Assignment 3

Statistical Machine Learning

- Mayank Chauhan, MT18008

About Dataset

Face dataset

Total images : 715

Total images after preprocessing : 710

No. of classes: 11

Original image size: 192 x 168

New image size: ***48 x 42***

Training set: 495

Test set: 215

Original dataset have 65 images per class.

Class labels: 0 to 10

Cifar dataset

Total images: 60,000

No. of classes: 10

Training set: 50000

Test set: 10000

Image size: 32 x 32

Training set contains 6000 images per class and test contains 1000 images per class.

Class labels : 0 – airplane, 1-automobile, 2-bird, 3-cat, 4-deer, 5-dog, 6-frog, 7-horse, 8-ship, 9-truck.

Classification algorithm

Gaussian Naive Bayes (GaussianNB)

Classifying Test set

	Face data	CIFAR
Accuracy	70.23%	26.83%

LDA Projected Data

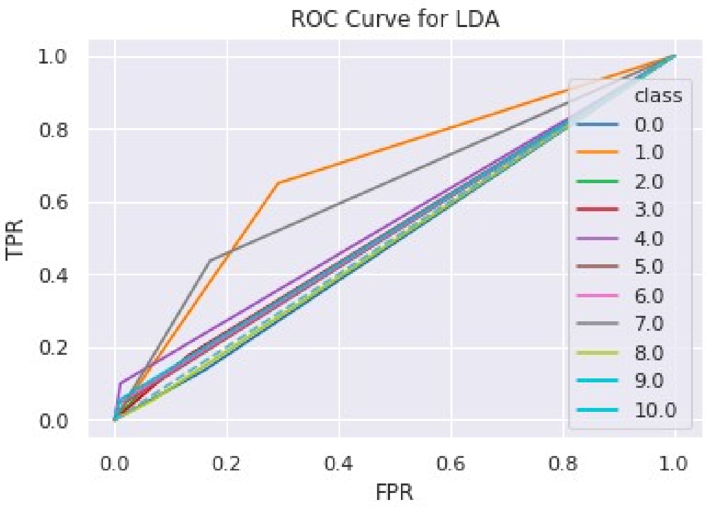
5-Fold Cross Validation

	Face data	CIFAR
Mean of accuracy	100%	34.91%
Standard deviation of accuracy	0	0.0039
Best model accuracy on test set	17.20%	27.79%

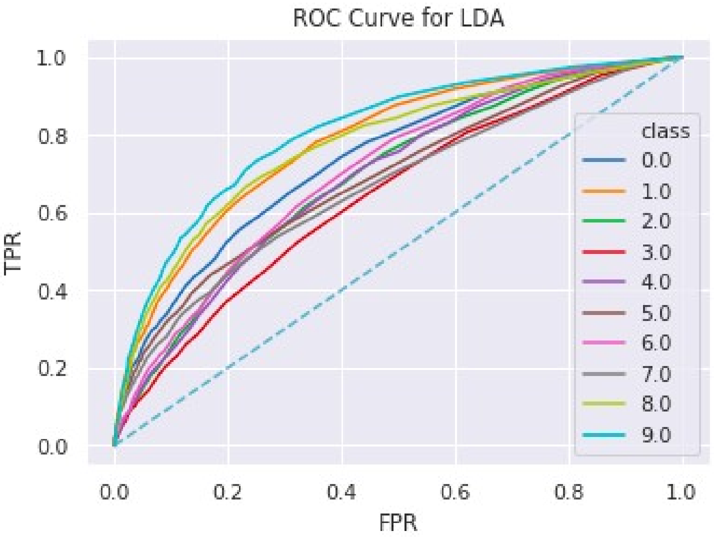
The cross validation accuracy over face data is 100% , because the classifier is able to learn about the data from the small no. of samples. Making the estimated error rate at 0%. But when tested on the test set, the error rate is 82.6%, which means we overly estimated over error rate using cross validation.

ROC curve

Face data

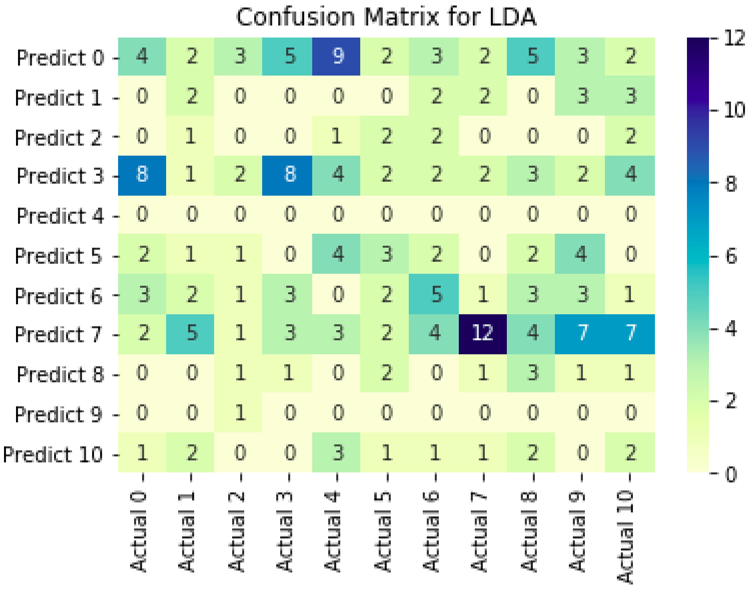


CIFAR



Confusion Matrix

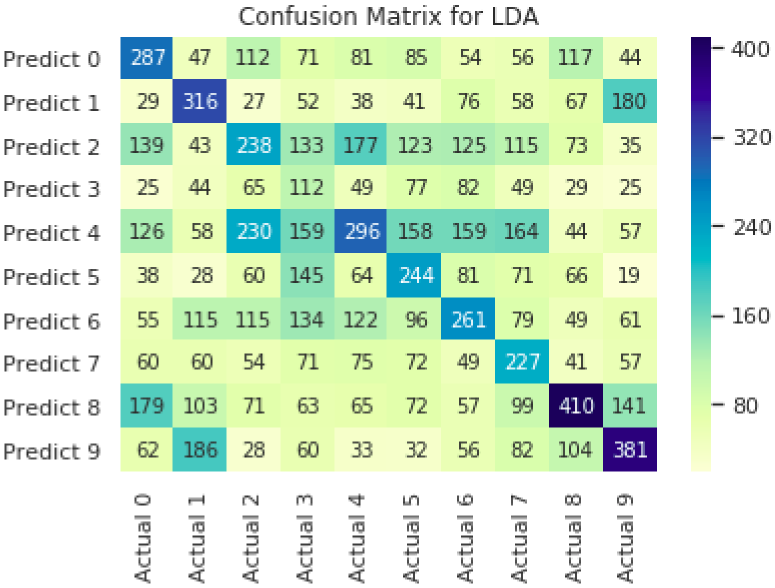
Face data



Confusion Matrix for LDA

	Actual 0	Actual 1	Actual 2	Actual 3	Actual 4	Actual 5	Actual 6	Actual 7	Actual 8	Actual 9	Actual 10
Predict 0	4	2	3	5	9	2	3	2	5	3	2
Predict 1	0	2	0	0	0	0	2	2	0	3	3
Predict 2	0	1	0	0	1	2	2	0	0	0	2
Predict 3	8	1	2	8	4	2	2	2	3	2	4
Predict 4	0	0	0	0	0	0	0	0	0	0	0
Predict 5	2	1	1	0	4	3	2	0	2	4	0
Predict 6	3	2	1	3	0	2	5	1	3	3	1
Predict 7	2	5	1	3	3	2	4	12	4	7	7
Predict 8	0	0	1	1	0	2	0	1	3	1	1
Predict 9	0	0	1	0	0	0	0	0	0	0	0
Predict 10	1	2	0	0	3	1	1	1	2	0	2

CIFAR



Confusion Matrix for LDA

	Actual 0	Actual 1	Actual 2	Actual 3	Actual 4	Actual 5	Actual 6	Actual 7	Actual 8	Actual 9
Predict 0	287	47	112	71	81	85	54	56	117	44
Predict 1	29	316	27	52	38	41	76	58	67	180
Predict 2	139	43	238	133	177	123	125	115	73	35
Predict 3	25	44	65	112	49	77	82	49	29	25
Predict 4	126	58	230	159	296	158	159	164	44	57
Predict 5	38	28	60	145	64	244	81	71	66	19
Predict 6	55	115	115	134	122	96	261	79	49	61
Predict 7	60	60	54	71	75	72	49	227	41	57
Predict 8	179	103	71	63	65	72	57	99	410	141
Predict 9	62	186	28	60	33	32	56	82	104	381

PCA Projected data

5-Fold Cross Validation

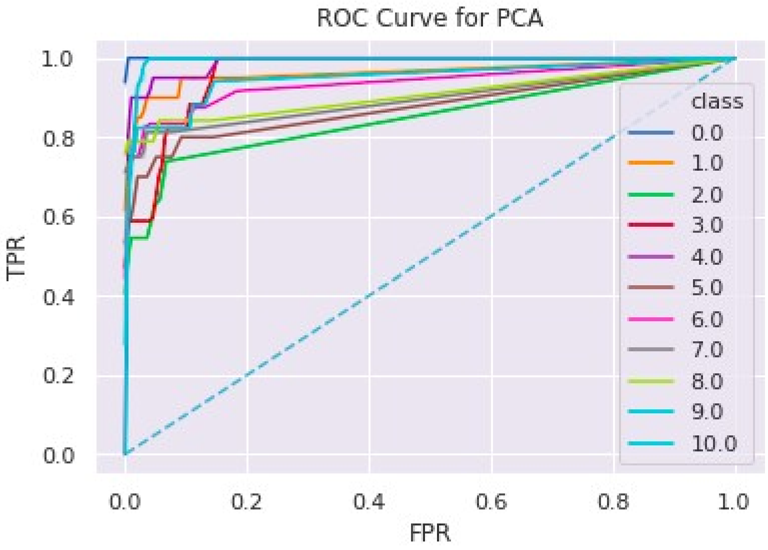
	Face data	CIFAR
Mean of accuracy	76.36%	27.31%
Standard deviation of accuracy	0.0375	0.0023
Best model accuracy on test set	84.18%	27.92%

By keeping 95% eigen energy we are getting 25 principal components for face data and for CIFAR 160 principal components. The face dataset have only images of faces, a lot features (pixels values) are common between images of different faces. Whereas, the images in CIFAR are very different between the classes. Therefore, the redundancy is very high in the face data and low in CIFAR.

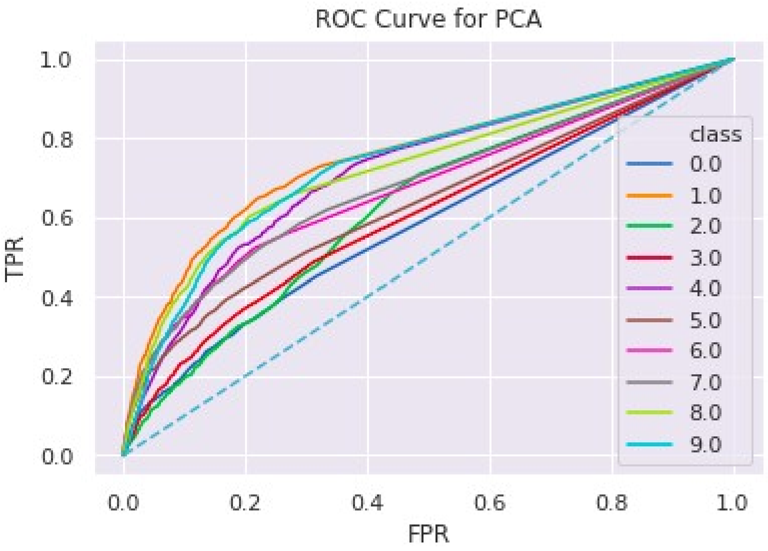
That’s why we have low accuracy in CIFAR.

ROC curve

Face data

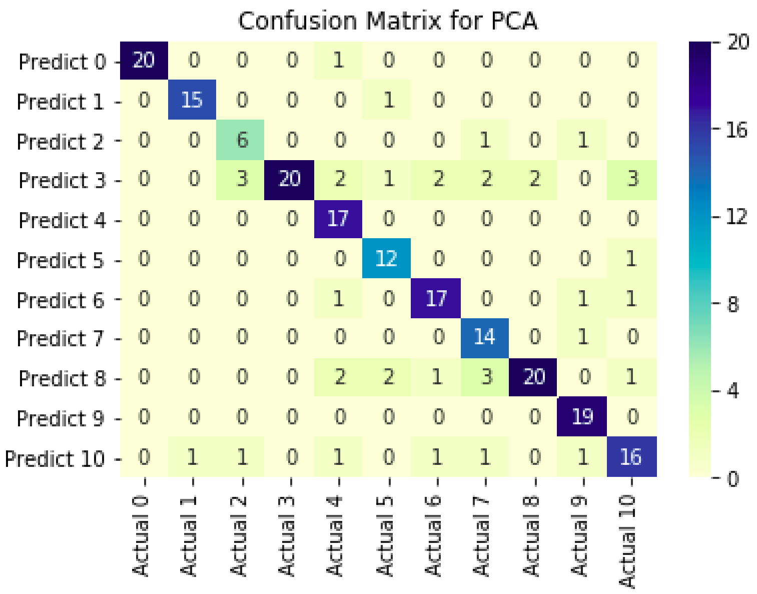


CIFAR



Confusion Matrix

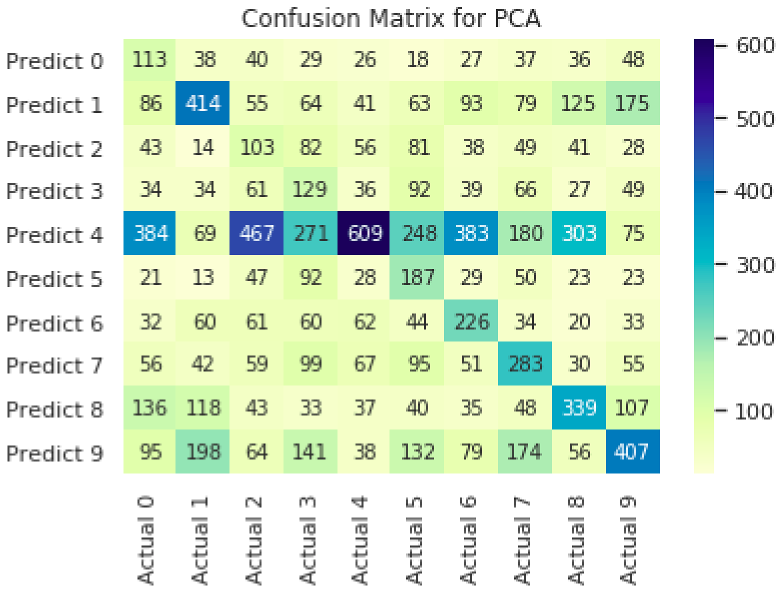
Face data



Confusion Matrix for PCA

	Actual 0	Actual 1	Actual 2	Actual 3	Actual 4	Actual 5	Actual 6	Actual 7	Actual 8	Actual 9	Actual 10
Predict 0	20	0	0	0	1	0	0	0	0	0	0
Predict 1	0	15	0	0	0	1	0	0	0	0	0
Predict 2	0	0	6	0	0	0	0	1	0	1	0
Predict 3	0	0	3	20	2	1	2	2	2	0	3
Predict 4	0	0	0	0	17	0	0	0	0	0	0
Predict 5	0	0	0	0	0	12	0	0	0	0	1
Predict 6	0	0	0	0	1	0	17	0	0	1	1
Predict 7	0	0	0	0	0	0	0	14	0	1	0
Predict 8	0	0	0	0	2	2	1	3	20	0	1
Predict 9	0	0	0	0	0	0	0	0	0	19	0
Predict 10	0	1	1	0	1	0	1	1	0	1	16

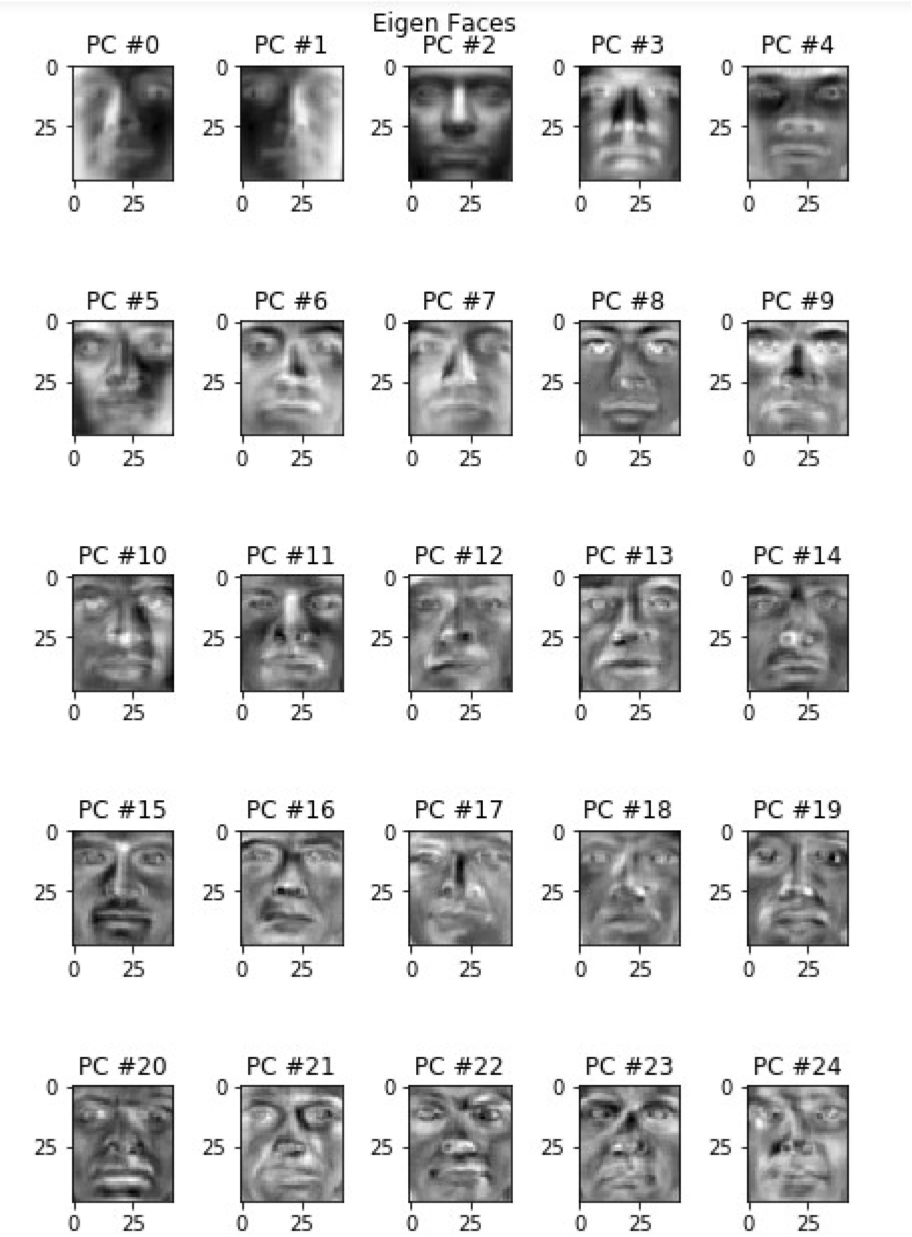
CIFAR



Confusion Matrix for PCA

	Actual 0	Actual 1	Actual 2	Actual 3	Actual 4	Actual 5	Actual 6	Actual 7	Actual 8	Actual 9
Predict 0	113	38	40	29	26	18	27	37	36	48
Predict 1	86	414	55	64	41	63	93	79	125	175
Predict 2	43	14	103	82	56	81	38	49	41	28
Predict 3	34	34	61	129	36	92	39	66	27	49
Predict 4	384	69	467	271	609	248	383	180	303	75
Predict 5	21	13	47	92	28	187	29	50	23	23
Predict 6	32	60	61	60	62	44	226	34	20	33
Predict 7	56	42	59	99	67	95	51	283	30	55
Predict 8	136	118	43	33	37	40	35	48	339	107
Predict 9	95	198	64	141	38	132	79	174	56	407

Visualizing eigenvectors for PCA



Each Principal Component is called eigen faces, becuase for this dataset we are able to visualize the meaning of each of the eigenvectors.

For 70% eigen energy

	Face data	CIFAR
Mean of accuracy	9.8%	30.01%
Standard deviation of accuracy	0.0250	0.0048
Best model accuracy on test set	18.6%	30.53%

For 90% eigen energy

	Face data	CIFAR
Mean of accuracy	63.23%	28.93%
Standard deviation of accuracy	0.0226	0.0033
Best model accuracy on test set	66.5%	28.6%

For 99% eigen energy

	Face data	CIFAR
Mean of accuracy	82.02%	27.69%
Standard deviation of accuracy	0.0390	0.0037
Best model accuracy on test set	90.23%	28.29%

The higher the eigen energy the larger no. of principal components we are keeping. Therefore, the accuracy increase drastically for face data. Because it has very few relevant Principal components which are able to get the most of the relevant information of the image. While, CIFAR accuracy don’t change much because it’s has very large no. of eigenvectors in for the specified eigen energies.

LDA on PCA projected & PCA on LDA

Accuracy	Face data	CIFAR
LDA on PCA Projected data	84.65%	31.71%
PCA on LDA Projected data	17.20%	27.82%

Applying LDA after applying PCA increases the accuracy in case of CIFAR, because the discriminative power is in the mean. Each class is very different from each other, so the mean is also differs a lot. So, the LDA was able to presevere the class separation. Whereas, the Face data has discriminatory information in the variance rather than mean, therefore we don’t get any benefit of applying LDA over PCA projected data.

Q2.

Adaboost

Base Classifier : Decision Tree Classifier with max_depth = 2 and max_leaf_nodes = 3.

Error rate with base classifier = 90.8%

Check whether base classifer is a Weak Classifer

No. of classes = 26

We know, a weak classifer is slightly better than a random classifer.

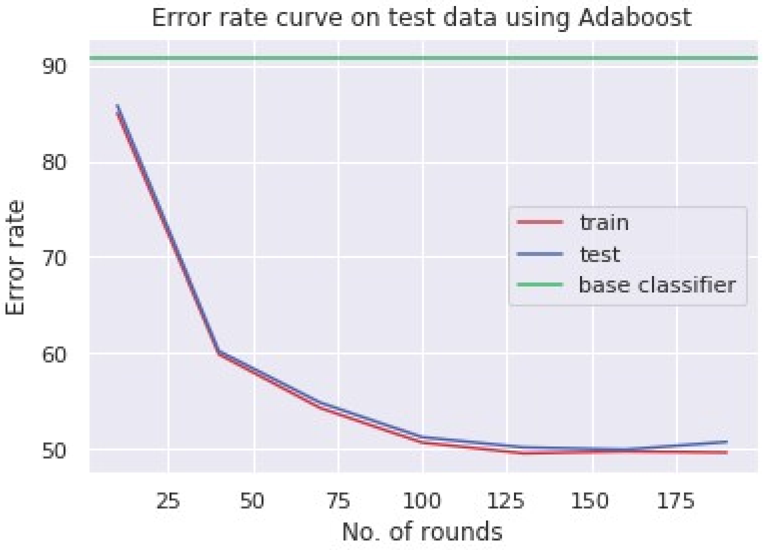
Random classifier accuracy = 1/26 = 3.84%

Base Classifer Accuracy has accuracy 9.2% which slightly bigger than the accuracy of a random classifer. Therefore, our base classifier is a weak classifier.

5-Fold Cross Validation

# Rounds = 10	Accuracy	Error rate
Mean	17.48%	82.52%
Standard deviation	2.638	2.638

Error rate Vs No. of rounds



The training error rate and test error rate both decrerases with increase in no. of rounds rapidly.

Classification accuracy using Adaboost

# Rounds	Accuracy
10	14.2%
40	39.75%
70	45.13%
100	49.3%

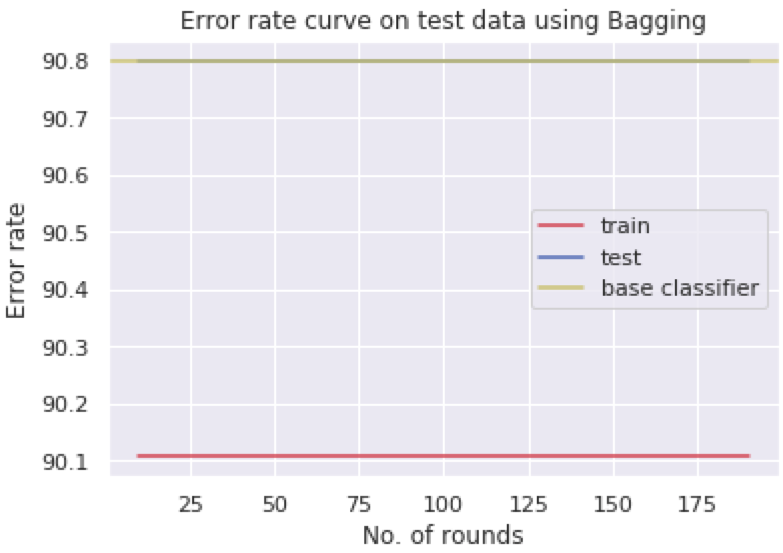
Weak classifiers suffers from high bias, we take an ensemble of many such classifers so that we can reduce the bias. The Adaboost is not overfitting the data, even after the algorithm for large no. Of rounds, which indicates that it is very resisitent to increasing the variance. Also, Adaboost is giving much better results than bagging because in this the trees are grown sequentially using the information from the preivously grown trees. Each new tree tries to emphasize the preivously misclassified data points in each iteration of the algorihtm.

Bagging

5-Fold Cross Validation

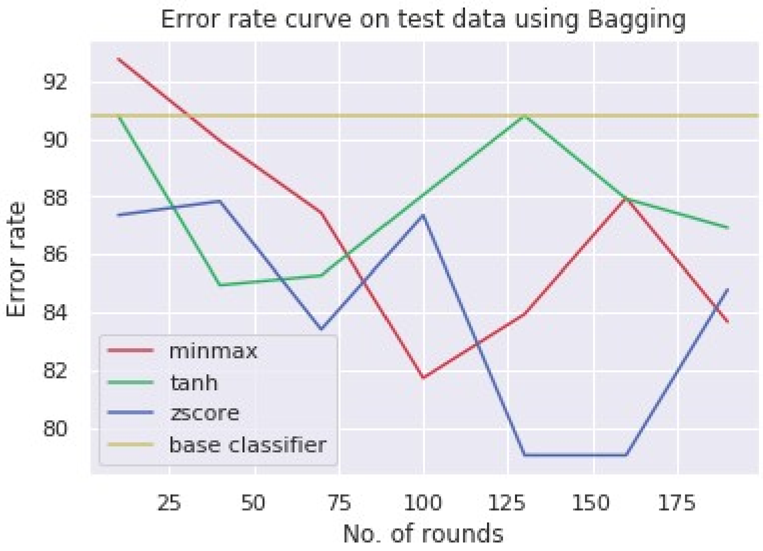
# Rounds = 10	Accuracy	Error rate
Mean	10.52%	89.47%
Standard deviation	1.1007	1.1007

Error rate Vs No. of rounds



Error rate = 90.8% on test data using Bagging with majority voting and without normalisation.

Score Normalisation



Accuracy

# Rounds	Minmax	Tanh	Zscore
10	7.23	12.83	12.16
40	16.08	11.78	15.21
70	11.11	13.08	12.16
100	14.13	9.2	12.16
130	16.33	15.93	19.58
160	18.33	17.63	20.96

Minmax have lower accuracy than other two normalisation technique. While, Zscore is giving max. accuracy.