MODULE 5
Feature Extraction

LESSON 8
Principal Components as Features

Keywords: Eigenvalues, Eigenvectors, Covariance Matrix, Dimensionality Reduction

Principal Component Analysis(PCA)

- Principal Component Analysis is a procedure by which the number of variables are reduced.
- The attempt is to find a smaller number of variables which are uncorrelated.
- The first principal component is the most important and accounts for as much of the variability as possible. The second principal comes next etc.
- The number of variables are reduced by projecting the data in the direction of maximum variance.
- The method involves finding the eigenvectors and the corresponding eigenvalues of the covariance matrix.
- If the eigenvectors are ordered in descending order of the eigenvalues, the first eigenvector gives the direction of the largest variance of the data.
- By excluding the directions giving very low eigenvalues, we can reduce the number of variables being considered.
- If $X = X_1, X_2, ..., X_n$ is the set of n patterns of dimension d, the sample mean of the data set is

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

The sample covariance matrix is

$C = (X - m)\ (X - m)^T$

- C is a symmetric matrix. The orthogonal basis can be calculated by finding the eigenvalues and eigenvectors.

- The eigenvectors g_i and the corresponding eigenvalues λ_i are solutions of the equation

  $$C * g_i = \lambda_i * g_i, i = 1, ..., d$$

- The eigenvector corresponding to the largest eigenvalue gives the direction of the largest variance of the data.

- By ordering the eigenvectors according to the eigenvalues, the directions along which there is maximum variance can be found.

- If E is the matrix consisting of eigenvectors as row vectors, we can transform the data X to get Y.

  $$Y = E(X - m)$$

- The original data X can be got from Y as follows :

  $$X = E^t Y + m$$

- Instead of using all d eigenvectors, the data can be represented by using the first k eigenvectors where $k < d$.

- If only the first k eigenvectors are used represented by E_K, then

  $$Y = E_K(X - m)$$

  and

  $$X' = E_K^t Y + m$$

- This reconstructed X' will not exactly match the original data X. This shows that some information is lost when only the first k eigenvectors are considered.
- By choosing the eigenvectors having largest eigenvalues, as little information as possible is lost.
- At the same time, the dimensionality of the data is reduced so as to simplify the representation.
- As an example, let us consider two patterns of class 1. The patterns are (2,2) and (3,1). Let us consider two patterns of class 2. The patterns are (5,4) and (7,4).

  The mean of the patterns will be

  $$mean = \begin{bmatrix} 4.25 \\ 2.75 \end{bmatrix}$$

  The covariance matrix is

  $$C = \begin{bmatrix} 5.25 & 2.96 \\ 2.96 & 2.25 \end{bmatrix}$$

  The eigenvalues of C are
  $\lambda_1 = 6.98$ and $\lambda_2 = 0.52$

  The first eigenvalue is very much larger than the second eigenvalue. The eigenvector corresponding to this is

  $$eigen_1 = \begin{bmatrix} 0.8633 \\ 0.5046 \end{bmatrix}$$

  The pattern (2,2) gets transformed to

  $$\begin{bmatrix} 0.8633 & 0.5045 \end{bmatrix} * \begin{bmatrix} -2.25 \\ -0.75 \end{bmatrix} = \text{-2.32}$$

Similarly, the patterns (3,1), (5,4) and (7,4) get transformed to -1.96, 1.71 and 2.57 .

When we try to get the original data from the transformed, some information gets lost. After transformation, pattern (2,2) becomes,

$$\begin{bmatrix} 0.8633 & 0.5046 \end{bmatrix} * (\text{-}2.32) + \begin{bmatrix} 4.25 \\ 2.75 \end{bmatrix} =$$

$$\begin{bmatrix} -2.00 & -1.17 \end{bmatrix} + \begin{bmatrix} 4.25 \\ 2.75 \end{bmatrix} =$$

$$\begin{bmatrix} 2.25 \\ 1.58 \end{bmatrix}$$

showing that there is some loss in information.

Assignment

1. Consider a two-class problem. What is the rank of the matrix B used in Fisher's discriminant?

2. Consider a two-class problem, where the training patterns are:
   $Class1: \ (2,2)^t, \ (4,3)^t, \ and \ (5,1)^t$
   $Class2: \ (1,3)^t, \ (5,5)^t, \ and \ (3,6)^t$
   Obtain the within class scatter matrix.

3. Obtain the direction V for the data given in problem 2.

4. Obtain the direction of the Fisher's discriminant for the data given in problem 2. Consider $X = (1,6)^t$. How do you classify it using the discriminant.

5. Consider the two-dimensional patterns: $(1,1)^t, \ (1,2)^t, \ (4,4)^t, \ (5,4)^t$. Obtain the sample covariance matrix. Obtain its eigenvalues and eigenvectors.

6. Use the first principal component obtained in problem 5 to transform the pattern $(1,1)^t$. What is the vector obtained if we try to reproduce the original data?

7. Solve problem 6 using both the principal components obtained in problem 5.

References :

V. S. Devi, M. N. Murty(2011) *Pattern Recognition: An Introduction*, Universities Press, Hyderabad.
R.O. Duda, P.E. Hart, D.G. Stork(2000) *Pattern Classification*, John Wiley and Sons., Singapore
Z. Nenadic(2007) Information discriminant analysis: Feature extraction with an information-theoretic objective, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8), pages 1394-1407.