DIMENSIONALITY REDUCTION

Background

$$D = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1^T & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2^T & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n^T & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

The goal of dimensionality reduction is to find a lower dimensional representation of the data matrix D to avoid the curse of dimensionality.

Given $n \times d$ data matrix, each point $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T$ is a vector in the ambient d-dimensional vector space spanned by the d standard basis vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_d$.

Background

Given any other set of d orthonormal vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d$ we can re-express each point $\boldsymbol{x}$ as

$$\boldsymbol{x} = a_1\boldsymbol{u}_1 + a_2\boldsymbol{u}_2 + \cdots + a_d\boldsymbol{u}_d$$

where $\boldsymbol{a} = (a_1, a_2, \ldots, a_d)^T$ represents the coordinates of $\boldsymbol{x}$ in the new basis. More compactly:

$$\boldsymbol{x} = \boldsymbol{U}\boldsymbol{a}$$

where $\mathbf{U}$ is the $d \times d$ matrix, whose ith column comprises the ith basis vector $\mathbf{u}_i$:

$$\mathbf{U} = \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_d \\ | & | & & | \end{pmatrix}$$

$\boldsymbol{U}$ is the $d \times d$ orthogonal matrix

$$\boldsymbol{U}^{-1} = \boldsymbol{U}^T$$

$$\boldsymbol{a} = \boldsymbol{U}^T\boldsymbol{x}$$

Background

we are interested in finding the optimal r-dimensional representation of D, with r ≪ d

➢ It is natural to ask whether we can find a reduced dimensionality subspace that still preserves the essential characteristics of the data
➢ In other words, given a point x, and assuming that the basis vectors have been sorted in decreasing order of importance

$$\mathbf{x}' = a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \cdots + a_r\mathbf{u}_r = \sum_{i=1}^{r} a_i\mathbf{u}_i$$

$$\mathbf{x}' = \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_r \\ | & | & & | \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_r \end{pmatrix} = \mathbf{U}_r\mathbf{a}_r$$

Here $\mathbf{x}'$ is the projection of $\mathbf{x}$ onto the first r basis vectors.

Background

$$\mathbf{a} = \mathbf{U}^T\mathbf{x} \qquad\qquad \mathbf{a}_r = \mathbf{U}_r^T\mathbf{x}$$

The r-dimensional projection of $\mathbf{x}$ is thus given as:

$$\boxed{\mathbf{x}' = \mathbf{U}_r\mathbf{U}_r^T\mathbf{x} = \mathbf{P}_r\mathbf{x}}$$

where $\mathbf{P}_r = \mathbf{U}_r\mathbf{U}_r^T = \sum_{i=1}^{r}\mathbf{u}_i\mathbf{u}_i^T$ is the *orthogonal projection matrix* for the subspace spanned by the first r basis vectors.

Background

The goal of dimensionality reduction is to seek an r-dimensional basis that gives the best possible approximation x_i' over all the points $x_i \in D$. Alternatively, we seek to minimize the error $\epsilon_i = x_i - x_i'$ over all the points.

PRINCIPAL COMPONENT ANALYSIS

PCA

➢ Principal Component Analysis (PCA) is a technique that seeks a r-dimensional basis that best captures the variance in the data.

➢ The direction with the largest projected variance is called the first principal component.

➢ The orthogonal direction that captures the second largest projected variance is called the second principal component, and so on.

PCA

$$\|\mathbf{u}\|^2 = \mathbf{u}^T\mathbf{u} = 1$$

$$\overline{\mathbf{D}} = \mathbf{D} - \mathbf{1}\cdot\mu^T$$

The projection of the centered point $\bar{\mathbf{x}}_i \in \overline{\mathbf{D}}$ on the vector $\mathbf{u}$

$$\mathbf{x}'_i = \left(\frac{\mathbf{u}^T\bar{\mathbf{x}}_i}{\mathbf{u}^T\mathbf{u}}\right)\mathbf{u} = (\mathbf{u}^T\bar{\mathbf{x}}_i)\mathbf{u} = a_i\mathbf{u}$$

choose the direction u such that the variance of the projected points is maximized

$$\sigma_{\mathbf{u}}^2 = \frac{1}{n}\sum_{i=1}^{n}(a_i - \mu_a)^2 \qquad \mu_a = \frac{1}{n}\sum_{i=1}^{n}a_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{u}^T(\bar{\mathbf{x}}_i) = \mathbf{u}^T\bar{\mu} = 0$$

PCA

$$\sigma_{\mathbf{u}}^2 = \frac{1}{n}\sum_{i=1}^{n}(a_i - \mu_a)^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^T\bar{\mathbf{x}}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{u}^T\left(\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\right)\mathbf{u} = \mathbf{u}^T\left(\frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\right)\mathbf{u}$$

$$\sigma_{\mathbf{u}}^2 = \mathbf{u}^T\Sigma\mathbf{u}$$

where Σ is the sample covariance matrix for the centered data $\bar{\mathbf{D}}$

$$\max_{\mathbf{u}} \quad \mathbf{u}^T\Sigma\mathbf{u}$$
$$\mathbf{u}^T\mathbf{u} = 1$$

PCA

Lagrangian multiplier

$$\max_{\mathbf{u}} J(\mathbf{u}) = \mathbf{u}^T \Sigma \mathbf{u} - \alpha(\mathbf{u}^T \mathbf{u} - 1)$$

$$\frac{\partial}{\partial \mathbf{u}} J(\mathbf{u}) = \mathbf{0}$$

$$\mathbf{u}^T \Sigma \mathbf{u} = \mathbf{u}^T \alpha \mathbf{u} = \alpha \mathbf{u}^T \mathbf{u} = \alpha$$

$$\frac{\partial}{\partial \mathbf{u}} \left(\mathbf{u}^T \Sigma \mathbf{u} - \alpha(\mathbf{u}^T \mathbf{u} - 1) \right) = \mathbf{0}$$

$$2\Sigma \mathbf{u} - 2\alpha \mathbf{u} = \mathbf{0}$$

$$\boxed{\Sigma \mathbf{u} = \alpha \mathbf{u}}$$

To maximize the projected variance $\sigma_{\boldsymbol{u}}^2$, we thus choose the largest eigenvalue λ_1 of Σ, and the dominant eigenvector $\boldsymbol{u}_1$ specifies the direction of most variance, also called the *first principal component*.

PCA

direction that maximizes the projected variance is also the one that minimizes the average squared error

$$MSE(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{\epsilon}_i\|^2 = \frac{1}{n}\sum_{i=1}^{n}\|\bar{\mathbf{x}}_i - \mathbf{x}'_i\|^2 = \frac{1}{n}\sum_{i=1}^{n}(\bar{\mathbf{x}}_i - \mathbf{x}'_i)^T(\bar{\mathbf{x}}_i - \mathbf{x}'_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\|\bar{\mathbf{x}}_i\|^2 - 2\bar{\mathbf{x}}_i^T\mathbf{x}'_i + (\mathbf{x}'_i)^T\mathbf{x}'_i\right) \qquad (7.15)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\|\bar{\mathbf{x}}_i\|^2 - 2\bar{\mathbf{x}}_i^T(\mathbf{u}^T\bar{\mathbf{x}}_i)\mathbf{u} + \left((\mathbf{u}^T\bar{\mathbf{x}}_i)\mathbf{u}\right)^T(\mathbf{u}^T\bar{\mathbf{x}}_i)\mathbf{u}\right), \text{ since } \mathbf{x}'_i = (\mathbf{u}^T\bar{\mathbf{x}}_i)\mathbf{u}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\|\bar{\mathbf{x}}_i\|^2 - 2(\mathbf{u}^T\bar{\mathbf{x}}_i)(\bar{\mathbf{x}}_i^T\mathbf{u}) + (\mathbf{u}^T\bar{\mathbf{x}}_i)(\bar{\mathbf{x}}_i^T\mathbf{u})\mathbf{u}^T\mathbf{u}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\|\bar{\mathbf{x}}_i\|^2 - (\mathbf{u}^T\bar{\mathbf{x}}_i)(\bar{\mathbf{x}}_i^T\mathbf{u})\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\|\bar{\mathbf{x}}_i\|^2 - \frac{1}{n}\sum_{i=1}^{n}\mathbf{u}^T(\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T)\mathbf{u}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\|\bar{\mathbf{x}}_i\|^2 - \mathbf{u}^T\left(\frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\right)\mathbf{u}$$
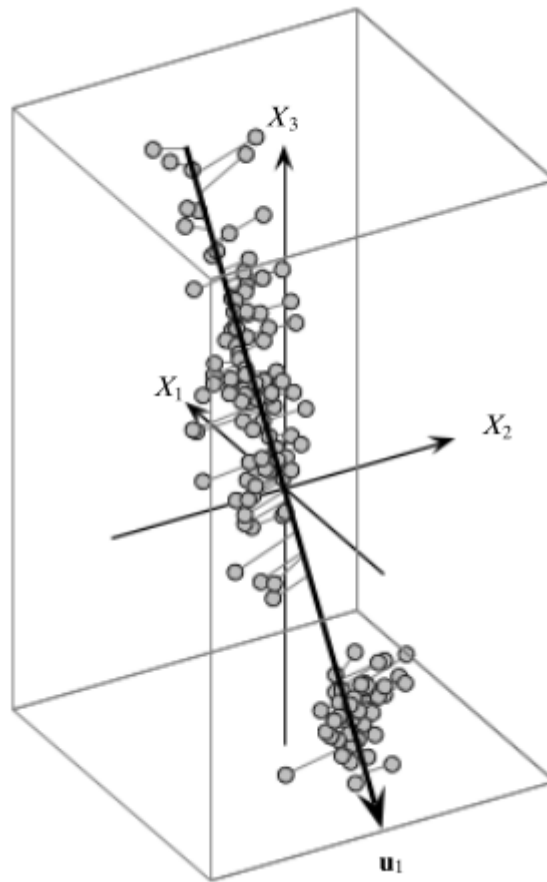
PCA

$$MSE = \sum_{i=1}^{n} \frac{\|\bar{\mathbf{x}}_i\|^2}{n} - \mathbf{u}^T \mathbf{\Sigma} \mathbf{u}$$

$$\text{var}(\mathbf{D}) = tr(\mathbf{\Sigma}) = \sum_{i=1}^{d} \sigma_i^2$$

$$MSE(\mathbf{u}) = \text{var}(\mathbf{D}) - \mathbf{u}^T \mathbf{\Sigma} \mathbf{u} = \sum_{i=1}^{d} \sigma_i^2 - \mathbf{u}^T \mathbf{\Sigma} \mathbf{u}$$

example

PCA

- Best 2-dimensional Approximation

- We already computed the direction with
  the most variance, namely u1, which is the eigenvector corresponding to the largest
  eigenvalue λ1 of

- We now want to find another direction v, which also maximizes the projected variance, but is orthogonal to u1.

$$\max_{\mathbf{v}} \quad \sigma_{\mathbf{v}}^2 = \mathbf{v}^T \Sigma \mathbf{v}$$

$$\mathbf{v}^T \mathbf{u}_1 = 0$$

$$\mathbf{v}^T \mathbf{v} = 1$$

second largest eigenvalue of °, with the second principal component being given by the corresponding eigenvector, that is, v = u2.

PCA

To find the best r-dimensional approximation to D, we compute the eigenvalues of Σ. Because Σ is positive semidefinite, its eigenvalues are non-negative

$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_r \geq \lambda_{r+1} \cdots \geq \lambda_d \geq 0$$

We select the r largest eigenvalues, and their corresponding eigenvectors to form the best r-dimensional approximation.

PCA

PCA (D, r)

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ // compute mean}$$

$$Z = D - 1 \cdot \mu^T \text{ // center the data}$$

$$\Sigma = \frac{1}{n}\left(Z^T Z\right) \text{ // compute covariance matrix}$$

$$(\lambda_1, \lambda_2, \ldots, \lambda_d) = \text{eigenvalues}(\Sigma) \text{ // compute eigenvalues}$$

$$U = \begin{pmatrix} u_1 & u_2 & \cdots & u_d \end{pmatrix} = \text{eigenvectors}(\Sigma) \text{ // compute}$$
eigenvectors

$$U_r = \begin{pmatrix} u_1 & u_2 & \cdots & u_r \end{pmatrix} \text{ // reduced basis}$$

$$A = \{a_i \mid a_i = U_r^T x_i, \text{for } i = 1, \ldots, n\} \text{ // reduced dimensionality}$$
data

LINEAR DISCRIMINANT ANALYSIS (LDA)

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

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

LDA

Let $\mathbf{D}_i$ denote the subset of points labeled with class c_i, i.e., $\mathbf{D}_i = \{\mathbf{x}_j | y_j = c_i\}$, and let $|\mathbf{D}_i| = n_i$ 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 $\mathbf{x}_i$ onto a unit vector $\mathbf{w}$ is given as
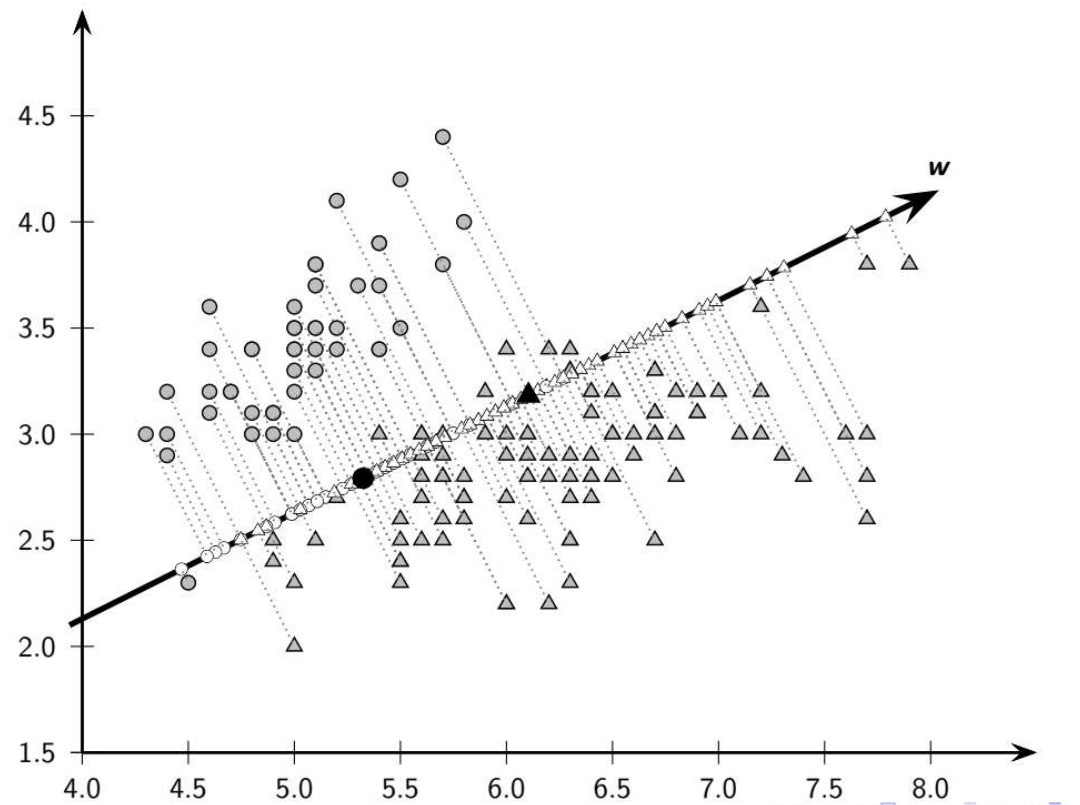
$$\mathbf{x}_i' = \left(\frac{\mathbf{w}^T \mathbf{x}_i}{\mathbf{w}^T \mathbf{w}} \right) \mathbf{w} = \left(\mathbf{w}^T \mathbf{x}_i \right) \mathbf{w} = a_i \mathbf{w}$$

where a_i specifies the offset or coordinate of $\mathbf{x}_i'$ along the line $\mathbf{w}$:
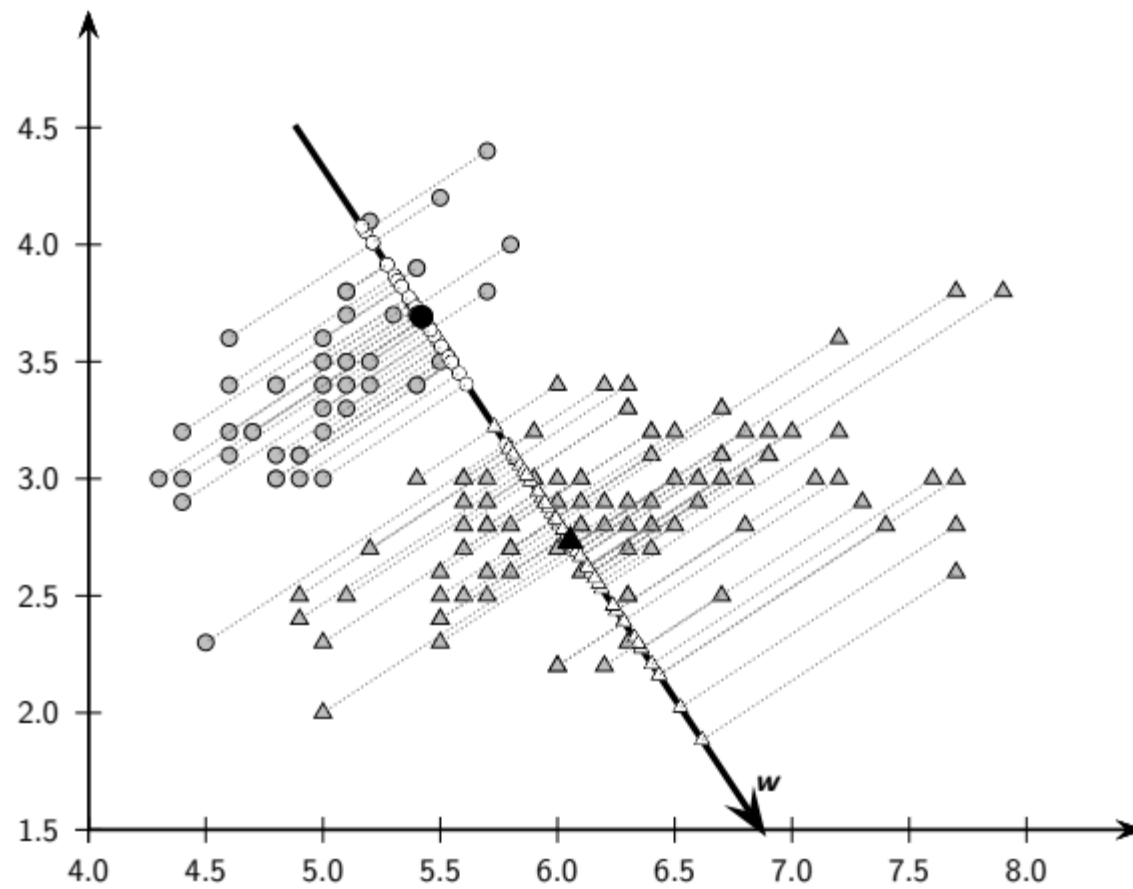
$$a_i = \mathbf{w}^T \mathbf{x}_i$$

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 $\mathbf{w}$).

LDA

LDA

LDA

To maximize the separation between the classes, it seems reasonable to maximize the difference between the projected means, $|m_1 - m_2|$.

$$m_1 = \frac{1}{n_1} \sum_{\mathbf{x}_i \in \mathbf{D}_1} a_i \qquad\qquad m_2 = \mathbf{w}^T \boldsymbol{\mu}_2$$

$$= \frac{1}{n_1} \sum_{\mathbf{x}_i \in \mathbf{D}_1} \mathbf{w}^T \mathbf{x}_i$$

$$= \mathbf{w}^T \left(\frac{1}{n_1} \sum_{\mathbf{x}_i \in \mathbf{D}_1} \mathbf{x}_i \right)$$

$$= \mathbf{w}^T \boldsymbol{\mu}_1$$

LDA

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

$$s_i^2 = \sum_{x_j \in D_i} (a_j - m_i)^2 = n_i \sigma_i^2$$

where σ_i^2 is the variance for class c_i.

LDA

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*:

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

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

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

LDA

$$s_1^2 = \sum_{\mathbf{x}_i \in \mathbf{D}_1} (a_i - m_1)^2$$

$$= \sum_{\mathbf{x}_i \in \mathbf{D}_1} (\mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \boldsymbol{\mu}_1)^2$$

$$= \sum_{\mathbf{x}_i \in \mathbf{D}_1} \left(\mathbf{w}^T (\mathbf{x}_i - \boldsymbol{\mu}_1) \right)^2$$

$$= \mathbf{w}^T \left(\sum_{\mathbf{x}_i \in \mathbf{D}_1} (\mathbf{x}_i - \boldsymbol{\mu}_1)(\mathbf{x}_i - \boldsymbol{\mu}_1)^T \right) \mathbf{w}$$

$$= \mathbf{w}^T \mathbf{S}_1 \mathbf{w}$$

where $\mathbf{S}_1$ is the *scatter matrix* for $\mathbf{D}_1$. Likewise, we can obtain

$$s_2^2 = \mathbf{w}^T \mathbf{S}_2 \mathbf{w}$$

LDA

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

where $\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2$ denotes the *within-class scatter matrix*

$$\max_{\mathbf{w}} \ J(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{Bw}}{\mathbf{w}^T\mathbf{Sw}}$$

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

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

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

$$\mathbf{S}^{-1}\mathbf{Bw} = \lambda\mathbf{S}^{-1}\mathbf{Sw}$$

$$(\mathbf{S}^{-1}\mathbf{B})\mathbf{w} = \lambda\mathbf{w}$$

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

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

LDA

LINEARDISCRIMINANT (D):

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

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

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

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

5 $\mathbf{S}_i \leftarrow \overline{\mathbf{D}}_i^T \overline{\mathbf{D}}_i, i = 1, 2$ // class scatter matrices

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

7 $\lambda_1, \mathbf{w} \leftarrow \mathrm{eigen}(\mathbf{S}^{-1}\mathbf{B})$ // compute dominant eigenvector