Local Fisher Discriminant Analysis for Supervised Dimensionality Reduction

Masashi Sugiyama

Tokyo Institute of Technology, Japan
Dimensionality Reduction

High dimensional data is not easy to handle: Need to reduce dimensionality

We focus on

- **Linear** dimensionality reduction:
  \[ z \in \mathbb{R}^R \quad z = T^T x \quad x \in \mathbb{R}^D \]
  \[ R \ll D \]

- **Supervised** dimensionality reduction:
  \[ (x, y) \quad y \in \{1, 2, \ldots, C\} \]
Within-Class Multimodality

One of the classes has several modes

- Medical checkup:
  - hormone imbalance (high/low) vs. normal
- Digit recognition:
  - even (0,2,4,6,8) vs. odd (1,3,5,7,9)
- Multi-class classification:
  - one vs. rest
Goal of This Research

We want to embed multimodal data so that

- Between-class separability is maximized
- Within-class multimodality is preserved
Fisher Discriminant Analysis (FDA)\(^5\)  

- **Within-class scatter matrix:**
  \[
  S^{(w)} = \sum_{c=1}^{C} \sum_{i: y_i = c} (\mathbf{x}_i - \mu_c)(\mathbf{x}_i - \mu_c)^\top
  \]

- **Between-class scatter matrix:**
  \[
  S^{(b)} = \sum_{c=1}^{C} n_c (\mu_c - \mu)(\mu_c - \mu)^\top
  \]

- **FDA criterion:**
  \[
  \max_{\mathbf{T}} \left[ \text{tr}\left( (\mathbf{T}^\top S^{(w)} \mathbf{T})^{-1} \mathbf{T}^\top S^{(b)} \mathbf{T} \right) \right]
  \]

  - Within-class scatter is made small
  - Between-class scatter is made large

Fisher (1936)
Interpretation of FDA

- **Pairwise expressions:**
  
  \[
  S^{(w)} = \frac{1}{2} \sum_{i,j=1}^{n} A_{i,j}^{(w)} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top
  \]

  \[
  A_{i,j}^{(w)} = \begin{cases} 
  1/n_c & (y_i = y_j = c) \\
  0 & (y_i \neq y_j)
  \end{cases}
  \]

  \[
  S^{(b)} = \frac{1}{2} \sum_{i,j=1}^{n} A_{i,j}^{(b)} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top
  \]

  \[
  A_{i,j}^{(b)} = \begin{cases} 
  1/n - 1/n_c & (y_i = y_j = c) \\
  1/n & (y_i \neq y_j)
  \end{cases}
  \]

- **Samples in the same class are made close**
- **Samples in different classes are made apart**

\( n_c \): Number of samples in class \( C \)  
\( n \): Total number of samples
Examples of FDA

\[ \mathbb{R}^2 \mapsto \mathbb{R}^1 \]

Simple
Label-mixed cluster
Multimodal

FDA does not take within-class multimodality into account

NOTE: FDA can extract only \( C-1 \) features since

\[
\text{rank}(S^{(b)}) = C - 1
\]

\( C \) : Number of classes
Locality Preserving Projection (LPP) 

He & Niyogi (NIPS2003)

- Locality matrix:
  \[ S^{(l)} = \frac{1}{2} \sum_{i,j=1}^{n} A_{i,j} (x_i - x_j)(x_i - x_j)^\top \]

- Affinity matrix:
  e.g., \[ A_{i,j} = \exp(-\|x_i - x_j\|^2) \]

- LPP criterion:
  \[ \min_T \left[ \text{tr}(T^\top S^{(l)} T) \right] \]
  subject to \[ T^\top X D X^\top T = I \]

- Nearby samples in original space are made close
- Constraint is to avoid \[ T = O \]
Examples of LPP

$L^2 \rightarrow L^1$

Simple
Label-mixed cluster
Multimodal

LPP does not take between-class separability into account (unsupervised)
Our Approach

We combine FDA and LPP

- Nearby samples in the same class are made close
- Far-apart samples in the same class are not made close
- Samples in different classes are made apart
Local Fisher Discriminant Analysis

\[
\max_T \begin{bmatrix} \text{tr}((T^\top \tilde{S}^{(w)} T)^{-1} T^\top \tilde{S}^{(b)} T) \end{bmatrix}
\]

- **Local** within-class scatter matrix:

\[
\tilde{S}^{(w)} = \frac{1}{2} \sum_{i,j=1}^{n} \hat{A}_{i,j}^{(w)} (x_i - x_j)(x_i - x_j)^\top
\]

\[
\hat{A}_{i,j}^{(w)} = \begin{cases} 
A_{i,j}/n_c & (y_i = y_j = c) \\
0 & (y_i \neq y_j)
\end{cases}
\]

- **Local** between-class scatter matrix:

\[
\tilde{S}^{(b)} = \frac{1}{2} \sum_{i,j=1}^{n} \hat{A}_{i,j}^{(b)} (x_i - x_j)(x_i - x_j)^\top
\]

\[
\hat{A}_{i,j}^{(b)} = \begin{cases} 
A_{i,j}(1/n - 1/n_c) & (y_i = y_j = c) \\
1/n & (y_i \neq y_j)
\end{cases}
\]
How to Obtain Solution

\[ T_{LFDA} = \arg \max_T \left[ \text{tr}((T^\top \tilde{S}^{(w)} T)^{-1} T^\top \tilde{S}^{(b)} T) \right] \]

Since LFDA has a similar form to FDA, solution can be obtained just by solving a generalized eigenvalue problem:

\[ \tilde{S}^{(b)} \varphi = \lambda \tilde{S}^{(w)} \varphi \]

\[ T_{LFDA} = (\tilde{\varphi}_1 | \tilde{\varphi}_2 | \cdots | \tilde{\varphi}_R) \]

\[ \tilde{\varphi}_1, \tilde{\varphi}_2, \cdots, \tilde{\varphi}_R \]

\[ \tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_D \]
Examples of LFDA

\[ \mathbb{R}^2 \rightarrow \mathbb{R}^1 \]

**Simple**

**Label-mixed cluster**

**Multimodal**

LFDA works well for all three cases!

Note: Usually \( \text{rank}(\tilde{S}^{(b)}) \gg C \) so LFDA can extract more than \( C \) features (cf. FDA)
Neighborhood Component Analysis (NCA)

Goldberger, Roweis, Hinton & Salakhutdinov (NIPS2004)

- Minimize leave-one-out error of a stochastic k-nearest neighbor classifier
- Obtained embedding is separable
- NCA involves non-convex optimization
  - There are local optima
- No analytic solution available
  - Slow iterative algorithm
- LFDA has analytic form of global solution
Maximally Collapsing Metric Learning (MCML)

- Idea is similar to FDA
  - Samples in the same class are close ("one point")
  - Samples in different classes are apart
- MCML involves non-convex optimization
- There exists a nice convex approximation
  - Non-global solution
- No analytic solution available
  - Slow iterative algorithm

Globerson & Roweis (NIPS2005)
Simulations

- Visualization of UCI data sets:
  - Letter recognition (D=16)
  - Segment (D=18)
  - Thyroid disease (D=5)
  - Iris (D=4)

- Extract 3 classes from original data

- Merge 2 classes

\[ \mathbb{R}^D \rightarrow \mathbb{R}^2 \]
## Summary of Simulation Results

<table>
<thead>
<tr>
<th></th>
<th>Lett</th>
<th>Segm</th>
<th>Thyr</th>
<th>Iris</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDA</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>x</td>
<td>No multi-modal</td>
</tr>
<tr>
<td>LPP</td>
<td>x</td>
<td>x</td>
<td>o</td>
<td>o</td>
<td>No label-separability</td>
</tr>
<tr>
<td>LFDA</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td></td>
</tr>
<tr>
<td>NCA</td>
<td>o</td>
<td>x</td>
<td>o</td>
<td>o</td>
<td>Slow, local optima</td>
</tr>
<tr>
<td>MCML</td>
<td>△</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>Slow, no multi-modal</td>
</tr>
</tbody>
</table>

- **O**: Separable and multimodality preserved
- **△**: Separable but no multimodality
- **x**: Multimodality preserved but no separability
- **x**: Slow, no multi-modal
Letter Recognition

Blue vs. Red
Segment

FDA

LPP

LFDA

Blue vs. Red

NCA

MCML
Thyroid Disease

Blue vs. Red
Iris

FDA

LPP

LFDA

NCA

MCML

Blue vs. Red
Kernelization

- LFDA can be non-linearized by kernel trick
  \[ \langle \phi(x_i), \phi(x_j) \rangle = K(x_i, x_j) \]

- FDA: Kernel FDA  
  Mika et al. (NNSP1999)

- LPP: Laplacian eigenmap  
  Belkin & Niyogi (NIPS2001)

- MCML: Kernel MCML  
  Globerson & Roweis (NIPS2005)

- NCA: not available yet?
Conclusions

- LFDA effectively combines FDA and LPP.
- LFDA is suitable for embedding multimodal data.
- Same as FDA, LFDA has analytic optimal solution thus computationally efficient.
- Same as LPP, LFDA needs to pre-specify affinity matrix.
- We used local scaling method for computing affinity, which does not include any tuning parameter.  
  Zelnik-Manor & Perona (NIPS2004)