Feature Selection – E186 Handout

Feature Selection

The main purpose of feature selection is to reduce the computational cost by using only m ($m < n$) features for recognition/classification purposes. These m features can be either directly chosen from the original n features, or generated as some linear combinations of the original features. To prevent the result from degrading, the features selected should keep as much separability information as possible.

Choose m features from n original ones

There are

$$C_n^m = \frac{n!}{(n-m)!m!}$$

ways to choose m features from n ones. We just need to find the m best ones to span an m-dimensional feature space in which any of the following separability criteria J is maximized.

- $$J_1 = \sum_{i \neq j} P_i P_j D_B(\omega_i, \omega_j)$$

  where P_i and P_j are the *a priori* probabilities for class ω_i and ω_j, respectively.

- $$J_2 = tr\ (S_W^{-1} S_B) = tr\ (S_{B/W})$$

  where, for convenience, $S_{B/W}$ is defined as

  $$S_{B/W} \triangleq S_W^{-1} S_B$$

Generate m features from n original ones

If the m features chosen optimally above do not produce satisfactory separability, we can try to generate some m new features as the linear combinations of the n old ones by a linear transform:

$$Y = A^T X$$

where A is a $n \times m$ matrix composed of m n-dimensional column vectors A_i:

$$A = [A_1, \cdots, A_m]$$

and Y is an m-dimensional vector whose m elements
$\{y_i = A_i^T X, \quad i = 1, \cdots, m\}$ are the new features.

First we recall that after a linear transform $Y = A^T X$, the mean vectors, the covariance matrices and the various scatter matrices become

$$M_i^{(Y)} = A^T M_i^{(X)} \qquad (i = 1, \cdots, c)$$

$$\Sigma_i^{(Y)} = A^T \Sigma_i^{(X)} A \qquad (i = 1, \cdots, c)$$

and

$$S_W^{(Y)} = A^T S_W^{(X)} A$$
$$S_B^{(Y)} = A^T S_B^{(X)} A$$
$$S_{B/W}^{(Y)} = A^T S_{B/W}^{(X)} A$$

We need to find the optimal matrix A which maximizes $J(A)$ in the m-dimensional feature space spanned by the new features $Y = A^T X$.

Optimal A for maximizing $tr(S_B)$

First we realize that the separability criterion $tr\ (S_B)$ in space $Y = A^T X$ can be expressed as:

$$\begin{aligned} J^{(Y)}(A) &= tr\ (S_B^{(Y)}) = tr\ (A^T\, S_B^{(X)}\, A) = tr\ \begin{bmatrix} A_1^T \\ \cdots \\ A_n^T \end{bmatrix} S_B^{(X)}[A_1, \cdots, A_n] \\ &= tr\ \begin{bmatrix} A_1^T \\ \cdots \\ A_n^T \end{bmatrix} [S_B^{(X)} A_1, \cdots, S_B^{(X)} A_n] = \sum_{i=1}^{n} (A_i^T S_B^{(X)} A_i) \end{aligned}$$

To find A which maximizes $tr(S_B^{(Y)})$ in space $Y = A^T X$, we solve the following optimization problem:

$$\begin{cases} J(A) \stackrel{\triangle}{=} tr(S_B) \rightarrow max \\ subject\ \ to \quad A_j^T A_j = 1 \quad (j = 0, \cdots, n-1) \end{cases}$$

Here we have further assumed that A is an orthogonal matrix (a justifiable constraint as orthogonal matrices conserve energy/information in the signal vector). This constrained optimization problem can be solved by Lagrange multiplier method:

$$\begin{aligned} & \frac{\partial}{\partial A_i}[J(A) - \sum_{j=0}^{m-1} \lambda_j (A_j^T A_j - 1)] = 0 \\ = & \frac{\partial}{\partial A_i}[\sum_{j=1}^{n} (A_j^T S_B^{(X)} A_j - \lambda_j A_j^T A_j + \lambda_j)] \\ = & \frac{\partial}{\partial A_i}[A_i^T S_B^{(X)} A_i - \lambda_i A_i^T A_i] \\ = & 2 S_B^{(X)} A_i - 2\lambda_i A_i = 0 \end{aligned}$$

We see that the column vectors of A must be the orthogonal eigenvectors of the symmetric matrix S_B:

$$S_B A_i = \lambda_i A_i \qquad (i = 1, \cdots, n)$$

i.e., the transform matrix must be

$$A = [A_1, \cdots, A_n] = \Phi = [\phi_1, \cdots, \phi_n]$$

Thus we have proved that the optimal feature selection transform is the principal component transform (KLT) which, as we have shown before, tends to compact most of the energy/information (representing separability here) into a small number of components. Therefore the m new features can be obtained by

$$Y = A_{m\times n}^T X = \begin{bmatrix} \phi_1 \\ \cdots \\ \phi_m \end{bmatrix}_{m\times n} X$$

and

$$J(A) = J(\Phi) = \sum_{i=1}^{m} \phi_i^T S_B \phi_i = \sum_{i=1}^{m} \lambda_i$$

Obviously, to maximize $J(A)$, we just need to choose the m eigenvectors ϕ_i's corresponding to the m largest eigenvalues of S_B:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq \cdots \geq \lambda_n$$

Optimal A for maximizing $tr(S_{B/W})$

To find the transform matrix A which maximizes the criterion

$$J(A) = tr\ [S_{B/W}^{(Y)}]$$

(instead of $tr(S_B)$ as shown above) in the new space $Y = A^T X$, we have to use a different approach from what was used previously. This is because $S_{B/W} = S_W^{-1} S_B$ is not necessarily symmetric, therefore the transform matrix A can no longer be assumed to be orthogonal.

We first simultaneously diagonalize the two scatter matrices S_W and S_B in the original X space. S_W can be diagonalized by its orthogonal eigenvector matrix Φ

$$\Phi^T S_W \Phi = \Lambda$$

where $\Lambda = diag(\lambda_1, \cdots, \lambda_n)$ is the eigenvalue matrix (all λ_i's are real and positive), or

$$\Lambda^{-1/2} \Phi^T S_W \Phi \Lambda^{-1/2} = I$$

Applying the same transform to S_B gives

$$\Lambda^{-1/2} \Phi^T S_B \Phi \Lambda^{-1/2} = K$$

where K is symmetric and can be diagonalized by its orthogonal eigenvector matrix Ψ:

$$\Psi^T K \Psi = \Psi^T \Lambda^{-1/2} \Phi^T S_W \Phi \Lambda^{-1/2} \Psi = \Theta$$

where $\Theta = diag(\theta_1, \cdots, \theta_n)$ is the eigenvalue matrix of K (all $\theta_i' s$ are real and positive).

We now define the transform matrix A as

$$A \stackrel{\triangle}{=} \Phi \Lambda^{-1/2} \Psi$$

(A is not orthogonal as $A^{-1} = \Psi^{-1} \Lambda^{1/2} \Phi^{-1} = \Psi^T \Lambda^{1/2} \Phi^T \neq A^T$) and apply it to X and get

$$Y = A^T X$$

In space Y, both the within-class and between-class scatter matrices are diagonalized:

$$\begin{cases} S_W^{(Y)} = A^T S_W A = I \\ S_B^{(Y)} = A^T S_B A = \Theta \end{cases}$$

and the separability criterion J becomes

$$J(A) = tr\ [S_{B/W}^{(Y)}] = tr\ [(S_W^{(Y)})^{-1} S_B^{(Y)}] = tr\ \Theta = \sum_{i=1}^{m} \theta_i$$

In the original X space, S_W and S_B can now expressed as:

$$\begin{cases} S_W = (A^T)^{-1} A^{-1} \\ S_B = (A^T)^{-1} \Theta A^{-1} \end{cases}$$

and

$$S_{B/W} = S_W^{-1} S_B = A A^T (A^T)^{-1} \Theta A^{-1} = A \Theta A^{-1}$$

i.e.,

$$S_{B/W} A = A \Theta$$

We see that Θ and A are just the eigenvalue and eigenvector matrices of $S_{B/W} = S_W^{-1} S_B$. If only m featues are to be used in space $Y = A^T X$, the criterion $J(A)$ can be maximized by a transform matrix A composed of the m eigenvectors corresponding to the m largest eigenvalues of $S_{B/W}$, i.e.:

$$A_{n \times m} = \Phi \Lambda^{-1/2} [\psi_1, \cdots, \psi_m]$$

where Φ and Λ are respectively the eigenvector and eigenvalue matrices of S_W, and ψ_i is the eigenvector corresponding to the ith largest eigenvalue θ_i of $\Lambda^{-1/2} \Phi^T S_B \Phi \Lambda^{-1/2}$.

Suboptimal feature selection

When the number of features n is large, solving the eigenvalue problem of the $n \times n$ matrix $S_{B/W}^{(X)}$ maybe very time consuming. To compromise, we can use other orthogonal transform such as DFT or WHT instead of KLT for the transform $Y = A^T X$.

Obviously DFT and WHT are not dependent on the feature selection criterion $S_{B/W}^{(X)}$. The reason why they can be used to replace KLT is that orthogonal transforms in general tend to decorrelate signals so that the energy/information (separability information here) is concentrated in a small number of components while others containing little. (However, this energy compaction is suboptimal compared to KLT.) We should choose the m rows of the n by n DFT or WHT matrix corresponding to the m largest $A_i^T S_{B/W}^{(X)} A_i$ values to achieve best feature selection effect.

Information conservation in feature selection

The percentage of separability information (energy) contained in the m-D space after feature selection can be found as

$$\begin{aligned} r &= \frac{\sum_{i=1}^m A_i^T S_{B/W} A_i}{\sum_{i=1}^n A_i^T S_{B/W} A_i} = \frac{\sum_{i=1}^m A_i^T S_{B/W} A_i}{tr\ A S_{B/W} A^T} \\ &= \frac{\sum_{i=1}^m A_i^T S_{B/W} A_i}{tr\ S_{B/W}} = \frac{\sum_{i=1}^m A_i^T S_{B/W} A_i}{\sum_{i=1}^n \lambda_i} \end{aligned}$$

where λ_i's are the eigenvalues of $S_{B/W}$. When KLT is used, the above can be further written as

$$r = \frac{\sum_{i=1}^m \phi_i^T S_{B/W} \phi_i}{\sum_{i=1}^n \lambda_i} = \frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^n \lambda_i}$$

as here $A_i = \phi_i \quad (i = 1, \cdots, m)$ are the eigenvectors of $S_{B/W}$ (corresponding to the m largest eigenvalues $\lambda_i \quad (i = 1, \cdots, m)$).