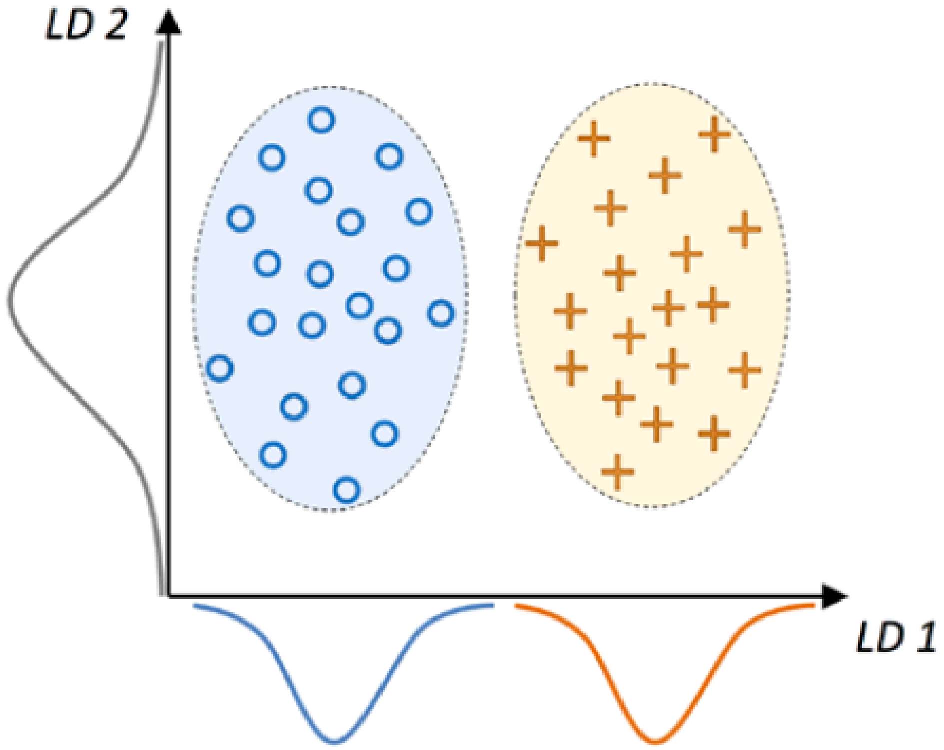
Linear Discriminant Analysis

生醫光電所 吳育德

Chap7. A First Course in Machine Learning, 2ed, Simon Rogers and Mark Girolami, 2017
CSC421/2516 Lecture 12, Roger Grosse and Jimmy Ba, Univ. of Toronto

The General Problem

- Given training samples of *i* classes, determine a set of optimal projection axes such that the set of projective feature vectors has the maximum between-class scatter and minimum within-class scatter simultaneously.



Projection of data $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{N_i}$ in a class $\boldsymbol{i}$ onto vector $\boldsymbol{w}$

- Given a set of N_i samples, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{N_i}$, each is d-dimensional

$$\underbrace{\boldsymbol{X}}_{N_i \times d} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_{N_i}^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N_i,1} & x_{N_i,2} & \cdots & x_{N_i,d} \end{bmatrix}$$

- Let's start to find a unit $\underbrace{\boldsymbol{w}}_{d\times 1}$ vector and project $\underbrace{\boldsymbol{x}_n}_{d\times 1}$ into 1-dim $\boldsymbol{w}^T\boldsymbol{x}_n$.

- Let $\overline{\boldsymbol{x}}_i = \frac{1}{N_i}\sum_{n=1}^{N_i} \boldsymbol{x}_n$

- The variance

$$\sigma_x^2 = \frac{1}{N_i}\sum_{n=1}^{N_i}(\boldsymbol{w}^T\boldsymbol{x}_n - \boldsymbol{w}^T\overline{\boldsymbol{x}}_i)^2$$

Projection of data $x_1, x_2, \ldots, x_{N_i}$ in a class *i* onto vector *w*

- The variance

$$\sigma_x^2 = \frac{1}{N_i}\sum_{n=1}^{N_i}(\boldsymbol{w}^T\boldsymbol{x}_n - \boldsymbol{w}^T\overline{\boldsymbol{x}}_i)^2 = \frac{1}{N_i}\sum_{n=1}^{N_i}\boldsymbol{w}^T(\boldsymbol{x}_n-\overline{\boldsymbol{x}}_i)(\boldsymbol{x}_n-\overline{\boldsymbol{x}}_i)^T\boldsymbol{w}$$

$$= \underbrace{\boldsymbol{w}^T}_{1\times d}\left(\frac{1}{N_i}\sum_{n=1}^{N_i}\underbrace{(\boldsymbol{x}_n-\overline{\boldsymbol{x}}_i)}_{d\times 1}\underbrace{(\boldsymbol{x}_n-\overline{\boldsymbol{x}}_i)^T}_{1\times d}\right)\underbrace{\boldsymbol{w}}_{d\times 1}$$

$$= \boldsymbol{w}^T \underbrace{\mathbf{C}_i}_{d\times d} \boldsymbol{w},$$

where $\mathbf{C}_i \equiv \frac{1}{N_i}\sum_{n=1}^{N_i}(\boldsymbol{x}_n-\overline{\boldsymbol{x}}_i)(\boldsymbol{x}_n-\overline{\boldsymbol{x}}_i)^T$

Linear (Fisher's) Discriminant Analysis (LDA) for 2 Classes

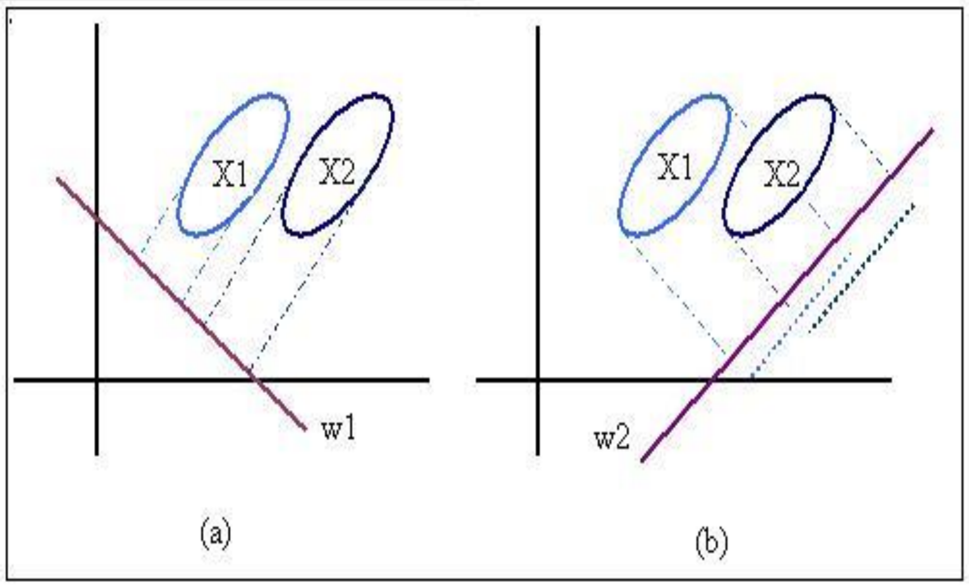
- Given data of two classes, we seek a projection that best separate the data

$$\max_{w} \frac{(\theta_1 - \theta_2)^2}{\sigma_1^2 + \sigma_2^2} = \max_{w} \frac{(\boldsymbol{w}^T \bar{\boldsymbol{x}}_1 - \boldsymbol{w}^T \bar{\boldsymbol{x}}_2)^2}{\boldsymbol{w}^T \boldsymbol{C}_1 \boldsymbol{w} + \boldsymbol{w}^T \boldsymbol{C}_2 \boldsymbol{w}} = \max_{w} \frac{\boldsymbol{w}^T (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T \boldsymbol{w}}{\boldsymbol{w}^T (\boldsymbol{C}_1 + \boldsymbol{C}_2) \boldsymbol{w}}$$

$$= \max_{w} \frac{\boldsymbol{w}^T S_b \boldsymbol{w}}{\boldsymbol{w}^T S_w \boldsymbol{w}} \equiv \max_{w} J(\boldsymbol{w})$$

where $S_b \equiv (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T$: between-class scatter matrix,

$S_w \equiv \boldsymbol{C}_1 + \boldsymbol{C}_2$: within-class scatter matrix



LDA for 2 Classes: The first derivation method

- $\max_{w} J(\boldsymbol{w}) = \max_{w} \frac{\boldsymbol{w}^T S_b \boldsymbol{w}}{\boldsymbol{w}^T S_w \boldsymbol{w}}$, We differentiate $J(\boldsymbol{w})$ w.r.t w and set it to 0
- $\frac{d(g^{-1}f)}{dx} = g^{-1}\frac{df}{dx} + f\frac{dg^{-1}}{dx} = g^{-1}\frac{df}{dx} - fg^{-2}\frac{dg}{dx} = 0 \Rightarrow g\frac{df}{dx} - f\frac{dg}{dx} = 0$

Therefore $\frac{dJ(\boldsymbol{w})}{d\boldsymbol{w}} = \frac{d}{d\boldsymbol{w}}\left(\frac{\boldsymbol{w}^T S_b \boldsymbol{w}}{\boldsymbol{w}^T S_w \boldsymbol{w}}\right) = 0$

$\Rightarrow (\boldsymbol{w}^T S_w \boldsymbol{w})\frac{d}{d\boldsymbol{w}}(\boldsymbol{w}^T S_b \boldsymbol{w}) - (\boldsymbol{w}^T S_b \boldsymbol{w})\frac{d}{d\boldsymbol{w}}(\boldsymbol{w}^T S_w \boldsymbol{w}) = 0$

$\Rightarrow (\boldsymbol{w}^T S_w \boldsymbol{w})2S_b\boldsymbol{w} - (\boldsymbol{w}^T S_b \boldsymbol{w})2S_w\boldsymbol{w} = 0$

$\Rightarrow \left(\frac{\boldsymbol{w}^T S_w \boldsymbol{w}}{\boldsymbol{w}^T S_w \boldsymbol{w}}\right) S_b\boldsymbol{w} - \left(\frac{\boldsymbol{w}^T S_b \boldsymbol{w}}{\boldsymbol{w}^T S_w \boldsymbol{w}}\right) S_w\boldsymbol{w} = 0$ (Dividing by $2\boldsymbol{w}^T S_w \boldsymbol{w}$)

$\Rightarrow S_b\boldsymbol{w} - J(\boldsymbol{w})S_w\boldsymbol{w} = 0$

$\Rightarrow S_w^{-1}S_b\boldsymbol{w} - J(\boldsymbol{w})\boldsymbol{w} = 0,$

i.e. Solving the generalized eigenvalue problem $S_w^{-1}S_b\boldsymbol{w} = J(\boldsymbol{w})\boldsymbol{w}$

$\Rightarrow$ The optimal $\boldsymbol{w}^* = S_w^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)$

LDA for 2 Classes: The second derivation method

- Let $\boldsymbol{w} = \boldsymbol{V}\boldsymbol{b}$

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T S_b \boldsymbol{w}}{\boldsymbol{w}^T S_w \boldsymbol{w}} = \frac{\boldsymbol{b}^T(\boldsymbol{V}^T S_b \boldsymbol{V})\boldsymbol{b}}{\boldsymbol{b}^T(\boldsymbol{V}^T S_w \boldsymbol{V})\boldsymbol{b}} = \frac{\boldsymbol{b}^T \boldsymbol{\Lambda} \boldsymbol{b}}{\boldsymbol{b}^T \boldsymbol{I} \boldsymbol{b}} = \frac{\boldsymbol{b}^T \boldsymbol{\Lambda} \boldsymbol{b}}{\boldsymbol{b}^T \boldsymbol{b}} = \frac{\boldsymbol{b}^T \boldsymbol{\Lambda} \boldsymbol{b}}{||\boldsymbol{b}||^2}$$

where we let

$$\boldsymbol{V}^T S_w \boldsymbol{V} = \boldsymbol{I}, \qquad \boldsymbol{V}^T S_b \boldsymbol{V} = \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{bmatrix}, \boldsymbol{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_d]$$

$\Rightarrow \lambda_i' s$ satisfy the generalized eigenvalue problem: $S_w^{-1} S_b \boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$, $\lambda_1 \geq \cdots \lambda_d$

$$\max_b \frac{\boldsymbol{b}^T \boldsymbol{\Lambda} \boldsymbol{b}}{||\boldsymbol{b}||^2} = \frac{\lambda_1 \boldsymbol{b}^T \boldsymbol{b}}{\boldsymbol{b}^T \boldsymbol{b}} = \lambda_1 \text{ with } \boldsymbol{b} = \boldsymbol{\phi}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \text{since } \boldsymbol{\Lambda}\boldsymbol{\phi}_1 = \lambda_1 \boldsymbol{\phi}_1$$

Find the Best Projection Vector w

$$\boldsymbol{w} = \boldsymbol{V}\boldsymbol{b} = \boldsymbol{V}\boldsymbol{\phi}_1 = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_d]\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix} = \boldsymbol{v}_1$$

Since $\boldsymbol{w}$ satisfies $S_w^{-1}S_b\boldsymbol{v}_1 = \lambda_1\boldsymbol{v}_1$ or $S_w^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T\boldsymbol{w} = \lambda_1\boldsymbol{w}$

$\Rightarrow \boldsymbol{w} = \frac{1}{\lambda_1}S_w^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T\boldsymbol{w} \cdots (1)$

Let $(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T\boldsymbol{w} = k \cdots (2) \Rightarrow k^2 = \boldsymbol{w}^T(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T\boldsymbol{w}$

Since we can constrain $\boldsymbol{w}$ to be a unit vector, i.e., $||\boldsymbol{w}||^2 = \boldsymbol{w}^T\boldsymbol{w} = 1$

$(1) \Rightarrow \frac{1}{\lambda_1}\boldsymbol{w}^T(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T(S_w^{-1})^T \times \frac{1}{\lambda_1}S_w^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T\boldsymbol{w} = 1$

$\Rightarrow \frac{1}{{\lambda_1}^2}k(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T(S_w^{-1})^T\frac{1}{\lambda_1}S_w^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)k = 1 \Rightarrow k^2 = \frac{{\lambda_1}^2}{||S_w^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)||^2}$

$(1) \Rightarrow \boldsymbol{w} = \frac{1}{\lambda_1}S_w^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)k = \frac{S_w^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)}{||S_w^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)||}$

Theorem: eigendecomposition

Let M be a real symmetric matrix with largest eigenvalue λ_1, then

$$\lambda_1 = \max_{\boldsymbol{u}} \boldsymbol{u}^T M \boldsymbol{u}, \left|\left|\boldsymbol{u}\right|\right| = 1$$

the maximum occurs when $\boldsymbol{u} = \boldsymbol{\phi}_1$, i.e. the unit eigenvector associated with λ_1 .

Proof:

Let $\{\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_n\}$ be the unit eigenvector associated with $\lambda_1 \neq 0 \geq \cdots, \geq \lambda_n$

$\Rightarrow M\boldsymbol{\phi}_i = \lambda_i \boldsymbol{\phi}_i, i = 1, \cdots, n$

$$\Rightarrow M\,[\boldsymbol{\phi}_1\ \boldsymbol{\phi}_2 \cdots \boldsymbol{\phi}_n] = [\boldsymbol{\phi}_1\ \boldsymbol{\phi}_2 \cdots \boldsymbol{\phi}_n] \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}, \text{or } M\boldsymbol{\Phi} = \boldsymbol{\Phi\Lambda}$$

$$\Rightarrow M = \boldsymbol{\Phi\Lambda\Phi}^{-1} = M^T = (\boldsymbol{\Phi\Lambda\Phi}^{-1})^T = (\boldsymbol{\Phi}^{-1})^T \boldsymbol{\Lambda\Phi}^T \Rightarrow \boldsymbol{\Phi\Phi}^T = \boldsymbol{I} = \boldsymbol{\Phi}^T\boldsymbol{\Phi}$$

$\Rightarrow \{\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_n\}$ forms a complete orthonormal basis in $\mathbb{R}^n$, $\boldsymbol{\phi}_i^T \boldsymbol{\phi}_j = \begin{cases} 1, \text{if } i = j \\ 0, \text{if } i \neq j \end{cases}$

Theorem: eigendecomposition

$\forall\, \boldsymbol{u} \in \mathbb{R}^n$, we can express $\boldsymbol{u} = \alpha_1 \boldsymbol{\phi}_1 + \cdots + \alpha_n \boldsymbol{\phi}_n,\ \ \boldsymbol{\phi}_i^T \boldsymbol{\phi}_j = \begin{cases} 1\,, \text{if } i = j \\ 0\,, \text{if } i \neq j \end{cases}$

$$\Rightarrow \boldsymbol{\phi}_j^T \boldsymbol{u} = \alpha_1 \boldsymbol{\phi}_j^T \boldsymbol{\phi}_1 + + \cdots + \alpha_n \boldsymbol{\phi}_j^T \boldsymbol{\phi}_n = \alpha_j$$

$$\Rightarrow \boldsymbol{u} = \boldsymbol{\phi}_1^T \boldsymbol{u} \boldsymbol{\phi}_1 + \cdots + \boldsymbol{\phi}_n^T \boldsymbol{u} \boldsymbol{\phi}_n = \sum_{i=1}^n \boldsymbol{\phi}_i^T \boldsymbol{u}\, \boldsymbol{\phi}_i, \quad \boldsymbol{u}^T \boldsymbol{u} = 1$$

$$\Rightarrow M\boldsymbol{u} = \sum_{i=1}^n \boldsymbol{\phi}_i^T \boldsymbol{u}\, M \boldsymbol{\phi}_i = \sum_{i=1}^n \boldsymbol{\phi}_i^T \boldsymbol{u}\, \lambda_i \boldsymbol{\phi}_i$$

$$\Rightarrow \boldsymbol{u}^T M \boldsymbol{u} = \sum_{i=1}^n \boldsymbol{\phi}_i^T \boldsymbol{u}\, \boldsymbol{\phi}_i^T \sum_{j=1}^n \lambda_j \boldsymbol{\phi}_j^T \boldsymbol{u}\, \boldsymbol{\phi}_j = \sum_{i=1}^n \sum_{j=1}^n \lambda_j \left(\boldsymbol{\phi}_i^T \boldsymbol{u}\right)\left(\boldsymbol{\phi}_j^T \boldsymbol{u}\right) \boldsymbol{\phi}_i^T \boldsymbol{\phi}_j$$

$$= \sum_{i=1}^n \lambda_i \left(\boldsymbol{\phi}_i^T \boldsymbol{u}\right)^2 \leq \sum_{i=1}^n \lambda_1 \left(\boldsymbol{\phi}_i^T \boldsymbol{u}\right)^2 = \lambda_1$$

since $1 = \boldsymbol{u}^T \boldsymbol{u} = \sum_{i=1}^n \boldsymbol{\phi}_i^T \boldsymbol{u}\, \boldsymbol{\phi}_i^T \sum_{j=1}^n \boldsymbol{\phi}_j^T \boldsymbol{u}\, \boldsymbol{\phi}_j \sum_{i=1}^n \sum_{j=1}^n \left(\boldsymbol{\phi}_i^T \boldsymbol{u}\right)\left(\boldsymbol{\phi}_j^T \boldsymbol{u}\right) \boldsymbol{\phi}_i^T \boldsymbol{\phi}_j$

$$= \sum_{i=1}^n \left(\boldsymbol{\phi}_i^T \boldsymbol{u}\right)^2$$

If we choose $\boldsymbol{u} = \boldsymbol{\phi}_1 \Rightarrow \boldsymbol{\phi}_1^T M \boldsymbol{\phi}_1 = \boldsymbol{\phi}_1^T \lambda_1 \boldsymbol{\phi}_1 = \lambda_1$

Corollary: eigendecomposition

Let M be a real symmetric matrix with largest eigenvalue λ_1, then

$$\lambda_1 = \max_{\boldsymbol{x}} \frac{\boldsymbol{x}^T M \boldsymbol{x}}{||\boldsymbol{x}||^2}, \qquad \boldsymbol{x} \neq 0$$

the maximum occurs when $\boldsymbol{x} = k\boldsymbol{\phi}_1$, $\boldsymbol{\phi}_1$ is the unit eigenvector associated with λ_1 and $k \in \mathbb{R}$.

Note: Let $\boldsymbol{u} = \frac{\boldsymbol{x}}{||\boldsymbol{x}||}$, we can rewrite $\frac{\boldsymbol{x}^T M \boldsymbol{x}}{||\boldsymbol{x}||^2}$ into $\boldsymbol{u}^T M \boldsymbol{u}$ with $||\boldsymbol{u}|| = 1$, the proof is the same as the previous theorem

Theorem: Generalized eigendecomposition

Let S_w and S_b be $n \times n$ real symmetric matrices . If S_w is positive definite, then there exists an $n \times n$ matrix V which achieves

$$\mathbf{V}^T S_w \mathbf{V} = \mathbf{I}, \qquad \mathbf{V}^T S_b \mathbf{V} = \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}, \mathbf{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n]$$

The real numbers $\lambda_1 \cdots \lambda_n$ satisfy the generalized eiegenvalue equation :
$S_w^{-1} S_b \boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i, \ \lambda_1 \geq \cdots \lambda_n$

where $\boldsymbol{v}_i$'s are the generalized eigenvectors
λ_i's are the generalized eigenvalues

Theorem: Generalized eigendecomposition

Proof:

Let $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_n$ and $r_1 \cdots r_n$ be the unit eigenvectors and eigenvalues of S_w

$\Rightarrow S_w \boldsymbol{\phi}_i = r_i \boldsymbol{\phi}_i, i = 1, \cdots, n$

$\Rightarrow S_w [\boldsymbol{\phi}_1\ \boldsymbol{\phi}_2 \cdots \boldsymbol{\phi}_n] = [\boldsymbol{\phi}_1\ \boldsymbol{\phi}_2 \cdots \boldsymbol{\phi}_M] \begin{bmatrix} r_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & r_n \end{bmatrix}$, or $S_w \boldsymbol{\Phi} = \boldsymbol{\Phi} \boldsymbol{R}$

$\Rightarrow \boldsymbol{\Phi}^T S_w \boldsymbol{\Phi} = \boldsymbol{\Phi}^T \boldsymbol{\Phi} \boldsymbol{R} = \boldsymbol{R},$ recall $\boldsymbol{\Phi}^T \boldsymbol{\Phi} = \boldsymbol{\Phi}\ \boldsymbol{\Phi}^T = \mathbf{I}$ if S_w is symmetric

Since S_w is positive, $r_i > 0, \forall i$, let's define $\boldsymbol{Z} = \begin{bmatrix} r_1^{-\frac{1}{2}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & r_n^{-\frac{1}{2}} \end{bmatrix}$

$\Rightarrow \boldsymbol{Z}^T \boldsymbol{\Phi}^T S_w \boldsymbol{\Phi} \boldsymbol{Z} = \boldsymbol{Z}^T \boldsymbol{R} \boldsymbol{Z} = \mathbf{I}$ (whitening)

Theorem: Generalized eigendecomposition

Note that

$$((\boldsymbol{\Phi Z})^T S_b(\boldsymbol{\Phi Z}))^T = (\boldsymbol{\Phi Z})^T S_b(\boldsymbol{\Phi Z})$$

$\Rightarrow (\boldsymbol{\Phi Z})^T S_b(\boldsymbol{\Phi Z})$ is symmetric and let's define $\mathbf{A} = (\boldsymbol{\Phi Z})^T S_b(\boldsymbol{\Phi Z})$

Let $\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n$ and $\lambda_1 \cdots \lambda_n$ be the unit eigenvectors and eigenvalues of $\mathbf{A}$

$$\Rightarrow \mathbf{A}[\boldsymbol{w}_1\ \boldsymbol{w}_2 \cdots \boldsymbol{w}_n] = [\boldsymbol{w}_1\ \boldsymbol{w}_2 \cdots \boldsymbol{w}_n]\begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}, \text{ or } \mathbf{AW} = \mathbf{W}\boldsymbol{\Lambda}$$

$$\Rightarrow \mathbf{A} = \mathbf{W\Lambda W}^{-1} = \mathbf{A}^T = (\mathbf{W\Lambda W}^{-1})^T = (\mathbf{W}^{-1})^T\boldsymbol{\Lambda}\mathbf{W}^{\mathrm{T}} \Rightarrow \mathbf{WW}^{\mathrm{T}} = \mathbf{I} = \mathbf{W}^{\mathrm{T}}\mathbf{W}$$

$\Rightarrow$ That is, **W** is a unitary (rotation) matrix such that $\mathbf{W}^T\mathbf{AW} = \boldsymbol{\Lambda}$

Theorem: Generalized eigendecomposition

We need to claim:

1. $\mathbf{V} = \boldsymbol{\Phi Z}\mathbf{W}$ such that $\mathbf{V}^T S_w \mathbf{V} = \mathbf{I}$

2. $S_w^{-1} S_b \boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$ or $S_w^{-1} S_b [\boldsymbol{v}_1 \cdots \boldsymbol{v}_n] = [\boldsymbol{v}_1 \cdots \boldsymbol{v}_d] \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}$

i.e., $S_w^{-1} S_b \mathbf{V} = \mathbf{V}\boldsymbol{\Lambda}$

---

1. $\mathbf{V}^T S_w \mathbf{V} = (\boldsymbol{\Phi Z}\mathbf{W})^T S_w (\boldsymbol{\Phi Z}\mathbf{W}) = \mathbf{W}^{\mathbf{T}} (\boldsymbol{Z}^T \boldsymbol{\Phi}^T S_w \boldsymbol{\Phi Z}) \mathbf{W} = \mathbf{W}^{\mathbf{T}} \mathbf{I} \mathbf{W} = \mathbf{I}$

2. $\mathbf{W}^T \mathbf{A} \mathbf{W} = \boldsymbol{\Lambda}$, $\mathbf{A} = (\boldsymbol{\Phi Z})^T S_b (\boldsymbol{\Phi Z})$

$\Rightarrow \mathbf{W}^T (\boldsymbol{\Phi Z})^T S_b (\boldsymbol{\Phi Z}) \mathbf{W} = \boldsymbol{\Lambda}$

$\Rightarrow \mathbf{V}^T S_b \mathbf{V} = \boldsymbol{\Lambda} = \mathbf{I}\boldsymbol{\Lambda} = \mathbf{V}^T S_w \mathbf{V}\boldsymbol{\Lambda}$

$\Rightarrow (\mathbf{V}^T)^{-1} \mathbf{V}^T S_b \boldsymbol{V} = (\mathbf{V}^T)^{-1} \mathbf{V}^T S_w \mathbf{V}\boldsymbol{\Lambda} = S_w \mathbf{V}\boldsymbol{\Lambda}$

$\Rightarrow S_w^{-1} S_b \mathbf{V} = \mathbf{V}\boldsymbol{\Lambda}$

Theorem: Generalized eigendecomposition

Note $(\mathbf{V}^T)^{-1}$ exits because $\mathbf{V}^T S_w \mathbf{V} = \mathbf{I}$

$\Rightarrow \det(\mathbf{V}^T S_w \mathbf{V}) = \det(\mathbf{I})$

$\Rightarrow \det(\mathbf{V}^T)\det(S_w)\det(\mathbf{V}) = 1$, note $\det(\mathbf{V}^T) = \det(\mathbf{V})$ and $\det(S_w) > 0$

$\Rightarrow \det(\mathbf{V}^T) = \sqrt{\frac{1}{\det(S_w)}} > 0$

Procedure for diagonalizing S_w (real symmetric and positive definite) and S_b (real symmetric) simultaneously is as follows :

1. Find λ_i by solving $\det(S_w^{-1}S_b - \lambda\mathbf{I}) = 0$ and find normalized $\boldsymbol{v}_i$ such that $S_w^{-1}S_b\boldsymbol{v}_i = \lambda_i\boldsymbol{v}_i, i = 1, \cdots, n$

2. Normalize $\boldsymbol{v}_i$ such that $\mathbf{V}^T S_w \mathbf{V} = \mathbf{I}$

LDA for Multiple Classes c

- The within-class scatter matrix S_w,

$$S_w = \sum_{i=1}^{c} \boldsymbol{C}_i$$

where $\boldsymbol{C}_i = \frac{1}{N_i}\sum_{n=1}^{N_i}(\boldsymbol{x}_n - \overline{\boldsymbol{x}}_i)(\boldsymbol{x}_n - \overline{\boldsymbol{x}}_i)^T$, $\overline{\boldsymbol{x}}_i = \frac{1}{N_i}\sum_{n=1}^{N_i}\boldsymbol{x}_n$

- The between-class scatter matrix S_b

$$S_b = \sum_{i=1}^{c} N_i\,(\overline{\boldsymbol{x}}_i - \overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}_i - \overline{\boldsymbol{x}})^T$$

where $\overline{\boldsymbol{x}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n$ is the overall mean, $N = N_1 + \cdots + N_c$

https://en.wikipedia.org/wiki/Linear_discriminant_analysis#Multiclass_LDA

Main steps to perform LDA

1. Standardize the *d*-dimensional dataset (*d* is the number of features).
   - Standardization shifts the mean of each feature so that it is centered at zero and each feature has a standard deviation of 1 (unit variance)
2. For each class, compute the *d*-dimensional mean vector.
3. Construct the S_b, and the S_w.
4. Compute the eigenvectors and corresponding eigenvalues of $S_w^{-1}S_b$.
5. Sort the eigenvalues by decreasing order to rank the corresponding eigenvectors.
6. Choose the *k* eigenvectors that correspond to the *k* largest eigenvalues to construct a $d \times k$-dimensional transformation matrix, W; the eigenvectors are the columns of this matrix.
7. Project the examples onto the new feature subspace using the transformation matrix, W.

Implementation trick

$$\underbrace{\boldsymbol{X}}_{N_i\times d} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_{N_i}^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N_i,1} & x_{N_i,2} & \cdots & x_{N_i,d} \end{bmatrix} \text{ or } \underbrace{\boldsymbol{X}^T}_{d\times N_i} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_{N_i}] = \begin{bmatrix} x_{1,1} & x_{2,1} & \cdots & x_{N_i,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{N_i,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,d} & x_{2,d} & \cdots & x_{N_i,d} \end{bmatrix}$$

$$\sum_{n=1}^{N_i} \underbrace{\boldsymbol{x}_n \boldsymbol{x}_n^T}_{(d\times 1)\times(1\times d)}$$

$$= \sum_{n=1}^{N_i} \begin{bmatrix} x_{n,1} \\ x_{n,2} \\ \vdots \\ x_{n,d} \end{bmatrix} [x_{n,1} \quad x_{n,2} \quad \cdots \quad x_{n,d}] = \sum_{n=1}^{N_i} \begin{bmatrix} x_{n,1}x_{n,1} & x_{n,1}x_{n,2} & \cdots & x_{n,1}x_{n,d} \\ x_{n,2}x_{n,1} & x_{n,2}x_{n,2} & \cdots & x_{n,2}x_{n,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,d}x_{n,1} & x_{n,d}x_{n,2} & \cdots & x_{n,d}x_{n,d} \end{bmatrix}$$

$$= \begin{bmatrix} x_{1,1} & x_{2,1} & \cdots & x_{N_i,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{N_i,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,d} & x_{2,d} & \cdots & x_{N_i,d} \end{bmatrix} \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N_i,1} & x_{N_i,2} & \cdots & x_{N_i,d} \end{bmatrix} = \underbrace{\boldsymbol{X}^T}_{d\times N_i} \underbrace{\boldsymbol{X}}_{N_i\times d}$$

https://en.wikipedia.org/wiki/Linear_discriminant_analysis#Multiclass_LDA

Homework # 4

Finish Homework 4 LDA_sklearn_to_do.jpynb using skylearn.

Deadline of Homework #4: 2022/10/17 3:30pm