CS 789 ADVANCED BIG DATA

FISHER LINEAR DISCRIMINANT ANALYSIS

* Some contents are adapted from Dr. Hung Huang, Dr. Vassilis Athitsos at UT Arlington, and Prof. Olga Veksler, Western university

Migon Kang, Ph.D.
Department of Computer Science, University of Nevada, Las Vegas
Data Representation vs. Classification

- PCA is for data representation minimizing information loss (developed in 1939)
  - Project data into new spaces of maximum variance
- However, good to classification?
Linear Discriminant Analysis

- Project data to a new space that discriminates classes
Linear Discriminant Analysis

- Suppose we have 2 classes and $d$-dimensional samples $x_1, \ldots, x_n$ where
  - $n_1$ samples come from the first class
  - $n_2$ samples come from the second class
- Consider projection on a line
- Let the line direction be given by unit vector $v$

- Scalar $v^T x_i$ is the distance of projection of $x_i$ from the origin
- Thus it $v^T x_i$ is the projection of $x_i$ into a one dimensional subspace
Linear Discriminant Analysis

- Thus the projection of sample $x_i$ onto a line in direction $v$ is given by $v^t x_i$
- How to measure separation between projections of different classes?
- Let $\bar{\mu}_1$ and $\bar{\mu}_2$ be the means of projections of classes 1 and 2
- Let $\mu_1$ and $\mu_2$ be the means of classes 1 and 2
- $|\bar{\mu}_1 - \bar{\mu}_2|$ seems like a good measure

\[ \bar{\mu}_1 = \frac{1}{n_1} \sum_{x_i \in C_1} v^t x_i = v^t \left( \frac{1}{n_1} \sum_{x_i \in C_1} x_i \right) = v^t \mu_1 \]

Similarly,
\[ \bar{\mu}_2 = v^t \mu_2 \]
Linear Discriminant Analysis

- How good is $|\tilde{\mu}_1 - \tilde{\mu}_2|$ as a measure of separation?
  - The larger $|\tilde{\mu}_1 - \tilde{\mu}_2|$, the better is the expected separation

- the vertical axes is a better line than the horizontal axes to project to for class separability
- however $|\tilde{\mu}_1 - \tilde{\mu}_2| > |\tilde{\mu}_1 - \tilde{\mu}_2|$
Linear Discriminant Analysis

- The problem with $|\bar{\mu}_1 - \bar{\mu}_2|$ is that it does not consider the variance of the classes.
Linear Discriminant Analysis
Linear Discriminant Analysis

- We need to normalize $|\bar{\mu}_1 - \bar{\mu}_2|$ by a factor which is proportional to variance.
- Have samples $z_1, \ldots, z_n$. Sample mean is $\mu_z = \frac{1}{n} \sum_{i=1}^{n} z_i$.
- Define their **scatter** as
  \[ s = \sum_{i=1}^{n} (z_i - \mu_z)^2 \]
- Thus scatter is just sample variance multiplied by $n$.
  - scatter measures the same thing as variance, the spread of data around the mean.
  - scatter is just on different scale than variance.

"larger scatter: \[ \bullet \bullet \bullet \bullet \bullet \]
"smaller scatter: \[ \bullet \bullet \]

Linear Discriminant Analysis

- Fisher Solution: normalize $|\bar{\mu}_1 - \bar{\mu}_2|$ by scatter
- Let $y_i = v^T x_i$, i.e. $y_i$'s are the projected samples
- Scatter for projected samples of class 1 is
  \[ s_1^2 = \sum_{y_i \in \text{Class 1}} (y_i - \bar{y}_1)^2 \]
- Scatter for projected samples of class 2 is
  \[ s_2^2 = \sum_{y_i \in \text{Class 2}} (y_i - \bar{y}_2)^2 \]
Linear Discriminant Analysis

- We need to normalize by both scatter of class 1 and scatter of class 2
- Thus Fisher linear discriminant is to project on line in the direction $v$ which maximizes

$$J(v) = \frac{(\mu_1 - \mu_2)^2}{\bar{s}_1^2 + \bar{s}_2^2}$$

- want projected means are far from each other

want scatter in class 1 is as small as possible, i.e. samples of class 1 cluster around the projected mean $\bar{\mu}_1$

want scatter in class 2 is as small as possible, i.e. samples of class 2 cluster around the projected mean $\bar{\mu}_2$
Linear Discriminant Analysis

\[ J(v) = \frac{(\mu_1 - \mu_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \]

- If we find \( v \) which makes \( J(v) \) large, we are guaranteed that the classes are well separated

Projected means are far from each other

\( \hat{\mu}_1 \)

small \( \tilde{s}_1 \) implies that projected samples of class 1 are clustered around projected mean

\( \hat{\mu}_2 \)

small \( \tilde{s}_2 \) implies that projected samples of class 2 are clustered around projected mean
Linear Discriminant Analysis

- Now define the within the class scatter matrix
  \[ S_w = S_1 + S_2 \]

- Recall that
  \[ \tilde{s}_1^2 = \sum_{y_i \in \text{Class 1}} (y_i - \tilde{\mu}_1)^2 \]

- Using \( y_i = \mathbf{v}^t \mathbf{x}_i \) and \( \tilde{\mu}_1 = \mathbf{v}^t \mu_1 \)
  \[
  \tilde{s}_1^2 = \sum_{y_i \in \text{Class 1}} (\mathbf{v}^t \mathbf{x}_i - \mathbf{v}^t \mu_1)^2 \\
  = \sum_{y_i \in \text{Class 1}} (\mathbf{v}^t (\mathbf{x}_i - \mu_1))^t (\mathbf{v}^t (\mathbf{x}_i - \mu_1)) \\
  = \sum_{y_i \in \text{Class 1}} ((\mathbf{x}_i - \mu_1)^t \mathbf{v})^t ((\mathbf{x}_i - \mu_1)^t \mathbf{v}) \\
  = \sum_{y_i \in \text{Class 1}} \mathbf{v}^t (\mathbf{x}_i - \mu_1)(\mathbf{x}_i - \mu_1)^t \mathbf{v} = \mathbf{v}^t S_1 \mathbf{v} \]
Linear Discriminant Analysis

- Similarly \( \tilde{s}_2^2 = v^t S_2 v \)
- Therefore \( \tilde{s}_1^2 + \tilde{s}_2^2 = v^t S_1 v + v^t S_2 v = v^t S_W v \)
- Define between the class scatter matrix
  \[ S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^t \]
- \( S_B \) measures separation between the means of two classes (before projection)
- Let’s rewrite the separations of the projected means
  \[ (\bar{\mu}_1 - \bar{\mu}_2)^2 = (v^t \mu_1 - v^t \mu_2)^2 \]
  \[ = v^t (\mu_1 - \mu_2)(\mu_1 - \mu_2)^t v \]
  \[ = v^t S_B v \]
Thus our objective function can be written:

$$J(v) = \frac{(\bar{\mu}_1 - \bar{\mu}_2)^2}{\hat{S}_1^2 + \hat{S}_2^2} = \frac{v^t S_B v}{v^t S_W v}$$

Minimize $J(v)$ by taking the derivative w.r.t. $v$ and setting it to 0.

$$\frac{d}{dv} J(v) = \frac{\left(\frac{d}{dv} v^t S_B v\right) v^t S_W v - \left(\frac{d}{dv} v^t S_W v\right) v^t S_B v}{(v^t S_W v)^2}$$

$$= \frac{(2 S_B v) v^t S_W v - (2 S_W v) v^t S_B v}{(v^t S_W v)^2} = 0$$
Linear Discriminant Analysis

- Need to solve \( \nu^t S_w \nu (S_B \nu) - \nu^t S_B \nu (S_w \nu) = 0 \)

\[
\Rightarrow \frac{\nu^t S_w \nu (S_B \nu)}{\nu^t S_w \nu} - \frac{\nu^t S_B \nu (S_w \nu)}{\nu^t S_w \nu} = 0
\]

\[
\Rightarrow S_B \nu - \frac{\nu^t S_B \nu (S_w \nu)}{\nu^t S_w \nu} = 0
\]

\[
\Rightarrow S_B \nu = \lambda S_w \nu
\]

\textit{generalized eigenvalue problem}
Linear Discriminant Analysis

\[ S_B v = \lambda S_W v \]

- If \( S_W \) has full rank (the inverse exists), can convert this to a standard eigenvalue problem
  \[ S_W^{-1} S_B v = \lambda v \]

- But \( S_B x \) for any vector \( x \), points in the same direction as \( \mu_1 - \mu_2 \)
  \[ S_B x = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^t x = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^t x = \alpha(\mu_1 - \mu_2) \]

- Thus can solve the eigenvalue problem immediately
  \[ v = S_W^{-1} (\mu_1 - \mu_2) \]

\[ S_W^{-1} S_B [S_W^{-1} (\mu_1 - \mu_2)] = S_W^{-1} [\alpha(\mu_1 - \mu_2)] = \alpha [S_W^{-1} (\mu_1 - \mu_2)] \]
Linear Discriminant Analysis

- **Data**
  - Class 1 has 5 samples $c_1 = [(1,2),(2,3),(3,3),(4,5),(5,5)]$
  - Class 2 has 6 samples $c_2 = [(1,0),(2,1),(3,1),(3,2),(5,3),(6,5)]$

- **Arrange data in 2 separate matrices**

  $c_1 = \begin{bmatrix} 1 & 2 \\ 5 & 5 \end{bmatrix} \quad c_2 = \begin{bmatrix} 1 & 0 \\ 6 & 5 \end{bmatrix}$

- Notice that PCA performs very poorly on this data because the direction of largest variance is not helpful for classification.
Linear Discriminant Analysis

- First compute the mean for each class
  \[ \mu_1 = \text{mean}(c_1) = \begin{bmatrix} 3 \\ 3.6 \end{bmatrix}, \quad \mu_2 = \text{mean}(c_2) = \begin{bmatrix} 3.3 \\ 2 \end{bmatrix} \]

- Compute scatter matrices \( S_1 \) and \( S_2 \) for each class
  \[ S_1 = 4 \cdot \text{cov}(c_1) = \begin{bmatrix} 10 & 8.0 \\ 8.0 & 7.2 \end{bmatrix}, \quad S_2 = 5 \cdot \text{cov}(c_2) = \begin{bmatrix} 17.3 & 16 \\ 16 & 16 \end{bmatrix} \]

- Within the class scatter:
  \[ S_W = S_1 + S_2 = \begin{bmatrix} 27.3 & 24 \\ 24 & 23.2 \end{bmatrix} \]
  - it has full rank, don’t have to solve for eigenvalues

- The inverse of \( S_W \) is \( S_W^{-1} = \text{inv}(S_W) = \begin{bmatrix} 0.39 & -0.41 \\ -0.41 & 0.47 \end{bmatrix} \)

- Finally, the optimal line direction \( \mathbf{v} \)
  \[ \mathbf{v} = S_W^{-1}(\mu_1 - \mu_2) = \begin{bmatrix} -0.79 \\ 0.89 \end{bmatrix} \]
Linear Discriminant Analysis

- Notice, as long as the line has the right direction, its exact position does not matter.

- Last step is to compute the actual 1D vector $y$. Let’s do it separately for each class.

\[
Y_1 = v^t c_1^t = \begin{bmatrix} -0.65 & 0.73 \end{bmatrix} \begin{bmatrix} 1 & \cdots & 5 \\ 2 & \cdots & 5 \end{bmatrix} = \begin{bmatrix} 0.81 & \cdots & 0.4 \end{bmatrix}
\]

\[
Y_2 = v^t c_2^t = \begin{bmatrix} -0.65 & 0.73 \end{bmatrix} \begin{bmatrix} 1 & \cdots & 6 \\ 0 & \cdots & 5 \end{bmatrix} = \begin{bmatrix} -0.65 & \cdots & -0.25 \end{bmatrix}
\]
Reference

- Original paper

- Prof. Olga Veksler, Western University

- Good example shows LDA step by step
  - [http://sebastianraschka.com/Articles/2014_python_lda.htm](http://sebastianraschka.com/Articles/2014_python_lda.htm)

- Very good explanation for equations
  - [https://www.youtube.com/watch?v=T26-kN54gJY](https://www.youtube.com/watch?v=T26-kN54gJY)