PR ASSIGNMENT - 3 (New Deadline: 17/04/2022)

Design of PCA, LDA
Deliverables for this assignment:
1. Programming Assignment (MATLAB or Python)
2. Code file and output screenshots for all.

---

1. Consider the 128- dimensional feature vectors (d=128) given in the "**face feature vectors.csv**" file. (2 classes, male and female)

a) Use PCA to reduce the dimension from d to d'. (Here d=128)
b) Display the eigenvalue based on increasing order, select the d' of the corresponding eigenvector which is the appropriate dimension d' (select d' S.T first 95% of λ values of the covariance matrix are considered).
c) Use d' features to classify the test cases (any classification algorithm taught in class like Bayes classifier, minimum distance classifier, and so on)

Dataset Specifications:

Total number of samples = 800
Number of classes = 2 (labeled as "male" and "female")
Samples from "1 to 400" belongs to class "male"
Samples from "401 to 800" belongs to class "female"
Number of samples per class = 400
Number of dimensions = 128

Use the following information to design classifier:

Number of test cases (first 10 in each class) = 20
Number of training feature vectors (remaining 390 in each class) = 780
Number of reduced dimensions = d' (map 128 to d' features)

2. For the same dataset (2 classes, male and female)

a) Use LDA to reduce the dimension from d to d'. (Here d=128)
b) Choose the direction W to reduce the dimension d' (select appropriate d').
c) Use d' features to classify the test cases (any classification algorithm will do, Bayes classifier, minimum distance classifier, and so on).

3. Give the comparative study for the above results w.r.t to classification accuracy in terms of the confusion matrix.

Steps for LDA

Input - data (X) of size n x d; where, n is number of samples, d is the number of dimensions(features)

Output - X' with size n x k, where k << d

STEPS:

1. Compute the within class(Sw) and between class scatter (S_B) matrices

(a) <u>Within Class Scatter Matrix</u>

We calculate the *within class scatter matrix* using the following formula

$$S_W = \sum_{i=1}^{c} S_i$$

where ***c*** is the total number of distinct classes and

$$S_i = \sum_{\boldsymbol{x} \in D_i}^{n} (\boldsymbol{x} - \boldsymbol{m}_i)\,(\boldsymbol{x} - \boldsymbol{m}_i)^T$$

$$\boldsymbol{m}_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in D_i}^{n} \boldsymbol{x}_k$$

where ***x*** is a sample (i.e. row) and ***n*** is the total number of samples with a given class.

For every class, we create a vector with the means of each feature.

(b) <u>Between Class Scatter Matrix</u>

Next, we calculate the *between class scatter matrix* using the following formula.

$$S_B = \sum_{i=1}^{c} N_i(\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^T$$

where N_i = number of samples in class 'i' (in the assignment shared, N_1=400 are for the male class and N_2=400 for the female class)

$$\boldsymbol{m}_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in D_i}^{n} \boldsymbol{x}_k$$

$$m = \frac{1}{n} \sum_{i}^{n} x_i$$

2. Compute the eigenvectors and corresponding eigenvalues for

$$S_W^{-1} S_B$$

3. Sort the eigenvalues and select the top ***k*** eigen vectors.

   Eigen vectors with the highest eigenvalues carry the most information about the distribution of the data. Thus, we sort the eigenvalues from highest to lowest and select the first ***k*** eigenvectors.

4. Create a new matrix, W containing eigenvectors that map to the ***k*** eigenvalues calculated in step 3 (size: d x k).
5. Obtain the new reduced features (k; LDA components) by taking the matrix multiplication of the data, X with the matrix from step 4.

   Multiply matrix X with matrix W:

$$X_{[n \times d]} W_{[d \times k]} \Rightarrow X'_{[n \times k]'}$$

6. Perform classification for the obtained reduced features X' and validate results with test data chosen randomly from the given data.

NOTE: Reduced dimension : $1 <= k < (c-1)$

Why c-1 ? (Refer the following link)

https://stats.stackexchange.com/questions/447499/why-is-the-number-of-components-in-linear-discriminant-analysis-bounded-by-the-n