Feature Engineering in Dimensionality Reduction

Sirajulhugh V M

January 11, 2023

1 Introduction

Dimensionality reduction is a way used to reduce the number of independent variables or features.Feature extraction is a technique commonly used for selecting the subset from the set of features and then generating(by means of some transformation) a new set of features from previously selected set of features, statistical formulas are used for this purpose. This is done when there are a lot of features to execute, the performance of the code becomes poor, especially for techniques like SVM and Neural networks in these cases it takes a long time to train the model.

2 Linear discriminant analysis (LDA)

LDA is dimensionality reduction technique used as a pre-processing step for pattern classification and machine learning application,LDA is similar to PCA,in LDA it maximise the seperation between multiple classes it also reduce number of features by doing so,the key difference between PCA and LDA is that the separation between each to which each feature belongs to is also maintained

In other words we can describe it as by its role,ie we need to project the feature space of N dimensional onto a smaller subspace K (where K is less than or equal to n-1)while maintaining the class discriminating information of the features as well.

2.1 Steps to LDA

For doing LDA for reducing the dimensionality we need to follow few steps.The first step is to calculate the difference between classes (i.e. the distance between the means of different classes), which is called the between-class variance or between-class matrix. The second step is to calculate the distance between the mean and the samples of each class, which is called the within-class variance or within-class matrix. The third step is to construct the lower dimensional space which maximizes the between-class variance and minimizes the within-class variance.

2.1.1 Between-class variance (S_B)

The between-class variance of the i th class (S_{bi}) represents the distance between the mean of the ith class (υ_i) and the total mean (υ). LDA technique searches for a lower-dimensional space, which is used to maximize the between-class variance, or simply maximize the separation distance between classes.To explain how the between-class variance or the between-class matrix (S_B) can be calculated, the following assumptions are made. Given the original data matrix X= x1,x2,..., x_n , where x i represents the ith sample, pattern, or observation and N is the total number of samples. Each sample is represented by M features ($x_i \epsilon R^m$). In other words, each sample is represented as a point in a M-dimensional space.

To calculate the between-class variance (S_B), the separation distance between different classes which is denoted by (m_im) will be calculated as follows:

$$(\mathrm{m}_i - m)^2 = (W^T \upsilon_i - W^T \upsilon)^2$$
$$= \mathrm{W}^T(\upsilon_i - \upsilon)(\upsilon_i - \upsilon)^T f \to 1$$

where m_i represents the projection of the mean of the ith class and it is calculated as follows, $m_i = W^T \upsilon_i$, where m is the projection of the total mean of all classes and it is calculated as follows,

$m = W^T v$, W represents the transformation matrix of LDA,$^T v_i(1 \times$M) represents the mean of the ith class and it is computed as in Equation (2), and $v(1 \times$ M) is the total mean of all classes and it can be computed as in Equation(3).

$$v = \frac{1}{n_j} \sum_{x_i \epsilon \omega_j} x_i \rightarrow 2$$
$$v = \frac{1}{N} \sum_{i=1}^{N} x_i = \sum_{i=1}^{c} \frac{n_i}{N} v_i \rightarrow 3$$

2.1.2 Within class variance (S_W)

The within class variance of the ith class (S_{Wi}) represents the difference between the mean and the samples of that class. LDA technique searches for a lower dimensional space, which is used to minimize the difference between the projected mean (m_i) and the projected samples of each class (W^T xi), or simply minimizes the within class variance . The within class variance of each class (S_{Wj}) is calculated as in Equation (4).

$$\sum_{x_i \epsilon, \omega, j=1...c} \left(W^T x_i - m_j\right)^2$$
$$\sum_{x_i \epsilon, \omega, j=1...c} \left(W^T x_{ij} - W^T v_j\right)^2$$
$$\sum_{x_i \epsilon, \omega, j=1...c} W^T \left(x_{ij} - v_j\right)^2 W$$
$$\sum_{x_i \epsilon, \omega, j=1...c} W^T \left(x_{ij} - v_j\right) \left(x_{ij} - v_j\right)^T W$$
$$\sum_{x_i \epsilon, \omega, j=1...c} W^T S_{wj} W \rightarrow 4$$

2.1.3 Constructing the lower dimensional space)

After calculating the between-class variance (S_B) and within-class variance (S_W), the transformation matrix (W) of the LDA technique can be calculated as in Equation (5), which is called Fisher's criterion. This formula can be reformulated as in Equation (6)

$$\arg\max_w \frac{W^T S_b W}{W^T S_w W} \rightarrow 5$$
$$\mathrm{s}_w W = \lambda S_B W \rightarrow 6$$

where λ represents the eigenvalues of the transformation matrix (W). The solution of this problem can be obtained by calculating the eigenvalues ($\lambda = \{\lambda_1, \lambda_2, ..., \lambda_M\}$) and eigenvectors (V = $\{v_1, v_2, ..., v_M\}$) of W = $S^{-1} S_B$, if S_W is non-singular . The eigenvalues are scalar values, while the eigenvectors are non-zero vectors, which satisfies the Equation (6) and provides us with the information about the LDA space. The eigenvectors represent the directions of the new space, and the corresponding eigenvalues represent the scaling factor

2.2 LDA . . . Two Classes - Example

lets Compute the Linear Discriminant projection for the following two dimensional data set. – Samples for class 1 : X1=(x1,x2)=$\{(4,2),(2,4),(2,3),(3,6),(4,4)\}$
– Sample for class 2 : X2=(x1,x2)=$\{(9,10),(6,8),(9,5),(8,7),(10,8)\}$
The class mean are:

$$\mu_1 = \frac{1}{N} \sum_{x \epsilon \omega_2} x = \frac{1}{5} \left[\begin{pmatrix} 4 \\ 2 \end{pmatrix} + \begin{pmatrix} 2 \\ 4 \end{pmatrix} + \begin{pmatrix} 2 \\ 3 \end{pmatrix} + \begin{pmatrix} 3 \\ 6 \end{pmatrix} + \begin{pmatrix} 4 \\ 4 \end{pmatrix} \right] = \begin{pmatrix} 3 \\ 3.8 \end{pmatrix}$$
$$\mu_2 = \frac{1}{N} \sum_{x \epsilon \omega_2} x = \frac{1}{5} \left[\begin{pmatrix} 9 \\ 10 \end{pmatrix} + \begin{pmatrix} 6 \\ 8 \end{pmatrix} + \begin{pmatrix} 9 \\ 5 \end{pmatrix} + \begin{pmatrix} 8 \\ 7 \end{pmatrix} + \begin{pmatrix} 10 \\ 8 \end{pmatrix} \right] = \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}$$

Covariance matrix of the first class:

$$\mathrm{S}_1 = \sum_{x \epsilon \omega_1} \left(x - \mu_1\right) \left(x - \mu_1\right)^T =$$

$$\left[\begin{pmatrix} 4 \\ 2 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix} \right]^2 + \left[\begin{pmatrix} 2 \\ 4 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix} \right]^2 + \left[\begin{pmatrix} 2 \\ 3 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix} \right]^2 + \left[\begin{pmatrix} 3 \\ 6 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix} \right]^2 + \left[\begin{pmatrix} 4 \\ 4 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.38 \end{pmatrix} \right]^2$$

$$= \begin{pmatrix} 1 & -.25 \\ -.25 & 2.2 \end{pmatrix}$$

Covariance matrix of the second class:

$$S_2 = \sum_{x \epsilon \omega_2} (x - \mu_2)(x - \mu_2)^T =$$

$$\left[\begin{pmatrix} 9 \\ 10 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 6 \\ 8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 9 \\ 5 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 8 \\ 7 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2 + \left[\begin{pmatrix} 10 \\ 8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}\right]^2$$

$$= \begin{pmatrix} 2.3 & -.05 \\ -.05 & 3.3 \end{pmatrix}$$

Within-class scatter matrix:

$$S_w = S_1 + S_2 = \begin{pmatrix} 1 & -.25 \\ -.25 & 2.2 \end{pmatrix} + \begin{pmatrix} 2.3 & -.05 \\ -.05 & 3.3 \end{pmatrix}$$

$$= \begin{pmatrix} 3.3 & -.3 \\ -.3 & 5.5 \end{pmatrix}$$

Between-class scatter matrix:

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

$$= \left[(\begin{pmatrix} 3 \\ 3.8 \end{pmatrix}) - (\begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix})\right]\left[(\begin{pmatrix} 3 \\ 3.8 \end{pmatrix}) - (\begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix})\right]^T$$

$$= \begin{pmatrix} -5.4 \\ -3.8 \end{pmatrix}\begin{pmatrix} -5.4 & -3.8 \end{pmatrix}$$

$$= \begin{pmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{pmatrix}$$

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

$$S_W^{-1} S_B w = \lambda w$$

$$\Rightarrow \left|S_W^{-1} S_B - \lambda I\right| = 0$$

$$\Rightarrow \left| \begin{pmatrix} 3.3 & -.3 \\ -.3 & 5.5 \end{pmatrix}^{-1} \begin{pmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = 0$$

$$\Rightarrow \left| \begin{pmatrix} .30 & -.01 \\ -.01 & .18 \end{pmatrix} \begin{pmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = 0$$

$$\Rightarrow \left| \begin{pmatrix} 9.22 - \lambda & 6.48 \\ 4.23 & 2.97 - \lambda \end{pmatrix} \right| = 0$$

$$\Rightarrow (9.22 - \lambda)(2.97 - \lambda) - 6.48 * 4.23 = 0$$

$$\Rightarrow \lambda^2 - 12.2\lambda = 0 \Rightarrow \lambda(\lambda - 12.2) = 0$$

$$\Rightarrow \lambda_1 = 0, \lambda_2 = 12.2$$

Hence,

$$\lambda_1 \Rightarrow \begin{pmatrix} 9.22 & 6.48 \\ 4.23 & 2.97 \end{pmatrix} w_1 = 0 \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

and,

$$\lambda_2 \Rightarrow \begin{pmatrix} 9.22 & 6.48 \\ 4.23 & 2.97 \end{pmatrix} w_1 = 12.2 \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

Thus,;

$$w_1 = \begin{pmatrix} -.57 \\ .81 \end{pmatrix} \quad and \quad w_2 = \begin{pmatrix} .90 \\ .41 \end{pmatrix}$$

$\Rightarrow$The optimal projection is the one that gives maximum λ=J(w)

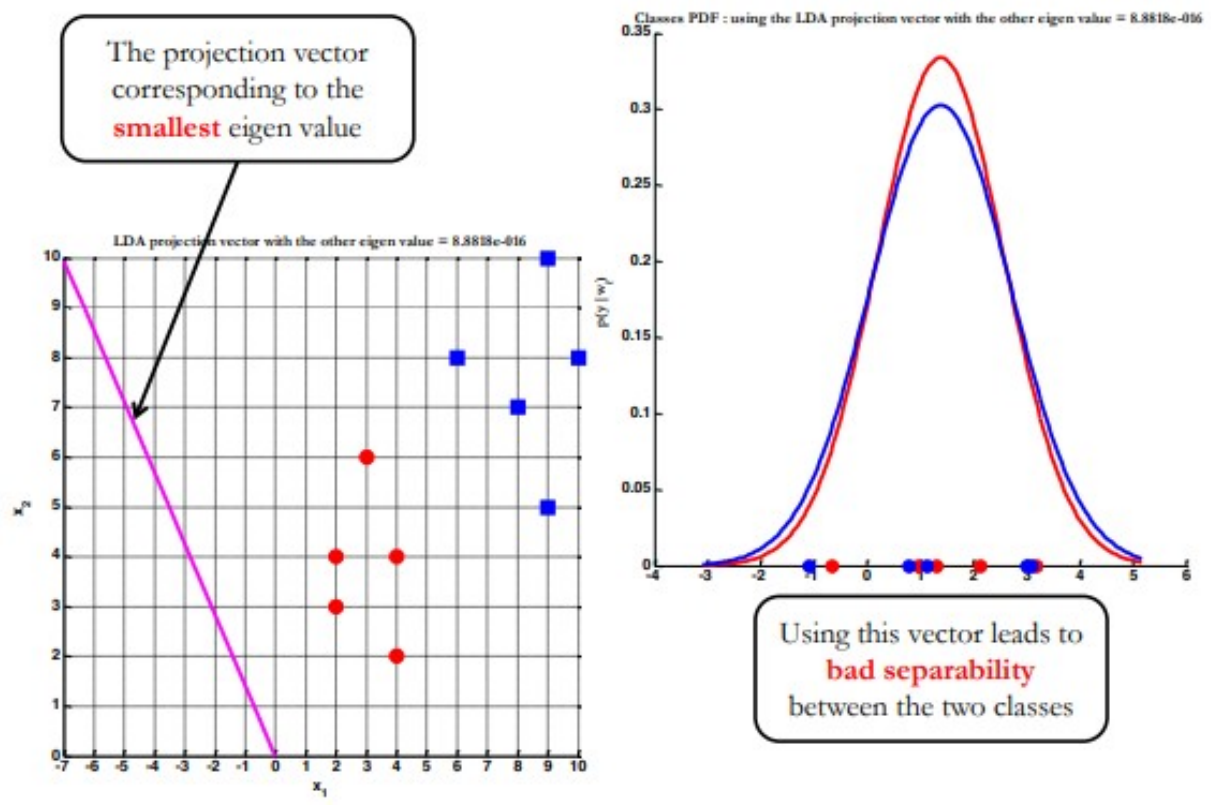

Figure 1: projection of smallerλ

$$\text{ie;} \quad w_2 = \begin{pmatrix} .90 \\ .41 \end{pmatrix}$$

We can also find the optimal projection using another method ie;

$$w^* = S_w^{-1}(\mu_1 - \mu_2) = \begin{pmatrix} 3.3 & -.3 \\ -.3 & 5.5 \end{pmatrix}^{-1} \left[\begin{pmatrix} 3 \\ 3.8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} \right]$$

$$= \begin{pmatrix} .30 & .01 \\ .01 & .18 \end{pmatrix} \begin{pmatrix} -5.4 \\ -3.8 \end{pmatrix}$$

$$= \begin{pmatrix} .90 \\ .41 \end{pmatrix}$$

LDA projectin of 2 euigen values are plotted in figure 1 and figure 2 respectively

3 Singular Value Decomposition(SVD)

Suppose A is an m $\times$ n matrix with rank r. The matrix AA^T will be m $\times$ m and have rank r. The matrix A^TA will be n $\times$ n and also have rank r. Both matrices A^TA and AA^T will be positive semi definite, and will there fore have r (possible repeated) positive eigenvalues, and r linearly independent corresponding eigenvectors. As the matrices are symmetric, these eigenvectors will be orthogonal, and we can choose them to be ortho-normal.

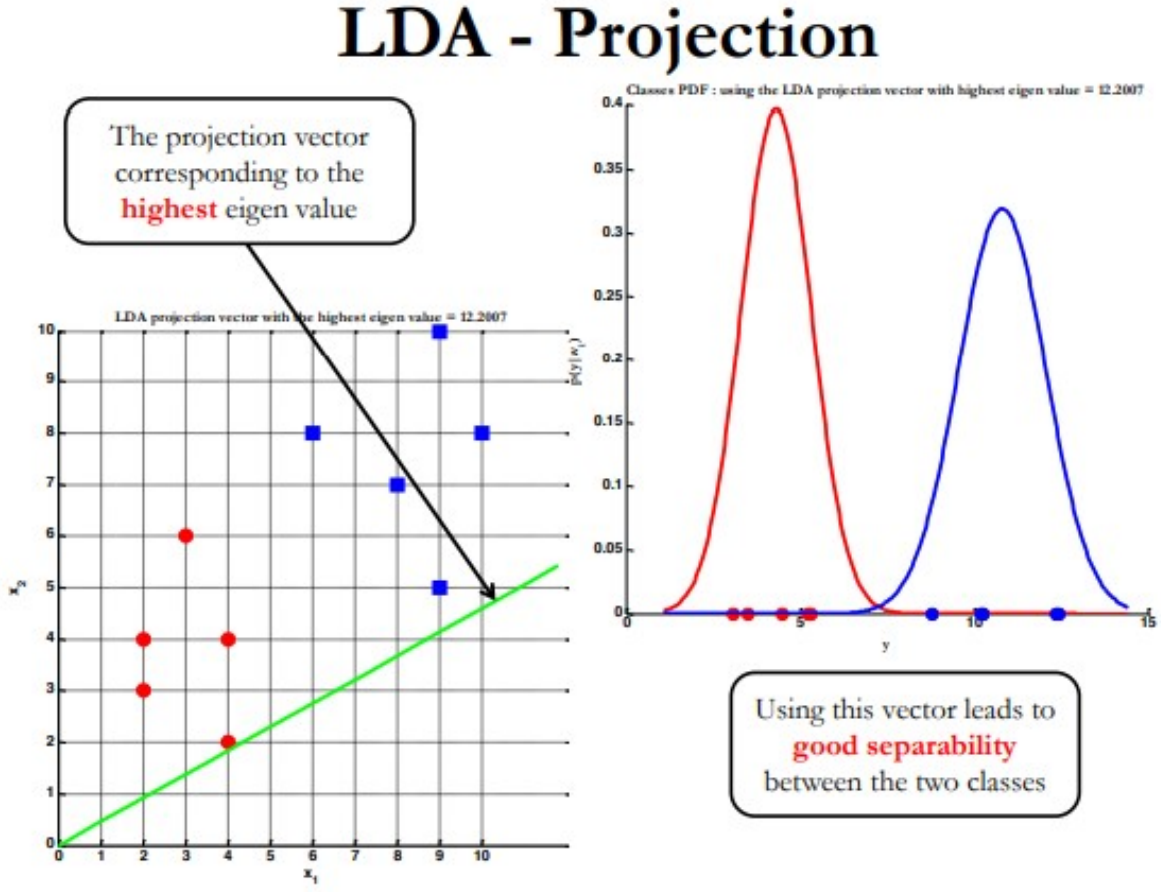

Figure 2: projection of higherλ

We call the eigenvectors of A^TA corresponding to its non-zero eigen-values v_1, . . . , v_r. These vectors will be in the row space of A. We call the eigenvectors of AA^T corresponding to its non-zero eigenvalues u_1, . . . , u_r.These vectors will be in the column space of A.

Now, these vectors have a remarkable relation. Namely,

$$\mathrm{A}v_1 = \sigma u_1, Av_2 = \sigma_2 u_2, ..., Av_r = \sigma_r u_r$$

where$\sigma_1, ...\sigma_r$ are positive numbers called the singualr values of the matrix A.
This relation lets us write

$$\mathrm{A}\big(v_1....v_r\big)=\big(u_1....u_r\big)\begin{pmatrix}\sigma_1 & & & \\ & . & & \\ & & . & \\ & & & . \\ & & & & \sigma_r\end{pmatrix}$$

This gives us a decomposition AV=U$\sum$

Noting that the columns of V are ortho normal we can right multiply both sides of this equality by V^T to get A = $U \sum V^T$. This is the singular value decomposition of A.

If we want to we can make V and U square. We just append orthonormal vectors v_{r+1}, . . . , v_n in the nullspace of A to V,and orthonormal vectors u_{r+1}, . . . , u_m in the left-nullspace of A to M. We'll still get AV = U$\sum$and A = U$\sum V^T$.

This singular value decomposition has a particularly nice representation if we carry through the multiplication of the matrices:

$$\mathrm{A}=\bigcup \sum V^T = u_1\sigma_1 v_1 + ... + u_r\sigma_r v_r^T$$

Each of these "pieces" has rank 1. If we order the singular values

$$\sigma_1 \geq \sigma_2.....\sigma_r$$

then the singular value decomposition gives A in r rank 1 pieces in order of importance.

3.1 SVD example

Let's find the singular value decompositions by taking a rank 2 unsymmetric matrix

$$\mathrm{A}=\begin{pmatrix}2 & 2\\ -1 & 1\end{pmatrix}$$

A is not symmetric, and there will be no orthogonal matrix Q that will make $Q^{-1}AQ$ diagonal. We need two different orthogonal matrices U and V .

We find these matrices with the singular value decomposition. So, we want to compute A^T A and its eigenvectors.

$$\mathrm{A}^T A = \begin{pmatrix}5 & 3\\ 3 & 5\end{pmatrix}$$

and so,

$$\begin{vmatrix}5-\lambda & 3\\ 3 & 5-\lambda\end{vmatrix}=(5-\lambda)^2 - 9 = \lambda^2 - 10\lambda + 16 = (\lambda - 8)(\lambda - 2)$$

So, A^T A has eigenvalues 8 and 2. The corresponding eigenvectors will be

$$\mathrm{v}_1 = \begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix} \qquad v_2 = \begin{pmatrix}\frac{-1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix}$$

Now, to find the vectors u1 and u2 we multiply v1 and v2 by A:

$$\mathrm{Av}_1 = \begin{pmatrix} 2 & 2 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} 2\sqrt{2} \\ 0 \end{pmatrix}$$

$$\mathrm{Av}_2 = \begin{pmatrix} 2 & 2 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} 0 \\ 2\sqrt{2} \end{pmatrix}$$

So, the unit vectors u1 and u2 will be:

$$\mathrm{u}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad u_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

The singular values will be $2\sqrt{2} = \sqrt{8}$ *and* $\sqrt{2}$. This gives us the singular value decomposition:

$$\begin{pmatrix} 2 & 2 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2\sqrt{2} & 0 \\ 0 & 2\sqrt{2} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$