MODULE 5
Feature Extraction

LESSON 7

Fisher's Discriminant

Keywords: Projection, Within Class, Between Class, Scatter

Feature Extraction

- Feature extraction refers to the process of identifying and combining certain features of patterns. Given a collection F of features, the extracted features are linear/non-linear combinations of elements of F.
- This is a pre-processing step of pattern recognition.
- Before pattern classification is carried, it is necessary to decide what attributes of the patterns are to be measured and recorded.
- The features chosen should be discriminating features of the pattern.
- It is an important stage of pre-processing, as the feature extraction influences the quality of the pattern classification carried out.
- Two methods explained in this module are :
    1. Fisher's Linear Discriminant : This is used in supervised classification. Supervised classification refers to classification being carried out where labeled training examples are available to learn the classifier.
    2. Principal Component Analysis : This is an unsupervised learning activity. Unsupervised learning activity uses the underlying distribution of the patterns to carry out classification. No labeled examples are available. A typical unsupervised learning activity is clustering.

Fisher's Linear Discriminant

- Fisher's Linear Discriminant is used to map a d-dimensional data to one dimensional data using a projection vector V such that it maps a vector X to a scalar V^tX; note that V^tX is a scalar.
- Classification is then carried out in the one dimensional space.
- Let us consider a two-class problem.

- If the mean of class 1 is m_1 and the mean of class 2 is m_2, and if v_1 is proportional to the variance of class 1 patterns and v_2 is proportional to the variance of class 2 patterns, then the Fisher's criterion is

  $J(V) \;=\; \frac{|m_1 - m_2|^2}{v_1^2 + v_2^2}$
  $J(V)$ is the criterion function to be maximized and it is necessary to find V such that that $J(V)$ is maximized.

- If there are n patterns, then the mean of the patterns is

$$m \;=\; \frac{1}{n}\sum_{k=1}^{n} X_k$$

- If there are k classes, then the mean of the patterns of Class i is

$$m_i \;=\; \frac{1}{n_i}\sum_{X_i \in Class\, i} X_i$$

- The between class scatter matrix is

$$B \;=\; \sum_{i=1}^{k} n_i(m_i - m)(m_i - m)^t$$

- The within class scatter matrix is

$$W \;=\; \sum_{i=1}^{k}\sum_{X_i \in Class\, i} (X_i - m_i)(X_i - m_i)^t$$

- The criterion function to be maximized is

  $J(V) \ = \ \frac{(V^t \ B \ V)}{(V^t \ W \ V)}$

  The problem of maximizing $J(V)$ can be written as the following constrained optimization problem.

  min $-\frac{1}{2}V^t * B * V$

  s.t. $V^t * W * V = 1$

  which corresponds to the Lagrangian

  L=$-\frac{1}{2}V^t * B * V + \frac{1}{2} * \lambda * (V^t * W * V - 1)$

  From the KKT conditions[1]

  we can see that the following equation has to hold.

  $B * V = \lambda * W * V$

  This means that

  $W^{-1} * B * V = \lambda * V$

- Let $v_1, v_2, ..., v_d$ gives the generalized eigenvectors of B and W, which gives a projection space of dimension d.

- We need to get a projection space of dimension $m < d$ giving the eigenvectors

[1]For information on Karush-Kuhn-Tucker(KKT) conditions refer Bazaraa, Sherali and Shetty, *Nonlinear Programming, Theory and Algorithms*, 3rd Edition, John Wiley and Sons, 2006.

$V_m \ = \ (v_1, v_2, ..., v_m)$

- The projection of vector X_i into a subspace of dimension m is

y $= W_d^t X$

Two-class problem

In this case,

$$m_1 = \frac{1}{n_1} \sum_{X_i \in C_1} X_i$$

$$m_2 = \frac{1}{n_2} \sum_{X_i \in C_2} X_i$$

$$B = n_1(m_1 - m)(m_1 - m)^t + n_2(m_2 - m)(m_2 - m)^t$$

$$W = \sum_{X_i \in C_1} (X_i - m_1)(X_i - m_1)^t + \sum_{X_i \in C_2} (X_i - m_2)(X_i - m_2)^t$$

$B * V \ = \lambda W * V$
This means that

$W^{-1} B V \ = \lambda V$

The matrix W can be a singular matrix in which case W^{-1} will not exist.

We know that $B * V = (m_2 - m_1)(m_2 - m_1)^t * V$
$= (m_2 - m_1)(V^t m_2 - V^t m_1)$
$= c * (m_2 - m_1)$

This is because $(V^t m_2 - V^t m_1) =$ c is a scalar.
This shows that B*V is in the direction of $(m_2 - m_1)$.

Since BV is in the direction of $(m_2 - m_1)$, the solution for V is :

$$V = W^{-1}(m_2 - m_1)$$

This will be the solution for V if W^{-1} exists. It gives the direction of V.

- By doing this, a d-dimensional problem is converted into a one-dimensional problem.

- As an example, consider six points namely (1,1,1), (2,1,2) and (2,1,3) of Class 1 and (2,5,6), (1,6,5) and (2,7,5) of Class 2.

  Then, mean of Class 1 is

$$m_1 = \begin{bmatrix} 1.67 \\ 1 \\ 2 \end{bmatrix}$$

  and the mean of Class 2 is

$$m_2 = \begin{bmatrix} 1.67 \\ 6 \\ 5.33 \end{bmatrix}$$

$$(m_2 - m_1) = \begin{bmatrix} 1.67 - 1.67 \\ 6 - 1 \\ 5.33 - 2 \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ 5 \\ 3.33 \end{bmatrix}$$

The within class scatter matrix will be

$$W = \begin{bmatrix} -0.67 \\ 0 \\ -1 \end{bmatrix} * \begin{bmatrix} -0.67 & 0 & -1 \end{bmatrix} + \begin{bmatrix} 0.33 \\ 0 \\ 0 \end{bmatrix} * \begin{bmatrix} 0.33 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0.33 \\ 0 \\ 1 \end{bmatrix}$$

$$* \begin{bmatrix} 0.33 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 0.33 \\ -1 \\ 0.66 \end{bmatrix} * \begin{bmatrix} 0.33 & -1 & 0.66 \end{bmatrix} + \begin{bmatrix} -0.67 \\ 0 \\ -0.33 \end{bmatrix} * \begin{bmatrix} -0.67 & 0 & -0.33 \end{bmatrix}$$

$$+$$

$$\begin{bmatrix} 0.33 \\ 1 \\ -0.33 \end{bmatrix} * \begin{bmatrix} 0.33 & 1 & -0.33 \end{bmatrix} = \begin{bmatrix} 1.33 & 0 & 1.33 \\ 0 & 2 & -1 \\ 1.33 & -1 & 2.65 \end{bmatrix}$$

The inverse of the within class matrix will be

$$W^{-1} = \frac{1}{2.18} \begin{bmatrix} 4.3 & -1.33 & -2.66 \\ -1.33 & 1.76 & 1.33 \\ -2.66 & 1.33 & 2.66 \end{bmatrix}$$

The direction is given by

$$V = W^{-1}(m_1 - m2) = \frac{1}{2.18} \begin{bmatrix} 4.3 & -1.33 & -2.66 \\ -1.33 & 1.76 & 1.33 \\ -2.66 & 1.33 & 2.66 \end{bmatrix} * \begin{bmatrix} 0 \\ 5 \\ 3.33 \end{bmatrix}$$

$$V = \frac{1}{2.18} \begin{bmatrix} -2.21 \\ 4.37 \\ 2.21 \end{bmatrix} =$$

$$\begin{bmatrix} -1.01 \\ 2.00 \\ 1.01 \end{bmatrix}$$

Now given a point $\begin{bmatrix} 4 \\ 1 \\ 1 \end{bmatrix}$, the projection is

-4.04 + 3.01 = -1.03

So, $\begin{bmatrix} 4 \\ 1 \\ 1 \end{bmatrix} \epsilon$ Class 1

Given $\begin{bmatrix} 1 \\ 4 \\ 4 \end{bmatrix}$, the projection is

-1.01+2+4.04 = 5.03

So $\begin{bmatrix} 1 \\ 4 \\ 4 \end{bmatrix} \epsilon$ Class 2.