San José State University

Math 250: Mathematical Data Visualization

Linear Discriminant Analysis (LDA)

Dr. Guangliang Chen

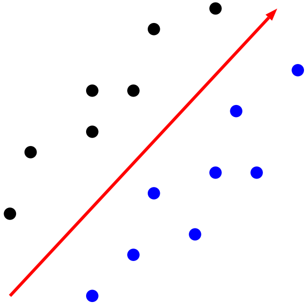
Outline

- The one-dimensional LDA problem
  - 2-class LDA
  - Multiclass extension
- The general LDA problem (to be added)
- Comparison between PCA and LDA (to be added)

Data representation vs data classification

PCA aims to find the most accurate data representation in a lower dimensional space spanned by the maximum-variance directions.

However, such directions might not work well for supervised tasks, where the data points have labels and (only) discriminative information needs to preserved in the data reduction step.



Representative, not discriminative

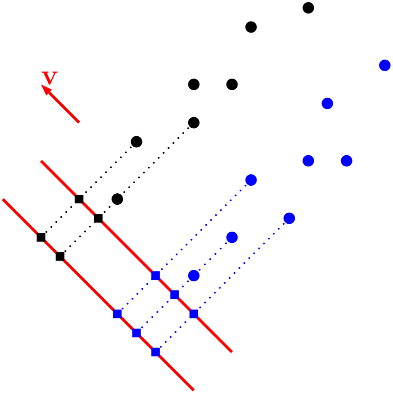
The two-class LDA problem

Problem. Given a labeled data set $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ consisting of two disjoint classes C_1, C_2, find a most discriminative line

$$\mathbf{x}(t) = t\mathbf{v} + \mathbf{b}, \quad t \in \mathbb{R}$$

where $\mathbf{v}, \mathbf{b} \in \mathbb{R}^d$ and $\|\mathbf{v}\| = 1$.

Note: Projections of the two classes onto parallel lines have "the same amount of separation".



Mathematical setup

This time we are going to focus on lines that pass through the origin:

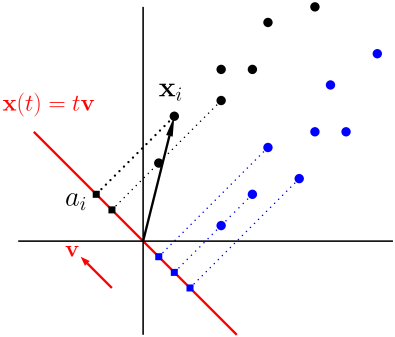
$$\mathbf{x}(t) = t\mathbf{v}, \quad t \in \mathbb{R}$$

where $\mathbf{v} \in \mathbb{R}^d$ is a unit vector.

The 1D projections of the data are

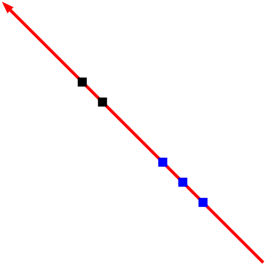
$$\mathbf{p}_i = \mathbf{v}\,\underbrace{\mathbf{v}^T\mathbf{x}_i}_{=a_i} = a_i\mathbf{v}, \quad i = 1, \ldots, n$$

Note that they also have labels.



Linear Discriminant Analysis (LDA)

Now the data look like this:



How do we quantify the separation between the two classes (in order to compare different directions $\mathbf{v}$ and select the best one)?

One (naive) idea is to measure the distance between the two class means in the 1D projection space: $|\mu_1 - \mu_2|$, where

$$\mu_1 = \frac{1}{n_1} \sum_{\mathbf{x}_i \in C_1} a_i = \frac{1}{n_1} \sum_{\mathbf{x}_i \in C_1} \mathbf{v}^T \mathbf{x}_i$$
$$= \mathbf{v}^T \cdot \frac{1}{n_1} \sum_{\mathbf{x}_i \in C_1} \mathbf{x}_i = \mathbf{v}^T \mathbf{m}_1$$

and similarly,

$$\mu_2 = \mathbf{v}^T \mathbf{m}_2, \quad \mathbf{m}_2 = \frac{1}{n_2} \sum_{\mathbf{x}_i \in C_2} \mathbf{x}_i.$$

Linear Discriminant Analysis (LDA)

That is, to solve the following problem

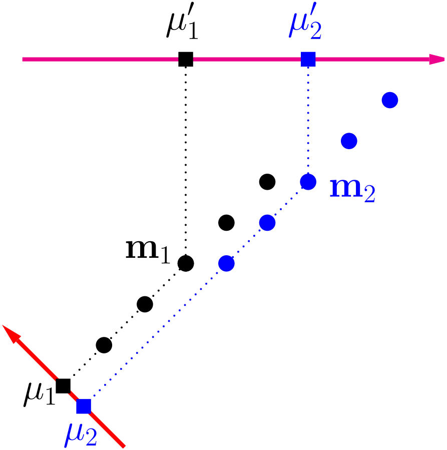
$$\max_{\mathbf{v}:\,\|\mathbf{v}\|=1} |\mu_1 - \mu_2|$$

where

$$\mu_j = \mathbf{v}^T \mathbf{m}_j,\ j = 1, 2.$$

However, this criterion does not work well (as shown in the right plot).

What else do we need to control?



It turns out that we should also pay attention to the variances of the projected classes:

$$s_1^2 = \sum_{\mathbf{x}_i \in C_1} (a_i - \mu_1)^2, \quad s_2^2 = \sum_{\mathbf{x}_i \in C_2} (a_i - \mu_2)^2$$

Ideally, the projected classes have both faraway means and small variances, which can be achieved through the following modified formulation:

$$\max_{\mathbf{v}:\|\mathbf{v}\|=1} \frac{(\mu_1 - \mu_2)^2}{s_1^2 + s_2^2}.$$

The optimal $\mathbf{v}$ should be such that

- $(\mu_1 - \mu_2)^2$: large
- s_1^2, s_2^2: both small

Mathematical derivation

First, we derive a formula for the distance between the projected centroids:

$$\begin{aligned}(\mu_1 - \mu_2)^2 &= (\mathbf{v}^T\mathbf{m}_1 - \mathbf{v}^T\mathbf{m}_2)^2 = (\mathbf{v}^T(\mathbf{m}_1 - \mathbf{m}_2))^2 \\ &= \mathbf{v}^T(\mathbf{m}_1 - \mathbf{m}_2) \cdot (\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{v} \\ &= \mathbf{v}^T\mathbf{S}_b\mathbf{v},\end{aligned}$$

where

$$\mathbf{S}_b = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \in \mathbb{R}^{d\times d}$$

is called the **between-class scatter matrix**.

Clearly, $\mathbf{S}_b$ is square, symmetric and positive semidefinite. Moreover, $\mathrm{rank}(\mathbf{S}_b) = 1$, which implies that it only has 1 positive eigenvalue!

Linear Discriminant Analysis (LDA)

Next, for each class $j = 1, 2$, the variance of the projections (onto $\mathbf{v}$) is

$$\begin{aligned} s_j^2 &= \sum_{\mathbf{x}_i \in C_j} (a_i - \mu_j)^2 = \sum_{\mathbf{x}_i \in C_j} (\mathbf{v}^T\mathbf{x}_i - \mathbf{v}^T\mathbf{m}_j)^2 \\ &= \sum_{\mathbf{x}_i \in C_j} \mathbf{v}^T(\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T\mathbf{v} \\ &= \mathbf{v}^T \left[\sum_{\mathbf{x}_i \in C_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T \right] \mathbf{v} = \mathbf{v}^T\mathbf{S}_j\mathbf{v}, \end{aligned}$$

where

$$\mathbf{S}_j = \sum_{\mathbf{x}_i \in C_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T \in \mathbb{R}^{d\times d}$$

is called the **within-class scatter matrix** for class j.

Linear Discriminant Analysis (LDA)

The total within-class scatter of the two classes in the projection space is

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

where

$$\mathbf{S}_w = \mathbf{S}_1 + \mathbf{S}_2 = \sum_{\mathbf{x}_i \in C_1}(\mathbf{x}_i - \mathbf{m}_1)(\mathbf{x}_i - \mathbf{m}_1)^T + \sum_{\mathbf{x}_i \in C_2}(\mathbf{x}_i - \mathbf{m}_2)(\mathbf{x}_i - \mathbf{m}_2)^T$$

is called the **total within-class scatter matrix** of the original data.

Remark. $\mathbf{S}_w \in \mathbb{R}^{d\times d}$ is also square, symmetric, and positive semidefinite.

Linear Discriminant Analysis (LDA)

Putting everything together, we have derived the following problem:

$$\max_{\mathbf{v}:\|\mathbf{v}\|=1} \frac{\mathbf{v}^T\mathbf{S}_b\mathbf{v}}{\mathbf{v}^T\mathbf{S}_w\mathbf{v}}$$

Theorem 0.1. Suppose $\mathbf{S}_w$ is nonsingular. The maximizer of the problem is given by the largest generalized eigenvector $\mathbf{v}_1$ of $(\mathbf{S}_b, \mathbf{S}_w)$, i.e.,

$$\mathbf{S}_b\mathbf{v}_1 = \lambda_1\mathbf{S}_w\mathbf{v}_1 \quad \Longleftrightarrow \quad \mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{v}_1 = \lambda_1\mathbf{v}_1$$

and the maximum is λ_1, the largest generalized eigenvalue of $(\mathbf{S}_b, \mathbf{S}_w)$

Remark. $\mathrm{rank}(\mathbf{S}_w^{-1}\mathbf{S}_b) = \mathrm{rank}(\mathbf{S}_b) = 1$, so λ_1 is the only nonzero (positive) eigenvalue that can be found. It represents the the largest amount of separation between the two classes along any single direction.

Computing

The following are different ways of finding the optimal direction $\mathbf{v}_1$:

- Slowest way (via three expensive steps):

  1. work really hard to invert the $d \times d$ matrix $\mathbf{S}_w$,

  2. do the matrix multiplication $\mathbf{S}_w^{-1}\mathbf{S}_b$,

  3. solve the eigenvalue problem $\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{v}_1 = \lambda_1\mathbf{v}_1$.

Linear Discriminant Analysis (LDA)

- A slight better way: Rewrite as a **generalized eigenvalue problem**

$$\mathbf{S}_b \mathbf{v}_1 = \lambda_1 \mathbf{S}_w \mathbf{v}_1,$$

  and then solve it through functions like *eigs(A,B)* in MATLAB.

- The smartest way is to rewrite as

$$\begin{aligned}\lambda_1 \mathbf{v}_1 &= \mathbf{S}_w^{-1} \underbrace{(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T}_{\mathbf{S}_b} \mathbf{v}_1 \\ &= \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \cdot \underbrace{(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{v}_1}_{\text{scalar}}\end{aligned}$$

Linear Discriminant Analysis (LDA)

This implies that

$$\mathbf{v}_1 \propto \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

and it can be computed from $\mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$ through rescaling!

Remark. Here, inverting $\mathbf{S}_w$ should still be avoided; instead, one should implement this by solving a linear system

$$\mathbf{S}_w \mathbf{x} = \mathbf{m}_1 - \mathbf{m}_2.$$

This can be done through $\mathbf{S}_w \setminus (\mathbf{m}_1 - \mathbf{m}_2)$ in MATLAB.

Two-class LDA: summary

The optimal discriminative direction is

$$\mathbf{v}^* = \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \quad \text{(plus normalization)}$$

It is the solution of

$$\max_{\mathbf{v}:\|\mathbf{v}\|=1} \frac{\mathbf{v}^T\mathbf{S}_b\mathbf{v}}{\mathbf{v}^T\mathbf{S}_w\mathbf{v}} \quad \longleftarrow \quad \frac{(\mu_1-\mu_2)^2}{s_1^2+s_2^2}$$

where

$$\mathbf{S}_b = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$$

$$\mathbf{S}_w = \mathbf{S}_1 + \mathbf{S}_2, \quad \mathbf{S}_j = \sum_{\mathbf{x}\in C_j} (\mathbf{x} - \mathbf{m}_j)(\mathbf{x} - \mathbf{m}_j)^T$$

An example

Consider the following labeled data:

- Class 1 has three points

$$(1,2),(2,3),(3,4.9),$$

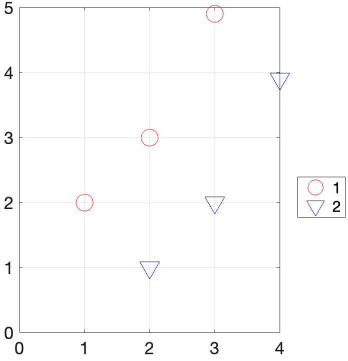
with mean $\mathbf{m}_1 = (2,3.3)^T$

- Class 2 has three points

$$(2,1),(3,2),(4,3.9),$$

with mean $\mathbf{m}_2 = (3,2.3)^T$

Find the optimal LDA direction.



Solution. By direct calculation,

$$
\begin{aligned}
\mathbf{S}_1 &= \left[\begin{pmatrix}1\\2\end{pmatrix} - \begin{pmatrix}2\\3.3\end{pmatrix}\right] \cdot \left[\begin{pmatrix}1 & 2\end{pmatrix} - \begin{pmatrix}2 & 3.3\end{pmatrix}\right] \\
&\quad + \left[\begin{pmatrix}2\\3\end{pmatrix} - \begin{pmatrix}2\\3.3\end{pmatrix}\right] \cdot \left[\begin{pmatrix}2 & 3\end{pmatrix} - \begin{pmatrix}2 & 3.3\end{pmatrix}\right] \\
&\quad + \left[\begin{pmatrix}3\\4.9\end{pmatrix} - \begin{pmatrix}2\\3.3\end{pmatrix}\right] \cdot \left[\begin{pmatrix}3 & 4.9\end{pmatrix} - \begin{pmatrix}2 & 3.3\end{pmatrix}\right] \\
&= \begin{pmatrix}1 & 1.3\\1.3 & 1.69\end{pmatrix} + \begin{pmatrix}0 & 0\\0 & 0.09\end{pmatrix} + \begin{pmatrix}1 & 1.6\\1.6 & 2.56\end{pmatrix} \\
&= \begin{pmatrix}2 & 2.9\\2.9 & 4.34\end{pmatrix}
\end{aligned}
$$

and similarly,

$$\mathbf{S}_2 = \begin{pmatrix} 2 & 2.9 \\ 2.9 & 4.34 \end{pmatrix}$$

It follows that the total within-class scatter matrix is

$$\mathbf{S}_w = \mathbf{S}_1 + \mathbf{S}_2 = \begin{pmatrix} 4 & 5.8 \\ 5.8 & 8.68 \end{pmatrix}$$

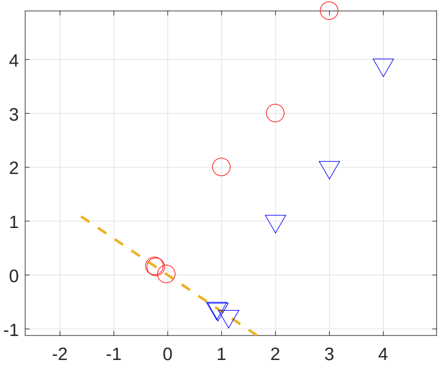
(Later, we will matricize the formula for $\mathbf{S}_w$ which is easier to use)

Thus, the optimal discriminative direction is

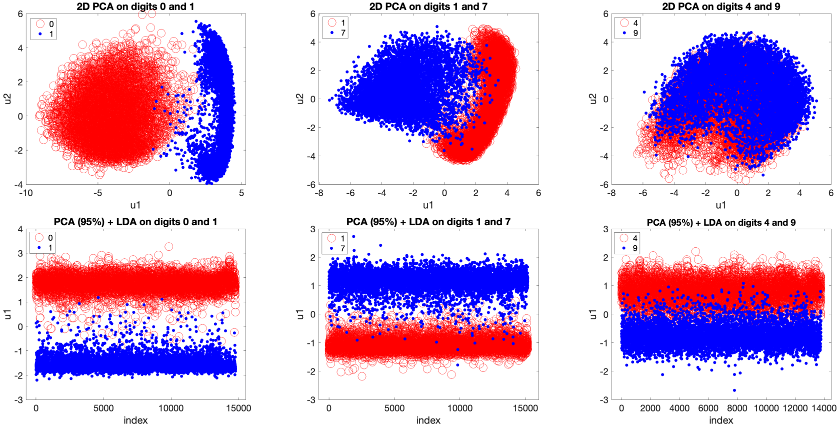
$$\mathbf{v} = \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2) = (-13.4074, 9.0741)^T \xrightarrow{\text{normalizing}} (-0.8282, 0.5605)^T$$

and the projection coordinates are

$$Y = [0.2928, 0.0252, 0.2619, -1.0958, -1.3635, -1.1267]$$



Experiment 1 (MNIST handwritten digits)



2D PCA on digits 0 and 1

2D PCA on digits 1 and 7

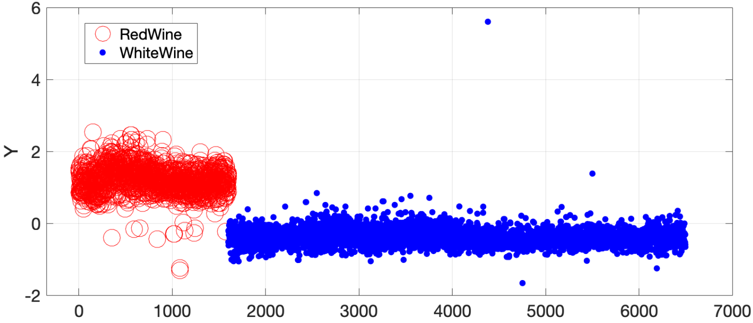
2D PCA on digits 4 and 9

PCA (95%) + LDA on digits 0 and 1

PCA (95%) + LDA on digits 1 and 7

PCA (95%) + LDA on digits 4 and 9

Experiment 2 (Wine quality data)



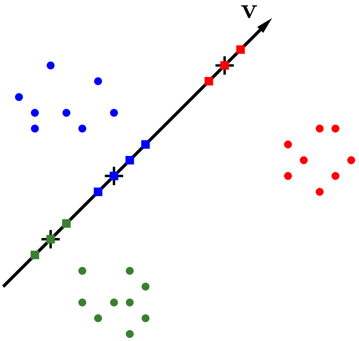
Multiclass extension

The previous procedure only applies to 2 classes. When there are $c \geq 3$ classes, what is the "most discriminative" direction?

It will be based on the same intuition that the optimal direction $\mathbf{v}$ should project the different classes such that

- classes are as tight as possible;
- their centroids are as far from each other as possible.

Both are actually about variances.



Mathematical derivation

For any unit vector $\mathbf{v}$, the tightness of the projected classes (of the training data) is still described by the total within-class scatter:

$$\sum_{j=1}^{c} s_j^2 = \sum \mathbf{v}^T \mathbf{S}_j \mathbf{v} = \mathbf{v}^T \left(\sum \mathbf{S}_j \right) \mathbf{v} = \mathbf{v}^T \mathbf{S}_w \mathbf{v}$$

where

$$\mathbf{S}_w = \sum_{j=1}^{c} \mathbf{S}_j, \quad \mathbf{S}_j = \sum_{\mathbf{x} \in C_j} (\mathbf{x} - \mathbf{m}_j)(\mathbf{x} - \mathbf{m}_j)^T$$

The matrix $\mathbf{S}_w \in \mathbb{R}^{d \times d}$ is called the total within-class scatter matrix. It is square, symmetric and positive semidefinite.

Linear Discriminant Analysis (LDA)

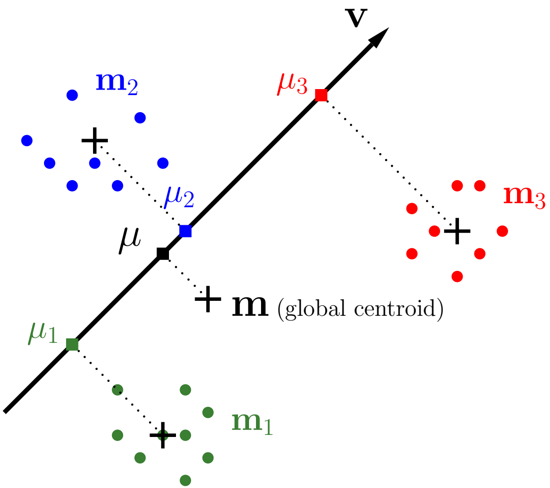
To make the class centroids μ_j (in the projection space) as far from each other as possible, we maximize the following between-class scatter:

$$\sum_{j=1}^{c} n_j(\mu_j - \mu)^2, \qquad \text{where} \quad \mu = \frac{1}{n}\sum_{j=1}^{c} n_j\mu_j \longleftarrow \text{weighted average}$$

Note that μ has the interpretation of the projection of the global centroid $(\mathbf{m})$ of the training data onto $\mathbf{v}$:

$$\mu = \frac{1}{n}\sum_{j=1}^{c} n_j\left(\mathbf{v}^T\mathbf{m}_j\right) = \mathbf{v}^T\left(\frac{1}{n}\sum_{j=1}^{c} n_j\mathbf{m}_j\right) = \mathbf{v}^T\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\right) = \mathbf{v}^T\mathbf{m}.$$

Linear Discriminant Analysis (LDA)



Linear Discriminant Analysis (LDA)

We simplify the between-class scatter (in the $\mathbf{v}$ space) as follows:

$$\begin{aligned}
\sum_{j=1}^{c} n_j(\mu_j - \mu)^2 &= \sum n_j (\mathbf{v}^T(\mathbf{m}_j - \mathbf{m}))^2 \\
&= \sum n_j \, \mathbf{v}^T(\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T \mathbf{v} \\
&= \mathbf{v}^T \left(\sum n_j (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T \right) \mathbf{v} \\
&= \mathbf{v}^T \mathbf{S}_b \mathbf{v},
\end{aligned}$$

where

$$\mathbf{S}_b = \sum_{j=1}^{c} n_j (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T \in \mathbb{R}^{d\times d}$$

is called the between-class scatter matrix. It is also square, symmetric, and positive semidefinite.

We have thus arrived at the following generalized Rayleigh quotient problem:

$$\max_{\mathbf{v}:\|\mathbf{v}\|=1} \frac{\mathbf{v}^T\mathbf{S}_b\mathbf{v}}{\mathbf{v}^T\mathbf{S}_w\mathbf{v}} \quad \longleftarrow \quad \frac{\sum n_j(\mu_j-\mu)^2}{\sum s_j^2}$$

Assuming $\mathbf{S}_w$ is nonsingular (positive definite), the solution is given by the largest generalized eigenvector $\mathbf{v}_1$ of $(\mathbf{S}_b, \mathbf{S}_w)$ (and also the largest eigenvector of $\mathbf{S}_w^{-1}\mathbf{S}_b$):

$$\mathbf{S}_b\mathbf{v}_1 = \lambda_1\mathbf{S}_w\mathbf{v}_1 \quad \Leftrightarrow \quad \mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{v}_1 = \lambda_1\mathbf{v}_1.$$

Linear Discriminant Analysis (LDA)

Remark. When $c = 2$, it can be verified that

$$\sum_{j=1}^{2} n_j(\mu_j - \mu)^2 = \frac{n_1 n_2}{n}(\mu_1 - \mu_2)^2$$

$$\sum_{j=1}^{2} n_j(\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T = \frac{n_1 n_2}{n}(\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T,$$

This shows that when there are only two classes, the weighted definitions of between-class scatter are just a scalar multiple of the unweighted definitions.

Therefore, the multiclass LDA is a generalization of the two-class LDA:

$$(\mu_1 - \mu_2)^2/(s_1^2 + s_2^2) \longrightarrow \sum n_j(\mu_j - \mu)^2 / \sum s_j^2$$

Computing

However, the formula $\mathbf{v}_1 \propto \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$ is no longer valid:

$$\lambda_1 \mathbf{v}_1 = \mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{v}_1 = \sum_{j=1}^{c} n_j \mathbf{S}_w^{-1}(\mathbf{m}_j - \mathbf{m}) \underbrace{(\mathbf{m}_j - \mathbf{m})^T \mathbf{v}_1}_{\text{scalar}}$$

which only shows that

$$\mathbf{v}_1 \in \mathrm{Span}\{\mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}), \ldots, \mathbf{S}_w^{-1}(\mathbf{m}_c - \mathbf{m})\}.$$

So we have to find $\mathbf{v}_1$ by solving a generalized eigenvalue problem:

$$\mathbf{S}_b\mathbf{v}_1 = \lambda_1 \mathbf{S}_w \mathbf{v}_1.$$

Next, we'll derive new formulas in matrix form for both matrices $\mathbf{S}_w, \mathbf{S}_b$.

Linear Discriminant Analysis (LDA)

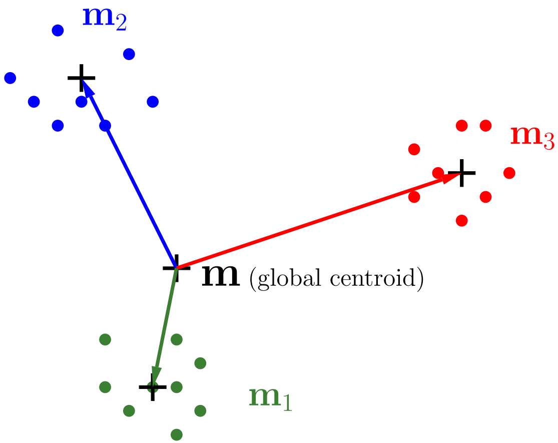
First, we have

$$
\begin{aligned}
\mathbf{S}_b &= \sum n_j(\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T \\
&= [\sqrt{n_1}(\mathbf{m}_1 - \mathbf{m}) \cdots \sqrt{n_c}(\mathbf{m}_c - \mathbf{m})] \cdot \begin{bmatrix} \sqrt{n_1}(\mathbf{m}_1 - \mathbf{m})^T \\ \vdots \\ \sqrt{n_c}(\mathbf{m}_c - \mathbf{m})^T \end{bmatrix} \\
&= \widetilde{\mathbf{M}}^T \widetilde{\mathbf{M}}
\end{aligned}
$$

where

$$
\widetilde{\mathbf{M}} = \begin{bmatrix} \sqrt{n_1}(\mathbf{m}_1 - \mathbf{m})^T \\ \vdots \\ \sqrt{n_c}(\mathbf{m}_c - \mathbf{m})^T \end{bmatrix} \in \mathbb{R}^{c \times d}
$$

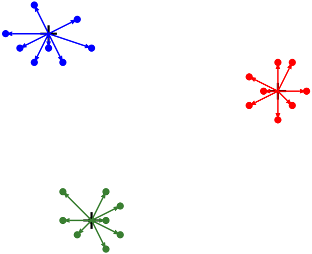
Linear Discriminant Analysis (LDA)



Next, let $\widetilde{\mathbf{x}}_i = \mathbf{x}_i - \mathbf{m}_j$ for each $\mathbf{x}_i \in C_j$, and define

$$\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1 \ldots \widetilde{\mathbf{x}}_n]^T \in \mathbb{R}^{n \times d},$$

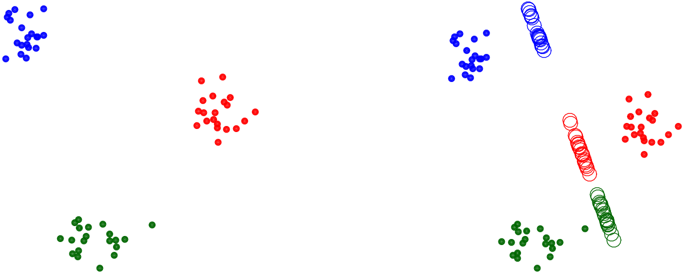
the locally centered data matrix.



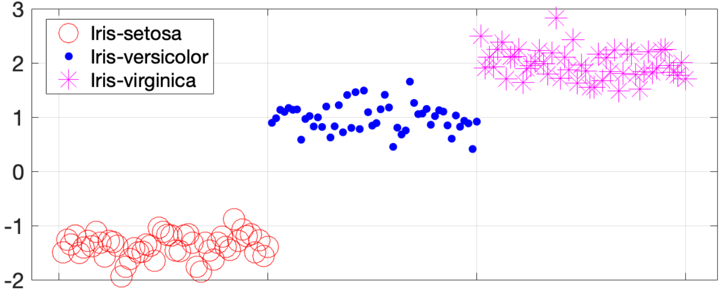
Then we can express $\mathbf{S}_w$ as follows:

$$\mathbf{S}_w = \sum_{j=1}^{c} \sum_{\mathbf{x}_i \in C_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T = \sum_{j=1}^{c} \sum_{\mathbf{x}_i \in C_j} \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^T = \sum_{i=1}^{n} \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^T$$
$$= \widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}}.$$

Simulation

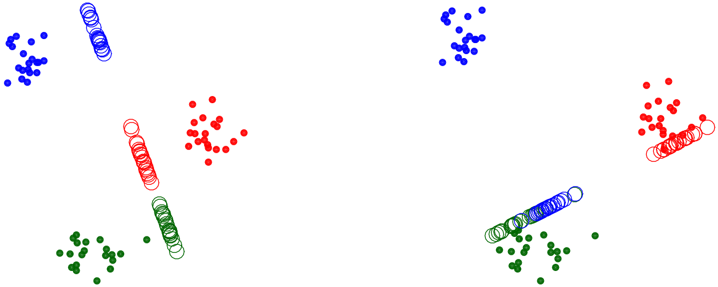


Experiment (Iris data)



Linear Discriminant Analysis (LDA)

What about the second eigenvector v_2?



How many discriminative directions can we find?

To answer this question, we need to count the nonzero eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{v} = \lambda\mathbf{v}$, since only their eigenvectors will be used as discriminative directions.

In the above equation, the within-class scatter matrix $\mathbf{S}_w$ is *assumed to be* nonsingular. However, the between-class scatter matrix $\mathbf{S}_b$ is of low rank.

To see this, first note that

$$\mathrm{rank}(\mathbf{S}_b) = \mathrm{rank}(\widetilde{\mathbf{M}}^T\widetilde{\mathbf{M}}) = \mathrm{rank}(\widetilde{\mathbf{M}})$$

Linear Discriminant Analysis (LDA)

Next, observe that the row of the matrix $\widetilde{\mathbf{M}}$ are linearly dependent:

$$\begin{aligned}
&\sqrt{n_1}\cdot\sqrt{n_1}(\mathbf{m}_1-\mathbf{m})+\cdots+\sqrt{n_c}\cdot\sqrt{n_c}(\mathbf{m}_c-\mathbf{m})\\
&=(n_1\mathbf{m}_1+\cdots n_c\mathbf{m}_c)-(n_1+\cdots+n_c)\mathbf{m}\\
&=n\mathbf{m}-n\mathbf{m}=\mathbf{0}.
\end{aligned}$$

As a result, $\operatorname{rank}(\widetilde{\mathbf{M}})\le c-1$.

It follows that $\operatorname{rank}(\mathbf{S}_w^{-1}\mathbf{S}_b)=\operatorname{rank}(\mathbf{S}_b)\le c-1$.

Therefore, LDA can only find at most $c-1$ discriminative directions.

Multiclass LDA algorithm

Input: Labeled data $\mathbf{X} \in \mathbb{R}^{n\times d}$ (with c classes)

Output: $\leq c-1$ discriminative directions and projections of $\mathbf{X}$ onto them

1. Find the class centroids $\{\mathbf{m}_j\}$ and center the data locally.
2. Compute the within-class and between-class scatter matrices, i.e., $\mathbf{S}_w = \widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}$ and $\mathbf{S}_b = \widetilde{\mathbf{M}}^T\widetilde{\mathbf{M}}$.
3. Solve the generalized eigenvalue problem $\mathbf{S}_b\mathbf{v} = \lambda\mathbf{S}_w\mathbf{v}$ to find all nonzero eigenvectors $\mathbf{V}_k = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$ (for some $k \leq c-1$)
4. Project the data $\mathbf{X}$ onto them $\mathbf{Y} = \mathbf{X} \cdot \mathbf{V}_k \in \mathbb{R}^{n\times k}$.

The singularity issue of $\mathbf{S}_w$

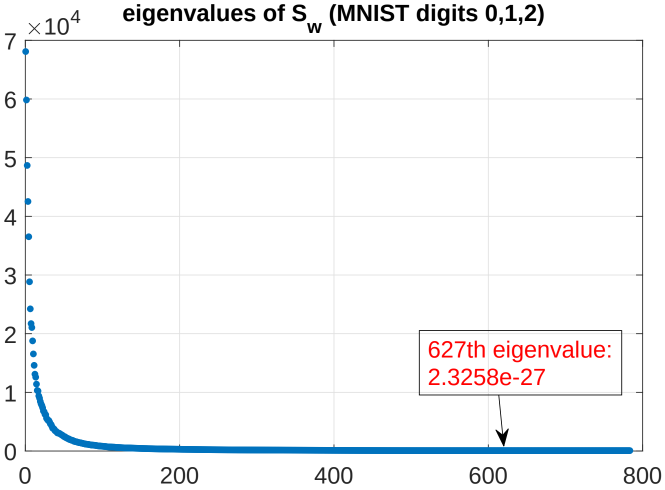
So far, we have assumed that the total within-class scatter matrix $\mathbf{S}_w$ is nonsingular, so that we can solve the LDA problem

$$\max_{\mathbf{v}:\|\mathbf{v}\|=1} \frac{\mathbf{v}^T\mathbf{S}_b\mathbf{v}}{\mathbf{v}^T\mathbf{S}_w\mathbf{v}} \qquad \text{via} \quad \mathbf{S}_b\mathbf{v} = \lambda\mathbf{S}_w\mathbf{v}.$$

However, in many cases (especially when having high dimensional data), the matrix $\mathbf{S}_w \in \mathbb{R}^{d\times d}$ is (nearly) singular.

The reason is often that the centered data points, i.e., the rows of $\widetilde{\mathbf{X}}$, do not fully span all d dimensions, thus making $\operatorname{rank}(\mathbf{S}_w) = \operatorname{rank}(\widetilde{\mathbf{X}}) < d$ (which implies that $\mathbf{S}_w$ is singular).

Linear Discriminant Analysis (LDA)



627th eigenvalue:
2.3258e-27

How do we fix it?

A common way is to first **apply global PCA** to reduce the dimensionality of the labeled data (all classes)

$$\mathbf{Y}_{\text{pca}} = \left(\mathbf{X} - [\mathbf{m} \ldots \mathbf{m}]^T\right) \cdot \mathbf{V}_{\text{pca}}$$

and then perform LDA on the reduced data:

$$\mathbf{Z}_{\text{lda}} = \mathbf{Y}_{\text{pca}} \cdot \mathbf{V}_{\text{lda}} \longleftarrow \text{learned from } \mathbf{Y}_{\text{pca}}$$

Linear Discriminant Analysis (LDA)

Two other methods are to

- Use **pseudoinverse** instead:

$$\mathbf{S}_w^{\dagger}\mathbf{S}_b\mathbf{v} = \lambda\mathbf{v} \qquad \longleftarrow \qquad \mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{v} = \lambda\mathbf{v}$$

- **Regularize** $\mathbf{S}_w$:

$$\begin{aligned}\mathbf{S}_w^{(\beta)} &= \mathbf{S}_w + \beta\mathbf{I}_d \\ &= \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T + \beta\mathbf{I}_d \\ &= \mathbf{Q}\left(\boldsymbol{\Lambda} + \beta\mathbf{I}_d\right)\mathbf{Q}^T \\ &= \mathbf{Q}\,\mathrm{diag}(\lambda_1+\beta,\ldots,\lambda_d+\beta)\,\mathbf{Q}^T\end{aligned}$$

  where $\beta > 0$ is parameter whose value needs to be tuned.

Other numerical issues

Similarly to PCA, one should pay attention to the following:

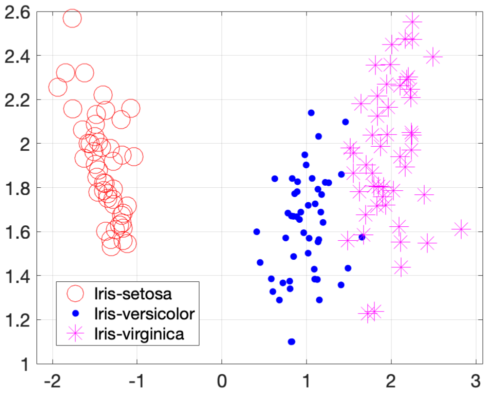
- **Feature scaling**: Standardize each feature to have mean zero and standard deviation 1, so that they are on comparable scales:

  $Xnorm = normalize(X)$;

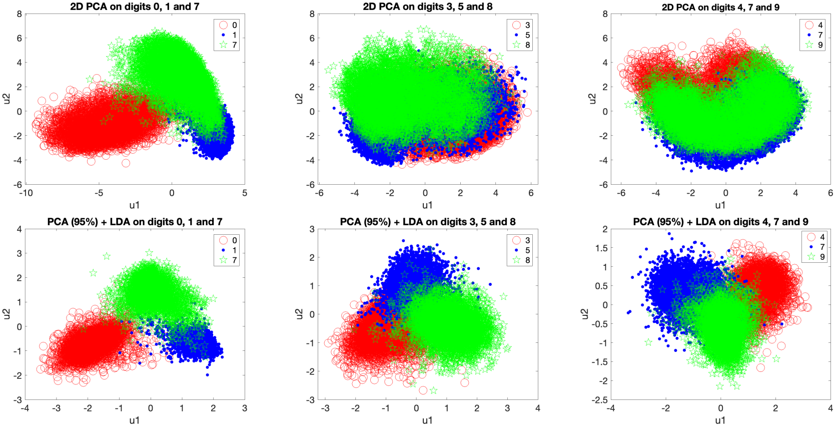
- **Out of sample extension**:

$$\mathbf{y}_0 = \mathbf{V}^T \mathbf{x}_0, \qquad \text{where} \quad \mathbf{V} \xleftarrow{\text{lda}} \mathbf{X}$$

Experiment (Iris data)



Experiment (MNIST handwritten digits)



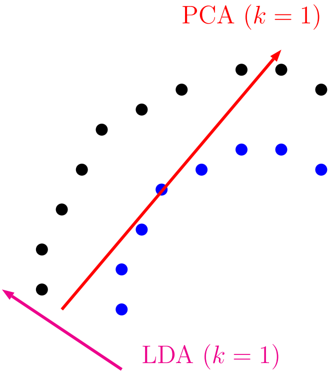
Comparison between PCA and LDA

	PCA	**LDA**
Model	nonparametric*	nonparametric*
Use labels?	no (unsupervised)	yes (supervised)
Criterion	variance	separation
#dimensions (k)	any	$\leq c-1$
Computing	SVD	generalized eigenvectors
Linear projection?	yes $(\mathbf{V}^T(\mathbf{x}-\bar{\mathbf{x}}))$	yes $(\mathbf{V}^T\mathbf{x})$
Nonlinear boundary	can handle**	cannot handle

*Both work the best with multivariate Gaussian samples

**In the case of nonlinearly separated classes, PCA often works better than LDA as the latter can only find at most $c-1$ directions (insufficient to preserve all the discriminative information in the data).

- LDA with $k=1$: does not work well;
- PCA with $k=1$: does not work well;
- PCA with $k=2$: can preserve all the nonlinear separation.


PCA $(k=1)$
LDA $(k=1)$