FISHER'S LINEAR DISCRIMINANT ANALYSIS

Fisher's Linear Discriminant Analysis (FDA) is a dimensionality reduction technique to ease classification.

1 Two Class Case

Consider the case of N points in d dimensions, with each point belonging to one of the two classes C_1 and C_2. The idea is to find the optimal direction to project the vector of these points to. Such a projection can be represented as,

$$y = \boldsymbol{w}^T \boldsymbol{x}$$

where $\boldsymbol{w}$ is a d dimensional vector defining the direction of projection, $\boldsymbol{x}$ is the vector being projected and y is a scalar representing the magnitude of projection.

The projection should serve two purposes as discussed in the following subsections:

1.1 Maximizing Between-Class Scatter

The class means should be projected as far apart as possible. Let $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ denote the mean of class C_1 and C_2 respectively.

$$\boldsymbol{m_1} = \frac{1}{N} \sum_{\boldsymbol{x_i} \in C_1} \boldsymbol{x_i}$$
$$\boldsymbol{m_2} = \frac{1}{N} \sum_{\boldsymbol{x_i} \in C_2} \boldsymbol{x_i}$$

If $\boldsymbol{m_1}$ is projected to m_1 and $\boldsymbol{m_2}$ is projected to m_2, then the between-scatter is defined by,

$$
\begin{aligned}
(m_1 - m_2)^2 &= (\boldsymbol{w}^T \boldsymbol{m_1} - \boldsymbol{w}^T \boldsymbol{m_2})^2 \\
&= (\boldsymbol{w}^T (\boldsymbol{m_1} - \boldsymbol{m_2}))^2 \\
&= \boldsymbol{w}^T (\boldsymbol{m_1} - \boldsymbol{m_2}) \boldsymbol{w}^T (\boldsymbol{m_1} - \boldsymbol{m_2}) \\
&= \boldsymbol{w}^T (\boldsymbol{m_1} - \boldsymbol{m_2})(\boldsymbol{m_1} - \boldsymbol{m_2})^T \boldsymbol{w} \\
&= \boldsymbol{w}^T \boldsymbol{S_B} \boldsymbol{w}
\end{aligned}
$$

where $\boldsymbol{S_B}$ represents the between-class scatter matrix.

1.2 Minimizing Within-Class Scatter

The projections of each class should be as condensed as possible. The within-class scatter of the transformed data belonging to class C_k is denoted by,

$$
\begin{aligned}
s_k^2 &= \sum_{i \in C_k} (y_i - m_k)^2 \\
&= \sum_{i \in C_k} (\boldsymbol{w}^T \boldsymbol{x_i} - \boldsymbol{w}^T \boldsymbol{m_k})^2 \\
&= \sum_{i \in C_k} (\boldsymbol{w}^T (\boldsymbol{x_i} - \boldsymbol{m_k}))^2 \\
&= \sum_{i \in C_k} \boldsymbol{w}^T (\boldsymbol{x_i} - \boldsymbol{m_k})(\boldsymbol{x_i} - \boldsymbol{m_k})^T \boldsymbol{w} \\
&= \boldsymbol{w}^T \boldsymbol{S_k} \boldsymbol{w}
\end{aligned}
$$

where,

$$
\boldsymbol{S_k} = \sum_{i \in C_k} (\boldsymbol{x_i} - \boldsymbol{m_k})(\boldsymbol{x_i} - \boldsymbol{m_k})^T
$$

In the case of two classes, total within class scatter is denoted by,

$$
\begin{aligned}
s_1^2 + s_2^2 &= \boldsymbol{w}^T \boldsymbol{S_1} \boldsymbol{w} + \boldsymbol{w}^T \boldsymbol{S_1} \boldsymbol{w} \\
&= \boldsymbol{w}^T (\boldsymbol{S_1} + \boldsymbol{S_2}) \boldsymbol{w} \\
&= \boldsymbol{w}^T \boldsymbol{S_W} \boldsymbol{w}
\end{aligned}
$$

where $\boldsymbol{S_W}$ represents the within-class scatter matrix.

1.3 Combining Minimization and Maximization

A reasonable way to simultaneously maximize the between-class scatter and minimize the within-class scatter is to maximize their fraction defined as follows,

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S_B} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S_W} \boldsymbol{w}}$$

Note that $J(\boldsymbol{w})$ is invariant under rescalings of the form $\boldsymbol{w} \Rightarrow \alpha \boldsymbol{w}$. This sets up the reformulation of this problem as per the following,

$$\begin{aligned} \text{maximize} \quad & \boldsymbol{w}^T \boldsymbol{S_B} \boldsymbol{w} \\ \text{s.t.} \quad & \boldsymbol{w}^T \boldsymbol{S_W} \boldsymbol{w} = 1 \end{aligned}$$

Using the concept of Lagrangian,

$$L(\boldsymbol{w}, \lambda) = \boldsymbol{w}^T \boldsymbol{S_B} \boldsymbol{w} - \lambda(\boldsymbol{w}^T \boldsymbol{S_W} \boldsymbol{w} - 1)$$

Differentiating w.r.t. $\boldsymbol{w}$,

$$\frac{\partial L(\boldsymbol{w}, \lambda)}{\partial \boldsymbol{w}} = 2\boldsymbol{S_B}\boldsymbol{w} - 2\lambda \boldsymbol{S_W} \boldsymbol{w}$$

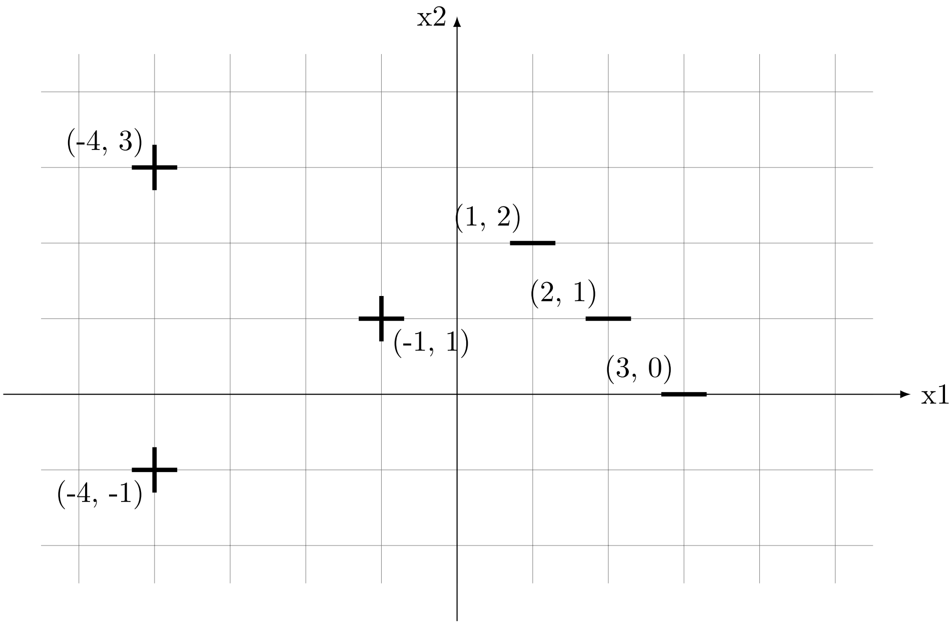
Equating the diffenrential to zero, the solution follows,

$$\boldsymbol{S_B}\boldsymbol{w} = \lambda \boldsymbol{S_W} \boldsymbol{w}$$

This is the generalized eigenvalue problem that can be solved easily. Note that since $\boldsymbol{S_B}$ is a product of two vectors and thus of rank one, the above equation will only yield one eigenvalue, eigenvector pair. The eigenvector is the sought projection direction.

2 Example

Consider the following set of six points in two dimesions distributed equally among the two classes,



The means for the respective classes are,

$$\boldsymbol{m_+} = \begin{pmatrix} -3 \\ 1 \end{pmatrix}$$
$$\boldsymbol{m_-} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

Now, the between-class scatter matrix is calculated as follows,

$$\begin{aligned} \boldsymbol{S_B} &= (\boldsymbol{m_+} - \boldsymbol{m_-})(\boldsymbol{m_+} - \boldsymbol{m_-})^T \\ &= \begin{pmatrix} -5 \\ 0 \end{pmatrix} \begin{pmatrix} -5 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 25 & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

And then, the within-class scatter matrix is calculated,

$$
\begin{aligned}
S_+ &= \sum_{i\in+} (\boldsymbol{x_i} - \boldsymbol{m_+})(\boldsymbol{x_i} - \boldsymbol{m_+})^T \\
&= \begin{pmatrix} 2 \\ 0 \end{pmatrix} \begin{pmatrix} 2 & 0 \end{pmatrix} + \begin{pmatrix} -1 \\ 2 \end{pmatrix} \begin{pmatrix} -1 & 2 \end{pmatrix} + \begin{pmatrix} -1 \\ -2 \end{pmatrix} \begin{pmatrix} -1 & -2 \end{pmatrix} \\
&= \begin{pmatrix} 6 & 0 \\ 0 & 8 \end{pmatrix} \\
S_- &= \sum_{i\in-} (\boldsymbol{x_i} - \boldsymbol{m_-})(\boldsymbol{x_i} - \boldsymbol{m_-})^T \\
&= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \end{pmatrix} + \begin{pmatrix} -1 \\ 1 \end{pmatrix} \begin{pmatrix} -1 & 1 \end{pmatrix} + \begin{pmatrix} 1 \\ -1 \end{pmatrix} \begin{pmatrix} 1 & -1 \end{pmatrix} \\
&= \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix} \\
\boldsymbol{S_W} &= \boldsymbol{S_+} + \boldsymbol{S_-} \\
&= \begin{pmatrix} 8 & -2 \\ -2 & 10 \end{pmatrix}
\end{aligned}
$$

The eigenvalue problem is formulated as,

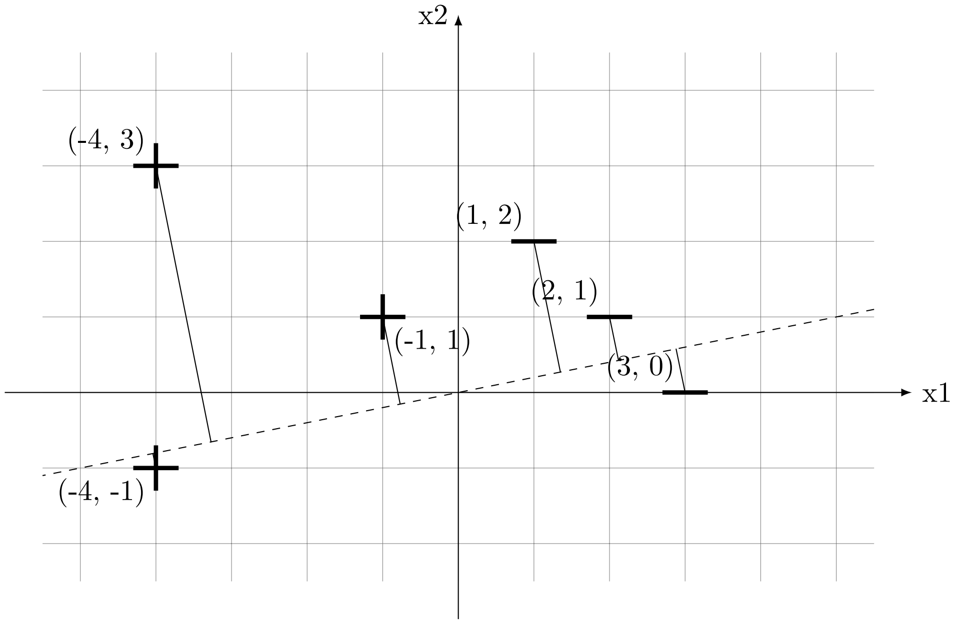
$$\boldsymbol{S_B w} = \lambda \boldsymbol{S_W w}$$

Equating the determinant of $\boldsymbol{S_B} - \lambda \boldsymbol{S_W}$ to zero,

$$
\begin{vmatrix} 25 - 8\lambda & 2\lambda \\ 2\lambda & -10\lambda \end{vmatrix} = 0
$$

$$\lambda = \frac{125}{38}$$

The corresponding eigenvector amounts to $\frac{1}{\sqrt{26}} \begin{pmatrix} 5 \\ 1 \end{pmatrix}$ which is the direction of projection. The below plot helps to visualize how this projection direction divides the two classes.



3 Multi-Class Case

In two class case dimensionality was reduced to 1. In K class case, it is reduced to $K-1$. Hence, instead of finding an optimal d-dimensional vector $\boldsymbol{w}$ in the two class case, a matrix of $K-1$ d-dimensional vectors $\boldsymbol{W}^T = \begin{pmatrix} \boldsymbol{w_1}^T \\ \boldsymbol{w_1}^T \\ \vdots \\ \boldsymbol{w_{K-1}}^T \end{pmatrix}$ must be found. The projection is defined as $\boldsymbol{y} = \boldsymbol{W}^T\boldsymbol{x}$ where $\boldsymbol{y}$ is a $K-1$ dimensional vector as expected.

The generalization for within-class matrix is as follows,

$$\boldsymbol{S_W} = \sum_{k=1}^{K} \sum_{i \in C_k} (\boldsymbol{x_i} - \boldsymbol{m_k})(\boldsymbol{x_i} - \boldsymbol{m_k})^T$$

For generalizing between-class matrix, consider the total scatter matrix $\boldsymbol{S_T}$ as follows,

$$
\begin{aligned}
\boldsymbol{S_T} &= \sum_{i=1}^{N} (\boldsymbol{x_i} - \boldsymbol{m})(\boldsymbol{x_i} - \boldsymbol{m})^T \\
&= \sum_{k=1}^{K} \sum_{i \in C_k} (\boldsymbol{x_i} - \boldsymbol{m})(\boldsymbol{x_i} - \boldsymbol{m})^T \\
&= \sum_{k=1}^{K} \sum_{i \in C_k} (\boldsymbol{x_i} - \boldsymbol{m_k} + \boldsymbol{m_k} - \boldsymbol{m})(\boldsymbol{x_i} - \boldsymbol{m_k} + \boldsymbol{m_k} - \boldsymbol{m})^T \\
&= \sum_{k=1}^{K} \sum_{i \in C_k} (\boldsymbol{x_i} - \boldsymbol{m_k})(\boldsymbol{x_i} - \boldsymbol{m_k})^T + \sum_{k=1}^{K} \sum_{i \in C_k} (\boldsymbol{m_k} - \boldsymbol{m})(\boldsymbol{m_k} - \boldsymbol{m})^T \\
&= \boldsymbol{S_W} + \sum_{k=1}^{K} N_k (\boldsymbol{m_k} - \boldsymbol{m})(\boldsymbol{m_k} - \boldsymbol{m})^T \\
&= \boldsymbol{S_W} + \boldsymbol{S_B}
\end{aligned}
$$

The maximization problem is then generalized as follows,

$$
J(\boldsymbol{W}) = \frac{det(\boldsymbol{W}^T \boldsymbol{S_B} \boldsymbol{W})}{det(\boldsymbol{W}^T \boldsymbol{S_W} \boldsymbol{W})}
$$

The solution to find the matrix $\boldsymbol{W}$ is to find the largest $K-1$ eigenvalues of the following equation and arrange the corresponding eigenvectors in a matrix.

$$
\boldsymbol{S_B w} = \lambda \boldsymbol{S_W w}
$$

Additionally, there are no more than $K-1$ non-zero eigenvectors to the above equation due to the redundancy of the matrix $\boldsymbol{S_B}$ as was seen in the two class case.