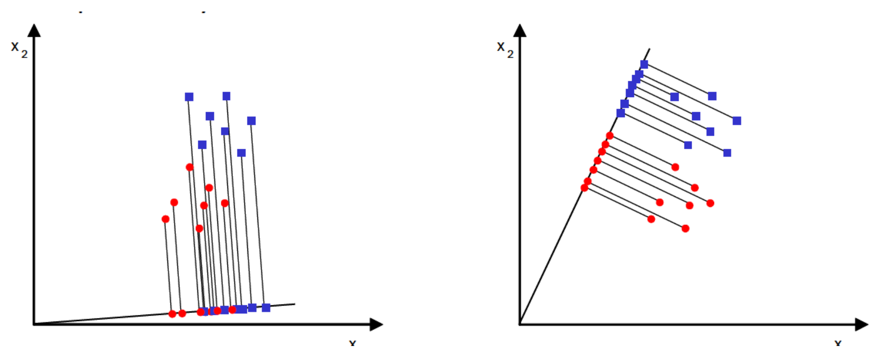
Fisherface based on LDA Note

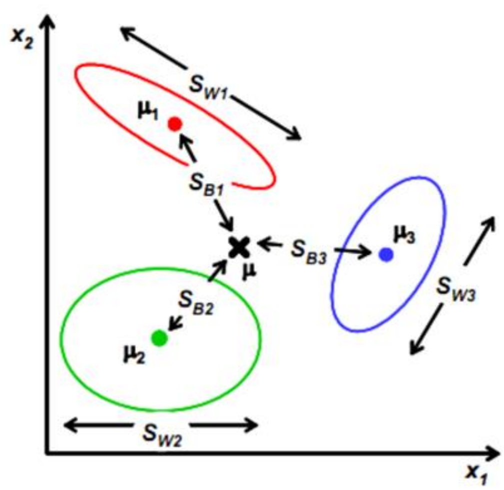
What is LDA?

Model: Mapping the original data to the low-dimensional space and classification.
Strategy:
Linear two classes:We seek to obtain a scalar y by projecting the samples x onto a line

$$y = w^T x$$



C classes:Find a suitable vector w, project the data onto w, and represent and distinguish the original data according to the projection point.



LDA, C classes

Fisher's LDA generalizes gracefully for C-class problems

– Instead of one projection y, we will now seek $(C-1)$ projections $[y_1, y_2, \dots y_{C-1}]$ by means of $(C-1)$ projection vectors w_i arranged by columns into a projection matrix $W = [w_1|w_2| \dots |w_{C-1}]$:

$$y_i = w_i^T x \Rightarrow y = W^T x$$

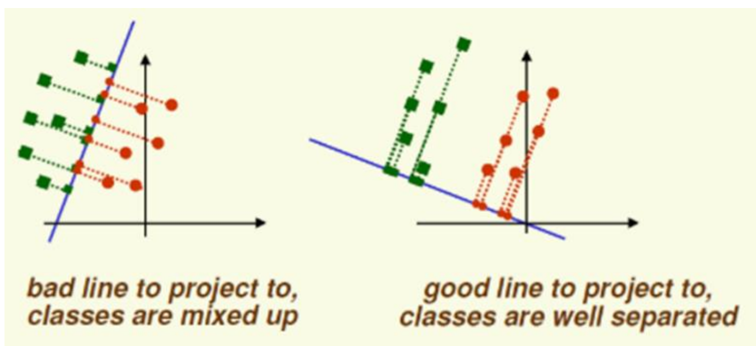
Algorithm：（LDA）Minimize Within-class scatter (S_W) while Maximize Between-class scatter(S_B).

Why LDA?

LDA Compare to PCA:

Common: all the dimension reduction project to a line.

Differences: LDA can reduce the amount of classification calculation because LDA takes the class labels as inputs.



How LDA achieved?

Assume we have a set of D-dimensional samples $x(1$, $x(2$, … $x(N$, $N1$ of which belong to class $\omega1$, and $N2$ to class $\omega2$– We seek to obtain a scalar y by projecting the samples x onto a line

$$y = w^T x$$

First calculate the mean (center point) of each type of data:

$$\mu_i = \frac{1}{N_i} \sum_{x \in \omega_i} x \quad (1)$$

The center of the data point projection onto w is:

$$\tilde{\mu}_i = \frac{1}{N_i}\sum_{y\in\omega_i} y = \frac{1}{N_i}\sum_{x\in\omega_i} w^T x = w^T \mu_i \quad (2)$$

How to judge the vector w is the best, it can be considered from two aspects: 1. The projection points obtained by different classifications should be separated as much as possible; 2. The points obtained after the same classification projection should be aggregated as much as possible.

So,We could then choose the distance between the projected means as our objective function.

$$J(w) = |\tilde{\mu}_1 - \tilde{\mu}_2| = |w^T(\mu_1 - \mu_2)| \quad (3)$$

However, the distance between projected means is not a good measure since it does not account for the standard deviation within classes.

To solve this problem, we can use Fisher's solution:

–For each class we define the scatter, an equivalent of the variance, as

$$\tilde{s}_i^2 = \sum_{y\in\omega_i}(y - \tilde{\mu}_i)^2 \quad (4)$$

The smaller the value, the more the projection point is aggregated.

Combining the two formulas(3)and(4), the first formula is the numerator and the other is the denominator:

$$J(w) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \quad (5)$$

The Fisher linear discriminant is defined as the linear function $w^T x$ that maximizes the criterion function.

To find the optimum $w*$, we must express $J(w)$ as a function of w–

First, we define a measure of the scatter in feature space x.

$$\begin{gathered} S_i = \sum_{x\in\omega_i}(x - \mu_i)(x - \mu_i)^T \\ S_1 + S_2 = S_W \end{gathered} \quad (6)$$

where S_W is called the within-class scatter matrix.

The scatter of the projection y can then be expressed as a function of the scatter matrix in feature space x.

$$\begin{aligned} \tilde{s}_i^2 &= \sum_{y\in\omega_i}(y - \tilde{\mu}_i)^2 = \sum_{x\in\omega_i}(w^T x - w^T\mu_i)^2 = \\ &= \sum_{x\in\omega_i} w^T(x - \mu_i)(x - \mu_i)^T w = w^T S_i w \end{aligned}$$

$$\tilde{s}_1^2 + \tilde{s}_2^2 = w^T S_W w \quad (7)$$

Similarly, the difference between the projected means can be expressed in terms of the means in the original feature space.

$$(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (w^T\mu_1 - w^T\mu_2)^2 = w^T \underbrace{(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T}_{S_B} w = w^T S_B w$$

The matrix S_B is called the between-class scatter. Note that, since S_B is the outer product of two vectors, its rank is at most one

We can finally express the Fisher criterion in terms of S_W and S_B as

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

To find the maximum of $J(w)$ we derive and equate to zero

$$\frac{d}{dw}[J(w)] = \frac{d}{dw}\left[\frac{w^T S_B w}{w^T S_W w}\right] = 0 \Rightarrow$$
$$[w^T S_W w]\frac{d[w^T S_B w]}{dw} - [w^T S_B w]\frac{d[w^T S_W w]}{dw} = 0 \Rightarrow$$
$$[w^T S_W w]2S_B w - [w^T S_B w]2S_W w = 0$$

Dividing by $w^T S_W w$

$$\left[\frac{w^T S_W w}{w^T S_W w}\right] S_B w - \left[\frac{w^T S_B w}{w^T S_W w}\right] S_W w = 0 \Rightarrow$$
$$S_B w - J S_W w = 0 \Rightarrow$$
$$S_W^{-1} S_B w - J w = 0$$

Solving the generalized eigenvalue problem ($S_W^{-1}S_B w = J w$) yields

$$w^* = \arg\max\left[\frac{w^T S_B w}{w^T S_W w}\right] = S_W^{-1}(\mu_1 - \mu_2)$$

LDA Example

Compute the LDA Projection for the following 2D dataset

$X_1 = \{(1,2), (3,5), (4,3), (5,6), (7,5)\}$

$X_2 = \{(6,2), (9,4), (10,1), (12,3), (13,6)\}$

Solution:

① Compute the class statistics S_1 and S_2, μ_1 and μ_2

$$S_1 = \sum_{\forall x_i \in w_1} (x_i - \mu_1)(x_i - \mu_1)^T$$

$$S_2 = \sum_{\forall x \in w_2} (x_i - \mu_2)(x_i - \mu_2)^T$$

						μ_1
$\omega_1 \Rightarrow$	1	3	4	5	7	4
	2	5	3	6	5	4.2
$x - \mu_1 \Rightarrow$	-3	-1	0	1	3	
	-2.2	0.8	-1.2	1.8	0.8	

$\Rightarrow \mu_1 = \begin{bmatrix} 4 \\ 4.2 \end{bmatrix}$

						μ_2
$\omega_2 \Rightarrow$	6	9	10	12	13	10
	2	4	1	3	6	3.2
$x - \mu_2 \Rightarrow$	-4	-1	0	2	3	
	-1.2	0.8	-2.2	-0.2	2.8	

$\Rightarrow \mu_2 = \begin{bmatrix} 10 \\ 3.2 \end{bmatrix}$

$$\begin{bmatrix}-3\\-2.2\end{bmatrix}(-3,-2.2)=\begin{bmatrix}9 & 6.6\\6.6 & 4.84\end{bmatrix}$$

$$\begin{bmatrix}-1\\0.8\end{bmatrix}(-1,0.8)=\begin{bmatrix}1 & -0.8\\-0.8 & 0.64\end{bmatrix}$$

$$\begin{bmatrix}0\\1.2\end{bmatrix}(0,1.2)=\begin{bmatrix}0 & 0\\0 & 1.44\end{bmatrix}$$

$$\begin{bmatrix}1\\1.8\end{bmatrix}(1,1.8)=\begin{bmatrix}1 & 1.8\\1.8 & 3.24\end{bmatrix}$$

$$\begin{bmatrix}3\\0.8\end{bmatrix}(3,0.8)=\begin{bmatrix}9 & 2.4\\2.4 & 0.64\end{bmatrix}$$

$$\Rightarrow S_1=\begin{bmatrix}20 & 10\\10 & 10.8\end{bmatrix}$$

$$\begin{bmatrix}-4\\-1.2\end{bmatrix}(-4,-1.2)=\begin{bmatrix}16 & 4.8\\4.8 & 1.44\end{bmatrix}$$

$$\begin{bmatrix}-1\\0.8\end{bmatrix}(-1,0.8)=\begin{bmatrix}1 & -0.8\\-0.8 & 0.64\end{bmatrix}$$

$$\begin{bmatrix}0\\-2.2\end{bmatrix}(0,-2.2)=\begin{bmatrix}0 & 0\\0 & 4.84\end{bmatrix}$$

$$\begin{bmatrix}2\\-0.2\end{bmatrix}(2,-0.2)=\begin{bmatrix}4 & -0.4\\-0.4 & 0.04\end{bmatrix}$$

$$\begin{bmatrix}3\\2.8\end{bmatrix}(3,2.8)=\begin{bmatrix}9 & 8.4\\8.4 & 7.84\end{bmatrix}$$

$$\Rightarrow S_2=\begin{bmatrix}30 & 12\\12 & 14.8\end{bmatrix}$$

② Compute Sw:

$$S_W=S_1+S_2=\begin{bmatrix}20 & 10\\10 & 10.8\end{bmatrix}+\begin{bmatrix}30 & 12\\12 & 14.8\end{bmatrix}=\begin{bmatrix}50 & 22\\22 & 25.6\end{bmatrix}$$

③ Compute S_W^{-1}

$$\boxed{\begin{aligned} A &= \begin{bmatrix} a & b \\ c & d \end{bmatrix} \\ A^{-1} &= \frac{1}{|A|}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \\ &= \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \end{aligned}}$$

$$\Rightarrow \quad S_W^{-1} = \frac{1}{\left|\begin{bmatrix} 50 & 22 \\ 22 & 25.6 \end{bmatrix}\right|}\begin{bmatrix} 25.6 & -22 \\ -22 & 50 \end{bmatrix}$$

$$= \frac{1}{796}\begin{bmatrix} 25.6 & -22 \\ -22 & 50 \end{bmatrix}$$

$$= \begin{bmatrix} 0.032 & -0.03 \\ -0.03 & 0.06 \end{bmatrix}$$

④ Compute w

$$w = S_W^{-1}(\mu_1 - \mu_2)$$

$$= \begin{bmatrix} 0.032 & -0.03 \\ -0.03 & 0.06 \end{bmatrix} \cdot \left(\overset{\mu_1}{\begin{bmatrix} 4 \\ 4.2 \end{bmatrix}} - \overset{\mu_2}{\begin{bmatrix} 10 \\ 3.2 \end{bmatrix}} \right)$$

$$= \begin{bmatrix} 0.032 & -0.03 \\ -0.03 & 0.06 \end{bmatrix} \cdot \begin{bmatrix} -6 \\ 1 \end{bmatrix}$$

$$= \boxed{\begin{bmatrix} -0.222 \\ 0.24 \end{bmatrix}} \approx \begin{bmatrix} -2.2 \\ 2.4 \end{bmatrix}$$