# 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

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

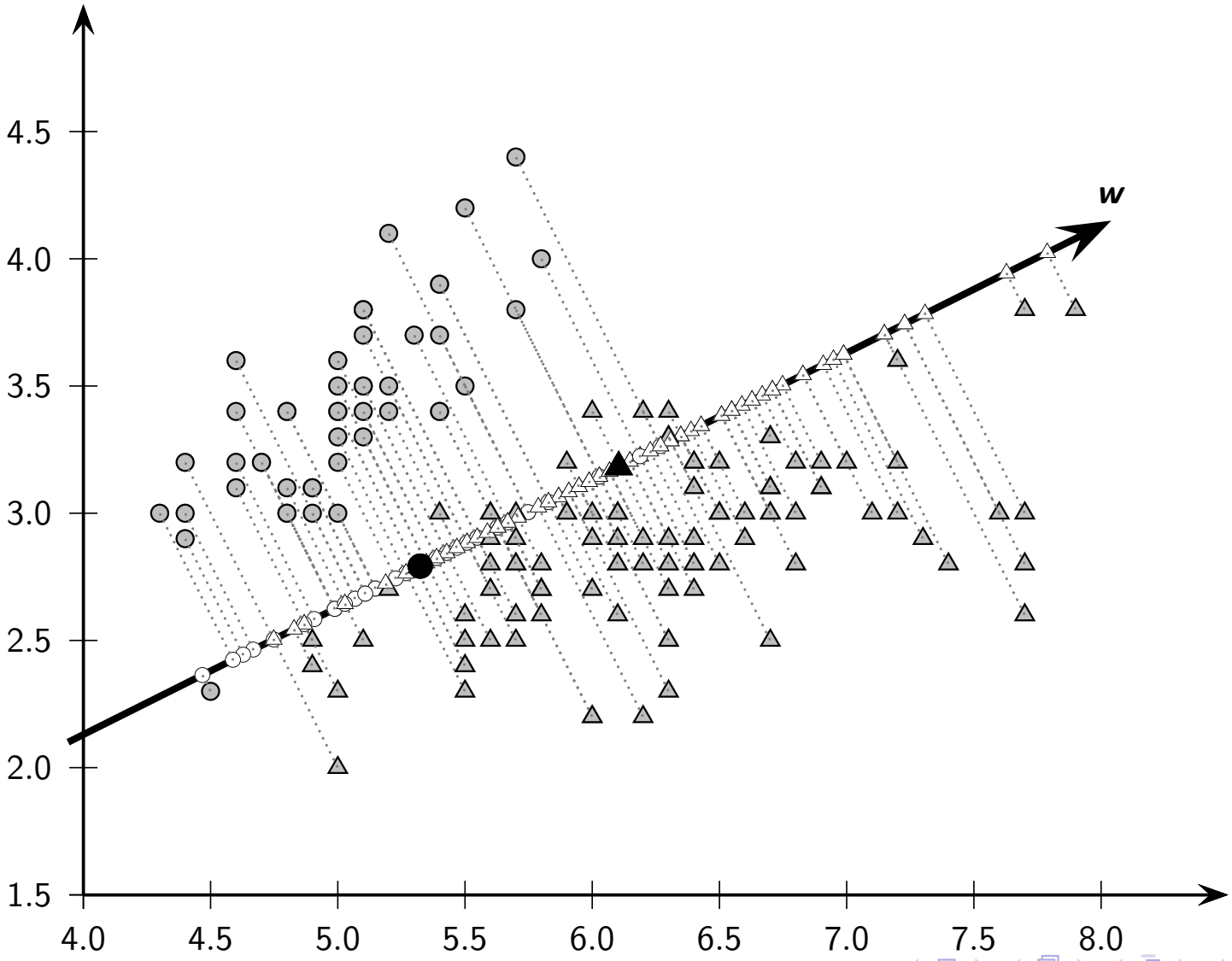
where $a_i$ specifies the offset or coordinate of $\boldsymbol{x}_i'$ along the line $\boldsymbol{w}$:

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

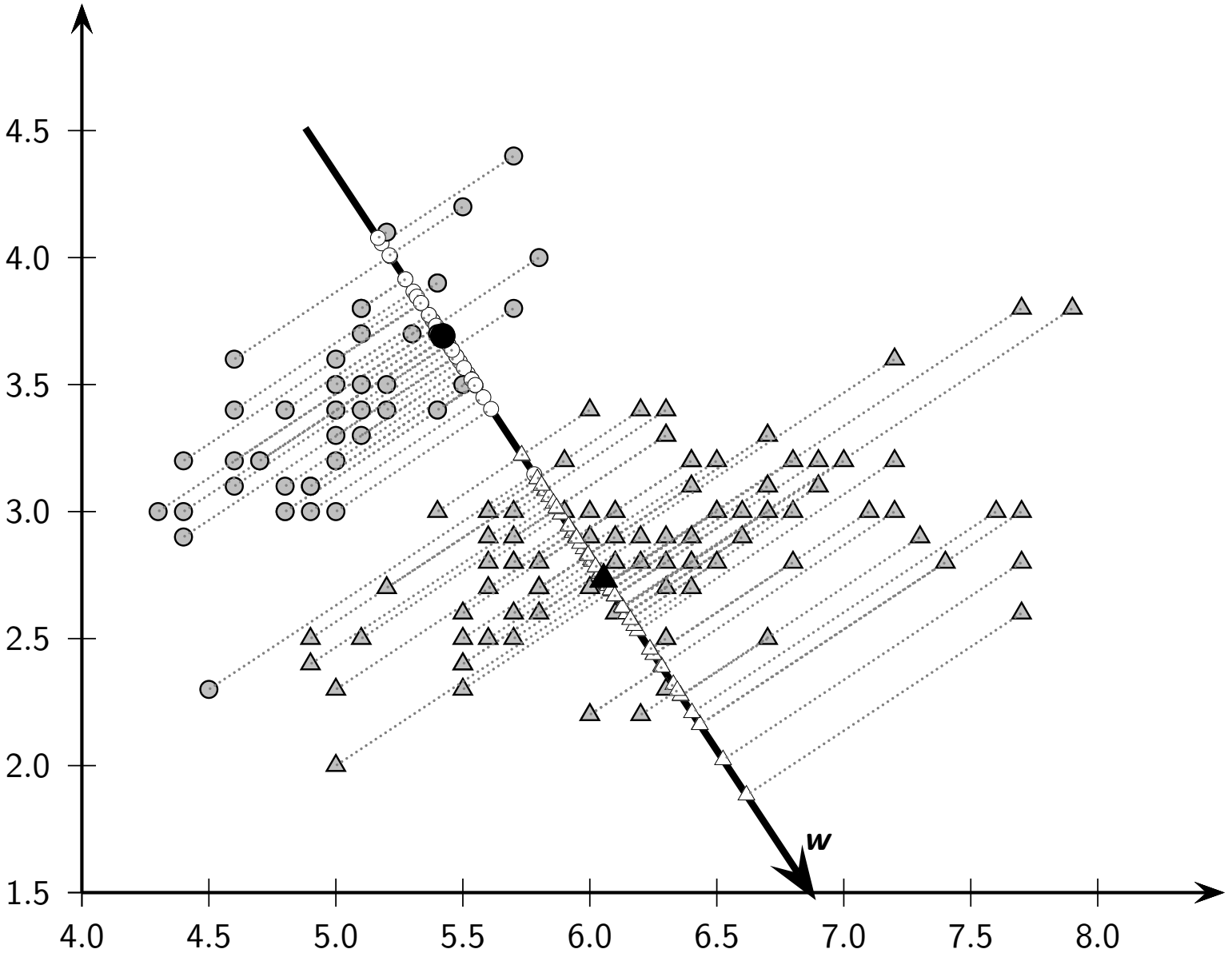
The set of $n$ scalars $\{a_1, a_2, \ldots, a_n\}$ represents the mapping from $\mathbb{R}^d$ to $\mathbb{R}$, that is, from the original $d$-dimensional space to a 1-dimensional space (along $\boldsymbol{w}$).

# 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:

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

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

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

where $\sigma_i^2$ 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*:

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

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

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

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 $\lambda$ should be chosen to be the largest generalized eigenvalue, and $\boldsymbol{w}$ to be the corresponding eigenvector.

# Linear Discriminant Algorithm

**LinearDiscriminant** $(\boldsymbol{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n)$**:**

1 $\boldsymbol{D}_i \leftarrow \{\boldsymbol{x}_j \mid y_j = c_i, j = 1, \ldots, n\}, i = 1, 2$ // class-specific subsets

2 $\boldsymbol{\mu}_i \leftarrow \text{mean}(\boldsymbol{D}_i), i = 1, 2$ // class means

3 $\boldsymbol{B} \leftarrow (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$ // between-class scatter matrix

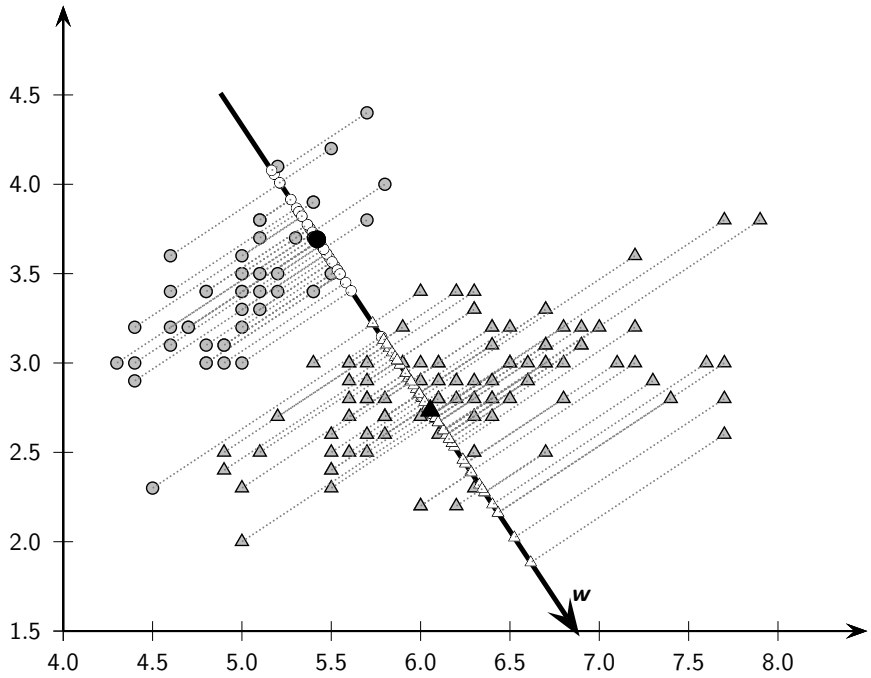
4 $\boldsymbol{Z}_i \leftarrow \boldsymbol{D}_i - 1_{n_i} \boldsymbol{\mu}_i^T, i = 1, 2$ // center class matrices

5 $\boldsymbol{S}_i \leftarrow \boldsymbol{Z}_i^T \boldsymbol{Z}_i, i = 1, 2$ // class scatter matrices

6 $\boldsymbol{S} \leftarrow \boldsymbol{S}_1 + \boldsymbol{S}_2$ // within-class scatter matrix

7 $\lambda_1, \boldsymbol{w} \leftarrow \text{eigen}(\boldsymbol{S}^{-1}\boldsymbol{B})$ // 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

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

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.

# 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.

# Kernel Discriminant Analysis

The goal of kernel LDA is to find the direction vector $\boldsymbol{w}$ in feature space that maximizes

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

It is well known that $\boldsymbol{w}$ can be expressed as a linear combination of the points in feature space

$$\boldsymbol{w} = \sum_{j=1}^{n} a_j \phi(\boldsymbol{x}_j)$$

The mean for class $c_i$ in feature space is given as

$$\boldsymbol{\mu}_i^{\phi} = \frac{1}{n_i} \sum_{\boldsymbol{x}_j \in \boldsymbol{D}_i} \phi(\boldsymbol{x}_j)$$

and the covariance matrix for class $c_i$ in feature space is

$$\Sigma_i^{\phi} = \frac{1}{n_i} \sum_{\boldsymbol{x}_j \in \boldsymbol{D}_i} \left(\phi(\boldsymbol{x}_j) - \boldsymbol{\mu}_i^{\phi}\right)\left(\phi(\boldsymbol{x}_j) - \boldsymbol{\mu}_i^{\phi}\right)^T$$

# Kernel Discriminant Analysis

The between-class scatter matrix in feature space is

$$\boldsymbol{B}_\phi = \left(\boldsymbol{\mu}_1^\phi - \boldsymbol{\mu}_2^\phi\right)\left(\boldsymbol{\mu}_1^\phi - \boldsymbol{\mu}_2^\phi\right)^T$$

and the within-class scatter matrix in feature space is

$$\boldsymbol{S}_\phi = n_1\Sigma_1^\phi + n_2\Sigma_2^\phi$$

$\boldsymbol{S}_\phi$ is a $d \times d$ symmetric, positive semidefinite matrix, where $d$ is the dimensionality of the feature space.

The best linear discriminant vector $\boldsymbol{w}$ in feature space is the dominant eigenvector, which satisfies the expression

$$\left(\boldsymbol{S}_\phi^{-1}\boldsymbol{B}_\phi\right)\boldsymbol{w} = \lambda\boldsymbol{w}$$

where we assume that $\boldsymbol{S}_\phi$ is non-singular.

# LDA Objective via Kernel Matrix: Between-class Scatter

The projected mean for class $c_i$ is given as

$$m_i = \boldsymbol{w}^T \boldsymbol{\mu}_i^{\phi} = \frac{1}{n_i} \sum_{j=1}^{n} \sum_{\boldsymbol{x}_k \in \boldsymbol{D}_i} a_j K(\boldsymbol{x}_j, \boldsymbol{x}_k) = \boldsymbol{a}^T \boldsymbol{m}_i$$

where $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)^T$ is the weight vector, and

$$\boldsymbol{m}_i = \frac{1}{n_i} \begin{pmatrix} \sum_{\boldsymbol{x}_k \in \boldsymbol{D}_i} K(\boldsymbol{x}_1, \boldsymbol{x}_k) \\ \sum_{\boldsymbol{x}_k \in \boldsymbol{D}_i} K(\boldsymbol{x}_2, \boldsymbol{x}_k) \\ \vdots \\ \sum_{\boldsymbol{x}_k \in \boldsymbol{D}_i} K(\boldsymbol{x}_n, \boldsymbol{x}_k) \end{pmatrix} = \frac{1}{n_i} \boldsymbol{K}^{c_i} 1_{n_i}$$

where $\boldsymbol{K}^{c_i}$ is the $n \times n_i$ subset of the kernel matrix, restricted to columns belonging to points only in $\boldsymbol{D}_i$, and $1_{n_i}$ is the $n_i$-dimensional vector all of whose entries are one.

The separation between the projected means is thus

$$(m_1 - m_2)^2 = \left(\boldsymbol{a}^T \boldsymbol{m}_1 - \boldsymbol{a}^T \boldsymbol{m}_2\right)^2 = \boldsymbol{a}^T \boldsymbol{M} \boldsymbol{a}$$

where $\boldsymbol{M} = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T$ is the between-class scatter matrix.

# LDA Objective via Kernel Matrix: Within-class Scatter

We can compute the projected scatter for each class, $s_1^2$ and $s_2^2$, purely in terms of the kernel function, as follows

$$s_1^2 = \sum_{\boldsymbol{x}_i \in \boldsymbol{D}_1} \left\| \boldsymbol{w}^T \phi(\boldsymbol{x}_i) - \boldsymbol{w}^T \boldsymbol{\mu}_1^{\phi} \right\|^2 = \boldsymbol{a}^T \left( \left( \sum_{\boldsymbol{x}_i \in \boldsymbol{D}_1} \boldsymbol{K}_i \boldsymbol{K}_i^T \right) - n_1 \boldsymbol{m}_1 \boldsymbol{m}_1^T \right) \boldsymbol{a} = \boldsymbol{a}^T \boldsymbol{N}_1 \boldsymbol{a}$$

where $\boldsymbol{K}_i$ is the $i$th column of the kernel matrix, and $\boldsymbol{N}_1$ is the class scatter matrix for $c_1$.

The sum of projected scatter values is then given as

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

where $\boldsymbol{N}$ is the $n \times n$ within-class scatter matrix.

# Kernel LDA

The kernel LDA maximization condition is

$$\max_{\boldsymbol{w}} J(\boldsymbol{w}) = \max_{\boldsymbol{a}} J(\boldsymbol{a}) = \frac{\boldsymbol{a}^T \boldsymbol{M} \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{N} \boldsymbol{a}}$$

The weight vector $\boldsymbol{a}$ is the eigenvector corresponding to the largest eigenvalue of the generalized eigenvalue problem:

$$\boldsymbol{M}\boldsymbol{a} = \lambda_1 \boldsymbol{N}\boldsymbol{a}$$

When there are only two classes $\boldsymbol{a}$ can be obtained directly:

$$\boldsymbol{a} = \boldsymbol{N}^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)$$

To normalize $\boldsymbol{w}$ to be a unit vector we scale $\boldsymbol{a}$ by $\frac{1}{\sqrt{\boldsymbol{a}^T \boldsymbol{K} \boldsymbol{a}}}$.

We can project any point $\boldsymbol{x}$ onto the discriminant direction as follows:

$$\boldsymbol{w}^T \phi(\boldsymbol{x}) = \sum_{j=1}^{n} a_j \phi(\boldsymbol{x}_j)^T \phi(\boldsymbol{x}) = \sum_{j=1}^{n} a_j K(\boldsymbol{x}_j, \boldsymbol{x})$$

# Kernel Discriminant Analysis Algorithm

**KernelDiscriminant** $(\boldsymbol{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n, K)$**:**

1 $\boldsymbol{K} \leftarrow \left\{K(\boldsymbol{x}_i, \boldsymbol{x}_j)\right\}_{i,j=1,\ldots,n}$ // compute $n \times n$ kernel matrix

2 $\boldsymbol{K}^{c_i} \leftarrow \left\{\boldsymbol{K}(j,k) \mid y_k = c_i, 1 \leq j,k \leq n\right\}, i = 1,2$ // class kernel matrix

3 $\boldsymbol{m}_i \leftarrow \frac{1}{n_i}\boldsymbol{K}^{c_i}\mathbf{1}_{n_i}, i = 1,2$ // class means

4 $\boldsymbol{M} \leftarrow (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T$ // between-class scatter matrix

5 $\boldsymbol{N}_i \leftarrow \boldsymbol{K}^{c_i}(\boldsymbol{I}_{n_i} - \frac{1}{n_i}\mathbf{1}_{n_i \times n_i})(\boldsymbol{K}^{c_i})^T,\ i = 1,2$ // class scatter matrices

6 $\boldsymbol{N} \leftarrow \boldsymbol{N}_1 + \boldsymbol{N}_2$ // within-class scatter matrix

7 $\lambda_1, \boldsymbol{a} \leftarrow \text{eigen}(\boldsymbol{N}^{-1}\boldsymbol{M})$ // compute weight vector

8 $\boldsymbol{a} \leftarrow \frac{\boldsymbol{a}}{\sqrt{\boldsymbol{a}^T\boldsymbol{K}\boldsymbol{a}}}$ // normalize $\boldsymbol{w}$ to be unit vector

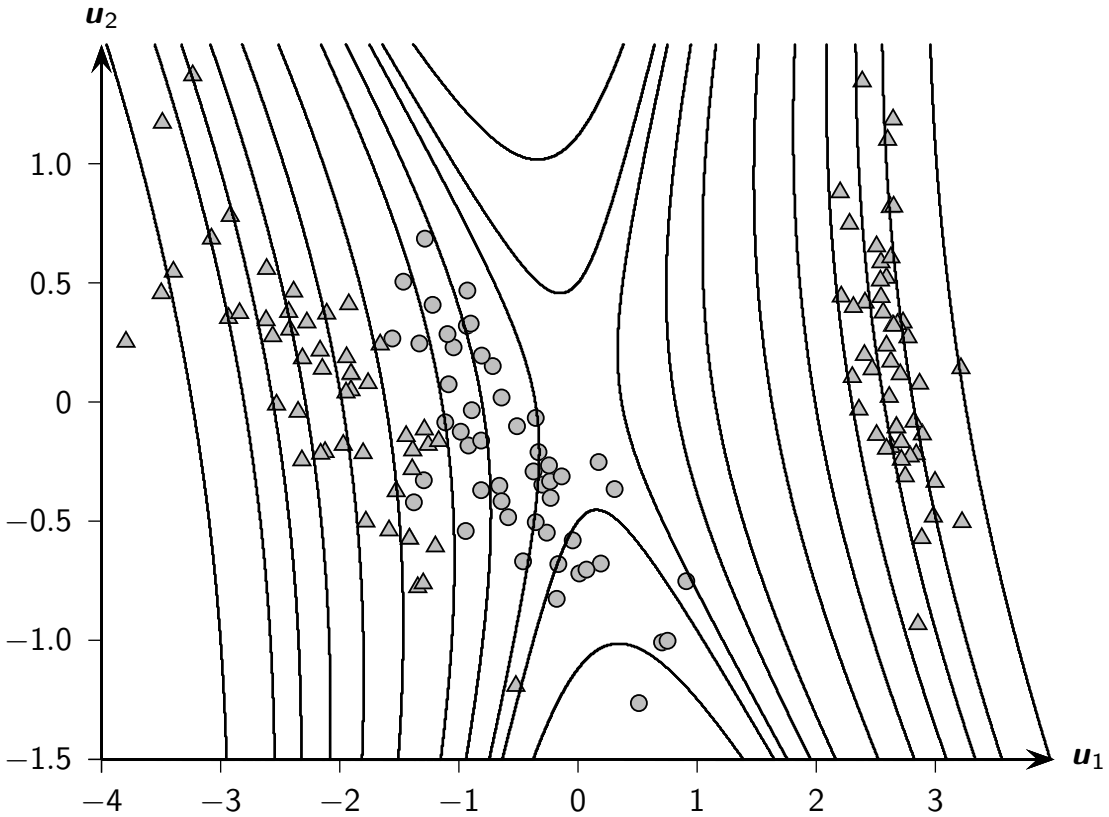
# Kernel Discriminant Analysis

## Quadratic Homogeneous Kernel

Iris 2D Data: $c_1$ (circles; `iris-virginica`) and $c_2$ (triangles; other two types).

Kernel Function: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T \boldsymbol{x}_j)^2$.

Contours of constant projection onto optimal kernel discriminant, i.e., points along both the curves have the same value when projected onto $\boldsymbol{w}$.



Projecting $\boldsymbol{x}_i \in \boldsymbol{D}$ onto $\boldsymbol{w}$, which separates the two classes.

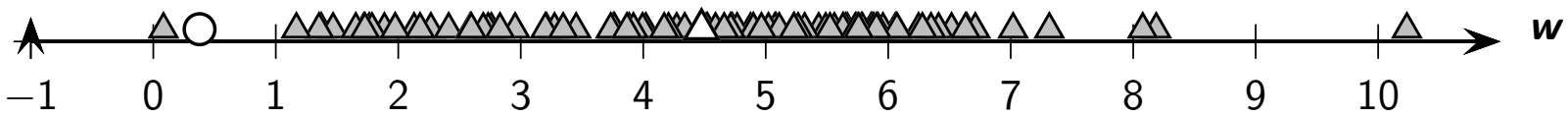
The projected scatters and means for both classes are as follows:

$$m_1 = 0.338 \qquad m_2 = 4.476$$
$$s_1^2 = 13.862 \qquad s_2^2 = 320.934$$

The value of $J(\boldsymbol{w})$ is given as

$$J(\boldsymbol{w}) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2} = \frac{(0.338 - 4.476)^2}{13.862 + 320.934} = \frac{17.123}{334.796} = 0.0511$$

which, as expected, matches $\lambda_1 = 0.0511$ from above.



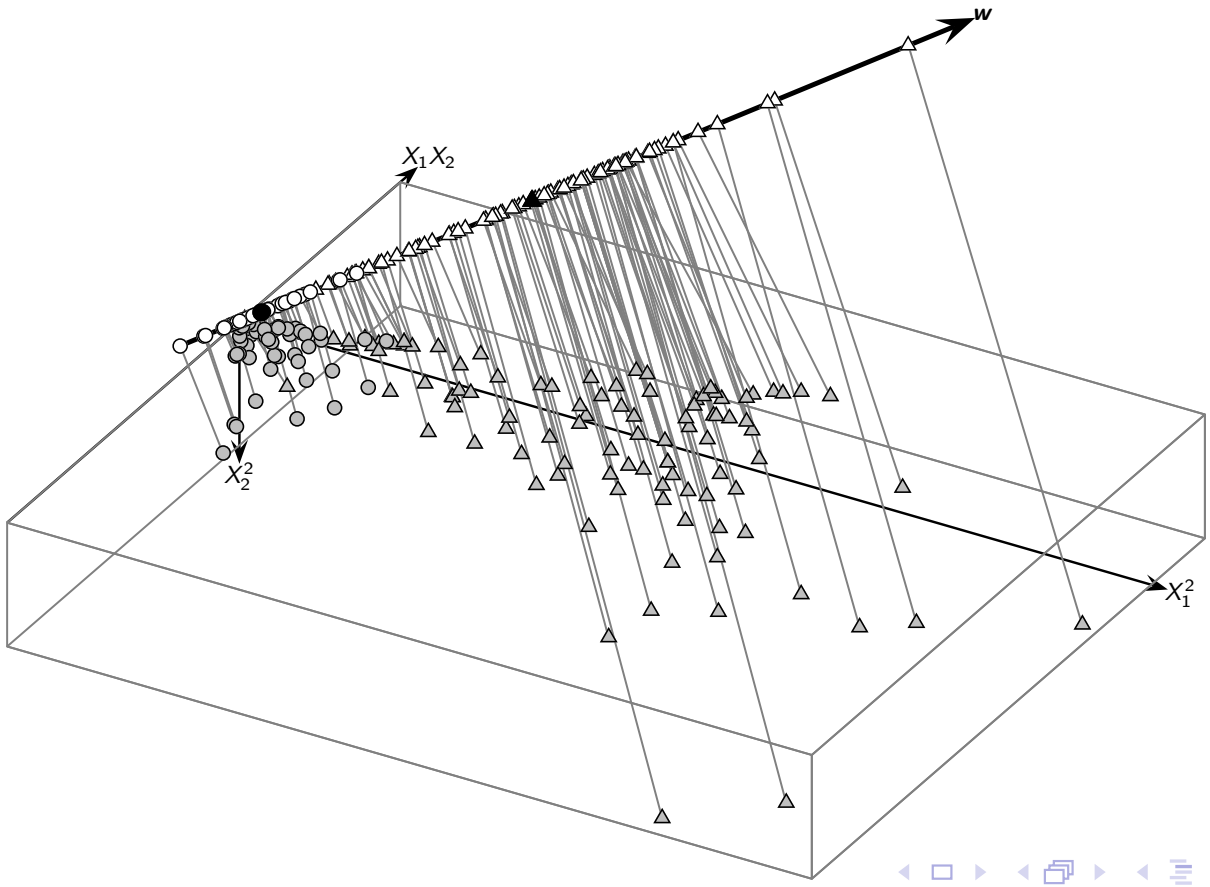
# Kernel Feature Space and Optimal Discriminant

It is not desirable or possible to obtain an explicit discriminant vector $\boldsymbol{w}$.

$\boldsymbol{x} = (x_1, x_2)^T \in \mathbb{R}^2$ is mapped to $\phi(\boldsymbol{x}) = (\sqrt{2}x_1x_2, x_1^2, x_2^2)^T \in \mathbb{R}^3$.

The projection of $\phi(\boldsymbol{x}_i)$ onto $\boldsymbol{w}$ is also shown, where

$$\boldsymbol{w} = 0.511x_1x_2 + 0.761x_1^2 - 0.4x_2^2$$



# 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