Prof. Marios Savvides

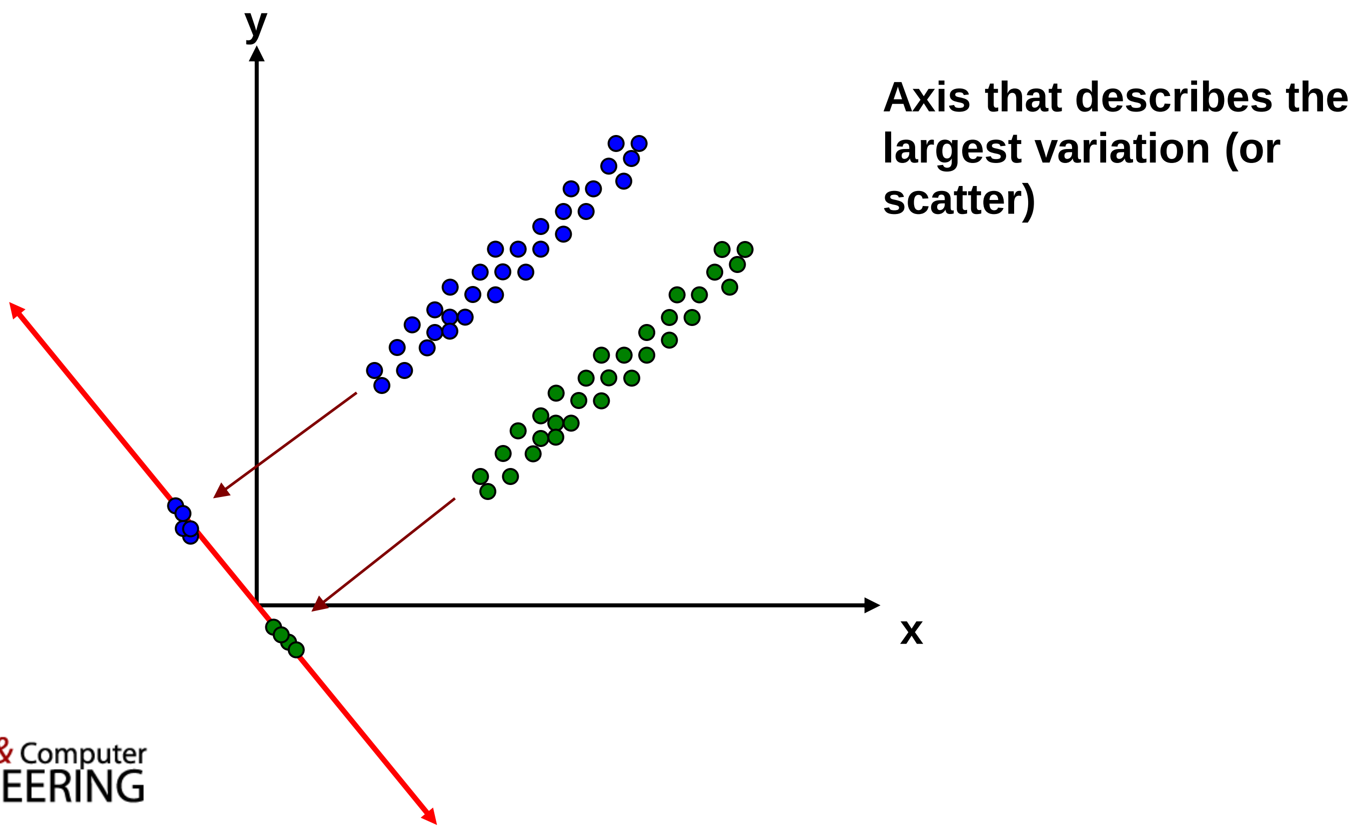
Pattern Recognition Theory

Lecture 9 : Linear Discriminant Analysis

Electrical & Computer ENGINEERING

What is LDA?
What are we trying to do?

- We want to find projections that separate the classes. (Assume unimodal Gaussian modes – maximize distance between two means and minimize variance =>will lead to minimize overall probability of error



Case 1: Simple 2 Class Problem

- We want to maximize the distance between the projected means:

e.g. maximize $|(\tilde{\mu}_1 - \tilde{\mu}_2)|^2$

Where $\tilde{\mu}_1$ is the projected mean $\boldsymbol{\mu}_1$ of class onto LDA direction vector $\mathbf{w}$, i.e.

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

and for class 2: $\tilde{\mu}_2 = \mathbf{w}^{\mathbf{T}}\boldsymbol{\mu}_2$ thus

$$|(\tilde{\mu}_1 - \tilde{\mu}_2)|^2 = |(\mathbf{w}^{\mathbf{T}}\boldsymbol{\mu}_1 - \mathbf{w}^{\mathbf{T}}\boldsymbol{\mu}_2)|^2$$

$$= \mathbf{w}^{\mathbf{T}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathbf{T}}\mathbf{w}$$

$$= \mathbf{w}^{\mathbf{T}}\mathbf{S}_{\mathbf{B}}\mathbf{w}$$

Between Class Scatter Matrix S_B

$$(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (\mathbf{w}^{\mathbf{T}}\boldsymbol{\mu}_1 - \mathbf{w}^{\mathbf{T}}\boldsymbol{\mu}_2)^2$$

$$= \mathbf{w}^{\mathbf{T}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathbf{T}}\mathbf{w}$$

$$= \mathbf{w}^{\mathbf{T}}\mathbf{S}_{\mathbf{B}}\mathbf{w}$$

We want to maximize $w^T S_B w$ where S_B is the between class scatter matrix defined as:

$$\mathbf{S}_{\mathbf{B}} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathbf{T}}$$

NOTE: S_B is rank 1. This will be useful later on to find closed form solution for 2-class LDA

We also want to minimize….

- The variance or scatter of the projected samples from each class (i.e. we want to make each class more compact or closer to its mean). The scatter from class 1 defined as s_1 is given as

$$\tilde{s}_1^{\,2} = \sum_{i=1}^{N_1} (\tilde{x}_i - \tilde{\mu}_1)^2$$

- Thus we want to minimize the scatter of class 1 and class 2 in projected space, i.e.

minimize the total scatter $\tilde{s}_1^{\,2} + \tilde{s}_2^{\,2}$

Fisher Linear Discriminant Criterion Function

- Objective [1]-We want to ***maximize*** the between class scatter defined as: $|(\tilde{\mu}_1 - \tilde{\mu}_2)|^2$
- [2]- We want to ***minimize*** the within-class scatter.

$$\tilde{s}_1^{\ 2} + \tilde{s}_2^{\ 2}$$

- Thus we define our objective function *J*(**w**) as the following ratio that we want to ***maximize*** in order to achieve [1] and [2]:

$$J(\mathbf{w}) = \frac{|(\tilde{\mu}_1 - \tilde{\mu}_2)|^2}{\tilde{s}_1^{\ 2} + \tilde{s}_2^{\ 2}}$$

Scatter s_1 of Class 1

- Thus we want to find the vector **w** that maximizes J(**w**).
- Lets expand on scatter

$$\tilde{s}_1^{\,2} = \sum_{i=1}^{N_1} (\tilde{x}_i - \tilde{\mu}_1)^2$$

$$= \sum_{i=1}^{N_1} (\mathbf{w}^{\mathbf{T}}\mathbf{x}_i - \mathbf{w}^{\mathbf{T}}\boldsymbol{\mu}_1)^2$$

$$= \sum_{i=1}^{N_1} \mathbf{w}^{\mathbf{T}}(\mathbf{x}_i - \boldsymbol{\mu}_1)(\mathbf{x}_i - \boldsymbol{\mu}_1)^{\mathbf{T}}\mathbf{w}$$

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

Same for Scatter s_2 of Class 2

$$\tilde{s}_2^{\;2} = \sum_{i=1}^{N_2} (\tilde{x}_i - \tilde{\mu}_2)^2$$

$$= \sum_{i=1}^{N_2} (\mathbf{w}^{\mathbf{T}} \mathbf{x}_i - \mathbf{w}^{\mathbf{T}} \boldsymbol{\mu}_2)^2$$

$$= \sum_{i=1}^{N_2} \mathbf{w}^{\mathbf{T}} (\mathbf{x}_i - \boldsymbol{\mu}_2)(\mathbf{x}_i - \boldsymbol{\mu}_2)^{\mathbf{T}} \mathbf{w}$$

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

Total Within-Class Scatter Matrix

- We want to minimize total *within*-class scatter. i.e.

$$\tilde{s}_1^{\,2} + \tilde{s}_2^{\,2}$$

- Which is equivalent to $\mathbf{w}^{\mathbf{T}}\mathbf{S}_{\mathrm{w}}\mathbf{w}$

$$\mathbf{S}_{\mathbf{w}} = \sum_{i=1}^{C}\sum_{j=1}^{Ni} (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^{\mathbf{T}}$$

C=2, N_i=No. of images in *i*th class

Solving LDA

- Maximize $J(\mathbf{w}) = \dfrac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathbf{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathbf{W}}\mathbf{w}}$
- We need to find the optimal **w** which will maximize the above ratio (or quotient).
- So what do we do now?

  Take derivative and solve for **w**

Solving LDA

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S_B}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S_W}\mathbf{w}}$$

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = \frac{2\mathbf{w}^{\mathrm{T}}\mathbf{S_W}\mathbf{w}\mathbf{S_B}\mathbf{w} - 2\mathbf{w}^{\mathrm{T}}\mathbf{S_B}\mathbf{w}\mathbf{S_W}\mathbf{w}}{(\mathbf{w}^{\mathrm{T}}\mathbf{S_W}\mathbf{w})^2} = 0$$

$$=> \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S_W}\mathbf{w}\mathbf{S_B}\mathbf{w} - \mathbf{w}^{\mathrm{T}}\mathbf{S_B}\mathbf{w}\mathbf{S_W}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S_W}\mathbf{w}} = 0$$

$$=> \mathbf{S_B}\mathbf{w} - J(\mathbf{w})\mathbf{S_W}\mathbf{w} = 0$$

Solving LDA

$$\mathbf{S_B w} - J(\mathbf{w})\mathbf{S_W w} = 0$$

$$\mathbf{S_B w} - \lambda \mathbf{S_W w} = 0$$

$$\mathbf{S_B w} = \lambda \mathbf{S_W w}$$

Generalized Eigenvalue problem

$$\mathbf{S_W}^{-1}\mathbf{S_B w} = \lambda \mathbf{w}$$

If S_w is non-singular and invertible

We want to maximize J(w) thus we want the eigenvector w with the largest eigenvalue!

Special Case: LDA Solution for 2 Class Problems

- Lets replace what $\mathbf{S}_B$ is for two classes and see how we can simplify to get a closed form solution

  i.e. a solution of the vector **w** for the 2-class case.
- We know that in two class case, there is only 1 **w** vector. Lets use this knowledge cleverly…

S_B is Rank 1

$$\mathbf{S_B} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\mathbf{T} = \mathbf{m}\mathbf{m}^\mathbf{T}$$

$$\mathbf{S_B} = \mathbf{m}\mathbf{m}^\mathbf{T} = \begin{bmatrix} | & | & | \\ m(1)\mathbf{m} & m(2)\mathbf{m} & m(N)\mathbf{m} \\ | & | & | \end{bmatrix}$$

S_B has only 1 linearly independent colum vector => Rank 1 matrix

2-Class LDA

$$\mathbf{S_B} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathbf{T}}$$

$$\mathbf{S_B w} = \lambda \mathbf{S_W w}$$

Basically in this generalized eigenvalue/eigenvector problem, the number of valid eigenvectors with non-zero eigenvalue is determined by the *MIN* rank of matrix $\mathbf{S}_{\mathrm{B}}$ and $\mathbf{S}_{\mathrm{W}}$.

So *in this case*, there is only 1 valid eigenvector with a non-zero eigenvalue! i.e. there is only one valid w vector solution.

2-Class LDA

Simplify the 2 class case:

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathbf{T}}\mathbf{w} = \lambda \mathbf{S}_{\mathbf{W}}\mathbf{w}$$

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathbf{T}}\mathbf{w} = scalar = \beta$$

Which gives

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\beta = \lambda \mathbf{S}_{\mathbf{W}}\mathbf{w}$$

2-Class LDA - Closed Form Solution

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\beta = \lambda \mathbf{S}_{\mathbf{W}} \mathbf{w}$$

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \tfrac{\lambda}{\beta} \mathbf{S}_{\mathbf{W}} \mathbf{w}$$

$$\mathbf{S}_{\mathbf{W}}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \tfrac{\lambda}{\beta} \mathbf{w}$$

Since we can normalize w, we don't have to worry about the constants

$$\mathbf{w} = \mathbf{S}_{\mathbf{W}}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

2-Class LDA - Closed Form Solution

Nice closed form solution for 2-class LDA.

$$\mathbf{w} = \mathbf{S_W}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

We know that this is the right solution as we showed before that there is only ONE solution, so the closed-form solution we found is the optimal one.

Multi-Class LDA

- What if we have more that 2 classes…what then?
- Answer is we need more than one **w** projection vector to provide separapability.
- Lets look at our math framework to see what changes.

Multi-Class LDA

- Maximize

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathbf{T}}\mathbf{S}_{\mathbf{B}}\mathbf{w}}{\mathbf{w}^{\mathbf{T}}\mathbf{S}_{\mathbf{W}}\mathbf{w}}$$

- Lets start with the Between-Class Scatter matrix for 2 class.

$$\mathbf{S}_{\mathbf{B}} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathbf{T}}$$

- However $\mathbf{S}_{\mathrm{B}}$ now is the between class scatter matrix for many classes. Now we want to make all the class means furthest from each other. One way is to push them as far away from their global mean.

$$\mathbf{S}_{\mathbf{B}} = \sum_{i=1}^{C} (\boldsymbol{\mu}_i - \boldsymbol{\mu}\ )(\boldsymbol{\mu}_i - \boldsymbol{\mu}\ )^{\mathbf{T}}$$

Multi-Class LDA

- Solution

$$\mathbf{S_B w} = \lambda \mathbf{S_W w}$$

Here, the number of valid eigenvectors are bound by the MINIMUM rank of matrix (S_B,S_W). In this case S_B is typically lowest rank which is sum of *C* outer-product matrices. (Since they subtract the global mean, the rank is *C*-1).

$$\mathbf{S_W}^{-1}\mathbf{S_B w} = \lambda \mathbf{w}$$

If S_w is non-singular and invertible.

So for *C* classes we have at most *C*-1 w vectors which we can project on to.

Dealing With High Dimensional Data

- What happens when we deal with high-dimensional data.
- If we have more dimensions *d* than data samples *N*, then we run into more problems.
- Sw is singular. It will still have at most *N*-*C* non-zero eigenvalues.
- (*N* is the total number of samples from all classes, *C* is the number of classes)

$$\mathbf{S_w} = \sum_{i=1}^{C}\sum_{j=1}^{Ni} (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^{\mathbf{T}}$$

Fisherfaces

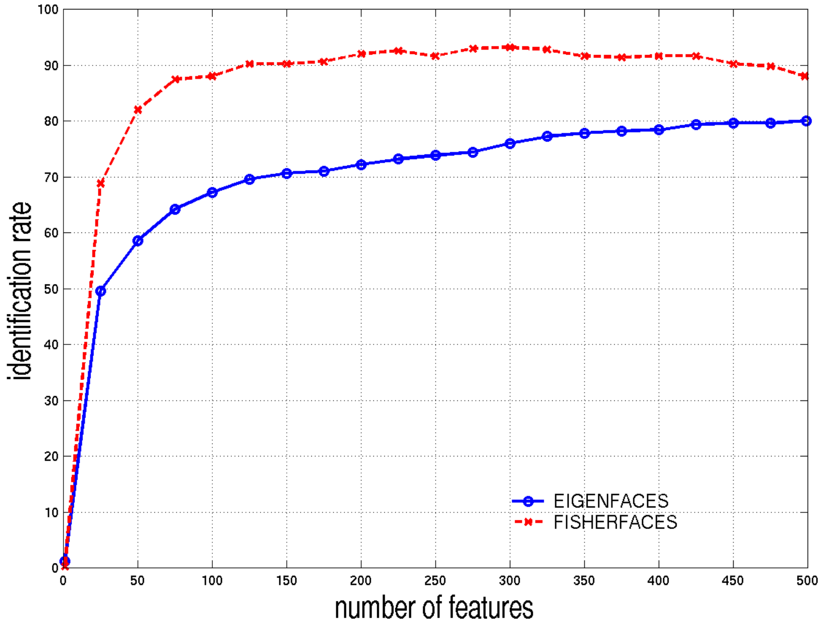
- Solution? Fisherfaces…..
- First do PCA and keep *N-C* eigenvectors. Project your data on to these *N-C* eigenvectors. ($\mathbf{S}_w$ will now be full rank)
- Then do LDA and compute the *C-1* projections in this lower-dimensional *N-C* subspace.
- PCA+LDA=Fisherfaces

Fisherfaces vs Eigenfaces

- Fisherfaces
  - Better discriminant for classification
  - The subject of a test image must be in the database
- Eigenfaces
  - No distinction between inter-class and intra-class faces
  - Optimal for representation but not for discrimination

Fisherfaces vs Eigenfaces

- FERET database



Best ID rates: eigenfaces - 80.0%, fisherfaces - 93.2%

Recap

- LDA vs PCA
- LDA Objective Function
- Closed Form Expression for 2 Class Case
- Multi-Class LDA
- Fisherfaces