Data Mining and Machine Learning: Fundamental Concepts and Algorithms

dataminingbook.info

Mohammed J. Zaki[1] Wagner Meira Jr.[2]

[1]Department of Computer Science
Rensselaer Polytechnic Institute, Troy, NY, USA

[2]Department of Computer Science
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Chap. 20: Linear Discriminant Analysis

Linear Discriminant Analysis

Given labeled data consisting of d-dimensional points $\boldsymbol{x}_i$ along with their classes y_i, the goal of linear discriminant analysis (LDA) is to find a vector $\boldsymbol{w}$ that maximizes the separation between the classes after projection onto $\boldsymbol{w}$.

The key difference between principal component analysis and LDA is that the former deals with unlabeled data and tries to maximize variance, whereas the latter deals with labeled data and tries to maximize the discrimination between the classes.

Projection onto a Line

Let $\boldsymbol{D}_i$ denote the subset of points labeled with class c_i, i.e., $\boldsymbol{D}_i = \{\boldsymbol{x}_j | y_j = c_i\}$, and let $|\boldsymbol{D}_i| = n_i$ denote the number of points with class c_i. We assume that there are only $k = 2$ classes.

The projection of any d-dimensional point $\boldsymbol{x}_i$ onto a unit vector $\boldsymbol{w}$ is given as

$$\boldsymbol{x}_i' = \left(\frac{\boldsymbol{w}^T \boldsymbol{x}_i}{\boldsymbol{w}^T \boldsymbol{w}}\right) \boldsymbol{w} = \left(\boldsymbol{w}^T \boldsymbol{x}_i\right) \boldsymbol{w} = a_i \boldsymbol{w}$$

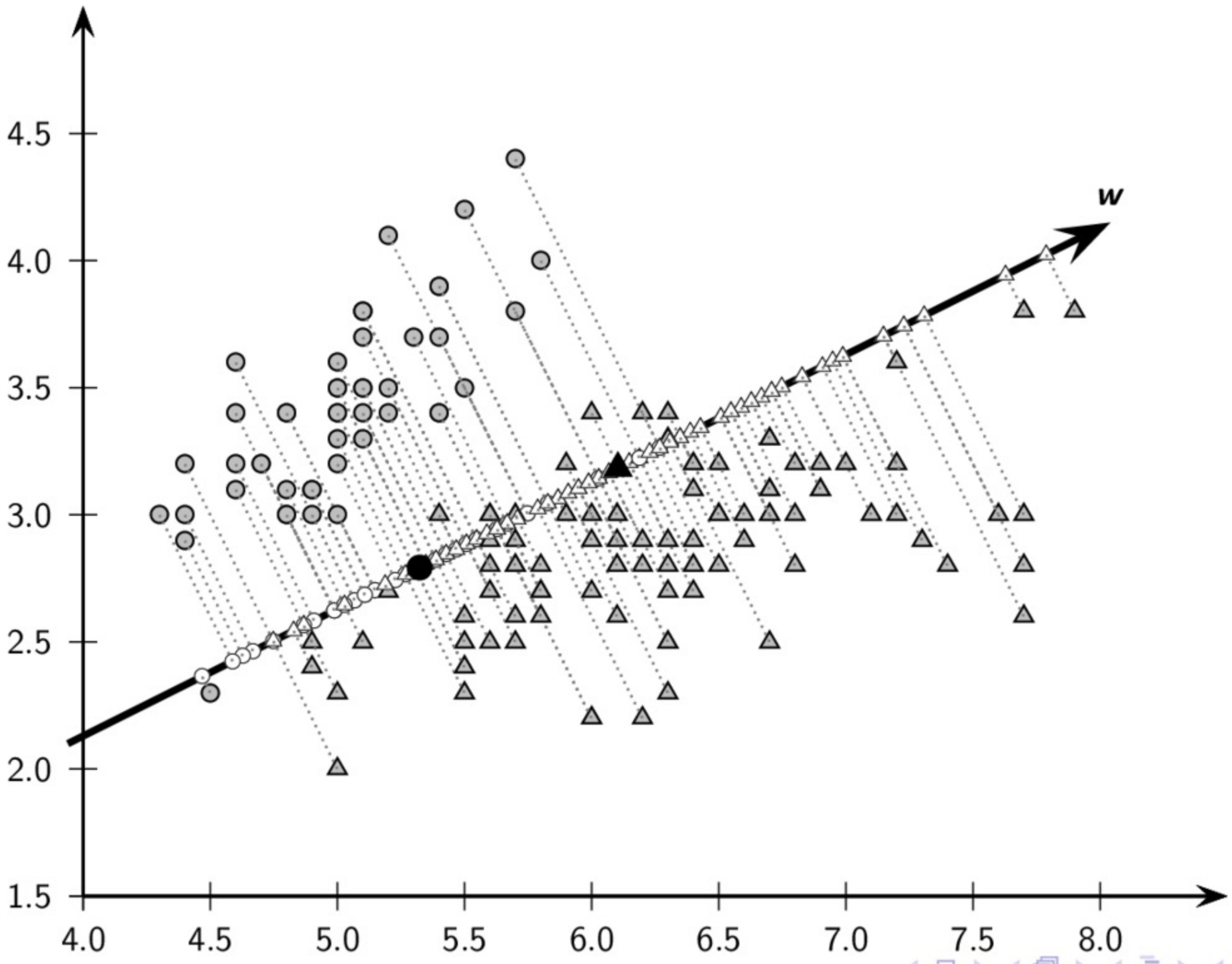
where a_i specifies the offset or coordinate of $\boldsymbol{x}_i'$ along the line $\boldsymbol{w}$:

$$a_i = \boldsymbol{w}^T \boldsymbol{x}_i$$

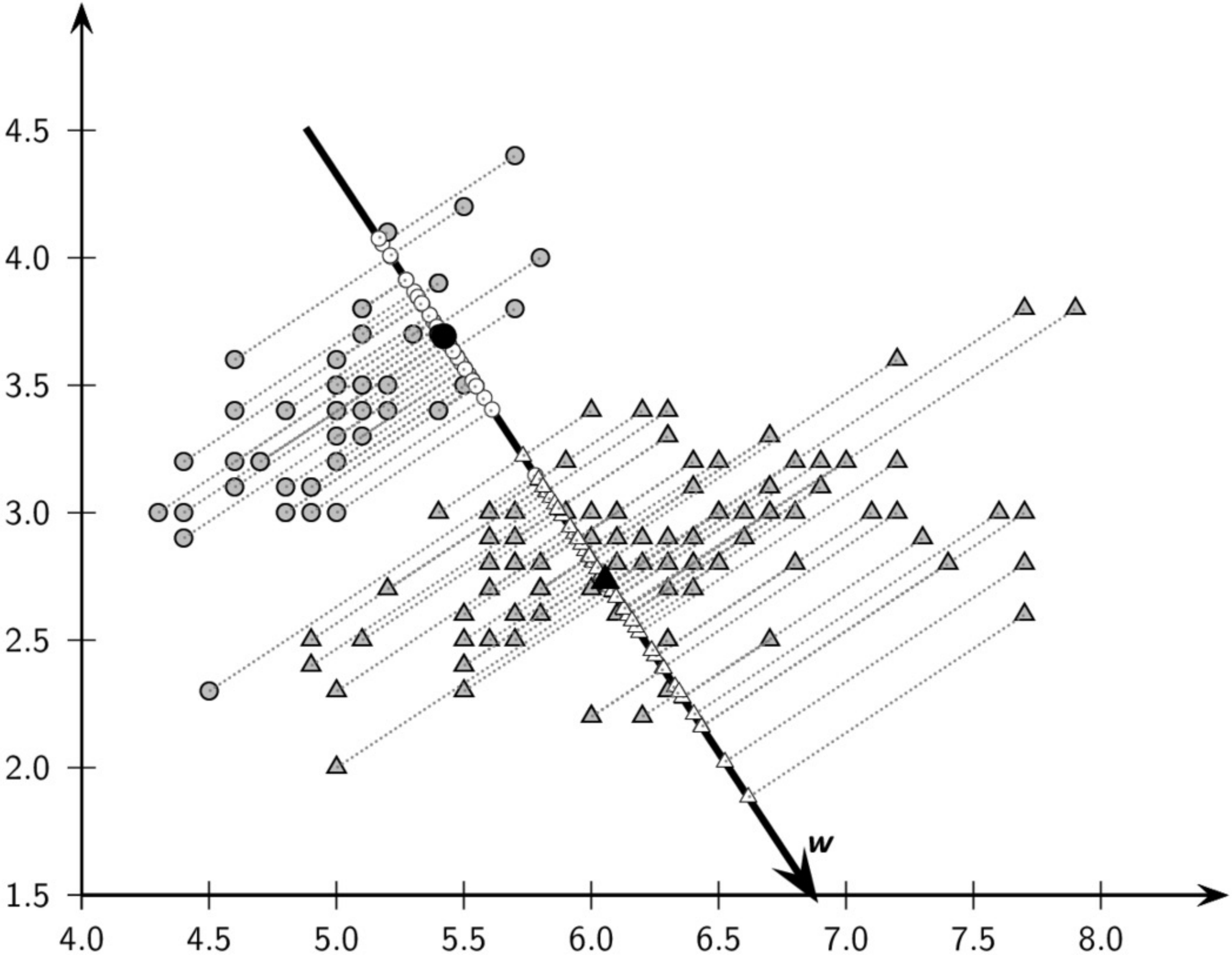
The set of n scalars $\{a_1, a_2, \ldots, a_n\}$ represents the mapping from $\mathbb{R}^d$ to $\mathbb{R}$, that is, from the original d-dimensional space to a 1-dimensional space (along $\boldsymbol{w}$).

Projection onto $\boldsymbol{w}$: Iris 2D Data

`iris-setosa` as class c_1 (circles), and the other two Iris types as class c_2 (triangles)



Iris 2D Data: Optimal Linear Discriminant Direction



Optimal Linear Discriminant

The mean of the projected points is given as:

$$m_1 = \mathbf{w}^T \boldsymbol{\mu}_1 \qquad\qquad m_2 = \mathbf{w}^T \boldsymbol{\mu}_2$$

To maximize the separation between the classes, we maximize the difference between the projected means, $|m_1 - m_2|$. However, for good separation, the variance of the projected points for each class should also not be too large. LDA maximizes the separation by ensuring that the *scatter* s_i^2 for the projected points within each class is small, where scatter is defined as

$$s_i^2 = \sum_{\mathbf{x}_j \in \mathbf{D}_i} (a_j - m_i)^2 = n_i \sigma_i^2$$

where σ_i^2 is the variance for class c_i.

Linear Discriminant Analysis: Fisher Objective

We incorporate the two LDA criteria, namely, maximizing the distance between projected means and minimizing the sum of projected scatter, into a single maximization criterion called the *Fisher LDA objective*:

$$\boxed{\max_{\boldsymbol{w}} \; J(\boldsymbol{w}) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}}$$

In matrix terms, we can rewrite $(m_1 - m_2)^2$ as follows:

$$(m_1 - m_2)^2 = \left(\boldsymbol{w}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right)^2 = \boldsymbol{w}^T\boldsymbol{B}\boldsymbol{w}$$

where $\boldsymbol{B} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$ is a $d \times d$ rank-one matrix called the *between-class scatter matrix*.

The projected scatter for class c_i is given as

$$s_i^2 = \sum_{\boldsymbol{x}_j \in \boldsymbol{D_1}} (\boldsymbol{w}^T\boldsymbol{x}_j - \boldsymbol{w}^T\boldsymbol{\mu}_i)^2 = \boldsymbol{w}^T\left(\sum_{\boldsymbol{x}_j \in \boldsymbol{D}_i} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T\right)\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{S}_i\boldsymbol{w}$$

where $\boldsymbol{S}_i$ is the *scatter matrix* for $\boldsymbol{D}_i$.

Linear Discriminant Analysis: Fisher Objective

The combined scatter for both classes is given as

$$s_1^2 + s_2^2 = \boldsymbol{w}^T \boldsymbol{S}_1 \boldsymbol{w} + \boldsymbol{w}^T \boldsymbol{S}_2 \boldsymbol{w} = \boldsymbol{w}^T (\boldsymbol{S}_1 + \boldsymbol{S}_2) \boldsymbol{w} = \boldsymbol{w}^T \boldsymbol{S} \boldsymbol{w}$$

where the symmetric positive semidefinite matrix $\boldsymbol{S} = \boldsymbol{S}_1 + \boldsymbol{S}_2$ denotes the *within-class scatter matrix* for the pooled data.
The LDA objective function in matrix form is

$$\boxed{\max_{\boldsymbol{w}} \; J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{B} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S} \boldsymbol{w}}}$$

To solve for the best direction $\boldsymbol{w}$, we differentiate the objective function with respect to $\boldsymbol{w}$; after simplification it yields the *generalized eigenvalue problem*

$$\boldsymbol{B} \boldsymbol{w} = \lambda \boldsymbol{S} \boldsymbol{w}$$

where $\lambda = J(\boldsymbol{w})$ is a generalized eigenvalue of $\boldsymbol{B}$ and $\boldsymbol{S}$. To maximize the objective λ should be chosen to be the largest generalized eigenvalue, and $\boldsymbol{w}$ to be the corresponding eigenvector.

Recall that if $f(x)$ and $g(x)$ are two functions then we have

$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right)=\frac{f'(x)g(x)-g'(x)f(x)}{g(x)^2}$$

$$\frac{d}{d\mathbf{w}}J(\mathbf{w})=\frac{2\mathbf{Bw}(\mathbf{w}^T\mathbf{Sw})-2\mathbf{Sw}(\mathbf{w}^T\mathbf{Bw})}{(\mathbf{w}^T\mathbf{Sw})^2}=\mathbf{0}$$

which yields

$$\mathbf{B}\,\mathbf{w}(\mathbf{w}^T\mathbf{Sw})=\mathbf{S}\,\mathbf{w}(\mathbf{w}^T\mathbf{Bw})$$

$$\mathbf{B}\,\mathbf{w}=\mathbf{S}\,\mathbf{w}\left(\frac{\mathbf{w}^T\mathbf{Bw}}{\mathbf{w}^T\mathbf{Sw}}\right)$$

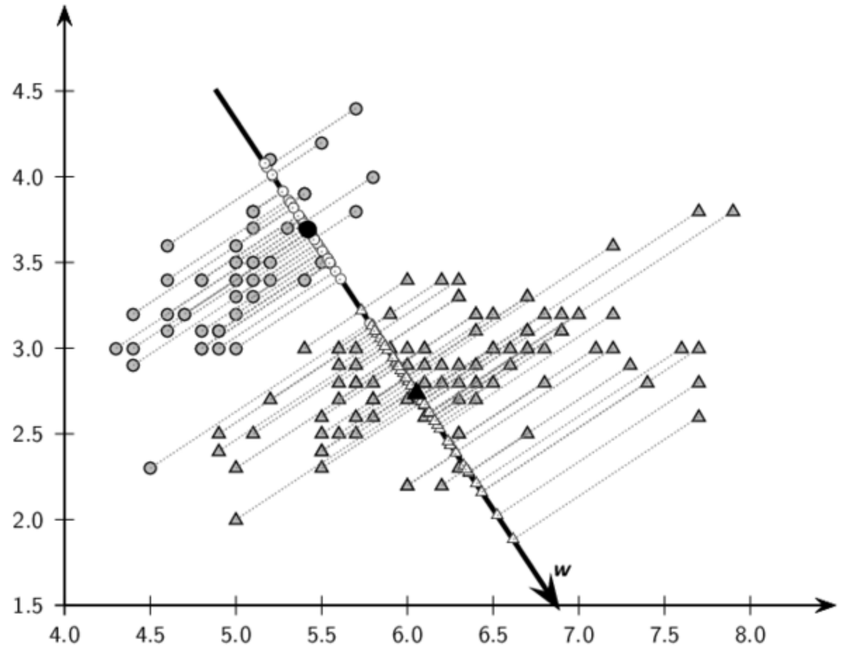
$$\mathbf{B}\,\mathbf{w}=J(\mathbf{w})\mathbf{Sw}$$

$$\mathbf{Bw}=\lambda\mathbf{Sw}$$

Linear Discriminant Algorithm

```
**LinearDiscriminant** ($\boldsymbol{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$):
$\boldsymbol{D}_i \leftarrow \left\{\boldsymbol{x}_j \mid y_j = c_i, j = 1, \dots, n\right\}, i = 1,2$ // class-specific subsets
$\boldsymbol{\mu}_i \leftarrow \text{mean}(\boldsymbol{D}_i), i = 1,2$ // class means
$\boldsymbol{B} \leftarrow (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$ // between-class scatter matrix
$\boldsymbol{Z}_i \leftarrow \boldsymbol{D}_i - 1_{n_i} \boldsymbol{\mu}_i^T, i = 1,2$ // center class matrices
$\boldsymbol{S}_i \leftarrow \boldsymbol{Z}_i^T \boldsymbol{Z}_i, i = 1,2$ // class scatter matrices
$\boldsymbol{S} \leftarrow \boldsymbol{S}_1 + \boldsymbol{S}_2$ // within-class scatter matrix
$\lambda_1, \boldsymbol{w} \leftarrow \text{eigen}(\boldsymbol{S}^{-1}\boldsymbol{B})$ // compute dominant eigenvector
```

Linear Discriminant Direction: Iris 2D Data



The between-class scatter matrix is

$$\boldsymbol{B} = \begin{pmatrix} 1.587 & -0.693 \\ -0.693 & 0.303 \end{pmatrix}$$

and the within-class scatter matrix is

$$\boldsymbol{S}_1 = \begin{pmatrix} 6.09 & 4.91 \\ 4.91 & 7.11 \end{pmatrix}$$

$$\boldsymbol{S}_2 = \begin{pmatrix} 43.5 & 12.09 \\ 12.09 & 10.96 \end{pmatrix}$$

$$\boldsymbol{S} = \begin{pmatrix} 49.58 & 17.01 \\ 17.01 & 18.08 \end{pmatrix}$$

The direction of most separation between c_1 and c_2 is the dominant eigenvector corresponding to the largest eigenvalue of the matrix $\boldsymbol{S}^{-1}\boldsymbol{B}$. The solution is

$$J(\boldsymbol{w}) = \lambda_1 = 0.11$$

$$\boldsymbol{w} = \begin{pmatrix} 0.551 \\ -0.834 \end{pmatrix}$$

Linear Discriminant Analysis: Two Classes

For the two class scenario, if $\boldsymbol{S}$ is nonsingular, we can directly solve for $\boldsymbol{w}$ without computing the eigenvalues and eigenvectors.

The between-class scatter matrix $\boldsymbol{B}$ points in the same direction as $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ because

$$
\begin{aligned}
\boldsymbol{B}\boldsymbol{w} =& \left((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\right)\boldsymbol{w} \\
=& (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\left((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{w}\right) \\
=& b(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)
\end{aligned}
$$

The generalized eigenvectors equation can then be rewritten as

$$\boldsymbol{w} = \frac{b}{\lambda}\boldsymbol{S}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

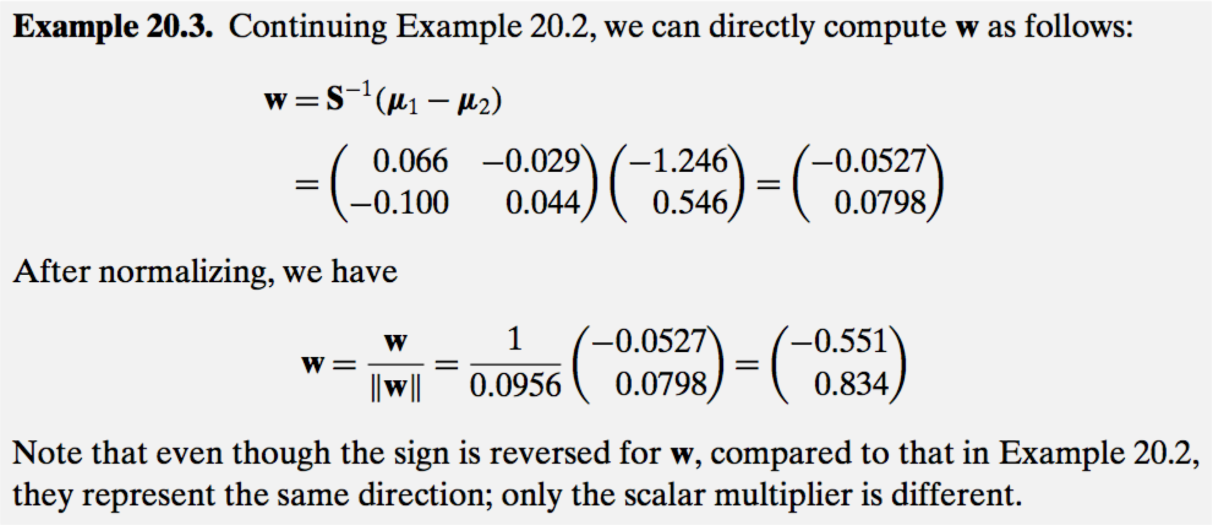
Because $\frac{b}{\lambda}$ is just a scalar, we can solve for the best linear discriminant as

$$\boxed{\boldsymbol{w} = \boldsymbol{S}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}$$

We can finally normalize $\boldsymbol{w}$ to be a unit vector.

$$\begin{aligned} \mathbf{B}\mathbf{w} &= \lambda \mathbf{S}\mathbf{w} \\ \mathbf{S}^{-1}\mathbf{B}\mathbf{w} &= \lambda \mathbf{S}^{-1}\mathbf{S}\mathbf{w} \\ (\mathbf{S}^{-1}\mathbf{B})\mathbf{w} &= \lambda \mathbf{w} \end{aligned} \tag{20.9}$$

Thus, if $\mathbf{S}^{-1}$ exists, then $\lambda = J(\mathbf{w})$ is an eigenvalue, and $\mathbf{w}$ is an eigenvector of the matrix $\mathbf{S}^{-1}\mathbf{B}$. To maximize $J(\mathbf{w})$ we look for the largest eigenvalue λ, and the corresponding dominant eigenvector $\mathbf{w}$ specifies the best linear discriminant vector.



Example 20.3. Continuing Example 20.2, we can directly compute $\mathbf{w}$ as follows:

$$\begin{aligned}\mathbf{w} &= \mathbf{S}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\\ &= \begin{pmatrix} 0.066 & -0.029 \\ -0.100 & 0.044 \end{pmatrix} \begin{pmatrix} -1.246 \\ 0.546 \end{pmatrix} = \begin{pmatrix} -0.0527 \\ 0.0798 \end{pmatrix}\end{aligned}$$

After normalizing, we have

$$\mathbf{w} = \frac{\mathbf{w}}{\|\mathbf{w}\|} = \frac{1}{0.0956} \begin{pmatrix} -0.0527 \\ 0.0798 \end{pmatrix} = \begin{pmatrix} -0.551 \\ 0.834 \end{pmatrix}$$

Note that even though the sign is reversed for $\mathbf{w}$, compared to that in Example 20.2, they represent the same direction; only the scalar multiplier is different.

Classification from LDA

- Training
  - training data $\{x_1,\ldots,x_n\}$
  - Compute best linear discriminant vector w
- Classifier
  - Test point x: project into direction w
    - coordinate $a = w^T x$
  - Test according to threshold
    - e.g. compare with average of projected class means

      $m = ½(w^T \mu_1 + w^T \mu_2)$
    - IF(a < m) THEN (class 1) ELSE (class 2)

Linear Discriminant Direction: Iris 2D Data

We can directly compute $\boldsymbol{w}$ as follows:

$$\begin{aligned}\boldsymbol{w} &= \boldsymbol{S}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\\ &= \begin{pmatrix} 0.066 & -0.029 \\ -0.100 & 0.044 \end{pmatrix} \begin{pmatrix} -1.246 \\ 0.546 \end{pmatrix} = \begin{pmatrix} -0.0527 \\ 0.0798 \end{pmatrix}\end{aligned}$$

After normalizing, we have

$$\boldsymbol{w} = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} = \frac{1}{0.0956} \begin{pmatrix} -0.0527 \\ 0.0798 \end{pmatrix} = \begin{pmatrix} -0.551 \\ 0.834 \end{pmatrix}$$

Note that even though the sign is reversed for $\boldsymbol{w}$, they represent the same direction; only the scalar multiplier is different.