Local Fisher Discriminant Analysis

Masashi Sugiyama

Presented by: Wittawat Jitkrittum
wittawat@gatsby.ucl.ac.uk

Gatsby Tea Talk

27 June 2014

About this Talk

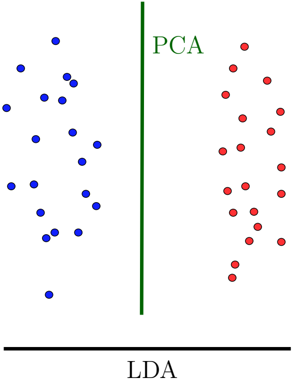
- **Local** Fisher discriminant analysis.
  - A modified version of linear discriminant analysis to handle multimodality:
- Only matrix algebra ...
- `Sugiyama, M.`
  `Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis.`
  `Journal of Machine Learning Research, vol.8 (May), pp.1027-1061, 2007.`
- `Sugiyama, M.`
  `Local Fisher discriminant analysis for supervised dimensionality reduction.`
  `ICML 2006`

Supervised Linear Dimensionality Reduction

- **Data matrix**: $X = (\boldsymbol{x}_1 | \cdots | \boldsymbol{x}_n) \in \mathbb{R}^{d \times n}$ where $\boldsymbol{x}_i \in \mathbb{R}^d$
- **Class labels**: $Y = (y_1, \ldots, y_n)$ where $y_i \in \{1, 2, \ldots, C\}$
- Find $T \in \mathbb{R}^{r \times d}$ to maximize some criterion $f(TX, Y)$.
- $r < d$
- T is a linear transform (hence the name).

Linear Discriminant Analysis (LDA)

- Also known as Fisher discriminant analysis
- T is found to maximize **Fisher's criterion**
  - between-class variance is maximized
  - within-class variance is minimized
- PCA is an unsupervised algorithm (does not see class labels).
- $T \in \mathbb{R}^{1 \times 2}$ in the plot



1d LDA for Two-class Problem

- Find the best direction $\boldsymbol{t}$ to maximize Fisher's criterion:

$$J(\boldsymbol{t}) = \frac{\boldsymbol{t}^\top S_b \boldsymbol{t}}{\boldsymbol{t}^\top S_w \boldsymbol{t}} = \frac{\text{between-class scatter}}{\text{within-class scatter}}$$

- Within-class scatter

$$\begin{aligned}
\sum_{c=1}^{C} \sum_{i:y_i=c} \left(\boldsymbol{t}^\top \boldsymbol{x}_i - \boldsymbol{t}^\top \boldsymbol{\mu}_c\right)^2 &= \sum_{c=1}^{C} \sum_{i:y_i=c} \boldsymbol{t}^\top \left(\boldsymbol{x}_i - \boldsymbol{\mu}_c\right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}_c\right)^\top \boldsymbol{t} \\
&= \boldsymbol{t}^\top \left[\sum_{c=1}^{C} \sum_{i:y_i=c} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_c\right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}_c\right)^\top\right] \boldsymbol{t} \\
&= \boldsymbol{t}^\top S_w \boldsymbol{t}
\end{aligned}$$

- Between-class scatter (difference of projected means)

$$\left(\boldsymbol{t}^\top \boldsymbol{\mu}_1 - \boldsymbol{t}^\top \boldsymbol{\mu}_2\right)^2 = \boldsymbol{t}^\top \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^\top \boldsymbol{t} = \boldsymbol{t}^\top S_b \boldsymbol{t}$$

Solution to 1d LDA

$$\boldsymbol{t}^* = \arg\max_{\boldsymbol{t}} \frac{\boldsymbol{t}^\top S_b \boldsymbol{t}}{\boldsymbol{t}^\top S_w \boldsymbol{t}}$$

Scale invariant. Equivalent to

$$\begin{aligned} \boldsymbol{t}^* &= \arg\max_{\boldsymbol{t}} \boldsymbol{t}^\top S_b \boldsymbol{t} \\ &\quad \text{subject to } \boldsymbol{t}^\top S_w \boldsymbol{t} = 1 \end{aligned}$$

Lagrangian

$$\begin{aligned} \mathcal{L} &= -\boldsymbol{t}^\top S_b \boldsymbol{t} + \lambda \left(\boldsymbol{t}^\top S_w \boldsymbol{t} - 1 \right) \\ \nabla_{\boldsymbol{t}} \mathcal{L} &= -2 S_b \boldsymbol{t} + 2\lambda S_w \boldsymbol{t} = 0 \\ \Rightarrow S_b \boldsymbol{t} &= \lambda S_w \boldsymbol{t} \end{aligned}$$

A generalized eigenvalue problem.

General LDA

$$\arg\max_{T} \operatorname{tr}\left(TS_bT^\top\right) = \sum_{i=1}^{r} \boldsymbol{t}_i^\top S_b \boldsymbol{t}_i$$

$$\text{subject to } TS_wT^\top = I$$

- Between-class scatter matrix

$$S_b = \sum_{c=1}^{C} n_c \left(\boldsymbol{\mu}_c - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_c - \boldsymbol{\mu}\right)^\top$$

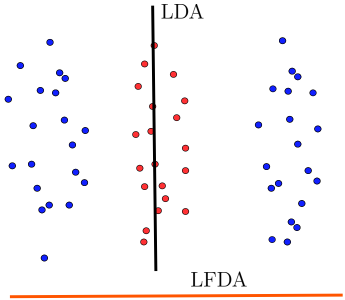
  where $n_c = \#\text{instances in class } c$, $\boldsymbol{\mu}_c = \frac{1}{n_c}\sum_{i:y_i=c} \boldsymbol{x}_i$ and $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$.
- Solution: $T := (\boldsymbol{t}_1|\cdots|\boldsymbol{t}_r)^\top$ where $\{\boldsymbol{t}_i\}_{i=1}^r$ are generalized eivenvectors

$$S_b\boldsymbol{t}_i = \lambda_i S_w \boldsymbol{t}_i$$

  with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$.

#1: Problem with Multimodality

- LDA cannot handle multimodal data e.g., blue class forms 2 clusters.
- Modified objective:
  - maximize between-class variance.
  - minimize within-class variance if class samples are close. Do not care if they are far away.
- ⇒ "**Local**" Fisher discriminant analysis
- Take locality of data into account



#2: Rank Deficiency of S_b

$$S_b = \sum_{c=1}^{C} n_c \left(\boldsymbol{\mu}_c - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_c - \boldsymbol{\mu}\right)^\top$$

- $S_b \in \mathbb{R}^{d \times d} =$ sum of C rank-one matrices. So, $\text{rank}\left(S_b\right) \leq C$.
- The C terms are dependent. In fact, $\text{rank}\left(S_b\right) \leq C - 1$.

$$S_b \boldsymbol{t}_i = \lambda_i S_w \boldsymbol{t}_i$$

- **Implications**:
  - $\lambda_1, \ldots, \lambda_{C-1}, \overbrace{\lambda_C, \ldots, \lambda_d}^{\text{always } 0}$. At most $C - 1$ non-zero eigenvalues.
  - At most $C - 1$ meaningful directions can be extracted.
  - For 2-class problems, only one direction can be extracted!

Basic Ideas of LFDA

- Rewrite S_w and S_b in a pairwise manner.
- Weight each pair according to a specified affinity matrix A.
- A captures the closeness of samples in the same class.
- LFDA solves both multimodality and rank problems.

Scatter Matrices Rewritten

$$S_b = \sum_{c=1}^{C} n_c (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^\top = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} B_{ij} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top$$

$$S_w = \sum_{c=1}^{C} \sum_{i:y_i=c} (\boldsymbol{x}_i - \boldsymbol{\mu}_c)(\boldsymbol{x}_i - \boldsymbol{\mu}_c)^\top = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top$$

where

$$B_{ij} = \begin{cases} 1/n - 1/n_c & \text{if } y_i = y_j = c, \\ 1/n & \text{if } y_i \neq y_j \end{cases}$$

$$W_{ij} = \begin{cases} 1/n_c & \text{if } y_i = y_j = c, \\ 0 & \text{if } y_i \neq y_j \end{cases}$$

- Proof. Expand $\boldsymbol{\mu}_c$ and rearrange terms

Local Scatter Matrices

$$\bar{S}_b = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{B}_{ij}\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^\top$$

$$\bar{S}_w = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{W}_{ij}\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^\top$$

where

$$\bar{B}_{ij} = \begin{cases} A_{ij}\,(1/n - 1/n_c) & \text{if } y_i = y_j = c, \\ 1/n & \text{if } y_i \neq y_j \end{cases}$$

$$\bar{W}_{ij} = \begin{cases} A_{ij}/n_c & \text{if } y_i = y_j = c, \\ 0 & \text{if } y_i \neq y_j \end{cases}$$

- Add $A \in \mathbb{R}^{n \times n}$, a pairwise affinity matrix capturing locality of data.
- $\bar{S}_b$ is typically not rank-deficient.

Local Fisher Discriminant Analysis

$$\arg\max_{T} \ \mathrm{tr}\left(T\bar{S}_b T^\top\right)$$

subject to $T\bar{S}_w T^\top = I$

Effects of LFDA

- Nearby pairs of the same class $\Rightarrow$ close
- Pairs of different classes $\Rightarrow$ apart
- Pairs of the same class but far apart $\Rightarrow$ don't care

Affinity matrix

- If $A_{ij} = 1$ for all in-class pairs, LFDA = LDA.
- To be useful, set $A_{ij} = 1$ only for nearby points.
- A_{ij} is only needed for in-class pairs. A is block diagonal.

Affinity Matrix Construction

Various choices from ([Belkin and Niyogi, 2003])

- ϵ-neighborhoods:

$$A_{ij} = 1 \text{ if } \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 < \epsilon$$

  May lead to several connected components

- k nearest neighbors (kNN)

$$A_{ij} = 1 \text{ if } \boldsymbol{x}_i \in \text{kNN}(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j \in \text{kNN}(\boldsymbol{x}_i)$$

- Gaussian kernel: $A_{ij} = \exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/2\sigma^2\right)$

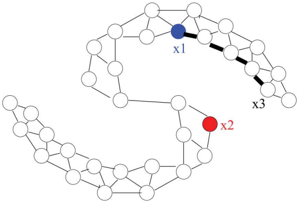


Image from [Zhu, 2007]

Equivalent Problem of LFDA

$$\begin{aligned} \arg\max_{T} \quad & \mathrm{tr}\left(T\,\bar{S}\,T^\top\right) \\ \text{subject to} \quad & T\bar{S}_w T^\top = I \end{aligned}$$

- $\bar{S} = \bar{S}_w + \bar{S}_b = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{M}_{ij}\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^\top$
- $\bar{M}_{ij} = \bar{B}_{ij} + \bar{W}_{ij} = \begin{cases} A_{ij}/n & \text{if } y_i = y_j \\ 1/n & \text{if } y_i \neq y_j \end{cases}$
- maximize between-class scatter = maximize global scatter
- It can be shown that $\bar{S} = X\bar{L}X^\top$ where $\bar{L} = \mathrm{diag}\left(\bar{M}\mathbf{1}\right) - \bar{M}$ (Laplacian matrix).
- $\bar{S}_w = X\bar{L}_w X^\top$ where $\bar{L}_w = \mathrm{diag}\left(\bar{W}\mathbf{1}\right) - \bar{W}$.

Kernel LFDA

$$\begin{aligned} \bar{S}\boldsymbol{t}_i &= \lambda_i \bar{S}_w \boldsymbol{t}_i \\ \Rightarrow X\bar{L}X^\top \boldsymbol{t}_i &= \lambda_i X\bar{L}_w X^\top \boldsymbol{t}_i \end{aligned}$$

- $\boldsymbol{t}_i$ must be in column space of X. So, $\boldsymbol{t}_i = X\boldsymbol{\alpha}_i$ for some $\boldsymbol{\alpha}_i \in \mathbb{R}^n$.

$$X\bar{L}X^\top X\boldsymbol{\alpha}_i = \lambda_i X\bar{L}_w X^\top X\boldsymbol{\alpha}_i$$

- Left multiply with $X^\top$. Replace $X^\top X$ with K (kernel matrix).

$$K\bar{L}K\boldsymbol{\alpha}_i = \lambda_i K\bar{L}_w K\boldsymbol{\alpha}_i$$

  where $K_{ij} = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$.
- Nonlinear embedding of $\boldsymbol{x}'$:

$$\underbrace{(\boldsymbol{\alpha}_1 | \cdots | \boldsymbol{\alpha}_r)^\top}_{r\times n} \underbrace{\big(k(\boldsymbol{x}_1, \boldsymbol{x}'), \ldots, k(\boldsymbol{x}_n, \boldsymbol{x}')\big)^\top}_{n\times 1}$$

Conclusions

- LDA gives T which
  - minimizes within-class variance
  - maximizes between-class variance
- LFDA extends LDA
  - capture locality of data with affinity matrix
- Kernelized version exists.

References I

Belkin, M. and Niyogi, P. (2003).
Laplacian eigenmaps for dimensionality reduction and data representation.
Neural Computation, 15:1373–1396.

Zhu, X. (2007).
Semi-supervised learning tutorial.