Linear Discriminant Analysis

Pattern Recognition And Machine Learning

Indian Institute of Technology Madras

November 1, 2019

Linear discriminant analysis

- Consider a two class problem
- FDA projects the data into a lower dimension where the data from both the classes are well separated
- The data points of the two classes $1, 2$ are given by

$$D_i = \{\bar{x}_{1i}, \bar{x}_{2i}, \cdots, \bar{x}_{ni}\} \quad \text{where, } i = \{1, 2\}$$

$$\bar{m}_i = \frac{1}{n_i} \sum_{\bar{x} \in D_i} \bar{x}$$

- The mean in the projected space is given by

$$\tilde{m}_i = \frac{1}{n_i} \sum_{\bar{y} \in D_i} \bar{y}$$

$$= \frac{1}{n_i} \sum_{\bar{y} \in D_i} \bar{w}^t \bar{x}$$

$$= \frac{1}{n_i} \bar{w}^t \sum_{\bar{y} \in D_i} \bar{x}$$

$$= \bar{w}^t \frac{1}{n_i} \sum_{\bar{y} \in D_i} \bar{x}$$

$$= \bar{w}^t m_i$$

- The difference in the projected mean is given by

$$|\tilde{m}_1 - \tilde{m}_2| = |\bar{w}^t (\bar{m}_2 - \bar{m}_2)|$$

- Form a scatter matrix

$$\tilde{S_1}^2 = \sum_{y \in y_i} (y - \tilde{m}_i)^2$$

- Pooled scatter matrix is given by $\frac{1}{n}({S_1}^2 + {S_1}^2)$
- A cost function is defined as

$$J(\bar{w}) = \frac{|\tilde{\bar{m}}_1 - \tilde{\bar{m}}_2|^2}{{S_1}^2 + {S_1}^2}$$

- To define an optimal $\bar{w}$, we define scatter matrices for each class on the input data

$$\begin{aligned} {S_i}^2 &= \sum_{\bar{x} \in D_i} (\bar{x} - \bar{m}_i)(\bar{x} - \bar{m}_i)^t \\ {\tilde{S}_i}^2 &= \sum_{\bar{x} \in D_i} \bar{W}^t (\bar{x} - \bar{m}_i)(\bar{x} - \bar{m}_i)^t \bar{W} \\ &= \bar{W}^t \Big[\sum_{\bar{x} \in D_i} (\bar{x} - \bar{m}_i)(\bar{x} - \bar{m}_i)^t \Big] \bar{W} \end{aligned}$$

$$S_1{}^2 + S_2{}^2 = \bar{W}^t S_w \bar{W}$$

In the same way

$$\begin{aligned}
|\tilde{\bar{m}}_1 - \tilde{\bar{m}}_2|^2 &= (\bar{w}^t \bar{m}_1 - \bar{w}^t \bar{m}_2)^2 \\
&= \bar{w}^t (\bar{m}_1 - \bar{m}_2)(\bar{m}_1 - \bar{m}_2)^t \bar{w} \\
\implies J(\bar{w}) &= \frac{\bar{w}^t S_B \bar{w}}{\bar{w}^t S_W \bar{w}}
\end{aligned}$$

Setting $\dfrac{\partial J(\bar{w})}{\partial \bar{w}} = 0$ yields

$$\begin{aligned}
S_B \bar{w} &= \lambda S_W \bar{w} \\
\lambda &= S_W^{-1} S_B \bar{w} \\
S_B \bar{w} &= (\bar{m}_1 - \bar{m}_2)(\bar{m}_1 - \bar{m}_2)^t \bar{w} \\
S_B \bar{w} &\propto \bar{m}_1 - \bar{m}_2 \\
\implies S_W \bar{w} &\propto \bar{m}_1 - \bar{m}_2
\end{aligned}$$

Hence $\bar{w} = S_W^{-}1(\bar{m}_1 - \bar{m}_2)$

Multiple class discriminant analysis

C-class problem

- For c classes, one against rest strategy is applied. $\frac{c(c-1)}{2}$ pairs and voting rule is applied for classification.
- In FLDA, data is projected from d-space to $c-1$ spaces ($d \geq c$).

Generalization

$$S_w = \sum_{i=1}^{c} S_i$$

$$S_i = \sum_{x \in D_i} (\bar{x} - \bar{m}_i)(\bar{x} - \bar{m}_i)^t$$

Generalization of between class scatter

$$\bar{m} = \frac{1}{n} \sum_{\bar{x}} \bar{x} = \frac{1}{n} \sum_{i=1}^{c} n_i \bar{m}_i$$

(all classes)

Define Total scatter $S_T = \sum_{\bar{x}} (\bar{x} - \bar{m})(\bar{x} - \bar{m})^t$
(all classes)

$$S_T = \sum_{i=1}^{c} \sum_{x \in D_i} (\bar{x} - \bar{m})(\bar{x} - \bar{m})^t$$

$$= \sum_{i=1}^{c} \sum_{x \in D_i} (\bar{x} - \bar{m}_i + \bar{m}_i - \bar{m})(\bar{x} - \bar{m}_i + \bar{m}_i - \bar{m})^t$$

$$= \sum_{i=1}^{c} \sum_{x \in D_i} (\bar{x} - \bar{m}_i)(\bar{x} - \bar{m}_i)^t + \sum_{i=1}^{c} \sum_{x \in D_i} (\bar{m}_i - \bar{m})(\bar{m}_i - \bar{m})^t$$

$$S_T = \sum_{i=1}^{c} \sum_{x \in D_i} (\bar{x} - \bar{m}_i)(\bar{m}_i - \bar{m})^t + \sum_{i=1}^{c} (\sum_{x \in D_i} \bar{x} - \sum_{x \in D_i} \bar{m}_i)(\bar{m}_i - \bar{m})^t$$

$$S_T = S_W + S_B$$

$$S_B = \sum_{i=1}^{c} (\bar{m}_i - \bar{m})(\bar{m}_i - \bar{m})^t$$

$y_i = \bar{w}_i^t \bar{x}$; $i = 1, 2, ..., c-1$
$\bar{y} = W^t \bar{x}$, where W is $c-1$ FLDA directions

Measure of scatter

$$J(W) = |\frac{W^t S_B W}{W^t S_w W}|$$

$$S_B \bar{w}_i = \lambda_i S_w \bar{w}_i$$

$$|S_B - \lambda_i S_w| = 0$$

Aside:

- FLDA is like unsupervised k-means
- Finding $c - 1$ directions is like putting hard boundaries between one class and all other classes

- In k-means clustering:
  - $S_k = \sum_{x \in D} (\bar{x} - \bar{m}_k)(\bar{x} - \bar{m}_k)^t$ is like within class scatter in FLDA (k is the cluster index)
  - $S_w = \sum_{k=1}^{K} S_k$ where K is the total number of clusters
  - $S_B = \sum_{k=1}^{K} N_k(\bar{m}_k - \bar{m})(\bar{m}_k - \bar{m})^t$
  - $\bar{m} = \frac{1}{N} \sum_{n=1}^{N} \bar{x}_n$
  - $S_T = \sum_{n=1}^{N} (\bar{x}_n - \bar{m})(\bar{x}_n - \bar{m})^t$
    $= S_w + S_B$
  - Trace of $S_w = \sum_{i=1}^{K}$ Trace of S_K.
    In k-means, after every iteration, we try to reduce trace of S_w.
    $Trace(S_T) = Trace(S_w) + Trace(S_B)$
    $Trace(S_T)$ is consatant $\implies$ if $Trace(S_w)$ decreases $Trace(S_B)$ increases.