ML&MEA (2024)
Lecture 12 - Linear Discriminant Analysis (LDA)

Quanying Liu

BME, SUSTech

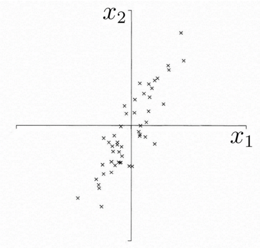
2024.4.16

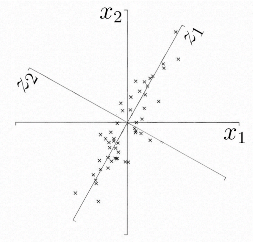
Content

1. Recap: PCA
2. LDA: model
3. LDA: solution
4. LDA: Eigendecomposition
5. Summary

What is Principal Component Analysis?

- (Wikipedia) **PCA** is the process of computing the principal components (PCs) and using them to perform a change of **basis** on the data. (转换坐标系: orthonormal basis)
- **Variance of samples**. The principal components are ordered by the variance of PCs.
- **Goals**: 最大化投影方差 $\Longleftrightarrow$ 最小化重构误差





PCA: maximizing projection variance

- To find the optimal vector $\mathbf{z}_1^*$
- Objective: maximize the variance after projection $\mathbf{z}_1^T S \mathbf{z}_1$
- Constraints: $\mathbf{z}_1^T \mathbf{z}_1 = 1$

  Let us denote λ as a Lagrange multiplier.
  The generalized Lagrangian function is:

$$\mathcal{L}(\mathbf{z}_1, \lambda) = -\mathbf{z}_1^T S \mathbf{z}_1 + \lambda(\mathbf{z}_1^T \mathbf{z}_1 - 1)$$

  The derivative of $\mathcal{L}$ is:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{z}_1} = -2S\mathbf{z}_1 + 2\lambda \mathbf{z}_1 = S\mathbf{z}_1 - \lambda \mathbf{z}_1 = 0$$

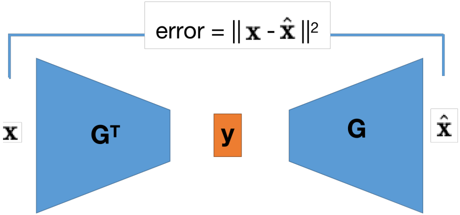
  $\mathbf{z}_1$ is an eigenvector of S, $\lambda = \lambda_1$ is the largest eigenvalue

PCA: minimizing reconstruction error

- **Data**: $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)^T \in \mathbb{R}^{N\times p}$
- **Projections** with PCA as encoder and decoder,

$$\text{Encoder:} \quad \mathbf{y}_i = G^T \mathbf{x}_i \in \mathbb{R}^q$$
$$\text{Decoder:} \quad \hat{\mathbf{x}}_i = G\mathbf{y}_i \quad \in \mathbb{R}^p$$

where $G = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_q] \in \mathbb{R}^{p\times q}$.



Reconstruction error

- Reconstruction error across N samples:

$$\mathcal{L}(W) = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 \tag{1}$$

- Substitute $G = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_q]$ to Eq. (1)

$$= \frac{1}{N}\sum_{i=1}^{N} \|ZZ^T\mathbf{x}_i - GG^T\mathbf{x}_i\|^2$$

$$= \frac{1}{N}\sum_{i=1}^{N} \|\sum_{k=q+1}^{p} \mathbf{z}_k\mathbf{z}_k^T\mathbf{x}_i\|^2 = \frac{1}{N}\sum_{i=1}^{N}(\sum_{k=q+1}^{p} \mathbf{z}_k^T\mathbf{x}_i)^2$$

$$= \sum_{k=q+1}^{p} \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_k^T\mathbf{x}_i)^2 = \sum_{k=q+1}^{p} \mathbf{z}_k^T {\color{red}S}\mathbf{z}_k$$

minimize error = maximize variance

Minimize reconstruction error（最小化重构误差）：

$$\mathbf{z}_k = \arg\min_{\mathbf{z}_k} \sum_{k=q+1}^{p} \mathbf{z}_k^T S \mathbf{z}_k$$

$$\text{s.t.} \quad \mathbf{z}_k^T \mathbf{z}_k = 1$$

Maximizing projection variance（最大化投影方差）：

$$\mathbf{z}_i = \arg\max_{\mathbf{z}_i} \mathbf{z}_i^T S \mathbf{z}_i$$

$$\text{s.t.} \quad \mathbf{z}_i^T \mathbf{z}_i = 1$$

SVD on the data matrix HX

- **Singular Value Decomposition** (SVD) on the matrix HX, where $H = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N^T\mathbf{1}_N, \quad H^T = H, \quad H^2 = H.$

$$HX = U\Sigma V^T \tag{2}$$

$$U^T U = \mathbf{I}_p, \quad V^T V = VV^T = \mathbf{I}_p, \quad \Sigma = diag([\sigma_1, \ \ldots, \ \sigma_p])$$

- Substitute Eq. (2) to the covariance matrix S, we derive

$$\begin{aligned} S &= \frac{1}{N}\mathbf{X}^T HH^T \mathbf{X} = \frac{1}{N}(HX)^T(HX) \\ &= \frac{1}{N}(U\Sigma V^T)^T U\Sigma V^T = \frac{1}{N}V\Sigma U^T U\Sigma V^T \\ &= \frac{1}{N}V\Sigma^2 V^T \Longrightarrow \text{Eigendecoposition of } S \end{aligned} \tag{3}$$

Relationship between PCA and SVD

表 1: Relationship between PCA and SVD

PCA	**SVD**
Eigendecomposition on S	SVD on HX
$S = Z\Lambda Z^T$	$HX = U\Sigma V^T$ $S = \frac{1}{N} V\Sigma^2 V^T$
$\Lambda = diag([\lambda_1, \lambda_2, \ldots, \lambda_p])$	$\Sigma^2 = diag([\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2])$

PCA might not be suitable for classification

- We usually use PCA for *dimensionality reduction*.
- **A major problem**: PCA may not be suitable for classification.
- Q: Why?

  PCA is based on the sample covariance S which characterizes the scatter of the entire data set, **regardless of the class label**.

  The projection axes chosen by PCA might not provide good discrimination between classes.
- Solution: $\longrightarrow$ Supervised feature extraction
- Today's method: Linear Discriminant Analysis (LDA)

LDA: 线性判别分析

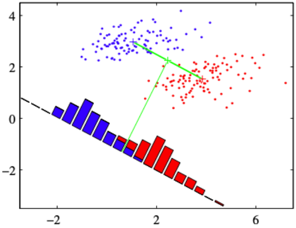
- (Wikipedia) **Linear discriminant analysis** (LDA), *normal discriminant analysis* (NDA), or *discriminant function analysis* finds a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for **dimensionality reduction** before later classification.

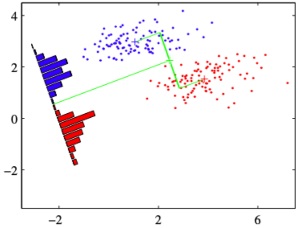
LDA: 线性判别分析

- (Wikipedia) **Linear discriminant analysis** (LDA), *normal discriminant analysis* (NDA), or *discriminant function analysis* finds a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for **dimensionality reduction** before later classification.
- **Goals**: 类间大，类内小

LDA: 线性判别分析

- (Wikipedia) **Linear discriminant analysis** (LDA), *normal discriminant analysis* (NDA), or *discriminant function analysis* finds a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for **dimensionality reduction** before later classification.
- **Goals**: 类间大，类内小





LDA: training data

- Data: $\{(\mathbf{x}_i,\ y_i)\}_{i=1}^{N}$

LDA: training data

- Data: $\{(\mathbf{x}_i,\ y_i)\}_{i=1}^{N}$

  $\mathbf{x}_i \in \mathbb{R}^p, \quad y_i \in \{+1,\ -1\}$

LDA: training data

- Data: $\{(\mathbf{x}_i,\ y_i)\}_{i=1}^{N}$
  $$\mathbf{x}_i \in \mathbb{R}^p, \quad y_i \in \{+1,\ -1\}$$
- Two Classes: $\mathcal{C}_1$ & $\mathcal{C}_2$

LDA: training data

- Data: $\{(\mathbf{x}_i,\ y_i)\}_{i=1}^{N}$

  $\mathbf{x}_i \in \mathbb{R}^p, \quad y_i \in \{+1,\ -1\}$
- Two Classes: $\mathcal{C}_1$ & $\mathcal{C}_2$
- Data in $\mathcal{C}_1$: $X_{\mathcal{C}_1} = \{\mathbf{x}_i | y_i = +1\}$

  sample size: $N_{\mathcal{C}_1}$

  mean: $\bar{\mathbf{x}}_{\mathcal{C}_1} = \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} \mathbf{x}_i, \quad \mathbf{x}_i \in X_{\mathcal{C}_1}$

  variance: $S_{\mathcal{C}_1} = \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} (\mathbf{x}_i - \bar{\mathbf{x}}_{\mathcal{C}_1})(\mathbf{x}_i - \bar{\mathbf{x}}_{\mathcal{C}_1})^T$

LDA: training data

- Data: $\{(\mathbf{x}_i,\ y_i)\}_{i=1}^{N}$

  $$\mathbf{x}_i \in \mathbb{R}^p, \quad y_i \in \{+1,\ -1\}$$

- Two Classes: $\mathcal{C}_1$ & $\mathcal{C}_2$
- Data in $\mathcal{C}_1$: $X_{\mathcal{C}_1} = \{\mathbf{x}_i | y_i = +1\}$

  sample size: $N_{\mathcal{C}_1}$

  mean: $\bar{\mathbf{x}}_{\mathcal{C}_1} = \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} \mathbf{x}_i, \quad \mathbf{x}_i \in X_{\mathcal{C}_1}$

  variance: $S_{\mathcal{C}_1} = \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} (\mathbf{x}_i - \bar{\mathbf{x}}_{\mathcal{C}_1})(\mathbf{x}_i - \bar{\mathbf{x}}_{\mathcal{C}_1})^T$

- Data in $\mathcal{C}_2$: $X_{\mathcal{C}_2} = \{\mathbf{x}_i | y_i = -1\}$

  sample size: $N_{\mathcal{C}_2}$

  mean: $\bar{\mathbf{x}}_{\mathcal{C}_2} = \frac{1}{N_{\mathcal{C}_2}} \sum_{i=1}^{N_{\mathcal{C}_2}} \mathbf{x}_i, \quad \mathbf{x}_i \in X_{\mathcal{C}_2}$

  variance: $S_{\mathcal{C}_2} = \frac{1}{N_{\mathcal{C}_2}} \sum_{i=1}^{N_{\mathcal{C}_2}} (\mathbf{x}_i - \bar{\mathbf{x}}_{\mathcal{C}_2})(\mathbf{x}_i - \bar{\mathbf{x}}_{\mathcal{C}_2})^T$

Projection

- Project all the training data to the direction, w.

Projection

- Project all the training data to the direction, w.
- The decision boundary is $\mathbf{w}^T\mathbf{x} = 0$, where w is orthogonal to the boundary.

Projection

- Project all the training data to the direction, w.
- The decision boundary is $\mathbf{w}^T\mathbf{x} = 0$, where w is orthogonal to the boundary.
- Data after projection: $z_i = \mathbf{w}^T\mathbf{x}_i$

Projection

- Project all the training data to the direction, w.
- The decision boundary is $\mathbf{w}^T\mathbf{x} = 0$, where w is orthogonal to the boundary.
- Data after projection: $z_i = \mathbf{w}^T\mathbf{x}_i$
- Q: what is the dimension of z_i ?

Two-class Data after projection

- Data after projection: $z_i = \mathbf{w}^T \mathbf{x}_i$

Two-class Data after projection

- Data after projection: $z_i = \mathbf{w}^T \mathbf{x}_i$
- Class 1: mean: $\bar{z}_1 = \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} \mathbf{w}^T \mathbf{x}_i, \quad \mathbf{x}_i \in X_{\mathcal{C}_1}$

$$\begin{aligned} \text{variance: } S_1 &= \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} (z_i - \bar{z}_1)(z_i - \bar{z}_1)^T \\ &= \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} (\mathbf{w}^T \mathbf{x}_i - \bar{z}_1)(\mathbf{w}^T \mathbf{x}_i - \bar{z}_1)^T \end{aligned} \tag{4}$$

Two-class Data after projection

- Data after projection: $z_i = \mathbf{w}^T \mathbf{x}_i$
- Class 1: mean: $\bar{z}_1 = \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} \mathbf{w}^T \mathbf{x}_i, \quad \mathbf{x}_i \in X_{\mathcal{C}_1}$

$$\begin{aligned} \text{variance: } S_1 &= \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} (z_i - \bar{z}_1)(z_i - \bar{z}_1)^T \\ &= \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} (\mathbf{w}^T \mathbf{x}_i - \bar{z}_1)(\mathbf{w}^T \mathbf{x}_i - \bar{z}_1)^T \end{aligned} \tag{4}$$

- Class 2: mean: $\bar{z}_2 = \frac{1}{N_{\mathcal{C}_2}} \sum_{i=1}^{N_{\mathcal{C}_2}} \mathbf{w}^T \mathbf{x}_i, \quad \mathbf{x}_i \in X_{\mathcal{C}_2}$

$$\begin{aligned} \text{variance: } S_2 &= \frac{1}{N_{\mathcal{C}_2}} \sum_{i=1}^{N_{\mathcal{C}_2}} (z_i - \bar{z}_2)(z_i - \bar{z}_2)^T \\ &= \frac{1}{N_{\mathcal{C}_2}} \sum_{i=1}^{N_{\mathcal{C}_2}} (\mathbf{w}^T \mathbf{x}_i - \bar{z}_2)(\mathbf{w}^T \mathbf{x}_i - \bar{z}_2)^T \end{aligned} \tag{5}$$

Between-class and Within-class

- Class 1: mean: $\bar{z}_1 = \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} \mathbf{w}^T \mathbf{x}_i, \quad \mathbf{x}_i \in X_{\mathcal{C}_1}$

  variance: $S_1 = \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} (\mathbf{w}^T \mathbf{x}_i - \bar{z}_1)(\mathbf{w}^T \mathbf{x}_i - \bar{z}_1)^T$

Between-class and Within-class

- Class 1: mean: $\bar{z}_1 = \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} \mathbf{w}^T \mathbf{x}_i, \quad \mathbf{x}_i \in X_{\mathcal{C}_1}$

$$\text{variance: } S_1 = \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} (\mathbf{w}^T \mathbf{x}_i - \bar{z}_1)(\mathbf{w}^T \mathbf{x}_i - \bar{z}_1)^T$$

- Class 2: mean: $\bar{z}_2 = \frac{1}{N_{\mathcal{C}_2}} \sum_{i=1}^{N_{\mathcal{C}_2}} \mathbf{w}^T \mathbf{x}_i, \quad \mathbf{x}_i \in X_{\mathcal{C}_2}$

$$\text{variance: } S_2 = \frac{1}{N_{\mathcal{C}_2}} \sum_{i=1}^{N_{\mathcal{C}_2}} (\mathbf{w}^T \mathbf{x}_i - \bar{z}_2)(\mathbf{w}^T \mathbf{x}_i - \bar{z}_2)^T$$

Between-class and Within-class

- Class 1: mean: $\bar{z}_1 = \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} \mathbf{w}^T \mathbf{x}_i, \quad \mathbf{x}_i \in X_{\mathcal{C}_1}$

  $$\text{variance:} \quad S_1 = \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} (\mathbf{w}^T \mathbf{x}_i - \bar{z}_1)(\mathbf{w}^T \mathbf{x}_i - \bar{z}_1)^T$$

- Class 2: mean: $\bar{z}_2 = \frac{1}{N_{\mathcal{C}_2}} \sum_{i=1}^{N_{\mathcal{C}_2}} \mathbf{w}^T \mathbf{x}_i, \quad \mathbf{x}_i \in X_{\mathcal{C}_2}$

  $$\text{variance:} \quad S_2 = \frac{1}{N_{\mathcal{C}_2}} \sum_{i=1}^{N_{\mathcal{C}_2}} (\mathbf{w}^T \mathbf{x}_i - \bar{z}_2)(\mathbf{w}^T \mathbf{x}_i - \bar{z}_2)^T$$

- Between-class(类间): $(\bar{z}_1 - \bar{z}_2)^2$

Between-class and Within-class

- Class 1: mean: $\bar{z}_1 = \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} \mathbf{w}^T \mathbf{x}_i, \quad \mathbf{x}_i \in X_{\mathcal{C}_1}$

$$\text{variance: } S_1 = \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} (\mathbf{w}^T \mathbf{x}_i - \bar{z}_1)(\mathbf{w}^T \mathbf{x}_i - \bar{z}_1)^T$$

- Class 2: mean: $\bar{z}_2 = \frac{1}{N_{\mathcal{C}_2}} \sum_{i=1}^{N_{\mathcal{C}_2}} \mathbf{w}^T \mathbf{x}_i, \quad \mathbf{x}_i \in X_{\mathcal{C}_2}$

$$\text{variance: } S_2 = \frac{1}{N_{\mathcal{C}_2}} \sum_{i=1}^{N_{\mathcal{C}_2}} (\mathbf{w}^T \mathbf{x}_i - \bar{z}_2)(\mathbf{w}^T \mathbf{x}_i - \bar{z}_2)^T$$

- Between-class(类间): $(\bar{z}_1 - \bar{z}_2)^2$
- Within-class(类内): $S_1 + S_2$

LDA: objective function (类间大，类内小)

- Between-class (类间): $(\bar{z}_1 - \bar{z}_2)^2$ ↑

LDA: objective function (类间大，类内小)

- Between-class (类间): $(\bar{z}_1 - \bar{z}_2)^2 \uparrow$
- Within-class (类内): $S_1 + S_2 \downarrow$

LDA: objective function (类间大，类内小)

- Between-class (类间): $(\bar{z}_1 - \bar{z}_2)^2 \uparrow$
- Within-class (类内): $S_1 + S_2 \downarrow$
- Formulate the objective function of LDA, $J(w)$, the goal is to maximize $J(w)$:

$$J(w) = \frac{(\bar{z}_1 - \bar{z}_2)^2}{S_1 + S_2}$$

LDA: objective function (类间大，类内小)

- Between-class (类间): $(\bar{z}_1 - \bar{z}_2)^2 \uparrow$
- Within-class (类内): $S_1 + S_2 \downarrow$
- Formulate the objective function of LDA, $J(w)$, the goal is to maximize $J(w)$:

$$J(w) = \frac{(\bar{z}_1 - \bar{z}_2)^2}{S_1 + S_2}$$

- 首先，化简分子

$$\begin{aligned}
(\bar{z}_1 - \bar{z}_2)^2 &= \left[\frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} \mathbf{w}^T \mathbf{x}_i - \frac{1}{N_{\mathcal{C}_2}} \sum_{i=1}^{N_{\mathcal{C}_2}} \mathbf{w}^T \mathbf{x}_i \right]^2 \\
&= \left[\mathbf{w}^T \left(\frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} \mathbf{x}_i - \frac{1}{N_{\mathcal{C}_2}} \sum_{i=1}^{N_{\mathcal{C}_2}} \mathbf{x}_i \right) \right]^2 \\
&= \left[\mathbf{w}^T (\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2}) \right]^2 = \mathbf{w}^T (\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2})(\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2})^T \mathbf{w}
\end{aligned}$$

LDA: objective function (类间大，类内小)

- Formulate the objective function of LDA, $J(w)$:

$$J(w) = \frac{(\bar{z}_1 - \bar{z}_2)^2}{S_1 + S_2}$$

- 然后，化简分母

$$\begin{aligned}
S_1 &= \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} (\mathbf{w}^T \mathbf{x}_i - \bar{z}_1)(\mathbf{w}^T \mathbf{x}_i - \bar{z}_1)^T \\
&= \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} (\mathbf{w}^T \mathbf{x}_i - \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} \mathbf{w}^T \mathbf{x}_i)(\mathbf{w}^T \mathbf{x}_i - \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} \mathbf{w}^T \mathbf{x}_i)^T \\
&= \mathbf{w}^T \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} (\mathbf{x}_i - \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} \mathbf{x}_i)(\mathbf{x}_i - \frac{1}{N_{\mathcal{C}_1}} \sum_{i=1}^{N_{\mathcal{C}_1}} \mathbf{x}_i)^T \mathbf{w} \\
&= \mathbf{w}^T S_{\mathcal{C}_1} \mathbf{w}
\end{aligned}$$

$$S_1 + S_2 = \mathbf{w}^T (S_{\mathcal{C}_1} + S_{\mathcal{C}_2}) \mathbf{w}$$

LDA: objective function (类间大，类内小)

- Formulate the objective function of LDA $J(w)$:

$$J(w) = \frac{(\bar{z}_1 - \bar{z}_2)^2}{S_1 + S_2}$$

- We derive the objective function as Eq. (6)

$$J(w) = \frac{\mathbf{w}^T(\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2})(\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2})^T\mathbf{w}}{\mathbf{w}^T(S_{\mathcal{C}_1} + S_{\mathcal{C}_2})\mathbf{w}} = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}} \tag{6}$$

where
$$S_b = (\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2})(\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2})^T$$
$$S_w = S_{\mathcal{C}_1} + S_{\mathcal{C}_2}$$

Gradient of the objective function in Eq. (6)

- The objective function in Eq. (6)

$$J(w) = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}} = (\mathbf{w}^T S_b \mathbf{w})(\mathbf{w}^T S_w \mathbf{w})^{-1}$$

- Differentiating the objective function *w.r.t.* $\mathbf{w}$, we derive

$$\begin{aligned}
\frac{\partial J(w)}{\partial \mathbf{w}} &= 2S_b\mathbf{w}(\mathbf{w}^T S_w \mathbf{w})^{-1} + (\mathbf{w}^T S_b \mathbf{w})(-1)(\mathbf{w}^T S_w \mathbf{w})^{-2} 2S_w\mathbf{w} \\
&= S_b\mathbf{w}(\mathbf{w}^T S_w \mathbf{w}) - (\mathbf{w}^T S_b \mathbf{w}) S_w \mathbf{w} = 0 \\
S_w\mathbf{w} &= \frac{(\mathbf{w}^T S_w \mathbf{w})}{(\mathbf{w}^T S_b \mathbf{w})} S_b \mathbf{w}
\end{aligned} \tag{7}$$

- We the optimal $\mathbf{w}^*$,

$$\mathbf{w}^* = S_w^{-1} \frac{(\mathbf{w}^T S_w \mathbf{w})}{(\mathbf{w}^T S_b \mathbf{w})} (\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2})(\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2})^T \mathbf{w}$$

We only care about the direction of $\mathbf{w}$

- Since the $\mathbf{w}$ can be re-scaled (*e.g.*, $|\mathbf{w}| = 1$), we only care about the direction of $\mathbf{w}$.
- Let's have a detailed look at the optimal $\mathbf{w}^*$,

$$\mathbf{w}^* = S_w^{-1} \frac{(\mathbf{w}^T S_w \mathbf{w})}{(\mathbf{w}^T S_b \mathbf{w})} (\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2})(\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2})^T \mathbf{w}$$

- We notice that
  - $\frac{(\mathbf{w}^T S_w \mathbf{w})}{(\mathbf{w}^T S_b \mathbf{w})}$ is a scalar.
  - $(\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2})^T \mathbf{w}$ is a scalar.

The direction of $\mathbf{w}^*$:

$$\mathbf{w}^* \propto S_w^{-1} (\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2}) \tag{8}$$

LDA: Eigendecomposition of $S_w^{-1} S_b$

Recall Eq. (7):

$$S_w \mathbf{w} = \frac{(\mathbf{w}^T S_w \mathbf{w})}{(\mathbf{w}^T S_b \mathbf{w})} S_b \mathbf{w}$$

Let's denote $\lambda = \frac{(\mathbf{w}^T S_b \mathbf{w})}{(\mathbf{w}^T S_w \mathbf{w})}$

$$\begin{aligned} S_w \mathbf{w} &= \frac{1}{\lambda} S_b \mathbf{w} \\ \lambda S_w \mathbf{w} &= S_b \mathbf{w} \\ \lambda \mathbf{w} &= S_w^{-1} S_b \mathbf{w} \end{aligned} \tag{9}$$

Now we can easily notice that $\mathbf{w}$ is the largest eigenvectors of $S_w^{-1} S_b$.

LDA: Eigendecomposition of $S_w^{-1} S_b$

Recall the objective function in Eq. (6):

$$J(\mathbf{w}) = \frac{(\mathbf{w}^T S_b \mathbf{w})}{(\mathbf{w}^T S_w \mathbf{w})}$$

Now we have

$$\lambda = \frac{(\mathbf{w}^T S_b \mathbf{w})}{(\mathbf{w}^T S_w \mathbf{w})}$$

Therefore, maximizing the objective function is actually to find the q-th largest eigenvalue of $S_w^{-1} S_b$.

LDA: algorithm

LDA: a supervised dimensionality reduction method

Given the **training data**: $\{(\mathbf{x}_i,\ y_i)\}_{i=1}^{N}$

$\mathbf{x}_i \in \mathbb{R}^p, \quad y_i \in \{+1,\ -1\}$

Algorithm

1. Calculate the sample mean of all training data: $\bar{\mathbf{x}}$, get $\mathbf{x} \longleftarrow \mathbf{x} - \bar{\mathbf{x}}$
2. Calculate the sample mean of each class: $\bar{\mathbf{x}}_{\mathcal{C}_i}$
3. Calculate the variance of each class: $S_{\mathcal{C}_i}$
4. Calculate $S_w = S_{\mathcal{C}_1} + S_{\mathcal{C}_2}$, $S_b = (\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2})(\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2})^T$
5. EigenDecomposition of $S_w^{-1} S_b$
6. Sort the eigenvalues in a descending order, and take the largest q eigenvalues and the corresponding eigenvectors
7. Get the projection matrix $W = [\mathbf{w}_1,\ \mathbf{w}_1, \ldots, \mathbf{w}_q]$

LDA: 类内小，类间大

- Between-class (类间): $(\bar{z}_1 - \bar{z}_2)^2$

LDA: 类内小，类间大

- Between-class (类间): $(\bar{z}_1 - \bar{z}_2)^2$
- Within-class (类内): $S_1 + S_2$

LDA: 类内小，类间大

- Between-class (类间): $(\bar{z}_1 - \bar{z}_2)^2$
- Within-class (类内): $S_1 + S_2$
- Objective function: $J(\mathbf{w}) = \frac{(\bar{z}_1 - \bar{z}_2)^2}{S_1 + S_2}$

LDA: 类内小，类间大

- Between-class (类间): $(\bar{z}_1 - \bar{z}_2)^2$
- Within-class (类内): $S_1 + S_2$
- Objective function: $J(\mathbf{w}) = \frac{(\bar{z}_1 - \bar{z}_2)^2}{S_1 + S_2}$
- We can derive the objective function as,

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}$$

Between-class scatter matrix: $S_b = (\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2})(\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2})^T$

Within-class scatter matrix: $S_w = S_{\mathcal{C}_1} + S_{\mathcal{C}_2}$

LDA: solution

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}$$

From $\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = 0$, we derive the solution of $\mathbf{w}^*$:

$$\mathbf{w}^* \propto S_w^{-1}(\bar{\mathbf{x}}_{\mathcal{C}_1} - \bar{\mathbf{x}}_{\mathcal{C}_2})$$

This is known as *Fisher's linear discriminant*.

From Eigen Decomposition Perspective:
$\mathbf{w}^*$ is the eigenvector of $S_w^{-1} S_b$.

LDA vs PCA

- **Similarity**:

LDA vs PCA

- **Similarity**:
  1. Both LDA and PCA reduce dimension.

LDA vs PCA

- **Similarity**:
  1. Both LDA and PCA reduce dimension.
  2. Both construct new features which are linear combination of original features.

LDA vs PCA

- **Similarity**:
  1. Both LDA and PCA reduce dimension.
  2. Both construct new features which are linear combination of original features.
  3. Both use Eigen Decomposition. (PCA: S; LDA: $S_w^{-1}S_b$)

LDA vs PCA

- **Similarity**:
  1. Both LDA and PCA reduce dimension.
  2. Both construct new features which are linear combination of original features.
  3. Both use Eigen Decomposition. (PCA: S; LDA: $S_w^{-1}S_b$)
- **Difference**:

LDA vs PCA

- **Similarity**:
  1. Both LDA and PCA reduce dimension.
  2. Both construct new features which are linear combination of original features.
  3. Both use Eigen Decomposition. (PCA: S; LDA: $S_w^{-1}S_b$)
- **Difference**:
  1. PCA is unsupervised learning, which does not consider class label. PCA finds components along maximum variability of the data.

LDA vs PCA

- **Similarity**:
  1. Both LDA and PCA reduce dimension.
  2. Both construct new features which are linear combination of original features.
  3. Both use Eigen Decomposition. (PCA: S; LDA: $S_w^{-1} S_b$)
- **Difference**:
  1. PCA is unsupervised learning, which does not consider class label. PCA finds components along maximum variability of the data.
  2. LDA is supervised, which considers the class label. LDA finds components to maximially separate the classes.