LEC 3: Fisher Discriminant Analysis (FDA)

— A Supervised Dimensionality Reduction Approach

Dr. Guangliang Chen

February 18, 2016

Outline

- Motivation:
  - PCA is unsupervised which does not use training labels
  - Variance is not always useful for classification
- FDA: a supervised dimensionality reduction approach
  - 2-class FDA
  - Multiclass FDA
- Comparison between PCA and FDA

Two-class FDA

See Prof. Olga Veksler's slides at

`http://www.csd.uwo.ca/~olga/Courses/CS434a_541a/Lecture8.pdf`

Two-class FDA (a summary)

The optimal discriminatory direction is

$$\mathbf{v}^* = \mathbf{S}_w^{-1}(\mu_1 - \mu_2) \quad \text{(plus normalization)}$$

It is the solution of

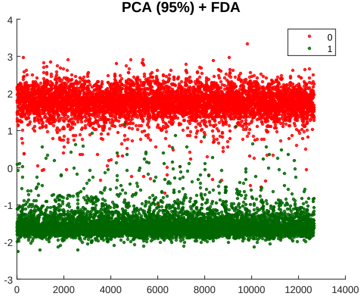
$$\max_{\mathbf{v}:\|\mathbf{v}\|=1} \frac{\mathbf{v}^T\mathbf{S}_b\mathbf{v}}{\mathbf{v}^T\mathbf{S}_w\mathbf{v}} \quad \longleftarrow \quad \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

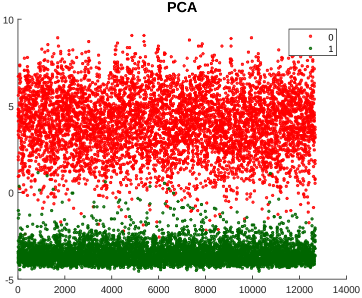
where

$$\mathbf{S}_b = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$
$$\mathbf{S}_w = \mathbf{S}_1 + \mathbf{S}_2, \quad \mathbf{S}_i = \sum_{\mathbf{x} \in \text{ Class } i} (\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^T$$

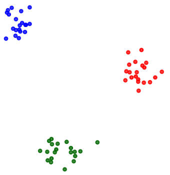
Experiment (2 digits)

MNIST handwritten digits 0 and 1





How to extend to $c \geq 3$ classes?



Let's start by finding the most discriminatory direction.

For any $\mathbf{v}$, the total within-class scatter in the $\mathbf{v}$ space is

$$\sum \tilde{s}_i^2 = \sum \mathbf{v}^T \mathbf{S}_i \mathbf{v} = \mathbf{v}^T \left(\sum \mathbf{S}_i \right) \mathbf{v} = \mathbf{v}^T \mathbf{S}_w \mathbf{v}$$

where the $\mathbf{S}_i$ are defined in the same way as before.

To define the between-class scatter in the $\mathbf{v}$ space, we need to introduce

- the global center of the training data

$$\mu = \frac{1}{n}\sum \mathbf{x}_i = \frac{1}{n}\sum n_i\mu_i,$$

- and its projection onto $\mathbf{v}$:

$$\tilde{\mu} = \mathbf{v}^T\mu = \frac{1}{n}\sum y_i = \frac{1}{n}\sum n_i\tilde{\mu}_i$$

The between-class scatter in the $\mathbf{v}$ space is defined as

$$\begin{aligned}\sum_i n_i(\tilde{\mu}_i - \tilde{\mu})^2 &= \sum_i n_i\, \mathbf{v}^T(\mu_i - \mu)(\mu_i - \mu)^T\mathbf{v} \\ &= \mathbf{v}^T \left(\sum_i n_i(\mu_i - \mu)(\mu_i - \mu)^T\right)\mathbf{v} \\ &= \mathbf{v}^T\mathbf{S}_b\mathbf{v}.\end{aligned}$$

We have thus arrived at the same kind of problem

$$\max_{\mathbf{v}:\|\mathbf{v}\|=1} \frac{\mathbf{v}^T\mathbf{S}_b\mathbf{v}}{\mathbf{v}^T\mathbf{S}_w\mathbf{v}} \quad \longleftarrow \quad \frac{\sum n_i(\tilde{\mu}_i - \tilde{\mu})^2}{\sum \tilde{s}_i^2}$$

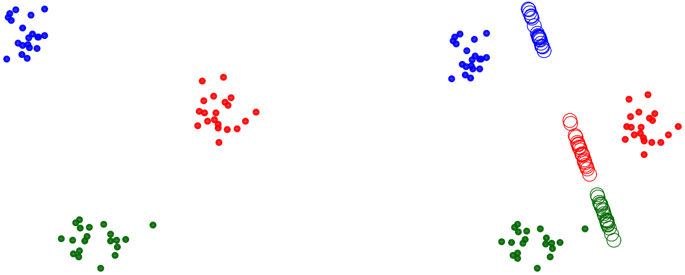
The solution is given by the largest eigenvector of $\mathbf{S}_w^{-1}\mathbf{S}_b$:

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

It is also a generalized eigenvector:

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

However, the formula $\mathbf{v}^* = \mathbf{S}_w^{-1}(\mu_1 - \mu_2)$ is no longer valid.



The connection to 2-class FDA

Proposition. When $c = 2$, we have

$$\sum_i n_i(\tilde{\mu}_i - \tilde{\mu})^2 = \frac{n_1 n_2}{n}(\tilde{\mu}_1 - \tilde{\mu}_2)^2$$

and

$$\mathbf{S}_b = \frac{n_1 n_2}{n}(\mu_2 - \mu_1)(\mu_2 - \mu_1)^T$$

.

This implies that the criterion $\frac{\sum_i n_i(\tilde{\mu}_i - \tilde{\mu})^2}{\sum_i \tilde{s}_i^2}$ is a generalization of that of the two-class FDA.

Proof: We prove the first identity below:

$$\begin{aligned}
\sum_i n_i(\tilde{\mu}_i - \tilde{\mu})^2 &= n_1 \left(\tilde{\mu}_1 - \frac{n_1\tilde{\mu}_1 + n_2\tilde{\mu}_2}{n}\right)^2 + n_2 \left(\tilde{\mu}_2 - \frac{n_1\tilde{\mu}_1 + n_2\tilde{\mu}_2}{n}\right)^2 \\
&= \frac{n_1 n_2^2}{n^2}(\tilde{\mu}_1 - \tilde{\mu}_2)^2 + \frac{n_2 n_1^2}{n^2}(\tilde{\mu}_2 - \tilde{\mu}_1)^2 \\
&= \frac{n_1 n_2}{n}(\tilde{\mu}_2 - \tilde{\mu}_1)^2.
\end{aligned}$$

The proof of the second identity is very similar:

$$\mathbf{S}_b = \sum n_i(\mu_i - \mu)(\mu_i - \mu)^T = \cdots$$

How many discriminatory directions can/should we use?

The answer is at most $c - 1$.

The discriminatory directions all satisfy the equation

$$\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{v} = \lambda\mathbf{v}.$$

with the corresponding eigenvalues representing the "magnitudes" of separation.

Therefore, we only need to count the number of nonzero eigenvectors.

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:

$$\begin{aligned}\mathbf{S}_b &= \sum n_i(\mu_i - \mu)(\mu_i - \mu)^T \\ &= \left[\sqrt{n_1}(\mu_1 - \mu) \cdots \sqrt{n_c}(\mu_c - \mu)\right] \cdot \begin{bmatrix} \sqrt{n_1}(\mu_1 - \mu)^T \\ \vdots \\ \sqrt{n_c}(\mu_c - \mu)^T \end{bmatrix}\end{aligned}$$

Observe that the columns of the left matrix are linearly dependent:

$$\sqrt{n_1} \cdot \sqrt{n_1}(\mu_1 - \mu) + \cdots + \sqrt{n_c} \cdot \sqrt{n_c}(\mu_c - \mu) = \mathbf{0}$$

and thus the column rank is at most $c - 1$.

Multiclass FDA: A summary

Input: c training classes

Output: At most $c-1$ discriminatory directions

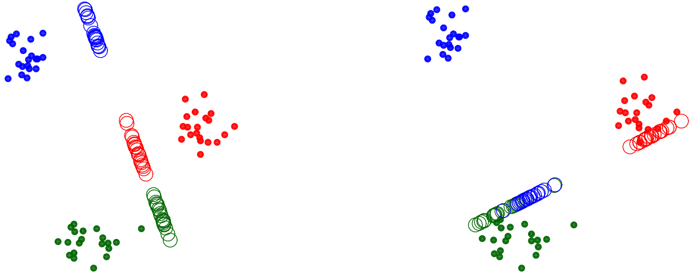
Steps:

1. Form $\mathbf{S}_w = \sum_i \sum_{\mathbf{x} \in \text{Class } i} (\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^T$ and $\mathbf{S}_b = \sum_i n_i (\mu_i - \mu)(\mu_i - \mu)^T$.

2. Solve the eigenvalue problem $\mathbf{S}_w^{-1} \mathbf{S}_b \mathbf{v} = \lambda \mathbf{v}$

3. Return all nonzero eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ $(k \leq c-1)$ in decreasing order.

Multiclass FDA Illustration



An important practical issue

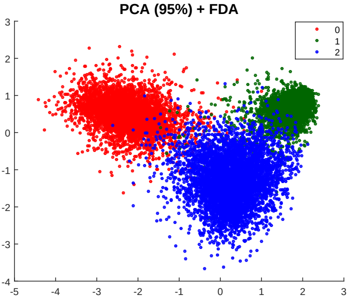
In the cases of high dimensional data, the within-class scatter matrix $\mathbf{S}_w \in \mathbb{R}^{d\times d}$ is often singular due to lack of observations (in certain dimensions).

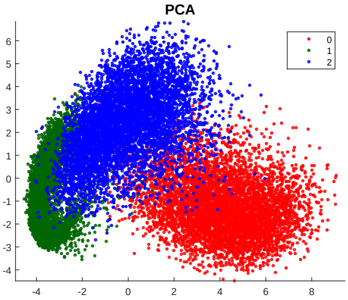
Two common fixes:

- Apply PCA before FDA.
- Regularize $\mathbf{S}_w$ to have $\mathbf{S}_w' = \mathbf{S}_w + \beta \mathbf{I}_d$

Experiment (3 digits)

MNIST handwritten digits 0, 1, and 2





Comparison between PCA and FDA

	PCA	**FDA**
Use labels?	no (unsupervised)	yes (supervised)
Criterion	variance	discriminatory
Linear separation?	yes	yes
Noninear separation?	no	no
#Dimensions	any	$\leq c-1$
Solution	SVD	eigenvalue problem

Remark. In the case of nonlinear separation, PCA (applied conservatively) often works better than FDA as the latter can only find at most $c-1$ directions (which are insufficient to preserve all the separation in the training data).

HW2b (due Friday, March 4)

First apply PCA 95% + FDA to all 10 classes of the MNIST digits and then do the following.

4 Apply the plain kNN classifier to the reduced data with $k = 1, \ldots, 10$ and display the test errors curve. Compare with that of PCA 50 + kNN (for each k). What is your conclusion?

5 Repeat Question 4 with local $k$means instead of kNN (everything else being the same).

Note. Be sure to project the test data onto the same PCA 95% and FDA bases learned on training data, in the same order!