Experiment 1 Report-Dimensionality Reduction

R Basheer Ahammad - 22EE65R19

February 8, 2023

1 PCA for dataset1 Observations

1. Projection of x_i onto eigen vectors to obtain y_i is given by

$$y_i{}^{(j)} = a_j{}^T x_i$$

2. The R_Y matrix is calculated in the following way and it also verified :

$$R_Y = E[YY^T] = A^T R_X A = \begin{bmatrix} 2.4 * 10^3 & -2.47 * 10^{-13} & -5.6 * 10^{-14} \\ -2.29 * 10^{-13} & 5.78 & 1.3^{-13} \\ 2.11 * 10^{-13} & 1.8 * 10^{-13} & 5.3 \end{bmatrix}$$

3. From the recovered data it is observed that it is enough to use only dominant eigen value's eigen vectors in order to recover $\hat{X}$

4. The eigen vectors corresponding to largest eigen vector is nothing but First Principal Component in PCA and eigen vector corresponding to second largest eigen value is called Second Principal Component in PCA and so on.

5. From the 3D plot of dataset 1 it is observed that the maximum direction in which the data is oriented is along the largest eigen value's eigen vector.

6. Mean square error for m=1,m=2 are 5.82, 0.04 respectively which also tells us that the mean square error decreases as we the value of m is increased i.e., more eigen vectors are taken to recover $\hat{X}$ which is eventually more close to original X

2 PCA for dataset 2 Observations

1. Matrix A is nothing but all the eigen vectors arranged in column by column which are obtained by eigen value decomposition of matrix R_X

$$R_X a_i = \lambda_i a_i$$

2. The eigen faces are nothing but the eigen vectors of matrix R_X of order $4096X4096$.The top 5 eigen vectors are the dominant vectors which are plotted individually by reshaping each vector into size of $64X64$ and plot it.

3. The value of m required to recover $\hat{X}$ such that $\epsilon_r <= 5\%$ is 123 which is obtained by following steps

   - Initialize m=0
   - Calculates the cumulative sum of an array 'e' and store it in err (which iterates upto total number of eigen values)
   - Keeps a count of the number of elements summed (m).
   - The loop continues to run until the cumulative sum exceeds 99% of the total sum of all elements in the array.

$$err > 0.95 * (sum\ \ of\ \ all\ \ eigen\ \ values)$$

- The final value of 'n' is returned.The final m value will give you the required number of principal components to recover $\hat{X}$

4. The values of m obtained by executing above algorithm such that the recovery is within 1%, 5%, 10%, 20% are 260, 123, 66, 27 and their respective mean square errors are 0.026, 0.089, 0.172, 0.318 respectively.The following steps are performed:

    - At first find out the projection Y based on number of principal components 'm' taken and is given by

    $$Y_{mXN} = A^T_{Dxm} X_{DXN}$$

    - Now find out the recovered data $\hat{X}$ with the help of below equation

    $$\hat{X}_{DXN} = A_{DXm} Y_{mXN}$$

    - Finally reshape each and every column of recovered matrix $\hat{X}$ into $64X64$ matrix and draw the image

3 SVD for dataset 1 Observations

1. Using SVD we can represent data set X=UΣV.where $U = XX^H, V = X^H X$ and Σ is the diagonal matrix formed by the singular values of X.
2. There are only r non zero eigen values,so they are r eigen vectors corresponding to this eigen values by removing remaining eigen vectors dosent effect our data set representation it means we can represent our X with the help of r eigen vectors only
3. $$X_r = U_r \cdot \Sigma_r \cdot V_r^T \tag{1}$$

    where U_r and V_r are matrices composed of the first r columns of U and V, respectively, and Σ_r is a diagonal matrix composed of the first r largest singular values of X.

    So, to find the equation of the subspace spanned by the first two dominant eigenvectors, we set $r = 2$, and obtain:

    $$X_2 = U_2 \cdot \Sigma_2 \cdot V_2^T \tag{2}$$

4. Euclidean distance formula between two matrices A and B of the same size:

    $$d(A, B) = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} (A_{ij} - B_{ij})^2} \tag{3}$$

    where n and m are the dimensions of the matrices A and B, respectively, and A_{ij} and B_{ij} are the elements in the i-th row and j-th column of matrices A and B, respectively. The Euclidean distance between two matrices is a measure of the difference between them.

5. By using above formulae we calculate the euclidian distance of our 10 sample images.
6. The equation for error between the signal recovered using k components and the corresponding original signal in the complete dataset can be written as

    $$\epsilon^2 = \frac{1}{N} \sum_{i=0}^{X-1} \sum_{j=0}^{D-1} \left| x_i^{(j)} - \hat{x}_i^{(j)} \right|^2 = \sum_{j=k}^{r-1} \lambda_j \tag{4}$$

7. The pair wise distance matrix D is obtained from taking the 10 samples from original data X and calculating the pairwise distances. All the diagonal elements in D are zeros because the distance between the same samples itself is zero
8. Pairwise matrix $\hat{D}$ is obtained from recovered data set $\hat{X}$.The values in D and $\hat{D}$ are not exactly same as $\hat{X}$ is recovered only from m number principal components

9. As per above equation the Frobenius norm of the difference between matrices D and $\hat{D}$:

$$F\left|D-\hat{D}\right| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(D_{ij}-\hat{D}_{ij})^2} \tag{5}$$

where n and m are the dimensions of matrices D and $\hat{D}$, respectively, D_{ij} and $\hat{D}_{ij}$ are the elements in the i-th row and j-th column of matrices D and from equation 3 and 5 we can say they are same. The frobenius norm difference obtained is 63.90

4 SVD for dataset 2 Observations

1. As discussed in previous section here also we calculate $U = XX^H, V = X^H X$ and Σ
2. let the rank of matrix V is r,we can also mention it as non zero eigen values of matrix V.
3. Here the dimenions are $X \in R^{4096\times 400}$,$U \in R^{4096\times 4096}$,$\Sigma \in R^{4096\times 400}$,$V \in R^{400\times 400}$
4. For the given data provided we get rank of matrix V is 400, so $r = 400$
5. $X = U_r \Lambda^{1/2} V_r^H$ also has the same dimensions as U,Σ and V
6. The values of m obtained by executing above algorithm such that the recovery is within 1%, 5%, 10%, 20% are 34, 1, 1, 1
7. The Frobenius norms obtained between original image '0' and their recovered images are 11.16920185, 8.90170479, 8.90170479, 8.90170479 and frobenius norm values obtained between original image '10' and their recovered images are 11.3537302 , 9.21290493, 9.21290493, 9.21290493 respectively