SMAI ASSIGNMENT 2 REPORT

BY 20171019

I. QUESTION 1

A. Eigen Faces

Eigenfaces is the name given to a set of eigenvectors when they are used in the computer vision problem of human face recognition. Eigenfaces are the eigenvectors of the co variance matrix of face images.

B. Eigen Value Spectrum

The eigenvectors/faces from the eigen value spectrum is taken whose sum of eigenvalues is 0.95 of the total sum and used for reconstruction.

This ratio is kept below 5 - 10 percent for efficient Reconstruction. (First k eigen vectors are taken)

$$ratio = \frac{(\sum_{i=k+1}^{d} \lambda_i)}{(\sum_{i=1}^{d} \lambda_i)} \tag{1}$$

These can "satisfactorily" reconstruct a person in these three datasets.

IIIT CFW - about 300 to 350 , I have taken 310
IMFDB - about 110 to 130 , I have taken 120
Yale faces dataset - about 50 to 70 , I have taken 60

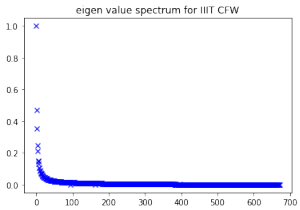

Fig. 1. Eigen value spectrum for IIIT CFW data

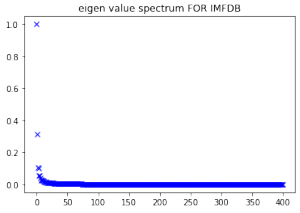

Fig. 2. Eigen value spectrum for IMFDB data

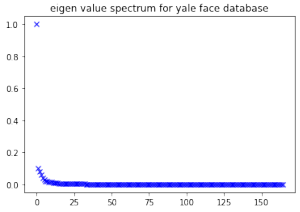

Fig. 3. Eigen value spectrum for Yale face data

C. Reconstruction

The images of the 3 datasets were reconstructed and their reconstruction errors were calculated.

IIIT CFW - 0.0677
IMFDB - 0.0333
Yale - 0.0514

Thus IIIT CFW dataset is hardest to represent with fewer eigen vectors/faces. This is because it required the most number of eigen vectors to preserve 95 percent variance , thus it has the highest reconstruction error. On the other hand IMFDB and yale datasets needed fewer eigen vectors and hence have lower losses and can be better represented with few vectors/faces.

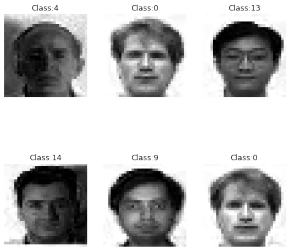

Fig. 4. yale dataset reconstruction

D. Hardest to represent

For every data set, we calculate the reconstruction loss for each class in the data set and conclude that the class with the highest loss will be the hardest to represent with few eigen vectors. With more eigen vectors though this error can be minimised.

IIIT CFW - class 5 Narendra Modi id hardest to represent
IMFDB - class 7 Amir Khan is hardest to represent
Yale face database - class 7 is hardest to represent

Empirically We can also say that the most difficult to represent classes were the ones which had error greater than 0.95 of the max reconstruction error.

II. QUESTION 2

We used different features and combined them with different classifiers to see which gave the best accuracy. Csv files have been attached in my project submission folder tabulating all the accuracy's and results. Usually combination of features with any classifier (PCA + LDA for example) and RESNET showed good accuracy while KPCA and other methods like KLDA had lesser accuracy values for all 3 datasets.

Also using classifier MLP with any feature gave better accuracy than LR, SVM and DT. LR seemed to be better when compared to SVM and DT which gave low accuracies with most features.

Plots of Accuracy vs classifier type that is MLP (multi layer perceptron), LR (Logistic regression) , SVM (Support vector machine) , DT (decision trees) for the 3 data sets are shown below for different features/combinations used.

NOTE : THE TEST TRAIN SPLIT WAS PERFORMED FIRST BEFORE FEATURE EXTRACTION AS TO MAKE SURE THAT WE DO NOT END UP TRAINING ON OUR TEST SET WHILE WE PERFORM FEATURE EXTRACTION. IF WE DID THE FEATURE SELECTION FIRST BEFORE SPLITTING AND TRAINING OUR DATA HIGH ACCURACIES WERE OBTAINED AS OUR MODEL HAS ALEREDY BEEN EXPOSED TO TEST DATA WHICH IT SHOULD NOT HAVE SEEN.

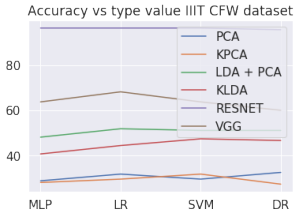

Fig. 5. IIIT CFW

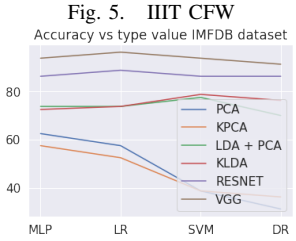

Fig. 6. IMFDB

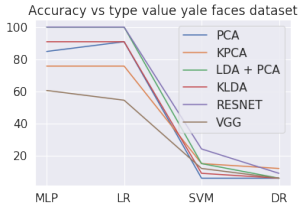

Fig. 7. Yale face dataset

III. QUESTION 3

A. t-SNE on original data

First we performed t-SNE on the original data sets and then for all the data sets combined. Here are the plots obtained (only 3D shown, rest in the notebook submitted)

There was more clustering in t-SNE as compared to the scatter plots obtained initially (in the notebook) using pca on original data. There was clustering seen in all three data sets and the combined form, but it was not very separated and can be clustered better.

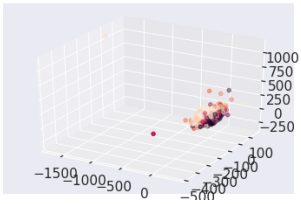

Fig. 8. IIIT CFW dataset

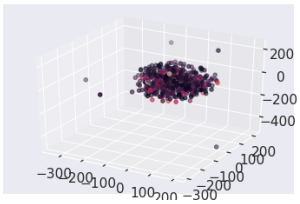

Fig. 9. IMFDB dataset

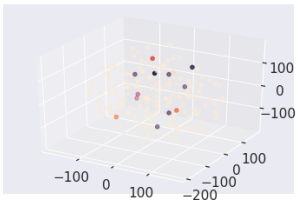

Fig. 10. Yale face dataset

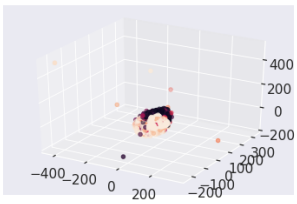

Fig. 11. Combined dataset

B. t-SNE on transformed data

Since original data did not give a satisfactory level of clustering, we tried t-SNE on data reduced by pca,lda and klda on all three data sets. These gave a got amount of clustering. especially klda and lda. When compared to 2d scatter plots of pca as obtained initially, t-SNE on reduced data clustered well and distinct classes can be seen. In IIIT CFW and IMFDB data points were clustered very well. It implies that the samples in the 3D dimen- sional space are still well-clustered in the reduced dimension, hence the feature transformation is an apt representation of the original data. Clusters were more close in the CFW data set though. But in yale data set it was not very well clustered. There was no apparent clustering/grouping of samples. This shows that the samples in the reduced dimension might not be an accurate representation of the samples in the original space.

Conclusions:-

1) In this way, t-SNE maps the multi-dimensional data to a lower dimensional space and attempts to find patterns in the data by identifying observed clusters based on similarity of data points with multiple features. It works better in terms of clustering than pca

2) When applied to reduced data by lda or pca or klda, t-SNE gives good clustering as shown in the plots below.

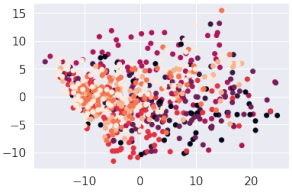

Fig. 12. 2d pca scatter plot

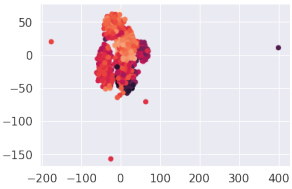

Fig. 13. t-SNE plot after LDA with CFW dataset

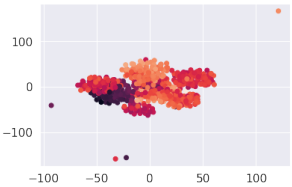

Fig. 14. t-SNE plot after LKDA with CFW dataset

IV. QUESTION 4

A. KNN formulation

Given a pair of face image and an arbitrary class ID. We can perform KNN on the image using the complete dataset. If the majority of the k-nearest neighbors of the image have the same class ID, the response is a "Yes" otherwise its a "No". This is how we can formulate KNN classifier on the 3 datasets. We first extract the features(using PCA, LDA, or other variants), and train a KNN clas- sifier. Then, given a data point X and the correspond- ing class ID, we find the predicted ID.

B. Metrics to analyze performance

$$Accuracy = \frac{correctpredictions}{totalpredictions} \tag{2}$$

$$Precision = \frac{truepositive}{truepositive + falsepositive} \tag{3}$$

$$Recall = \frac{truepositive}{truepositive + falsenegetive} \tag{4}$$

$$F1score = \frac{2 * precision * recall}{precision + recall} \tag{5}$$

We can also use k fold cross validation error and confusion matrix to analyze performances.

C. KNN results

We used different features and combined them with different k values in the KNN classifier to see which gave the best accuracy. Csv files have been attached in my project submission folder tabulating all the accuracy's and results. Accuracy did not change much with K value but reduced a little as K value increased. Again combination of features with knn classifier (PCA + LDA for example) and RESNET showed good accuracy while KPCA and other methods like KLDA had lesser accuracy values for all 3 datasets.

Plots of Accuracy vs K value ie 3 7 69 99 for the 3 data sets are shown below for different features/combinations used.

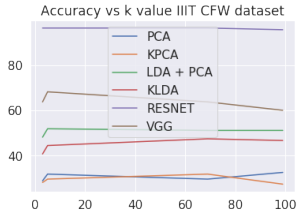

Fig. 15. IIIT CFW data

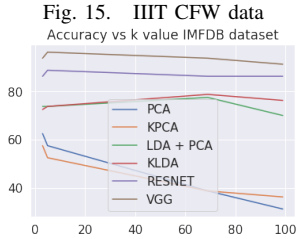

Fig. 16. IMFDB dataset

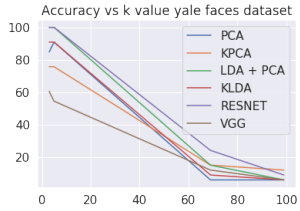

Fig. 17. Yale face dataset

V. EXTENSION

A. Problem statement

The problem statement taken by me is Gender classification. We combine the CFW and IMFDB datasets to form a new data set. Females are in one class with label 1 and males are in another class with label 0.

B. Applications

Given a photo/image my model predicts the gender with good accuracy. some applications are

1) Customization in apps

Apps can give customized results for males and females based on gender identification which is done by such gender classifiers installed in them.

2) Gender classification can be used for authentication

authentication devices that authenticate based on images can use gender classifiers for better accuracy etc.

3) Ads and marketing

Gender identification can be used to better sell and market goods on online shopping sites etc. Also personalized ads can be used based on gender classifier results to market products better to the respective target audience.

C. Model Pipeline

1) We first concatenate the CFW and the IMFDB data sets to get our gender classification data set

2) Then we split the data into the training and testing set as 80-20

3) We then did dimension reduction by the 3 following features. PCA LDA and PCA+LDA

4) Then the classifier was trained on the train set and validated on he test set

5) Accuracies and k fold cross validation score obtained for the 3 cases were found

6) All plots like pca isomap tsne etc were obtained and analyzed

D. Results

1) ACCURACY

Accuracy with PCA 0.75

Accuracy with LDA 0.6

Accuracy with PCA+LDA 0.79

PCA + LDA gave the highest accuracy as expected

2)K fold validation score

k-fold validation with PCA 0.001068

k-fold validation with LDA 0.000223

k-fold validation with PCA+LDA 0.001684

3)Correct and Wrong classification example

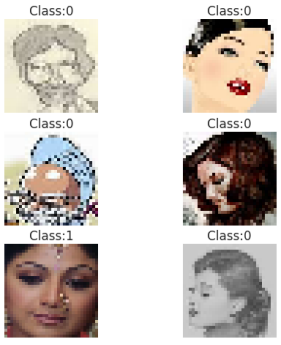

Fig. 18. Example

4) ALL 2d and 3d plots

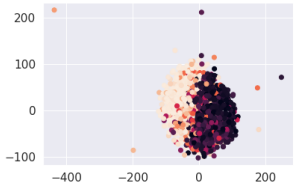

Fig. 19. 2d tSNE plot

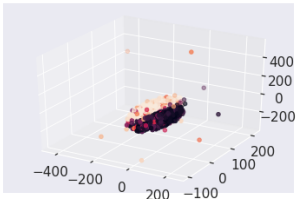

Fig. 20. 3d tSNE plot

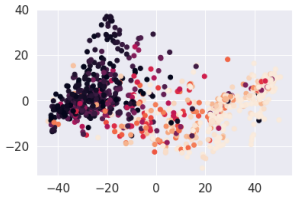

Fig. 21. 2d isomap plot

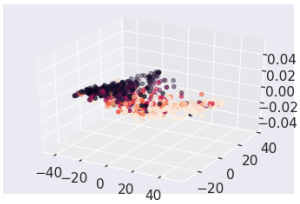

Fig. 22. 3d isomap plot

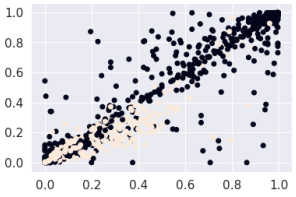

Fig. 23. 2d pca plot

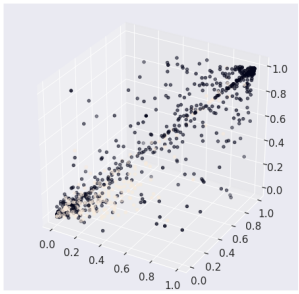

Fig. 24. 3d pca plot