dimension BOVW extracted from scene image data is projected on the single direction that leads to high separability and then classification task is carried out.

3.3.1 Performance analysis of GMM based classifier for scene image data

Table 7 shows the confusion matrix for k=2 and the table 7 shows the total performance of the GMM based classifier for different values of k.

Table 7

Confusion matrix when BOVW of scene image data is considered for GMM based classifier using LDA.

		Actual			
		Coast	Kennel	V court	Total
Prediction	Coast	39	12	7	48
	Kennel	6	32	3	41
	V court	5	6	40	51
	Total	50	50	50	

Table 8

Performance Analysis of GMM based classifier for scene image data with different no of mix(k):

Sl. No	k	Accuracy	Recall(Coast,Kennel,VC,Mean)	Precesion(Coast,Kennel,VC,Mean)	F-measure(Coast,Kennel,VC)
1	1	74	0.67,0.78,0.78,0.75	0.78,0.64,0.80,0.74	0.72,0.70,0.72
2	2	74	0.67,0.78,0.78,0.75	0.78,0.64,0.80,0.74	0.72,0.70,0.72
3	4	46	0.84,0.54,0,0.46	0.41,0.56,-,-	0.55,0.55,-

3.4 Observation

For artificial linearly separable data-set, the system classifies test data perfectly fine. We get 90.40% accuracy for non-linearly separable artificial data. GMM with k(1 or 2) Gaussians using reduced BOVW feature for scene image data yields the accuracy of 74% which is much better than 32 dimentional BOVW we used earlier in second assignment. But it performing slightly less than the GMM using PCA based reduced feature. System remains unconverged with larger values of k.

4 Perceptron-based classifier on artificial linearly separable data:

Perceptron based classifier is applied to artificial linearly separable data set. At training period optimal weights of separating line between two class is obtained. At the time of testing, discriminant function is investigated and depending on maximum voting scheme, decision is taken.

4.0.1 Performance Analysis of Perceptron-based classifier for artificial linearly separable data

Table 9 shows the confusion matrix for Perceptron-based classifier for artificial linearly separable data. Fig4 shows the decision boundary obtained by perceptron based classifier.

Table 9

Confusion matrix when artifical linearly separable data is considered for GMM based classifier.

		Actual			
		C1	C2	C3	Total
Prediction	C1	125	1	0	126
	C2	0	124	0	124
	C3	0	0	125	125
	Total	125	125	125	

$$Accuracy\ of\ the\ system = \frac{No\ of\ examples\ correctly\ classified}{Total\ no\ of\ test\ samples} \times 100\% = 99.73\% \tag{34}$$

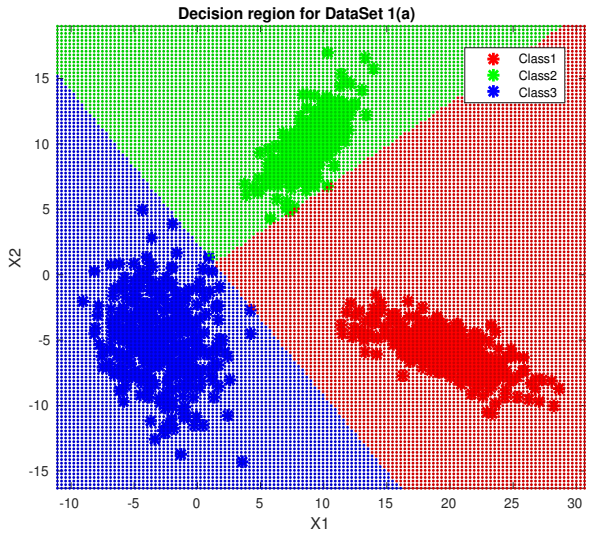

Fig. 4. Decision region plot of perceptron based classifier for all the three classes together with the training data superimposed.

4.0.2 Observation

We get accuracy of 99.73% in perceptron based classification experiment. The classifier always make a linear decision boundary but there is a high probability of lying the decision line close to one of the two classes as initial weight

vector is taken arbitrary and hence it rise to missclassifications. Perceptron learning always converges to a solution after finite iteration provided data is linearly seperable.

5 SVM based classifier using different kernel on different data set

SVM based classifier is used with different kernels on the three data-set.

5.1 SVM on artificial linearly separable data

Table 8 shows the confusion matrix for SVM based classifier using RBF kernel on data set1(a).And Figure ,figure 6, figure 7, figure 8 and figure 9 shows the supports vectors and decision regions superimposed with test examples.

Table 10

Confusion matrix when artificial linearly separable data is considered for SVM based classifier(kernel function is linear) on data set 1(a).

		Actual			
		C1	C2	C3	Total
Prediction	C1	125	0	0	125
	C2	0	125	0	125
	C3	0	0	125	125
	Total	125	125	125	

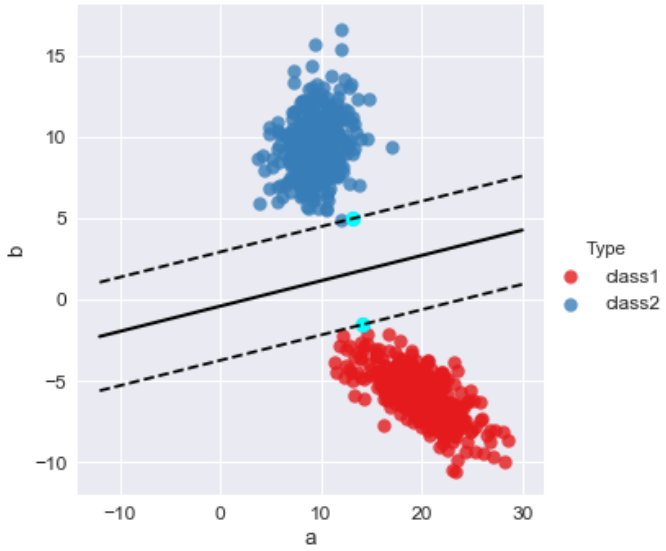

Fig. 5. Visualization of support vectors between class1 and class2 for 1(a) data set.

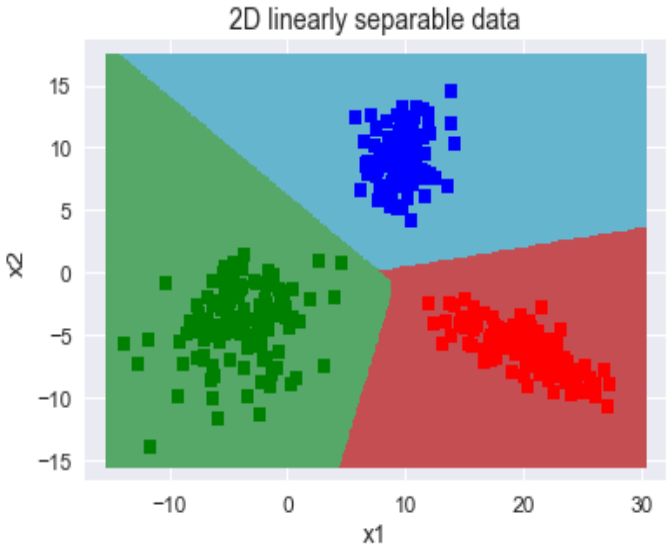

Fig. 6. Decision region plot for all the three classes together with the test data superimposed(linear kernel,C=32).

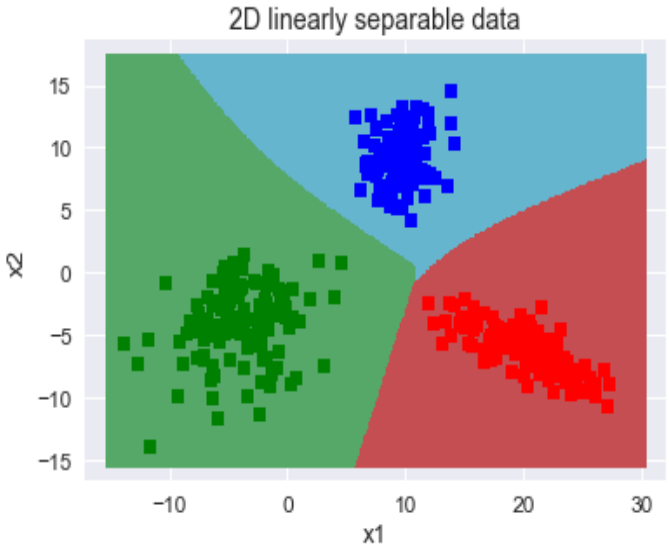

Fig. 7. Decision region plot for all the three classes together with the test data superimposed(polynomial kernel,C=1,P=2).

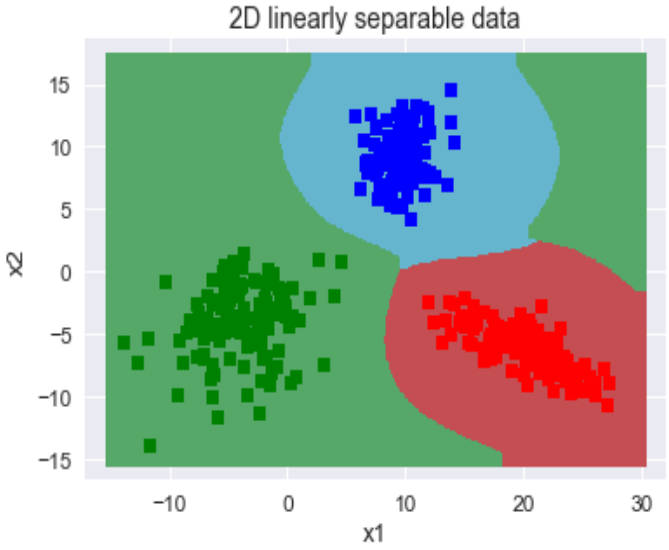

Fig. 8. Decision region plot for all the three classes together with the test data superimposed(RBF kernel,C=1).

5.2 SVM on artificial non-linearly separable data

Table 11 shows the confusion matrix for SVM based classifier using RBF kernel on data set1(b).And Figure ,figure 10, figure 11, figure 12 and figure 13 shows the supports vectors and decision regions superimposed with test examples.

Table 11

Confusion matrix when artificial non-linearly separable data is considered for SVM based classifier(kernel function is RBF,C=1,γ=32) on data set 1(b).

		Actual			
		C1	C2	C3	Total
Prediction	C1	125	0	0	125
	C2	0	125	0	125
	C3	0	0	125	125
	Total	125	125	125	

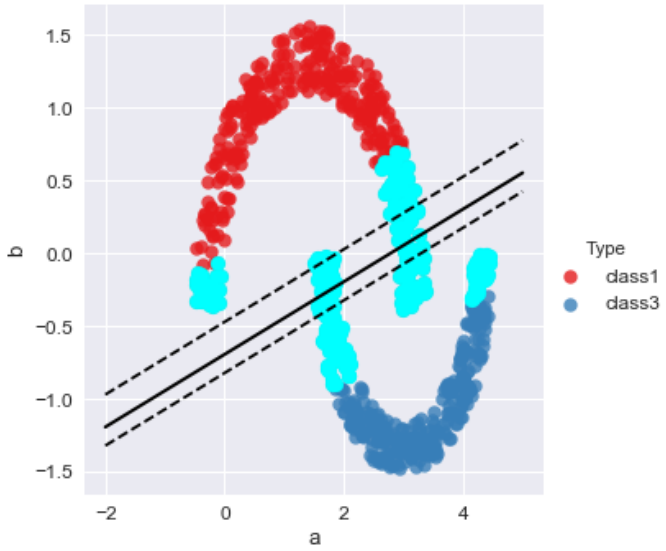

Fig. 9. Visualization of support vectors between class1 and class3 for 1(b) data set

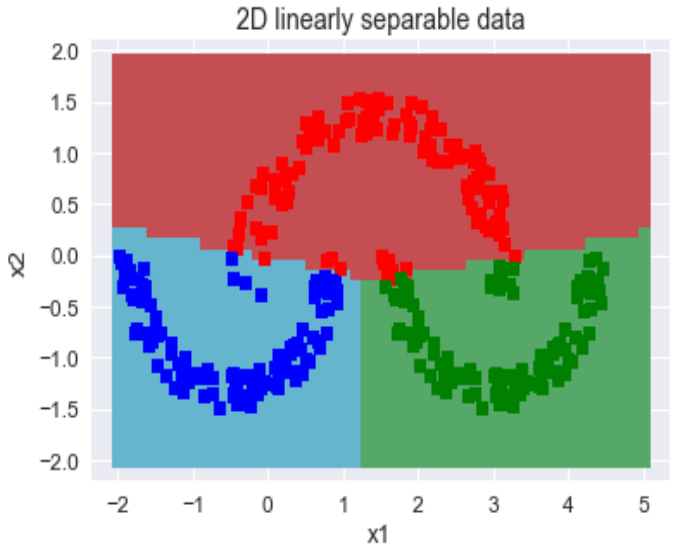

Fig. 10. Decision region plot for all the three classes together with the test data superimposed(Linear kernel,C=1).

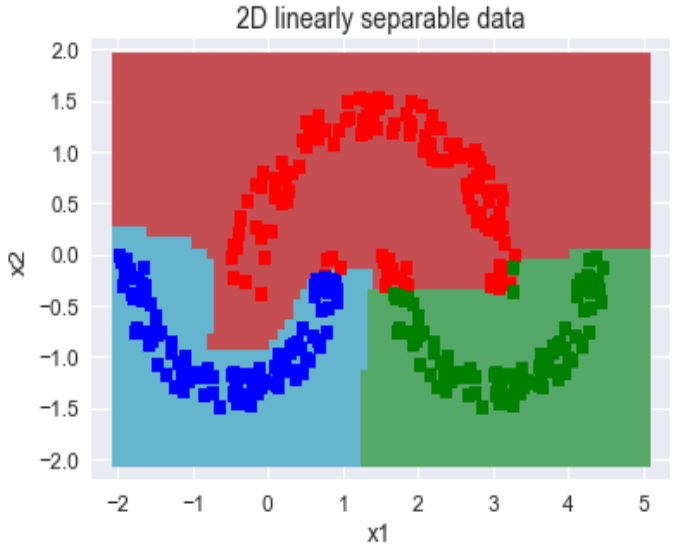

Fig. 11. Decision region plot for all the three classes together with the test data superimposed(Polynomial kernel,C=1,γ =1.5,P=5).

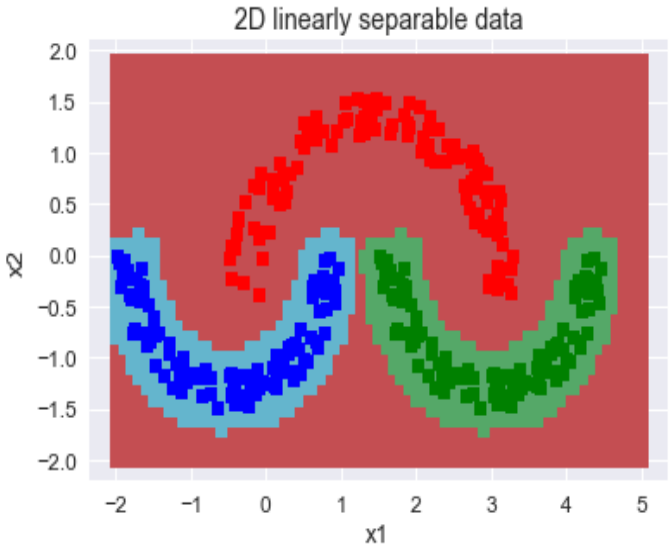

Fig. 12. Decision region plot for all the three classes together with the test data superimposed(RBF kernel,C=1,γ =1).

5.3 SVM on scene image data

32 dimension BOVW extracted from scene image data and then classification task is carried out. Table 12 shows the confusion matrix for SVM based classifier using RBF kernel on scene image data .

Table 12

Confusion matrix when BOVW of scene image data is considered for SVM based classifier using RBF kernel

		Actual			
		Coast	Kennel	V court	Total
Prediction	Coast	36	2	0	38
	Kennel	4	32	0	25
	V court	12	25	60	87
	Total	50	50	50	

Table 13

Performance Analysis of SVM based classifier on different data-set with different parameters(trade off parameter C,degree P,γ)

Sl. No	Data set	kernel	Accuracy	Mean Recall	Mean Precesion	Mean F-measure
1	1(a)	L($C = 32, \gamma = auto$)	100	1	1	1
2	1(a)	P($C = 32, \gamma = auto, P = 3$)	100	1	1	1
3	1(a)	G($C = 32, \gamma = 0.031$)	100	1	1	1
4	1(b)	L($C = 1, \gamma = auto$)	92.20	0.93	0.93	0.92
5	1(b)	P($C = 1, \gamma = 1.54, P = 5$)	94.4	0.95	0.94	0.94
6	1(b)	G($C = 1, \gamma = 32$)	100	1	1	1
7	2	L($C = 0.01, \gamma = auto$)	85.6	0.8	0.85	0.82
8	2	P($C = 2, \gamma = auto, P = 2$)	82.01	0.76	0.81	0.78
9	2	G($C = 1, \gamma = 0.0009$)	85.6	0.71	0.86	0.76

5.4 Observation

In both, bayes classifier and SVM we are getting 100% accuracy on artificial linearly separable data set. For artificially non-linearly separable data using bayes classifier, we were getting 80.1% accuracy. For both GMM($k > 1$ in assignment 2) and SVM based classifier yields the accuracy of 100%. The best accuracy given by kNN based classifier(in assignment 3) is 75.33% with comparision to 85.6% accuracy given by SVM on scene image data set. In Table 13, results are observed for different dataset using differnet kernels for classification. It was observed that radial basis function (RBF or Gaussian) gave best performance because of its nonlinear behaviour. Here, C and γ parameters have been altered on trail and error basis for getting better results, C parameter gives trade-off between smooth decision boundary and classifying training data correctly. γ parameter decides the complexity of decision boundary.

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.