Computational Data Mining

Part 14: LDA

Instructor: Zahra Narimani

Department of Computer Science and Information Technology,
Institute for Advanced Studies in Basic Sciences, Zanjan, Iran

Outline

- Dimensionality reduction
- Linear Discriminant Analysis

Dimensionality Reduction techniques

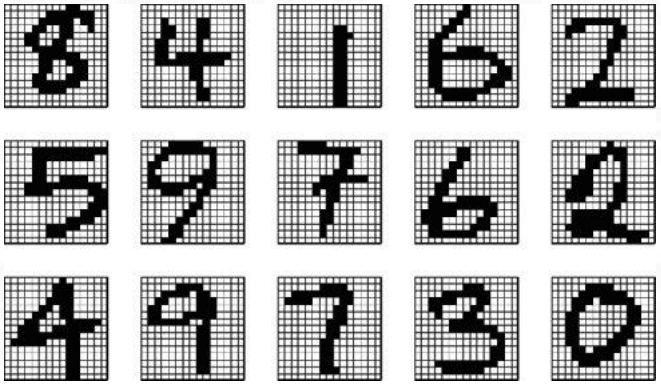
- Feature Selection
- Feature Extraction

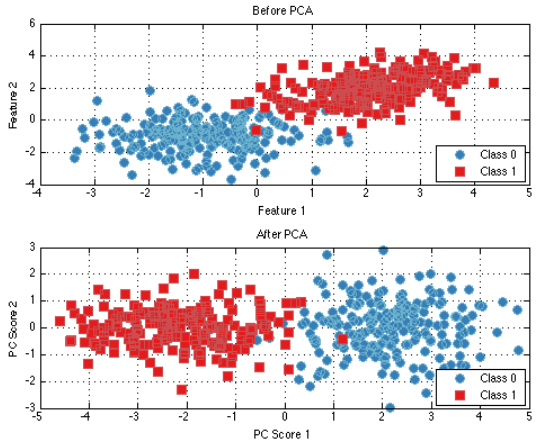
PCA

- http://setosa.io/ev/principal-component-analysis/

High dimensional data in linear regression

- Assume the data is high dimensional
- We should reduce the dimensions
  - **Does it help to use PCA if our goal is to classify data?**
  - **Example: reduce dimension in image data, in order to use the data with lower dimensionality for recognition of handwritten digits.**





Handwritten digit recognition vs. Image Compression

- In both we want to reduce the dimensionality of the data
- The first problem is supervised, the second is unsupervised

Supervised vs. unsupervised learning

- Applications in which the training data comprises examples of the input vectors along with their corresponding target vectors are known as supervised learning problems.
  - Cases such as the digit recognition example (classification problems)
  - If the desired output consists of one or more continuous variables, then the task is called regression. Cases such as determining the elasticity properties of a spring by attaching different weights to it and measuring its length
- In other pattern recognition problems, the training data consists of a set of input vectors **x** without any corresponding target values.
- The goal in such unsupervised learning problems may be to discover groups of similar examples within the data, where it is called clustering, or to determine the distribution of data within the input space, known as density estimation, or to project the data from a high-dimensional space down to two or three dimensions for the purpose of visualization.

Semi supervised learning

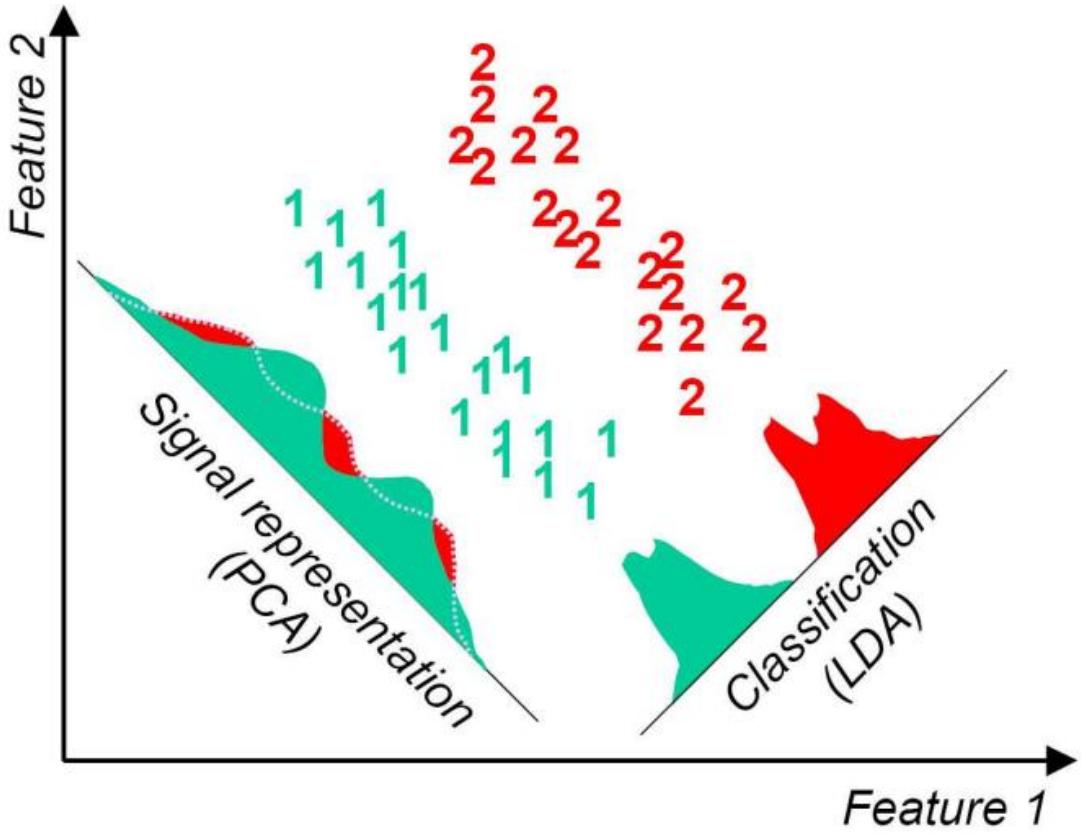
- Problems where you have a large amount of input data (X) and only some of the data is labeled (Y) are called semi-supervised learning problems.
- These problems sit in between both supervised and unsupervised learning.
- A good example is a photo archive where only some of the images are labeled, (e.g. dog, cat, person) and the majority are unlabeled.

LDA

- Used as a dimensionality reduction technique
- Used in pre-processing step for pattern classification
- Has the goal to project a dataset onto a lower-dimensional space
- **Differs from PCA**
- In addition to finding the component axises with LDA we are interested in the axes that maximize the separation between multiple classes

LDA

- Project a feature space (a dataset n-dimensional samples) onto a small subspace k (k<=n-1) while maintaining the class-discriminatory information
- Both PCA and LDA are linear transformation techniques used for dimensional reduction.
- PCA is described as unsupervised but LDA is supervised because of the relation to the dependent variable.


Feature 2
Feature 1
Signal representation (PCA)
Classification (LDA)

LDA vs. PCA

LDA theory

- Maximize the ratio of between-class variance to the within-class variance

LDA steps

1. Calculate the separability between classes (between-class variance or between-class matrix)
2. Calculate the distance between the mean and samples of each class (within-class variance or within-class matrix)
3. Construct a lower-dimensional space which maximizes the between-class variance and minimizes the within-class variance

Calculate the between-class variance (S_B)

- Between-class variance of the i^{th} class (S_{bi}) represents the distance between the mean of this class (μ_i) and the total mean (μ).
- Assumptions:

$$X = \{x_1, x_2, \ldots, x_N\}$$
$$(x_i \in \mathcal{R}^M)$$

$$c = 3$$
$$X = [\omega_1, \omega_2, \omega_3]$$

$$n_1 = n_2 = n_3 = 5$$

$$N = \sum_{i=1}^{3} n_i$$

The between-class variance is equal to:

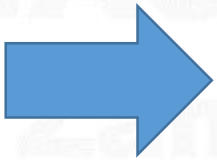
$$(m_i - m)^2 = (W^T \mu_i - W^T \mu)^2$$
$$= W^T (\mu_i - \mu)(\mu_i - \mu)^T W$$

W is the transformation matrix of LDA
m is the projection of mean

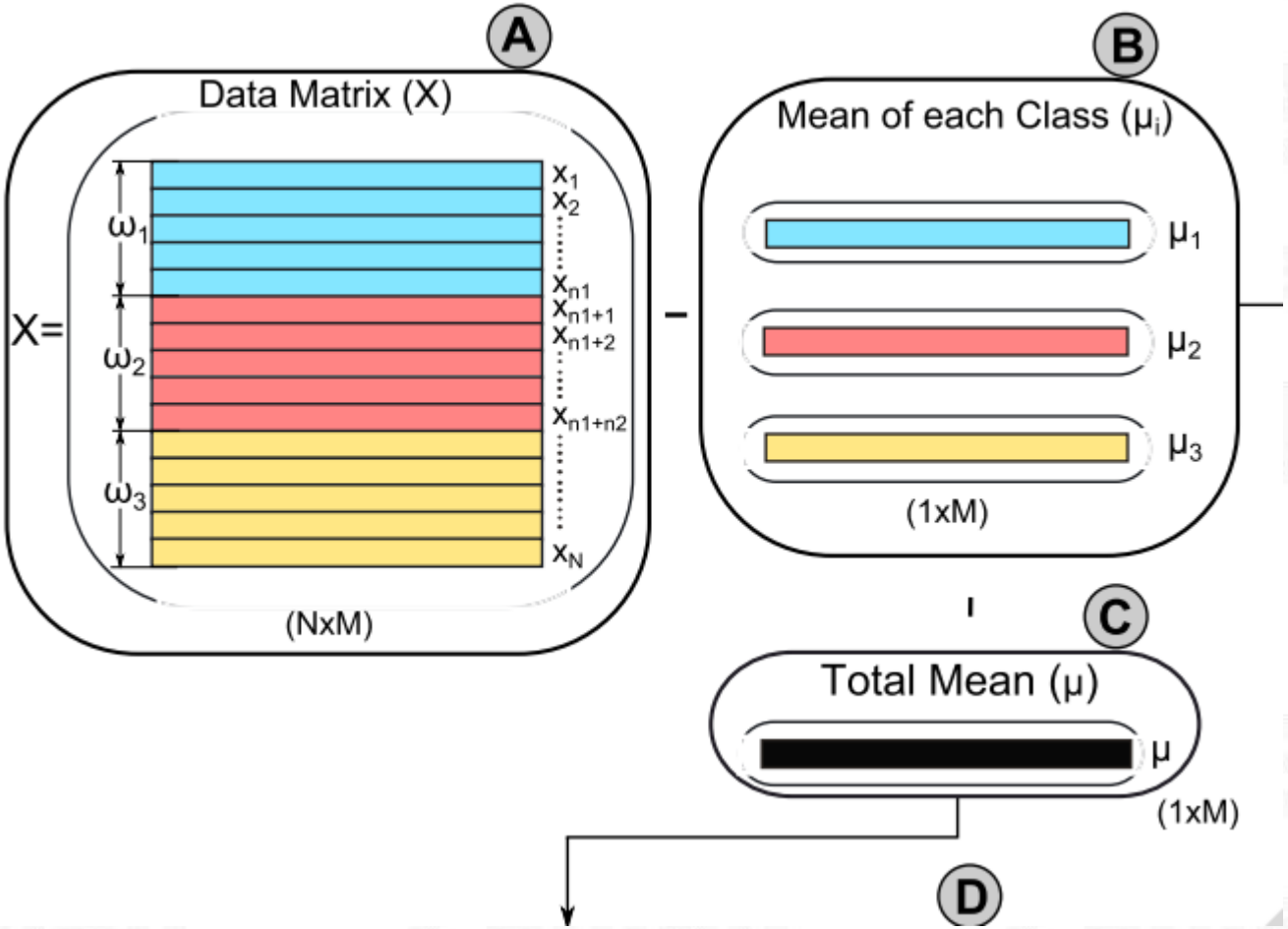
$$\boldsymbol{\mu}_j = \frac{1}{\mathrm{N}_j} \sum_{\mathbf{x} \in \omega_i} x_i \qquad \boldsymbol{\mu} = \frac{1}{N} \sum x_i$$

$$(m_i - m)^2 = (W^T \mu_i - W^T \mu)^2$$

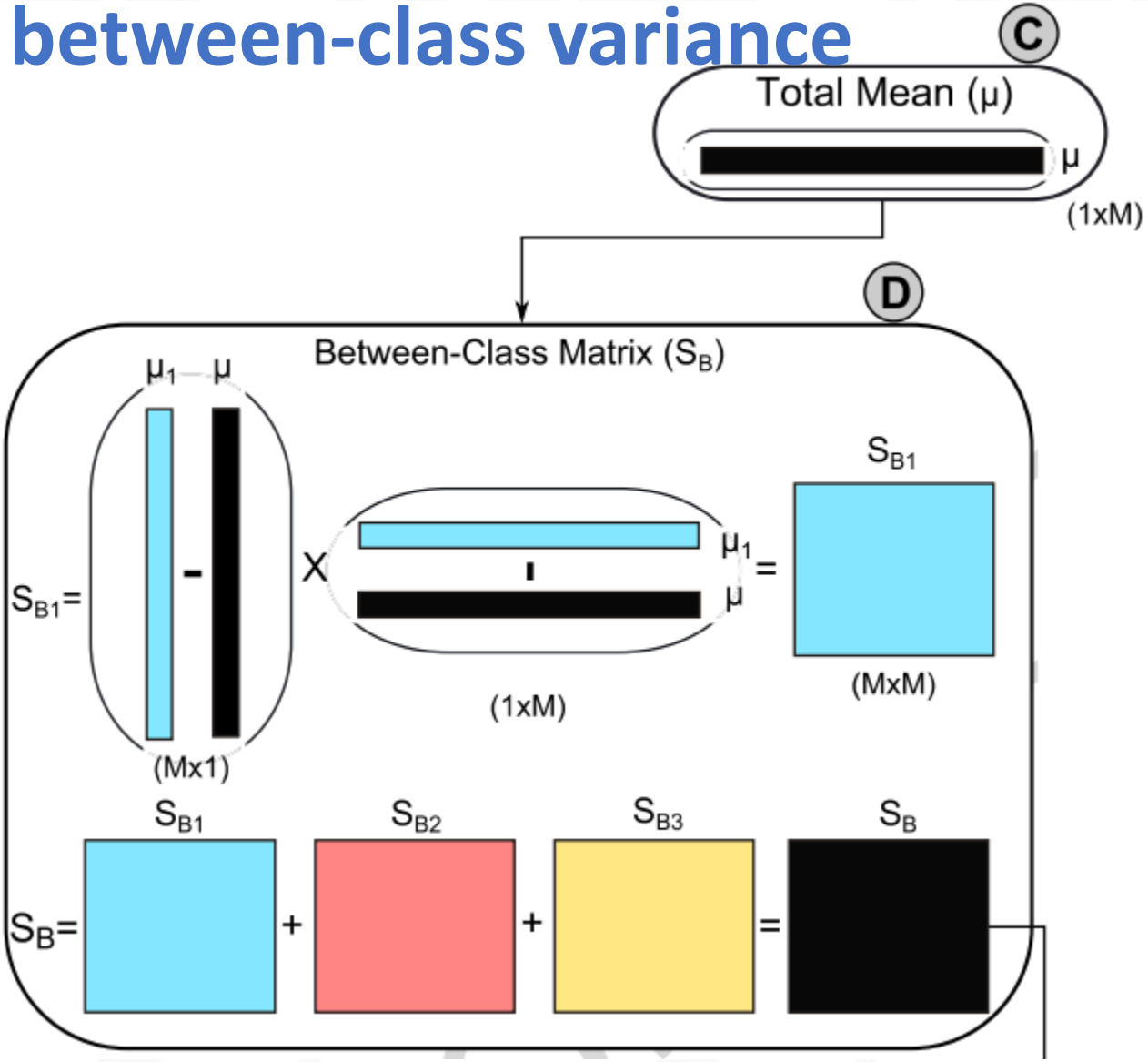
$$= W^T (\mu_i - \mu)(\mu_i - \mu)^T W$$



$$(m_i - m)^2 = W^T S_{Bi} W$$


A
Data Matrix (X)
X=
ω1
ω2
ω3
x1
x2
xn1
xn1+1
xn1+2
xn1+n2
xN
(NxM)
B
Mean of each Class (μi)
μ1
μ2
μ3
(1xM)
C
Total Mean (μ)
μ
(1xM)
D

Total between-class variance



Calculate the within-class variance (S_W)

- Difference between the mean and samples of that class (S_{wi})
- LDA goal:
- Search for a lower dimensional space which is used to minimize the difference between the projected mean (m_i) and the projected samples ($W^T x_i$)

- $S_{Wj} = d_j^T \times d_j = \sum_{i=1}^{n_j} (x_{ij} - \mu_j).(x_{ij} - \mu_j)^T$
- d_j is the centering data of the j th class, i.e. $d_j = \omega_j - \mu_j = \{x_i\}_{i=1..nj} - \mu_j$

Total within-class variance

$$\sum_{x_i \in \omega_j, j=1,\ldots,c} (W^T x_i - m_j)^2$$

$$= \sum_{x_i \in \omega_j, j=1,\ldots,c} (W^T x_{ij} - W^T \mu_j)^2$$

$$= \sum_{x_i \in \omega_j, j=1,\ldots,c} W^T (x_{ij} - \mu_j)^2 W$$

$$= \sum_{x_i \in \omega_j, j=1,\ldots,c} W^T (x_{ij} - \mu_j)(x_{ij} - \mu_j)^T W$$

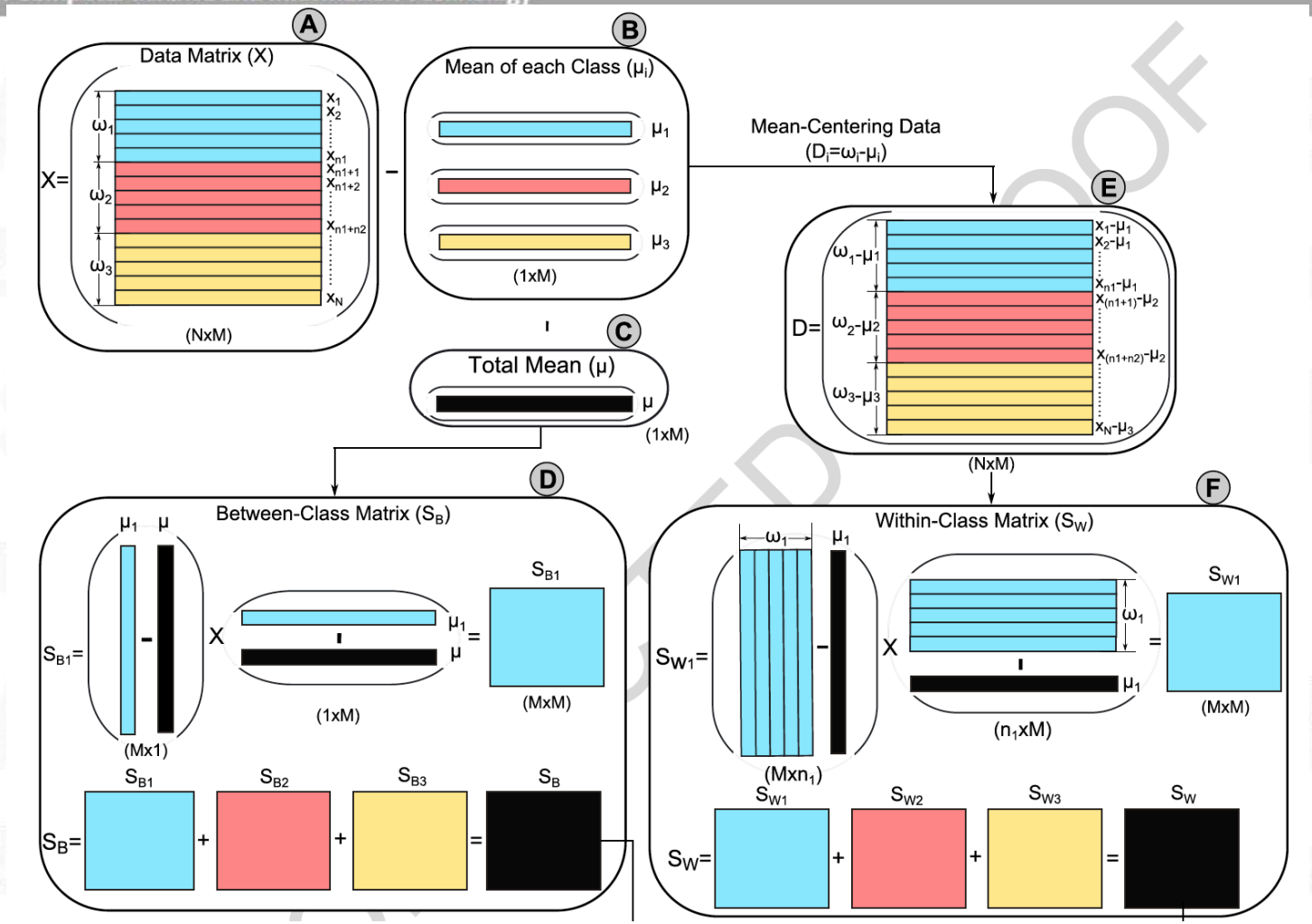
$$= \sum_{x_i \in \omega_j, j=1,\ldots,c} W^T S_{W_j} W$$

$$S_W = \sum_{i=1}^{3} S_{Wi}$$

$$= \sum_{x \in \omega_1} (x_i - \mu_1)(x_i - \mu_1)^{\mathrm{T}}$$

$$+ \sum_{x \in \omega_2} (x_i - \mu_1)(x_i - \mu_1)^{\mathrm{T}}$$

$$+ \sum_{x \in \omega_3} (x_i - \mu_1)(x_i - \mu_1)^{\mathrm{T}}$$



Constructing the lower dimensional space

- Equation to maximize:

$$\arg\max_{W} \frac{W^T S_B W}{W^T S_W W}$$

Fisher criterion

- We have to maximize:

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

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

$$\frac{d}{d\mathbf{w}}[\mathbf{J(w)}] = \frac{d}{d\mathbf{w}}\left[\frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}\right] = 0$$

$$[\mathbf{w}^T\mathbf{S}_W\mathbf{w}]\frac{d[\mathbf{w}^T\mathbf{S}_B\mathbf{w}]}{d\mathbf{w}} - [\mathbf{w}^T\mathbf{S}_B\mathbf{w}]\frac{d[\mathbf{w}^T\mathbf{S}_W\mathbf{w}]}{d\mathbf{w}} = 0$$

$$[\mathbf{w}^T\mathbf{S}_W\mathbf{w}]\,2\mathbf{S}_B\mathbf{w} - [\mathbf{w}^T\mathbf{S}_B\mathbf{w}]2\mathbf{S}_W\mathbf{w} = 0$$

$$\xrightarrow{\times \frac{1}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}} \left[\frac{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}\right] 2\mathbf{S}_B\mathbf{w} - \left[\frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}\right] 2\mathbf{S}_W\mathbf{w} = 0$$

$$\mathbf{S}_B\mathbf{w} - \mathbf{J}\,\mathbf{S}_W\mathbf{w} = 0$$

generalized eigenvalue problem

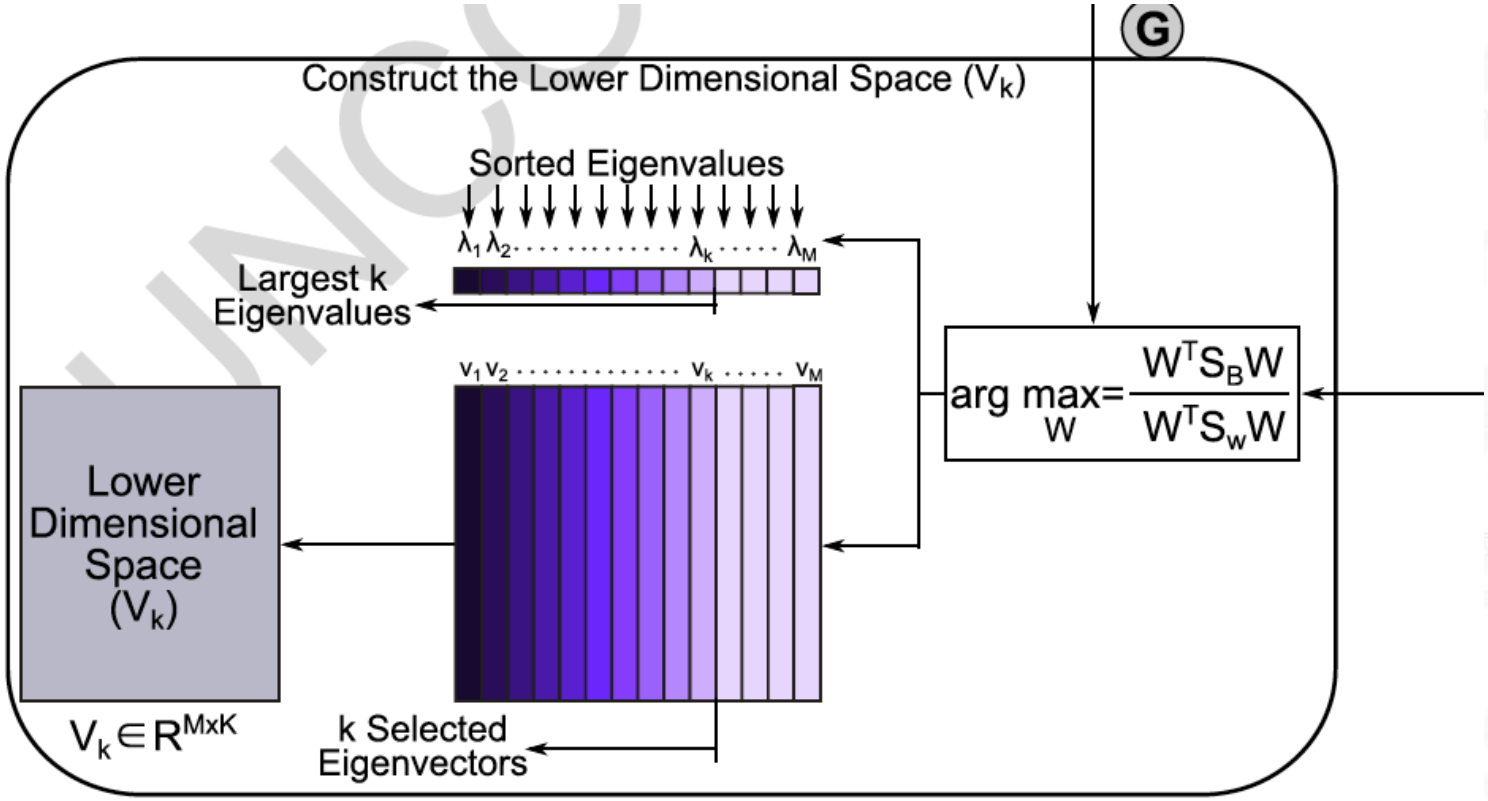
$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} - \mathbf{J}\,\mathbf{w} = 0 \qquad \mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} = \mathbf{J}\,\mathbf{w} \qquad \mathbf{A}\,\mathbf{v} = \lambda\mathbf{v}$$

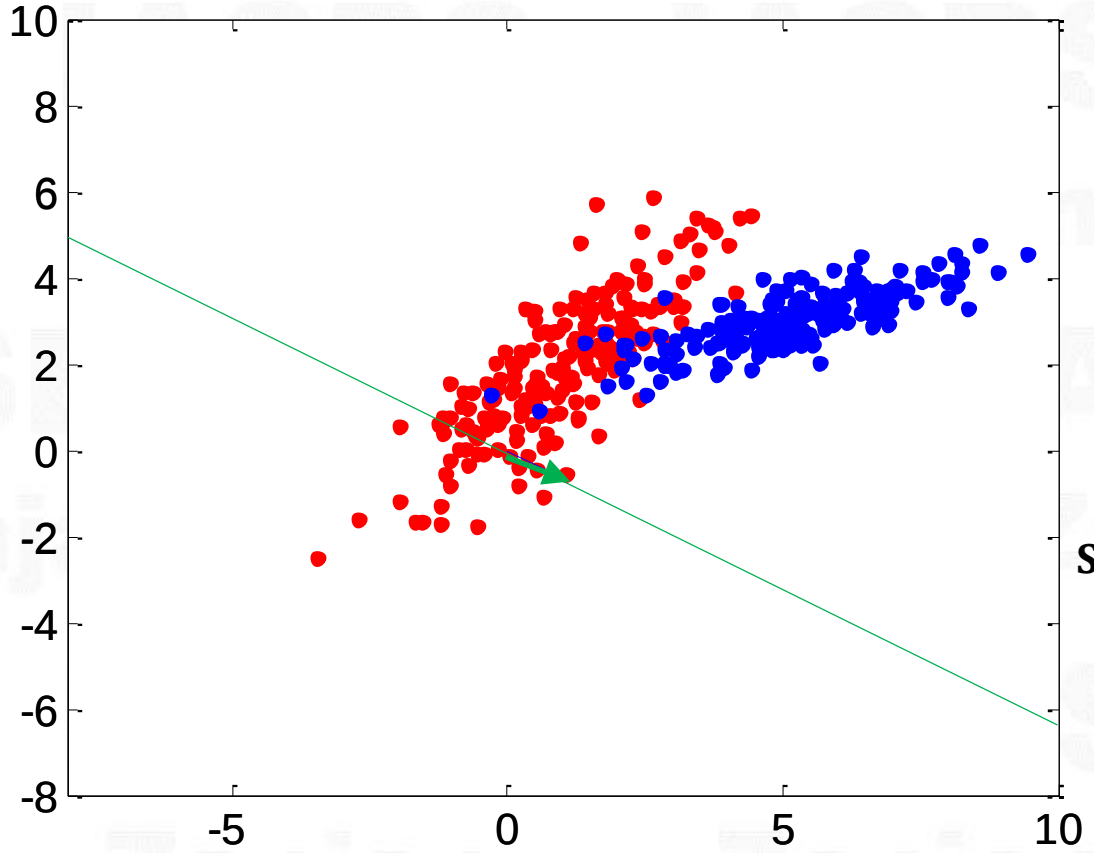
$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} = \mathbf{J}\,\mathbf{w} \qquad \mathbf{w}^* = \underset{w}{\operatorname{argmax}}\left\{\frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}\right\}$$

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

To perform **LDA**:

1. Calculate within class scatter matrix $\mathbf{S}_W$
2. Calculate between class scatter matrix $\mathbf{S}_B$
3. Calculate eigenvector of $\mathbf{S}_W^{-1}\mathbf{S}_B$
4. Project data point along new direction $\mathbf{w}^T\mathbf{X}$
5. New data points should be project on $\mathbf{w}$ direction, then can be classified according to projected means


G
Construct the Lower Dimensional Space (V_k)
Sorted Eigenvalues
λ_1 λ_2 ..... λ_k ..... λ_M
Largest k Eigenvalues
v_1 v_2 ..... v_k ..... v_M
$\arg\max_W \frac{W^T S_B W}{W^T S_w W}$
Lower Dimensional Space (V_k)
$V_k \in R^{MxK}$
k Selected Eigenvectors



Class (1×400)

X
(2×400)

$$\mu_1 = \begin{bmatrix} 1.03 \\ 2.04 \end{bmatrix} \qquad \mu_2 = \begin{bmatrix} 5.13 \\ 3.03 \end{bmatrix}$$

$$\mathbf{S}_1 = \begin{bmatrix} 435.1 & 450.9 \\ 450.9 & 643.7 \end{bmatrix} \qquad \mathbf{S}_2 = \begin{bmatrix} 585.9 & 217.6 \\ 217.6 & 111 \end{bmatrix}$$

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

$$\mathbf{S}_B = (\boldsymbol{\mu}_i - \boldsymbol{\mu}\ )(\boldsymbol{\mu}_i - \boldsymbol{\mu}\ )^T$$

$$\mathbf{S}_W^{-1}\mathbf{S}_B = \begin{bmatrix} 0.027 & 0.006 \\ -0.019 & -0.004 \end{bmatrix} \xrightarrow{\text{Eigen analysis}} \mathbf{w} = \begin{bmatrix} 0.81 \\ -0.58 \end{bmatrix}$$

Projection matrix
(orientation)

w^T X (2×400) → Y (1×400)

Fisher's LDA for C-class:

The generalization of the within-class scatter matrix is

$$\mathbf{S}_W = \sum_{i=1}^{C} \mathbf{S}_i \qquad \mathbf{S}_i = \sum_{\mathbf{x}\in\omega_i} (\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^T \qquad \boldsymbol{\mu}_i = \frac{1}{N_i}\sum_{\mathbf{x}\in\omega_i} \mathbf{x}$$

The generalization for the between-class scatter matrix is

$$\mathbf{S}_B = \sum_{i=1}^{C} (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \qquad \boldsymbol{\mu} = \frac{1}{N}\sum_{\forall \mathbf{x}} \mathbf{x}$$

$$\mathbf{J}(\mathbf{W}) = \frac{det(\tilde{\mathbf{S}}_B)}{det(\tilde{\mathbf{S}}_W)} = \frac{\det(\mathbf{W}^T\mathbf{S}_B\mathbf{W})}{\det(\mathbf{W}^T\mathbf{S}_W\mathbf{W})}$$

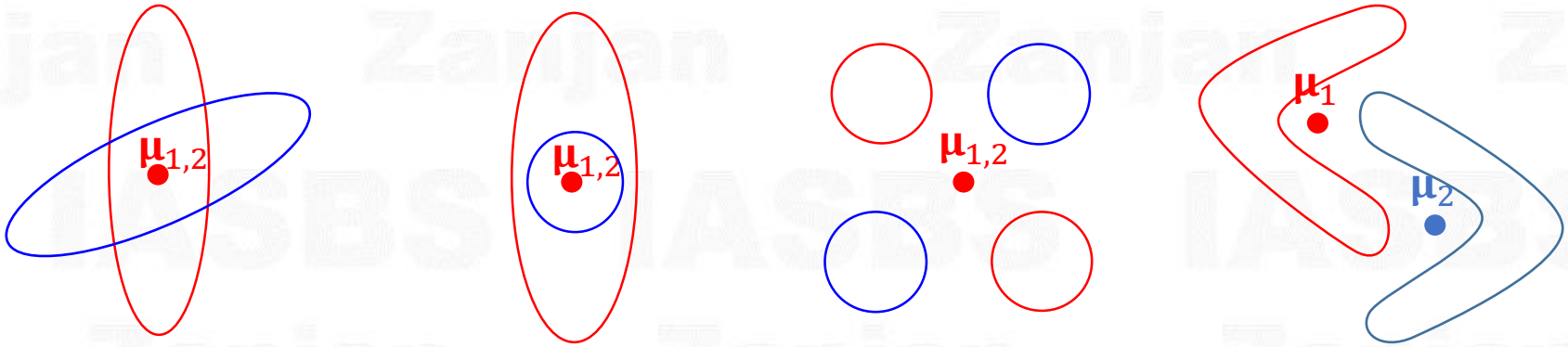
$$\mathbf{Y} = \mathbf{W}^T\mathbf{X} \qquad \mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{W} = \mathbf{J}\,\mathbf{W}$$

The projections with maximum class separability information are the eigenvectors corresponding to the largest eigenvalues of $\mathbf{S}_W^{-1}\mathbf{S}_B$

Limitations of LDA

LDA is a parametric method (it assumes Gaussian likelihoods)

- If the distributions are significantly non-Gaussian, the LDA projections may not preserve complex structure in the data needed for classification
- LDA will also fail if discriminatory information is not in the mean but in the variance of the data



In practice, $\mathbf{S}_w$ is often singular

- PCA before LDA
- Using Pseudo inverse

Linear Discriminant Analysis

Any Question?