Linear Discriminant Analysis

Component Analysis

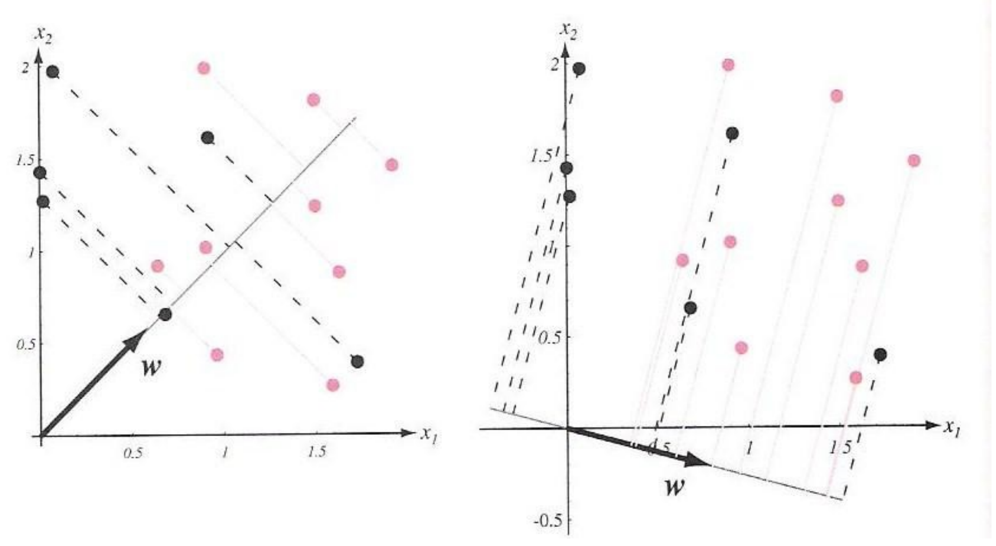
- **Component analysis is a technique that combines features to reduce the dimension of the feature space.**
- **Linear combinations are simple to compute and tractable.**
- **Project a high dimensional space onto a lower dimensional space.**
- **Classical approaches for finding the optimal transformation:**
  - **Principal Components Analysis (PCA): projection that best represents the data in a least-square sense.**
  - **Discriminant Analysis (LDA): projection that best separates the data in a least-squares sense.**

Discriminant Analysis

- **Discriminant analysis seeks directions that are efficient for discrimination.**
- **Consider the problem of projecting data from *d* dimensions onto a line with the hope that we can optimize the orientation of the line to minimize error.**
- **Consider a set of n d-dimensional samples $x_1,\ldots,x_n$ in the**
  - **subset** D1 **labeled** ω_1 **and the subset** D_2 **labeled** ω_2**.**
- **Define a transformation of x:** $y = \mathbf{w}^t \mathbf{x}$

  and a corresponding set of n **samples** $y_1, \ldots, y_n$ **divided into** Y_1 **and** Y_2**.**

- **Our challenge is to find w that maximizes separation.**
- **This can be done by considering the ratio of the between-class scatter to the within-class scatter.**



Separation of the Means and Scatter

- **Define a sample mean for class i:** $\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}$
- **The sample mean for the projected points are:**

$$\tilde{m}_i = \frac{1}{n_i} \sum_{y \in Y_i} y = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{w}^t \mathbf{x} = \mathbf{w}^t \mathbf{m}_i$$

The sample mean for the projected points is just the projection of the mean (which is expected since this is a linear transformation).

- **It follows that the distance between the projected means is:**

$$\left\| \tilde{m}_1 - \tilde{m}_2 \right\| = \left\| \mathbf{w}^t \mathbf{m}_1 - \mathbf{w}^t \mathbf{m}_2 \right\|$$

- **Define a scatter for the projected samples:**

$$\tilde{s}_i^{\,2} = \sum_{y \in Y_i} (y - \tilde{m}_i)^2$$

Fisher Linear Discriminant and Scatter

- **The Fisher linear discriminant maximizes the criteria:** $J(w) = \frac{\left\| \tilde{m}_1 - \tilde{m}_2 \right\|^2}{\tilde{s}_1^{\,2} + \tilde{s}_2^{\,2}}$
- **Define a matrix S_i for i class**

$$\mathbf{S}_i = \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$$

- **The total with class covariance matrix S_w, is:**

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

- **We can write the scatter for the projected samples as:**

$$\tilde{s}_i^{\,2} = \sum_{\mathbf{x} \in D_i} \left(\mathbf{w}^t \mathbf{x} - \mathbf{w}^t \mathbf{m}_i \right)^2$$

$$= \sum_{\mathbf{x} \in D_i} \mathbf{w}^t (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t \mathbf{w} = \mathbf{w}^t \mathbf{S}_i \mathbf{w}$$

- **Therefore, the sum of the scatters can be written as:**

$$\tilde{s}_1^{\,2} + \tilde{s}_2^{\,2} = \mathbf{w}^t \mathbf{S}_W \mathbf{w}$$

Separation of the Projected Means

- **The separation of the projected means obeys:**

$$\begin{aligned}\tilde{m}_1 - \tilde{m}_2 &= \left(\mathbf{w}^t\mathbf{m}_1 - \mathbf{w}^t\mathbf{m}_2\right)^2 \\ &= \mathbf{w}^t(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\mathbf{w} \\ &= \mathbf{w}^t\mathbf{S}_B\mathbf{w}\end{aligned}$$

- **Where the between class covaince matrix, $\mathbf{S}_\mathbf{B}$, is given by:**

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$$

- **$\mathbf{S}_\mathbf{B}$, the between-class scatter, is symmetric and positive definite**

- **The criterion function, J(w), can be written as:**

$$J(\mathbf{w}) = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}}$$

Linear Discriminant Analysis

This ratio is well-known as the generalized Rayleigh quotient and has the well-known property that the vector, w, that maximizes J(), must satisfy:
The solution is:

$$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$$

The solution is

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

This solution maps the d-dimensional problem to a one-dimensional problem (in this case).

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

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

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

$$\Rightarrow\;\left[w^T S_W w\right]2S_B w-\left[w^T S_B w\right]2S_W w=0$$

- Dividing by $w^T S_W w$

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

$$\Rightarrow\;S_B w-JS_W w=0\;\Rightarrow$$

$$\Rightarrow\;S_W^{-1}S_B w-Jw=0$$

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

$$\boxed{w^*=\underset{w}{\operatorname{argmax}}\left\{\frac{w^T S_B w}{w^T S_W w}\right\}=S_W^{-1}(\mu_1-\mu_2)}$$

  - This is know as **Fisher's Linear Discriminant** (1936), although it is not a discriminant but rather a specific choice of direction for the projection of the data down to one dimension

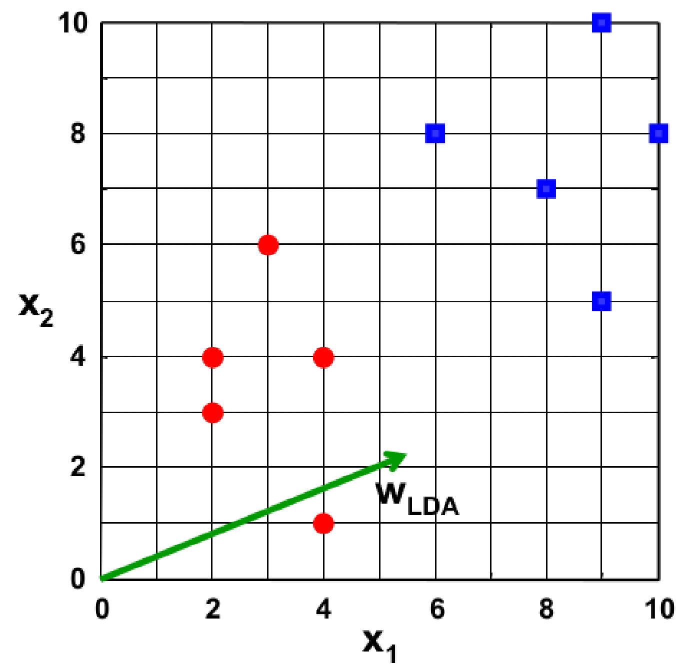
- **Compute the Linear Discriminant projection for the following two-dimensional dataset**
  - $X1=(x_1,x_2)=\{(4,1),(2,4),(2,3),(3,6),(4,4)\}$
  - $X2=(x_1,x_2)=\{(9,10),(6,8),(9,5),(8,7),(10,8)\}$
- **SOLUTION (by hand)**
  - The class statistics are:

$$S_1=\begin{bmatrix}0.80 & -0.40\\ -0.40 & 2.60\end{bmatrix};\ S_2=\begin{bmatrix}1.84 & -0.04\\ -0.04 & 2.64\end{bmatrix}$$

$$\mu_1=\begin{bmatrix}3.00 & 3.60\end{bmatrix};\qquad \mu_2=\begin{bmatrix}8.40 & 7.60\end{bmatrix}$$

  - The within- and between-class scatter are

$$S_B=\begin{bmatrix}29.16 & 21.60\\ 21.60 & 16.00\end{bmatrix};\ S_W=\begin{bmatrix}2.64 & -0.44\\ -0.44 & 5.28\end{bmatrix}$$



  - The LDA projection is then obtained as the solution of the generalized eigenvalue problem

$$S_W^{-1}S_B v=\lambda v\Rightarrow\left|S_W^{-1}S_B-\lambda I\right|=0\Rightarrow\begin{vmatrix}11.89-\lambda & 8.81\\ 5.08 & 3.76-\lambda\end{vmatrix}=0\Rightarrow\lambda=15.65$$

$$\begin{bmatrix}11.89 & 8.81\\ 5.08 & 3.76\end{bmatrix}\begin{bmatrix}v_1\\ v_2\end{bmatrix}=15.65\begin{bmatrix}v_1\\ v_2\end{bmatrix}\Rightarrow\begin{bmatrix}v_1\\ v_2\end{bmatrix}=\begin{bmatrix}0.91\\ 0.39\end{bmatrix}$$

  - Or directly by

$$w^*=S_W^{-1}(\mu_1-\mu_2)=\begin{bmatrix}-0.91 & -0.39\end{bmatrix}^T$$

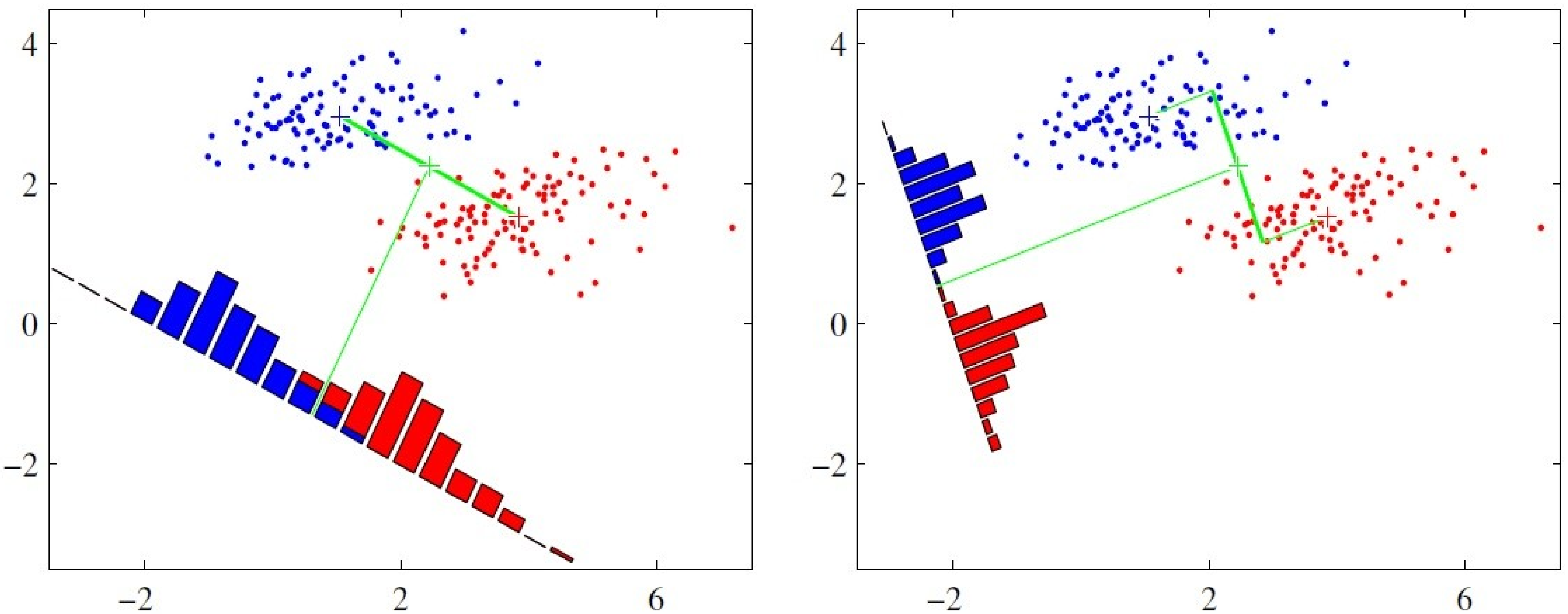


Figure — The left plot shows samples from two classes (depicted in red and blue) along with the histograms resulting from projection onto the line joining the class means. Note that there is considerable class overlap in the projected space. The right plot shows the corresponding projection based on the Fisher linear discriminant, showing the greatly improved class separation.