Machine Learning Assignment

Reshma Sajeevan

January 11, 2023

1 Introduction

In machine learning, feature extraction projects an initial set of features from a high dimensional space into a reduced set of features from a low dimensional space while still illustrating the data with adequate precision. Feature selection algorithms could be linear or non-linear. Non-linear methods assume that the data of interest remain on a fixed nonlinear diverse within the higher dimensional space and Linear methods perform a linear mapping of the data to a lower dimensional space. Most important linear algebra concepts are the:

1. Singular Vector Decomposition (SVD)
2. Linear Discriminant Analysis (LDA)

2 Singular Vector Decomposition (SVD)

The Singular Value Decomposition (SVD) method is a well-known technique for decomposing a matrix into a large number of component matrices. This method is valuable since it reveals many of the interesting and helpful characteristics of the initial matrix. We can use SVD to discover the optimal lower-rank approximation to the matrix, determine the rank of the matrix, or test a linear system's sensitivity to numerical error.
SVD is similar to Principal Component Analysis (PCA) but PCA assumes that input is a square matrix
General formula of SVD is:

$$S = U\Sigma V^T \tag{1}$$

where,

- S-is original matrix we want to decompose
- U-is left singular matrix (columns are left singular vectors).
- Σ -is a diagonal matrix containing singular (Eigen)values
- V-is right singular matrix (columns are right singular vectors).

2.1 Steps Involved In SVD explained with an example

Example: 1) Compute SVD for Given square matrix $\begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix}$

STEP 1 : Compute its transpose $\mathrm{A}^T A$ *and* A^T

$$sinceA^T = \begin{bmatrix} 4 & 3 \\ 0 & -5 \end{bmatrix}$$
$$Then\ A^T A = \begin{bmatrix} 4 & 3 \\ 0 & -5 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix}$$
$$i.e, A^T A = \begin{bmatrix} 25 & -15 \\ -15 & 25 \end{bmatrix}$$

$STEP\ 2 : Determine\ the\ eigenvalues\ of\ A^{-T} A\ and\ sort\ these\ in\ descending\ order\ , in\ the\ absolute\ sense.$ $Square\ roots\ these\ to\ obtain\ the\ singular\ values\ of\ A.$

$$|(A^T A) - \lambda| = \begin{bmatrix} 25 - \lambda & -15 \\ -15 & 25 - \lambda \end{bmatrix}$$
$$i.e, A^{-1} A - \lambda = 0, then\ we\ can\ get\ characteristic\ equation\ as$$

$$(\lambda)^2 - 50\lambda + 400$$

$from\ this\ equation\ we\ can\ find\ Eigen\ values\ which\ is\ \lambda_1 = 10\ and\ \lambda_2 = 40$

$now\ Singular\ Values\ \rightarrow s_1 = \sqrt{10}\ = 3.162\ and$

$s_2 = \sqrt{40}\ = 6.324$

$STEP3 :\ Construct\ diagonal\ matrix\ S\ by\ placing\ singular\ values\ in\ descending\ order\ along\ its\ diagonal.$ $Compute\ its\ inverse\ S^{-1}$

$$S = \begin{bmatrix} 6.324 & 0 \\ 0 & 3.162 \end{bmatrix} and\ \mathrm{S}^{-1} = \begin{bmatrix} 0.1581 & 0 \\ 0 & 0.3162 \end{bmatrix}$$

STEP 4 : Use the ordered eigenvalues from step 2 and compute the eigenvectors of $\mathrm{A}^T A. Place\ these\ eigenvectors\ along\ the\ columns\ of\ V\ and\ compute\ its\ transpose,\ V^T.$

$For\ \lambda_1 = 40$

$$(A^T A - \lambda_1 * I) = \begin{bmatrix} 25 - \lambda & -15 \\ -15 & 25 - \lambda \end{bmatrix}$$
$$i.e,\ (A^T A - \lambda_1 * I) = \begin{bmatrix} -15 & -15 \\ -15 & -15 \end{bmatrix}$$
$$Then(A^T A - \lambda * I) * X_1 = \begin{bmatrix} -15 & -15 \\ -15 & -15 \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$-15x_1 + (-15x_2) = 0$$

$$-15x_1 + (-15x_2) = 0$$

$Solving\ for\ the\ x_2\ either\ of\ the\ equation:\ x_2 = (-x_1)$

$$x_1 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_1 \end{bmatrix}$$

$Dividing\ by\ its\ length$

$$L = \sqrt{x_1^2 + x_2^2}\ = x_1 * \sqrt{2}$$

$$x_1 = \begin{bmatrix} x_1/L \\ -x_1/L \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 0.7071 \\ 0.7071 \end{bmatrix}$$

$For \lambda_2 = 10$

$$(A^T A - \lambda_2 * I) = \begin{bmatrix} 25-\lambda & -15 \\ -15 & 25-\lambda \end{bmatrix}$$

$$i.e,\ (A^T A - \lambda_2 * I) = \begin{bmatrix} 15 & -15 \\ -15 & 15 \end{bmatrix}$$

$$Then(A^T A - \lambda_2 * I) * X_2 = \begin{bmatrix} 15 & -15 \\ -15 & 15 \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$15x_1 + -15x_2 = 0$$

$$-15x_1 + 15x_2 = 0$$

$solving\ for\ x_2\ either\ of\ the\ the\ equation:\ x_2 = x_1$

$$x_1 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_1 \end{bmatrix}$$

$Dividing\ by\ its\ length$

$$L = \sqrt{x_1^2 + x_2^2}\ = x_1 * \sqrt{2}$$

$$x_1 = \begin{bmatrix} x_1/L \\ x_1/L \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 0.7071 \\ 0.7071 \end{bmatrix}$$

$$V = \begin{bmatrix} x_1 & x_2 \end{bmatrix} = \begin{bmatrix} 0.7071 & 0.7071 \\ 0.7071 & 0.7071 \end{bmatrix}$$

$$Eigen\ Vector = V^T = \begin{bmatrix} 0.7071 & -0.7071 \\ 0.7071 & 0.7071 \end{bmatrix}$$

$STEP5:\ Compute\, U as\, U = AVS^{-1}. To\, complete\, the\, proof,\ compute\, the\, full\, SVD\, using\, A = USV^T$

$U = AVS^{-1} = \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix} \begin{bmatrix} 0.7071 & 0.7071 \\ 0.7071 & 0.7071 \end{bmatrix} \begin{bmatrix} 0.1581 & 0 \\ 0 & 0.3162 \end{bmatrix}$

$U = AVS^{-1} = \begin{bmatrix} 0.4472 & 0.8944 \\ 0.8944 & -0.4472 \end{bmatrix}$

$A = USV^T = \begin{bmatrix} 0.4472 & 0.8944 \\ 0.8944 & -0.4472 \end{bmatrix} \begin{bmatrix} 6.324 & 0 \\ 0 & 3.162 \end{bmatrix} \begin{bmatrix} 0.7071 & -0.7071 \\ 0.7071 & 0.7071 \end{bmatrix}$

$A = USV^T = \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix}$

The orthogonal nature of the V and U matrices is evident by inspecting their eigenvectors.
This can be demonstrated by computing dot products between column vectors.
All dot products are equal to zero.

3 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is one of the commonly used dimensional reduction techniques in machine learning to solve more than two-class classification problems. It is also known as Normal Discriminant Analysis (NDA) or Discriminant Function Analysis (DFA).
LDA is a dimensionality reduction technique, used as a preprocessing step for pattern classification and machine learning applications. LDA is similar to PCA but LDA in addition finds the axes that maximizes the separation between multiple choices.

3.1 Steps Involved In LDA explained along with an example

Example: Compute the Linear Discriminant projection for the following two dimensional data-set.
Samples for class 1 : X1=(x1,x2)=(4,2),(2,4),(2,3),(3,6),(4,4) and Sample for class 2 : X2=(x1,x2)=(9,10),(6,8),(9,5),(8,7),(10,8)
STEP 1: Computing the d-dimensional mean vectors

$\mu_1 = 1/N_1 \sum_{x \epsilon w1} x$ =1/5 *[$\begin{bmatrix} 4 \\ 2 \end{bmatrix} + \begin{bmatrix} 2 \\ 4 \end{bmatrix} + \begin{bmatrix} 2 \\ 3 \end{bmatrix} + \begin{bmatrix} 3 \\ 6 \end{bmatrix} + \begin{bmatrix} 4 \\ 4 \end{bmatrix}$] = $\begin{bmatrix} 3 \\ 3.8 \end{bmatrix}$

$\mu_2 = 1/N_2 \sum_{x \epsilon w2} x$ =1/5 *[$\begin{bmatrix} 9 \\ 10 \end{bmatrix} + \begin{bmatrix} 6 \\ 8 \end{bmatrix} + \begin{bmatrix} 9 \\ 5 \end{bmatrix} + \begin{bmatrix} 8 \\ 7 \end{bmatrix} + \begin{bmatrix} 10 \\ 8 \end{bmatrix}$] = $\begin{bmatrix} 8.4 \\ 7.6 \end{bmatrix}$

class means are: $\boxed{\bar{\mu}_1 = \begin{bmatrix} 3 \\ 3.8 \end{bmatrix} \; and \; \bar{\mu}_2 = \begin{bmatrix} 8.4 \\ 7.6 \end{bmatrix}}$

STEP 2 : Computing the Scatter Matrices

within-class scatter matrix S_W $is\ computed\ by\ the\ following\ equation\ S_W = \sum_{i=1}^{c} S_i$

where $S_i = \sum_{x \epsilon D_i}^{n} (x - \mu_i) - (x - \mu_i)^T$ Covariance matrix of the first class:

$$S_i = \sum_{x \epsilon w1}^{n} (x - \mu_1) - (x - \mu_1)^T = \left[\begin{bmatrix}4\\2\end{bmatrix} - \begin{bmatrix}3\\3.8\end{bmatrix}\right]^2 + \left[\begin{bmatrix}2\\4\end{bmatrix} - \begin{bmatrix}3\\3.8\end{bmatrix}\right]^2 + \left[\begin{bmatrix}2\\3\end{bmatrix} - \begin{bmatrix}3\\3.8\end{bmatrix}\right]^2 + \left[\begin{bmatrix}3\\6\end{bmatrix} - \begin{bmatrix}3\\3.8\end{bmatrix}\right]^2 + \left[\begin{bmatrix}4\\4\end{bmatrix} - \begin{bmatrix}3\\3.8\end{bmatrix}\right]^2 = \begin{bmatrix}1 & -0.25\\-0.25 & 2.2\end{bmatrix}$$

$$\boxed{\bar{c}ovariance\ matrix\ of\ the\ first\ class\ S_1 \ = \ cov(X1)}$$

$$S_2 = \sum_{x \epsilon w2}^{n} (x - \mu_2) - (x - \mu_2)^T = \left[\begin{bmatrix}9\\10\end{bmatrix} - \begin{bmatrix}8.4\\7.6\end{bmatrix}\right]^2 + \left[\begin{bmatrix}6\\8\end{bmatrix} - \begin{bmatrix}8.4\\7.6\end{bmatrix}\right]^2 + \left[\begin{bmatrix}9\\5\end{bmatrix} - \begin{bmatrix}8.4\\7.6\end{bmatrix}\right]^2 + \left[\begin{bmatrix}8\\7\end{bmatrix} - \begin{bmatrix}8.4\\7.6\end{bmatrix}\right]^2 + \left[\begin{bmatrix}10\\8\end{bmatrix} - \begin{bmatrix}8.4\\7.6\end{bmatrix}\right]^2 = \begin{bmatrix}2.3 & -0.05\\-0.05 & 3.3\end{bmatrix}$$

$$\boxed{\bar{c}ovariance\ matrix\ of\ the\ first\ class\ S_2 \ = \ cov(X2)}$$

$$Within-class\ scatter\ matrix: S_W = \sum_{i=1}^{c} S_i = S_1 + S_2 = \begin{bmatrix}1 & -0.25\\-0.25 & 2.2\end{bmatrix} + \begin{bmatrix}2.3 & -0.05\\-0.05 & 3.3\end{bmatrix} = \begin{bmatrix}3.3 & -0.3\\-0.3 & 5.5\end{bmatrix}$$

$$\boxed{\bar{W}ithin-class\ scatter\ matrix: S_W \ = \begin{bmatrix}3.3 & -0.3\\-0.3 & 5.5\end{bmatrix}}$$

$Now\ compute\ Between-class\ scatter\ matrix: S_B$

$$S_B = (\mu_1 - \mu_2) * (\mu_1 - \mu_2)^T = \left[\begin{bmatrix}3\\3.8\end{bmatrix} - \begin{bmatrix}8.4\\7.6\end{bmatrix}\right] * \left[\begin{bmatrix}3\\3.8\end{bmatrix} - \begin{bmatrix}8.4\\7.6\end{bmatrix}\right]^T$$

$$= \begin{bmatrix}-5.4\\-3.8\end{bmatrix} * \begin{bmatrix}-5.4 & -3.8\end{bmatrix} = \begin{bmatrix}29.16 & 20.52\\20.52 & 14.44\end{bmatrix}$$

$Solving\ the\ generalized\ eigenvalue\ problem\ for\ the\ matrix\ S_W^{-1} * S_B$

$The\ LDA\ projection\ is\ then\ obtained\ as\ the\ solution\ of\ the\ generalized\ eigen\ value\ problem$

$i.e,\ S_W^{-1} * S_B = \lambda_W$

$|S_W^{-1} * S_B - \lambda_W| = 0$

$$= \ \left| \begin{bmatrix}3.3 & -0.3\\-0.3 & 5.5\end{bmatrix}^{-1} * \begin{bmatrix}29.16 & 20.52\\20.52 & 14.44\end{bmatrix} - \lambda * \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} \right|$$

$$= \left| \begin{bmatrix} 0.3045 & 0.0166 \\ 0.0166 & 0.1827 \end{bmatrix} * \begin{bmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{bmatrix} - \lambda * \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

$$= \left| \begin{bmatrix} 9.2213 - \lambda & 6.483 \\ 4.2339 & 2.9794 - \lambda \end{bmatrix} \right| = 0$$

$= (9.2213 - \lambda)(2.9794 - \lambda) - (6.483)(4.2339) = 0$
$= (\lambda)^2) - 12.2007 * \lambda = 0$
$i.e, \lambda_1 = 0 \ and \ \lambda_2 = 12.0027$

$STEP\ 4 : Selecting\ linear\ discriminant\ for\ the\ new\ feature\ subspace$

$Hence, \lambda_1 = 0$

$$\begin{bmatrix} 9.2213 & 6.483 \\ 4.2339 & 2.9794 \end{bmatrix} * W_1 = \lambda_1 * \begin{bmatrix} w1 \\ w2 \end{bmatrix} = \begin{bmatrix} 9.2213 & 6.483 \\ 4.2339 & 2.9794 \end{bmatrix} * W_1 = 0 * \begin{bmatrix} w1 \\ w2 \end{bmatrix}$$

$$and \begin{bmatrix} 9.2213 & 6.483 \\ 4.2339 & 2.9794 \end{bmatrix} * W_2 = \lambda_2 * \begin{bmatrix} w1 \\ w2 \end{bmatrix}$$

$$= \begin{bmatrix} 9.2213 & 6.483 \\ 4.2339 & 2.9794 \end{bmatrix} * W_1 = 12.0027 * \begin{bmatrix} w1 \\ w2 \end{bmatrix}$$

$$Thus W_1 = \begin{bmatrix} -0.5755 \\ 0.8178 \end{bmatrix}$$

$$and \ \boxed{\bar{W}_2 = \begin{bmatrix} 0.9088 \\ 0.4173 \end{bmatrix} = W^*}$$

$The\ optimal\ projection\ is\ the\ one\ that\ given\ maximum\ \lambda\ =\ J(w)$

$STEP\ 5 :\ Transforming\ the\ samples\ onto\ the\ new\ subspace$