FISHER LINEAR DISCRIMINANT

Prof. Dan A. Simovici

UMB
Fisher linear discriminant (FLD) seeks to find projections on a line such that the projections of examples from different samples are well separated.

Poor separation of projections

Good separation of projections
Recall that for a sequence $U = (u_1, \ldots, u_m)$ of vectors in $\mathbb{R}^n$ we denoted by $\bar{U} \in \mathbb{R}^n$ the mean

$$\bar{U} = \frac{1}{m} \sum_{i=1}^{m} u_i = \frac{1}{m} (u_1 \cdots u_m) 1_m,$$

where $X = \begin{pmatrix} u_1' \\ \vdots \\ u_m' \end{pmatrix} \in \mathbb{R}^{m \times n}$.

Equivalently, we have the row vector $\bar{U}' = \frac{1}{m} 1_m' X$.

Let $d$ be a unit vector; this defines the line used to project the vectors on.
Suppose that \( U \) is partitioned into two sequences, \( U = (U_1, U_2) \) that contain \( m_1 \) and \( m_2 \) vectors, respectively. Let \( \tilde{U}_1, \tilde{U}_2 \) be the means of the sequences.

\[
\mu_1(d) = d' \tilde{U}_1 \quad \text{and} \quad \mu_2(d) = d' \tilde{U}_2
\]

are the projections of the means of the sequences on \( d \).

Is \( |\mu_1(d) - \mu_2(d)| \) a good separation measure between the set of projections?
Not necessarily!

The main problem is that \( \tilde{\mu}_1(d) \) and \( \tilde{\mu}_2(d) \) ignore the variances within the classes.

\[
\tilde{\mu}_1(1,0) - \tilde{\mu}_2(1,0) < \tilde{\mu}_1(0,1) - \tilde{\mu}_2(0,1),
\]

Even though

the separation on the vertical axis is better than the one on the horizontal axis.
Let $X \in \mathbb{R}^{m \times n}$ be a sample matrix and let $\hat{X} = H_m X$ be its centered form. The scatter matrix of $X$ is the matrix $S(X) = \hat{X}' \hat{X} \in \mathbb{R}^{n \times n}$. The scatter of $X$ in direction $d \in \mathbb{R}^n$ is the number $s(X)_d$ defined by $s(X)_d^2 = d' S(X) d$. We have

$$s(X)_d^2 = d' S(X) d = d' \hat{X}' \hat{X} d = (\hat{X} d)' \hat{X} d.$$
Let $X_1 \in \mathbb{R}^{m_1 \times n}$ and $X_2 \in \mathbb{R}^{m_2 \times n}$ be two data matrices that refer to $m_1$ and $m_2$ experiments and the same set of $n$ variables. The **within class scatter matrix** of the two data sets is the matrix $S_w(X_1, X_2) = S(X_1) + S(X_2)$. Note that this is a correct definition because we have both $S(X_1), S(X_2) \in \mathbb{R}^{n \times n}$.

The total scatter in direction $d$ is

$$s(X_1)_d^2 + s(X_1)_d^2 = d'S(X_1)d + d'S(X_2)d.$$
The *between class scatter matrix* of $X_1$ and $X_2$ is

$$S_b(X_1, X_2) = (\tilde{U}_1 - \tilde{U}_2)(\tilde{U}_1 - \tilde{U}_2)' \in \mathbb{R}^{n \times n}.$$ 

Again, this is correctly defined because $\tilde{U}_1$ and $\tilde{U}_2$ belong to $\mathbb{R}^n$. $S_b$ measures the separation between the means of the two classes.
Note that

\[ d' S_b d = d' (\tilde{U}_1 - \tilde{U}_2)(\tilde{U}_1 - \tilde{U}_2)'d \]
\[ = (d' \tilde{U}_1 - d' \tilde{U}_2)(\tilde{U}_1' d - \tilde{U}_2' d)' \]
\[ = (\mu_1(d) - \mu_2(d))^2. \]

The objective function is

\[ J(d) = \frac{(\mu_1(d) - \mu_2(d))^2}{s(X_1)_d^2 + s(X_1)_d^2} = \frac{(\mu_1(d) - \mu_2(d))^2}{d' S(X_1)d + d' S(X_2)d} = \frac{d' S_b d}{d' S_w d}, \]

which should be maximized.
We have

\[ J'(d) = \frac{(d'S_b d)'(d'S_w d) - (d'S_b d)(d'S_w d)'}{(d'S_w d)^2} \]

\[ = \frac{(2S_b d)(d'S_w d) - (2S_w d)(d'S_b d)}{(d'S_w d)^2} = 0 \]

So, we have

\[(2S_b d)(d'S_w d) - (2S_w d)(d'S_b d) = 0,\]

or

\[ S_b d = S_w d \frac{d'S_b d}{d'S_w d}, \]

which amounts to solving a generalized eigenvalue problem \( S_b d = \lambda S_w d, \)

where \( \lambda = \frac{d'S_b d}{d'S_w d}. \)

If \( S_w \) has full rank, this amounts to a standard eigenvalue problem

\[ S_w^{-1} S_b d = \lambda d. \]
Fisher discriminant method consists of finding a direction $\mathbf{d}$ such that

- $\mu_1(\mathbf{d}) - \mu_2(\mathbf{d})$ is maximal, and
- $s(X_1)_d^2 + s(X_1)_d^2$ is minimal.

This is obtained by choosing $\mathbf{d}$ to be an eigenvector of the matrix $S_w^{-1}S_b$: classes will be well separated.
We have

\[ \mathbf{S}_b \mathbf{d} = (\tilde{\mathbf{U}}_1 - \tilde{\mathbf{U}}_2)(\tilde{\mathbf{U}}_1 - \tilde{\mathbf{U}}_2)' \mathbf{d} \]
\[ = (\tilde{\mathbf{U}}_1 - \tilde{\mathbf{U}}_2) \left( (\tilde{\mathbf{U}}_1 - \tilde{\mathbf{U}}_2)' \mathbf{d} \right), \]

which implies that \( \mathbf{S}_b \mathbf{d} \) is colinear with \( \tilde{\mathbf{U}}_1 - \tilde{\mathbf{U}}_2 \) for any vector \( \mathbf{d} \).

Therefore, the eigenvalue problem has the immediate solution

\[ \mathbf{d} = \mathbf{S}_w^{-1}(\tilde{\mathbf{U}}_1 - \tilde{\mathbf{U}}_2) \]
Example

Consider the following subsets of $\mathbb{R}^2$:

$$S_1 = \left\{ \left( \begin{array}{c} 4 \\ 7 \end{array} \right), \left( \begin{array}{c} 4 \\ 13 \end{array} \right), \left( \begin{array}{c} 3 \\ 17 \end{array} \right), \left( \begin{array}{c} 7 \\ 11 \end{array} \right), \left( \begin{array}{c} 10 \\ 13 \end{array} \right), \left( \begin{array}{c} 16 \\ 17 \end{array} \right) \right\},$$

$$S_2 = \left\{ \left( \begin{array}{c} 13 \\ 7 \end{array} \right), \left( \begin{array}{c} 16 \\ 10 \end{array} \right), \left( \begin{array}{c} 20 \\ 11 \end{array} \right), \left( \begin{array}{c} 23 \\ 11 \end{array} \right), \left( \begin{array}{c} 22 \\ 13 \end{array} \right), \left( \begin{array}{c} 25 \\ 13 \end{array} \right) \right\}.$$
Data matrices of the two data sets are

\[ X_1 = \begin{pmatrix} 4 & 7 \\ 4 & 13 \\ 3 & 17 \\ 7 & 11 \\ 10 & 13 \\ 16 & 17 \end{pmatrix} \quad \text{and} \quad X_2 = \begin{pmatrix} 13 & 7 \\ 16 & 10 \\ 20 & 11 \\ 23 & 11 \\ 22 & 13 \\ 25 & 13 \end{pmatrix} \]

The corresponding mean vectors are

\[ \tilde{U}_1' = \frac{1}{6} \cdot 1'_6 X_1 = (7.33 \ 13.00) \quad \text{and} \quad \tilde{U}_2' = \frac{1}{6} \cdot 1'_6 X_2 = (19.83 \ 10.83) \]
The Centered Data Matrices

We have

\[ \hat{X}_1 = H_6 X_1 = \begin{pmatrix} 0.833 & -0.167 & -0.167 & -0.167 & -0.167 & -0.167 \\ -0.167 & 0.833 & -0.167 & -0.167 & -0.167 & -0.167 \\ -0.167 & -0.167 & 0.833 & -0.167 & -0.167 & -0.167 \\ -0.167 & -0.167 & -0.167 & 0.833 & -0.167 & -0.167 \\ -0.167 & -0.167 & -0.167 & -0.167 & 0.833 & -0.167 \\ -0.167 & -0.167 & -0.167 & -0.167 & -0.167 & 0.833 \end{pmatrix} \]

\[ X_1 = \begin{pmatrix} -3.333 & -6 \\ -3.333 & 0 \\ -4.333 & 4 \\ -0.333 & -2 \\ 2.666 & 0 \\ 8.666 & 4 \end{pmatrix} \]
\[ \hat{x}_2 = H_6 x_2 = \begin{pmatrix} 0.833 & -0.167 & -0.167 & -0.167 & -0.167 & -0.167 \\ -0.167 & 0.833 & -0.167 & -0.167 & -0.167 & -0.167 \\ -0.167 & -0.167 & 0.833 & -0.167 & -0.167 & -0.167 \\ -0.167 & -0.167 & -0.167 & 0.833 & -0.167 & -0.167 \\ -0.167 & -0.167 & -0.167 & -0.167 & -0.167 & 0.833 \end{pmatrix} x_2 \]

\[ = \begin{pmatrix} -6.833 & -3.833 \\ -3.833 & -0.833 \\ 0.167 & 0.167 \\ 3.167 & 0.167 \\ 2.167 & 2.167 \\ 5.167 & 2.167 \end{pmatrix}. \]
The Scatter Matrices

The scatter matrix of \( X_1 \) is the matrix \( S(X_1) = \hat{X}_1' \hat{X}_1 \in \mathbb{R}^{2 \times 2} \) given by

\[
S(X_1) = \hat{X}_1' \hat{X}_1 = \begin{pmatrix}
123.333 & 38 \\
38.000 & 72
\end{pmatrix}
\]

and

\[
S(X_2) = \hat{X}_2' \hat{X}_2 = \begin{pmatrix}
102.833 & 45.833 \\
45.833 & 24.833
\end{pmatrix}.
\]

The within class scatter matrix is

\[
S_w(X_1, X_2) = S(X_1) + S(X_2) = \begin{pmatrix}
226.167 & 83.833 \\
83.833 & 96.833
\end{pmatrix}.
\]
The between class scatter matrix

Since

\[ \tilde{U}_1' = \frac{1}{6} \cdot 1_6' X_1 = (7.33 \ 13.00) \] and \[ \tilde{U}_2' = \frac{1}{6} \cdot 1_6' X_2 = (19.83 \ 10.83) \]

we have

\[ \tilde{U}_1 - \tilde{U}_2 = \begin{pmatrix} -12.5 \\ 2.167 \end{pmatrix} \]

which implies

\[ S_b = (\tilde{U}_1 - \tilde{U}_2)(\tilde{U}_1 - \tilde{U}_2)' = \begin{pmatrix} 156.250 & -27.083 \\ -27.083 & 4.694 \end{pmatrix}. \]
The inverse of the matrix

\[ S_w = \begin{pmatrix} 226.167 & 83.833 \\ 83.833 & 96.833 \end{pmatrix} \]

is

\[ S_w^{-1} = \begin{pmatrix} 0.0065 & -0.0056 \\ -0.0056 & 0.0152 \end{pmatrix} \]

The optimum direction of the projection is

\[ d = S_w^{-1}(\tilde{U}_1 - \tilde{U}_2) \]
\[ = \begin{pmatrix} -0.0936 \\ 1.034 \end{pmatrix}. \]
The projections of the data on the optimum direction are:

\[ X_1d = \begin{pmatrix} 0.3495 \\ 0.9699 \\ 1.4772 \\ 0.4823 \\ 0.4083 \\ 0.2604 \end{pmatrix} \]

and

\[ X_2d = \begin{pmatrix} -0.4929 \\ -0.4635 \\ -0.7345 \\ -1.0153 \\ -0.7149 \\ -0.9957 \end{pmatrix} \]
Example

Consider a training data set that consists of the following sets of examples in $\mathbb{R}^2$:

$$S_1 = \left\{ \left( \frac{5}{4}, \frac{10}{11} \right), \left( \frac{4}{11}, \frac{8}{11} \right), \left( \frac{4}{7} \right) \right\},$$

$$S_2 = \left\{ \left( \frac{6}{3}, \frac{10}{7} \right), \left( \frac{10}{6}, \frac{10}{11} \right) \right\}.$$
Data matrices of the two data sets are

\[
X_1 = \begin{pmatrix} 5 & 4 \\ 10 & 15 \\ 4 & 11 \\ 8 & 11 \\ 4 & 7 \end{pmatrix} \quad \text{and} \quad X_2 = \begin{pmatrix} 6 & 3 \\ 10 & 3 \\ 10 & 7 \\ 14 & 6 \\ 13 & 11 \end{pmatrix}
\]

The corresponding mean vectors are

\[
\tilde{\mu}_1' = \frac{1}{5} \times 1'_5 X_1 = (6.2 \ 9.6) \quad \text{and} \quad \tilde{\mu}_2' = \frac{1}{5} \times 1'_5 X_2 = (10.6 \ 6).
\]
The Centered Data Matrices

We have

\[ \hat{X}_1 = H_5 X_1 = \begin{pmatrix} 0.8 & -0.2 & -0.2 & -0.2 & -0.2 \\ -0.2 & 0.8 & -0.2 & -0.2 & -0.2 \\ -0.2 & -0.2 & 0.8 & -0.2 & -0.2 \\ -0.2 & -0.2 & -0.2 & 0.8 & -0.2 \\ -0.2 & -0.2 & -0.2 & -0.2 & 0.8 \end{pmatrix} \]

\[ X_1 = \begin{pmatrix} -1.2 & -5.6 \\ 3.8 & 5.4 \\ -2.2 & 1.4 \\ 1.8 & 1.4 \\ -2.2 & -2.6 \end{pmatrix} \]

and

\[ \hat{X}_2 = H_5 X_2 = \begin{pmatrix} 0.8 & -0.2 & -0.2 & -0.2 & -0.2 \\ -0.2 & 0.8 & -0.2 & -0.2 & -0.2 \\ -0.2 & -0.2 & 0.8 & -0.2 & -0.2 \\ -0.2 & -0.2 & -0.2 & 0.8 & -0.2 \\ -0.2 & -0.2 & -0.2 & -0.2 & 0.8 \end{pmatrix} \]

\[ X_2 = \begin{pmatrix} -4.6 & -3.0 \\ -0.6 & -3.0 \\ 3.4 & 0 \\ 2.4 & 5.0 \end{pmatrix} \]
The scatter matrix of $X_1$ is the matrix $S(X_1) = \hat{X}_1'\hat{X}_1 \in \mathbb{R}^{2 \times 2}$ given by

$$S(X_1) = \hat{X}_1'\hat{X}_1 = \begin{pmatrix} 28.8 & 32.4 \\ 32.4 & 71.2 \end{pmatrix}$$

and

$$S(X_2) = \hat{X}_2'\hat{X}_2 = \begin{pmatrix} 39.2 & 27 \\ 27 & 44 \end{pmatrix}.$$ 

The within class scatter matrix is

$$S_w(X_1, X_2) = S(X_1) + S(X_2) = \begin{pmatrix} 68 & 59.4 \\ 59.4 & 115.2 \end{pmatrix}.$$
The between class scatter matrix

Since

\[ \tilde{U}'_1 = \frac{1}{5} \times 1'_5 X_1 = (6.2 \ 9.6) \] and \( \tilde{U}'_2 = \frac{1}{5} \times 1'_5 X_2 = (10.6 \ 6) \),

we have

\[ \tilde{U}_1 - \tilde{U}_2 = \begin{pmatrix} 6.2 \\ 9.6 \end{pmatrix} - \begin{pmatrix} 10.6 \\ 6 \end{pmatrix} = \begin{pmatrix} -4.4 \\ 3.6 \end{pmatrix}, \]

which implies

\[ S_b = (\tilde{U}_1 - \tilde{U}_2)(\tilde{U}_1 - \tilde{U}_2)' \]

\[ = \begin{pmatrix} -4.4 \\ 3.6 \end{pmatrix} \begin{pmatrix} -4.4 & 3.6 \end{pmatrix} = \begin{pmatrix} 19.36 & -15.84 \\ -15.84 & 12.96 \end{pmatrix}. \]
The inverse of the matrix

\[ S_w = \begin{pmatrix} 68 & 59.4 \\ 59.4 & 115.2 \end{pmatrix} \]

is

\[ S_w^{-1} = \begin{pmatrix} 0.0268 & -0.0138 \\ -0.0138 & 0.0158 \end{pmatrix} \]

The optimum direction of the projection is

\[ d = S_w^{-1}(\tilde{U}_1 - \tilde{U}_2) \]

\[ = \begin{pmatrix} 0.0268 & -0.0138 \\ -0.0138 & 0.0158 \end{pmatrix} \begin{pmatrix} 6.2 \\ 9.6 \end{pmatrix} = \begin{pmatrix} 0.0334 \\ 0.0661 \end{pmatrix}. \]
The projections of the data on the optimum directions are:

\[ X_1d = \begin{pmatrix} 0.4316 \\ 1.3258 \\ 0.8607 \\ 0.9945 \\ 0.5964 \end{pmatrix} \quad \text{and} \quad X_2d = \begin{pmatrix} 0.3989 \\ 0.5327 \\ 0.7971 \\ 0.8648 \\ 1.1618 \end{pmatrix} \]
Example

Consider the following subsets of $\mathbb{R}^2$:

\[
S_1 = \left\{ \begin{pmatrix} 1 \\ 3 \\ \end{pmatrix}, \begin{pmatrix} 1 \\ 5 \\ \end{pmatrix}, \begin{pmatrix} 4 \\ 7 \\ \end{pmatrix}, \begin{pmatrix} 4 \\ 9 \\ \end{pmatrix} \right\},
\]

\[
S_2 = \left\{ \begin{pmatrix} 3 \\ 2 \\ \end{pmatrix}, \begin{pmatrix} 2.5 \\ 0.5 \\ \end{pmatrix}, \begin{pmatrix} 6 \\ 6 \\ \end{pmatrix}, \begin{pmatrix} 7.5 \\ 4.5 \\ \end{pmatrix} \right\}.
\]
Data matrices of the two data sets are

\[ X_1 = \begin{pmatrix} 1 & 3 \\ 1 & 5 \\ 4 & 7 \\ 4 & 9 \end{pmatrix} \quad \text{and} \quad X_2 = \begin{pmatrix} 3 & 2 \\ 2.5 & 0.5 \\ 6 & 6 \\ 7.5 & 4.5 \end{pmatrix} \]

The corresponding mean vectors are

\[ \tilde{U}_1' = \frac{1}{4} \ast 1'_4 X_1 = (2.5 \ 6) \quad \text{and} \quad \tilde{U}_2' = \frac{1}{4} \ast 1'_4 X_2 = (4.75 \ 3.25) \]
The Centered Data Matrices

We have

\[ \hat{X}_1 = H_4 X_1 = \begin{pmatrix} 0.75 & -0.25 & -0.25 & -0.25 \\ -0.25 & 0.75 & -0.25 & -0.25 \\ -0.25 & -0.25 & 0.75 & -0.25 \\ -0.25 & -0.25 & -0.25 & 0.75 \end{pmatrix} X_1 \]

\[ = \begin{pmatrix} -1.50 & -3 \\ -1.50 & -1 \\ 1.50 & 1 \\ 1.50 & 3 \end{pmatrix} \]
The Centered Data Matrices

We have

\[ \hat{X}_2 = H_4 X_1 = \begin{pmatrix} 0.75 & -0.25 & -0.25 & -0.25 \\ -0.25 & 0.75 & -0.25 & -0.25 \\ -0.25 & -0.25 & 0.75 & -0.25 \\ -0.25 & -0.25 & -0.25 & 0.75 \end{pmatrix} \begin{pmatrix} -1.75 & -1.25 \\ -2.25 & -2.75 \\ 1.25 & 2.75 \\ 2.75 & 1.25 \end{pmatrix} \]
The scatter matrix of $X_1$ is the matrix $S(X_1) = \hat{X}_1'\hat{X}_1 \in \mathbb{R}^{2\times2}$ given by

$$S(X_1) = \hat{X}_1'\hat{X}_1 = \begin{pmatrix} 9 & 12 \\ 12 & 20 \end{pmatrix}$$

and

$$S(X_2) = \hat{X}_2'\hat{X}_2 = \begin{pmatrix} 17.25 & 15.25 \\ 15.25 & 18.25 \end{pmatrix}.$$ 

The within class scatter matrix is

$$S_w(X_1, X_2) = S(X_1) + S(X_2) = \begin{pmatrix} 26.25 & 27.25 \\ 27.25 & 38.25 \end{pmatrix}.$$
The between class scatter matrix

Recall:

\[ \tilde{U}_1' = \frac{1}{4} \times 1_4' X_1 = (2.5 \ 6) \] and \[ \tilde{U}_2' = \frac{1}{4} \times 1_4' X_2 = (4.75 \ 3.25) \]

Thus, we have

\[ \tilde{U}_1 - \tilde{U}_2 = \begin{pmatrix} -2.25 \\ 2.75 \end{pmatrix} \]

which implies

\[
S_b = (\tilde{U}_1 - \tilde{U}_2)(\tilde{U}_1 - \tilde{U}_2)' \\
= \begin{pmatrix} -2.25 \\ 2.75 \end{pmatrix} \begin{pmatrix} -2.25 & 2.75 \end{pmatrix} = \begin{pmatrix} 5.0625 & -6.1875 \\ -6.1875 & 7.5625 \end{pmatrix}.
\]
The inverse of the matrix

\[ S_w = \begin{pmatrix} 26.25 & 27.25 \\ 27.25 & 38.25 \end{pmatrix}. \]

is

\[ S_w^{-1} = \begin{pmatrix} 0.1463 & -0.1042 \\ -0.1042 & 0.1004 \end{pmatrix}. \]

The optimum direction of the projection is

\[ \mathbf{d} = S_w^{-1}(\tilde{U}_1 - \tilde{U}_2) = \begin{pmatrix} -0.6157 \\ 0.5195 \end{pmatrix}. \]
The projections of the data on the optimum direction are:

\[
X_1d = \begin{pmatrix}
-0.6080 \\
0.4130 \\
-0.4130 \\
0.6080
\end{pmatrix}
\quad \text{and} \quad
X_2d = \begin{pmatrix}
0.4393 \\
-0.0186 \\
0.6343 \\
-1.0550
\end{pmatrix}
\]
Multiple Discriminant Analysis

- this is a generalization of Fisher’s discriminant to \( c \) classes;
- reduces the dimensionality of the data to \( c - 1 \).

Let \( V \) be the projection matrix on a \((c - 1)\)-dimensional space. Notations:
- \( \tilde{U}_i \): the mean of the \( n_i \) examples of class \( C_i \),

\[
\tilde{U}_i = \frac{1}{n_i} \sum \{x \mid x \in C_i\}.
\]

for \( 1 \leq i \leq c \);
- \( \tilde{U} \): the mean of the total mean of samples;
Scatter matrices

- within class scatter matrix: $S_w = \sum_{1 \leq i \leq c} \hat{X}_i \hat{X}_i'$
- between class scatter matrix: $S_b = \sum_{1 \leq i \leq c} n_i (\tilde{U}_i \tilde{U})(\tilde{U}_i \tilde{U})'$ (maximum rank is $c - 1$);
- objective function is

$$J(V) = \frac{V' S_b V}{V' S_w V}.$$
\[ J(V) = \frac{V'S_b V}{V'S_w V}. \]

- solve the generalized eigenvalue problem \( S_b v = \lambda S_w v \), which has at most \( c - 1 \) distinct eigenvalues;
- let \( v_1, \ldots, v_{c-1} \) are the respective eigenvectors;
- the optimal projection matrix to a subspace of dimension \( k \) is given by the eigenvectors that correspond to the largest \( k \) eigenvalues.