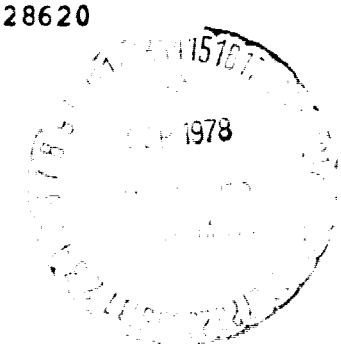
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

FEATURE SELECTION FOR BEST MEAN SQUARE APPROXIMATION OF CLASS DENSITIES

BY CHARLES PETERS

REPORT #71 JULY 1978



PREPARED FOR
EARTH OBSERVATION DIVISION, JSC
UNDER
CONTRACT NAS-9-15543

HOUSTON, TEXAS 77004

FEATURE SELECTION FOR BEST MEAN SQUARE APPROXIMATION OF CLASS DENSITIES

Charles Peters*
Department of Mathematics
University of Houston
Houston, Texas

Report #71

July 1978

* Author partially supported by NASA/Johnson Space Center under contract NAS-9-15000.

FEATURE SELECTION FOR BEST MEAN SQUARE APPROXIMATION OF CLASS DENSITIES

ABSTRACT. A criterion for linear feature selection is proposed which is based on mean square approximation of class density functions. It is shown that for the widest possible class of approximants, the criterion reduces to Devijver's Bayesian distance. For linear approximants the criterion is equivalent to well known generalized Fisher criteria.

Pattern recognition Feature selection

Discriminant analysis Pattern class separability

Feature Selection for Best Mean Square Approximation of Class Densities

1. Introduction

The purpose of this note is to describe a general mean square approach to linear feature selection which connects certain generalized Fisher criteria in discriminant analysis with a measure of pattern class separation introduced by Devijver[3]. The former are typical of those criteria which utilize only low order information about the pattern class distributions, while the latter requires that the class distributions be known, or at least accurately estimated.

Let X denote a random vector in real n-space R^n which arises from one of m pattern classes $\Pi_1, \ldots, \Pi_m$ having known prior probabilities $\alpha_1, \ldots, \alpha_m$, where $\alpha_i > 0$ and $\sum_{i=1}^{m} \alpha_i = 1$. Let $F_j(x)$ denote the $j^{\underline{th}}$ class conditional distribution function of X and let $F(x) = \sum_{i=1}^{m} \alpha_i F_i(x)$ denote the mixture distribution. For a given measureable transformation $T:R^n \to R^k$ let $G_j(y, T)$ and $G(y, T)$ denote, respectively, the $j^{\underline{th}}$ class conditional distribution and mixture distribution of the random variable $Y = TX$. We let $f_j(x)$ (resp. $g_j(y, T)$) denote the class conditional densities of X (resp. Y) with respect to their corresponding mixture distributions; i.e.,

$$f_j = \frac{dF_j}{dF} \quad \text{and} \quad g_j(., T) = \frac{dG_j(., T)}{dG(., T)}.$$

We will restrict our attention to the set of linear transformations T of rank k, and assume that each pattern class Π_i has a mean μ_i and positive definite covariance matrix Ω_i. Let $\mu = \sum_{i=1}^{m} \alpha_i \mu_i$ and let

$$S_B = \sum_{i=1}^{m} \alpha_i (\mu_i - \mu)(\mu_i - \mu)^T$$

$$S_W = \sum_{i=1}^{m} \alpha_i \Omega_i$$

and $$S = S_W + S_B$$

denote the between class scatter matrix, the average within class scatter matrix, and the total scatter matrix respectively.

A number of interesting feature selection criteria can be formulated using only the parameters μ, μ_i, Ω_i, S, S_W, S_B; e.g., the criteria proposed by Kittler and Young [8], Foley and Sammon[4], Fukanaga and Koontz[6], and the discrete analogue of the modified Karhunen-Loeve expansion of Chien and Fu[1]. The modified K-L expansion minimizes an entropy function, and also best represents the pattern vector X in an overall least squares sense; however, its value for discrimination has been questioned by several authors (see Kittler[7]). Fukanaga[5] considers several criteria of the generalized Fisher type, including

$$J_k(T) = \mathrm{tr}(T^T S_W T)^{-1}(T^T S_B T).$$

Thus, according to this criterion, the best $k \times n$ matrix T of rank k is one which maximizes $J_k(T)$. The solution is any T which is row equivalent to a $k \times n$ matrix whose rows are linearly independent principal eigenvectors (i.e., corresponding to the largest eigenvalues) of $S_W^{-1} S_B$. We also consider a modification

$$J_k'(T) = \mathrm{tr}\ (T^T S T)^{-1}(T^T S_B T)$$

which admits the same maximizing T.

The Bayesian distance corresponding to the pattern classes $\Pi_1, \ldots, \Pi_m$, as defined by Devijver[3], is

$$B_n = \sum_{i=1}^{m} \alpha_i^2 \, E\,[f_i(X)^2]$$

$$= \sum_{i=1}^{m} \alpha_i^2 \int_{R^n} f_i(x)^2 \, dF(x).$$

Its transformed value is

$$B_k(T) = \sum_{i=1}^{m} \alpha_i^2 \int_{R^k} g_i(y, T)^2 \, dG(y, T).$$

Devijver proves a number of interesting inequalities relating B_n to the Bayes probability of misclassification, the Bhattacharrya coefficient, and other measures of class separation. In addition, he notes that Cover and Hart[2] have shown that $1-B_n$ is the asymptotic error rate of the nearest neighbor classifier.

2. Mean Square Optimality of Bayesian Distance

For a given $k \times n$ matrix T of rank k, let $L_2(T)$ denote the set of all measureable functions $\varphi : R^k \to R^1$ such that $\int_{R^k} \varphi(y)^2 \, dG(y, T) < \infty$ and let C_T be a given closed linear subspace of $L_2(T)$. Our general approach to linear feature selection is to choose that T, if possible, which minimizes

$$R(T) = \sum_{i=1}^{m} \beta_i \min_{\varphi_i \in C_T} \int_{R^n} [\varphi_i(Tx) - f_i(x)]^2 \, dF(x),$$

where the β_i are positive weights. That is, we attempt to find a T which produces a set of approximations $\varphi_i(Tx)$ to the class densities $f_i(x)$ which is best in an overall mean square sense. Given such approximations we may classify observations of X according to the pseudo-Bayes rule: decide that X is from class Π_i if $\alpha_i \varphi_i(TX) > \alpha_j \varphi_j(TX)$ for each $j \neq i$. Since we are interested in classification accuracy, it seems appropriate to choose weights β_i which reflect the relative importance of the classes in the mixture distribution; e.g., $\beta_i = \alpha_i$ for all i or $\beta_i = \alpha_i^2$ for all i. For the remainder of this section we choose $\beta_i = \alpha_i^2$ and $C_T = L_2(T)$.

Proposition 1: For $\beta_i = \alpha_i^2$, i = 1, ..., m and $C_T = L_2(T)$ for each T,

$$R(T) = B_n - B_k(T).$$

Proof: Observe that $g_i(y, T) \in L_2(T)$, since it is bounded by α_i^{-1}. Moreover, for each $\varphi \in L_2(T)$,

$$\int_{R^n} \varphi(Tx)[g_i(Tx, T) - f_i(x)] \, dF(x)$$

$$= \int_{R^n} \varphi(Tx) \, g_i(Tx, T) \, dF(x) - \int_{R^n} \varphi(Tx) \, dF_i(x)$$

$$= \int_{R^k} \varphi(y) \, g_i(y, T) \, dG(y, T) - \int_{R^k} \varphi(y) \, d \, G_i(y, T)$$

$$= 0.$$

Therefore,

$$\min_{\varphi \in L_2(T)} \int_{R^n} [\varphi(Tx) - f_i(x)]^2 \, dF(x)$$

$$= \int_{R^n} [g_i(Tx, T) - f_i(x)]^2 \, dF(x)$$

$$= \int_{R^n} f_i(x)^2 \, dF(x) - \int_{R^n} g_i(y, T)^2 \, dG(y, T).$$

The assertion of the proposition follows on multiplying by α_i^2 and summing over i.

We may summarize by saying that if there exists a $k \times n$ matrix T_0 of rank k which maximizes $B_k(T)$, then the functions $g_1(T_0x, T_0), \ldots, g_m(T_0x, T_0)$ constitute the best mean square approximation to the class densities $f_1(x), \ldots, f_m(x)$ attainable through a linear compression of the data into k dimensions. Since $B_k(QT) = B_k(T)$ for each nonsingular $k \times k$ matrix Q and each $k \times n$ matrix T of rank k, $B_k(T)$ has a maximum if and only if it has a maximum on the compact set $\{T \mid TT^T = I_{k \times k}\}$. In particular, if $B_k(T)$ is continuous, it has a maximum.

3. Best Linear Approximation of Class Densities

In this section we let C_T be the set of functions $\varphi(y) = w + b^T y$, where w is a real number and $b \in R^k$. For simplicity, we use the notation $\varphi(y) = a^T v(y)$, where $a = \begin{pmatrix} w \\ b \end{pmatrix} \in R^{k+1}$ and $v(y) = \begin{pmatrix} 1 \\ y \end{pmatrix} \in R^{k+1}$. For given T, $a_i = \begin{pmatrix} w_i \\ b_i \end{pmatrix}$ minimizes

$$\int_{R^n} [a^T v(Tx) - f_i(x)]^2 \, dF(x)$$

$$= a^T \left[\int_{R^n} v(Tx)v(Tx)^T \, dF(x)\right]a$$

$$-2a^T \int_{R^n} v(Tx) \, dF_i(x) + \int_{R^n} f_i(x)^2 \, dF(x)$$

$$= a^T \left(\begin{array}{c|c} 1 & \mu^T T^T \\ \hline T\mu & TWT^T \end{array}\right) a \; - 2(1 \mid \mu_i^T T^T)a$$

$$+ \int_{R^n} f_i(x)^2 \, dF(x)$$

if and only if

$$\left(\begin{array}{c|c} 1 & \mu^T T^T \\ \hline T\mu & TWT^T \end{array}\right)\left(\begin{array}{c} w_i \\ \hline b_i \end{array}\right) = \left(\begin{array}{c} 1 \\ \hline T\mu_i \end{array}\right)$$

where $W = E[XX^T] = S + \mu\mu^T$. Solving this system gives

$$w_i = 1 - \mu^T T^T (TST^T)^{-1} T(\mu_i - \mu)$$

and

$$b_i = (TST^T)^{-1} T(\mu_i - \mu).$$

The corresponding squared error of approximation is

$$-1 + (\mu_i - \mu)^T T^T (TST^T)^{-1} T(\mu_i - \mu) + \int_{R^n} f_i(x)^2 \, dF(x).$$

Therefore, the criterion to be minimized is

$$R(T) = -\sum_{i=1}^{m} \beta_i (\mu i - \mu)^T T^T (TST^T)^{-1} T(\mu_i - \mu)$$

+ terms independent of T.

That is, we want to maximize

$$\hat{R}(T) = \text{trace } (TST^T)^{-1} T\hat{S}_B T^T,$$

where

$$\hat{S}_B = \sum_{i=1}^{m} \beta_i (\mu_i - \mu)(\mu_i - \mu)^T.$$

The solution is $T = QT_0$, where T_0 is a $k \times n$ matrix whose rows are linearly independent principal eigenvalues of, $S^{-1}\hat{S}_B$ and Q is an arbitrary nonsingular $k \times k$ matrix. In particular, for $\beta_i = \alpha_i$ we obtain the same solution given by Fukanaga's criterion,

$$\text{trace } (TS_w T^T)^{-1} (TS_B T^T).$$

4. Concluding Remarks

An equivalent set of criteria for feature selection are expressions such as

$$\bar{R}(T) = \sum_{i=1}^{m} \bar{\beta}i \min_{\varphi \epsilon C_T} \int_{R^n} [\varphi(Tx) - \alpha_i f_i(x)]^2 \, dF(x)$$

in which the posterior probabilities $\alpha_i f_i(x)$ of the classes are approximated. If each $\overline{\beta}_i$ is chosen to be 1 $\overline{R}(T)$ is the same as $R(T)$ with $\beta_i = \alpha_i^2$. This, together with Proposition 1 and the relationship between Bayesian distance and the probability of error, seems to indicate that the choice $\beta_i = \alpha_i^2$ is a good one.

In some cases it may be numerically feasible to use $B_k(T)$ as a feature selection criterion when assumptions about the parametric form of the class distributions are made. For example, if each class distribution F_i is multivariate normal, then $B_k(T)$ reduces to an expression which is continuously differentiable in T and which, moreover, can be approximated by sample averages over an unlabeled sample from the mixture distribution. Thus descent algorithms might be successfully employed in maximizing $B_k(T)$.

Finally, we remark that there is no reason in principle why $B_k(T)$ cannot be regarded as a criterion for nonlinear feature extraction. Indeed, Proposition 1 remains true when T is any measureable transformation from R^n onto R^k.

5. Acknowledgements

The author was partially supported by NASA Johnson Space Center under contract NAS-9-15000.

REFERENCES

1. Y. T. Chien and K. S. Fu. On the generalized Karhunen-Loeve expansion. IEEE Trans. IT-15, 518-520, (1967).

2. T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. IEEE Trans. IT-13, 21-27, (1967).

3. P. A. Devijver. On a new class of bounds on Bayes risk in multihypothesis pattern recognition. IEEE Trans. C-23, 70-80, (1974).

4. D. H. Foley and J. W. Sammon, Jr. An optimal set of discriminant vectors. IEEE Trans. C-24, 281-289, (1975).

5. K. Fukanaga. Introduction to Statistical Pattern Recognition, Academic Press, New York, (1972).

6. K. Fukanaga and W. L. G. Koontz. Application of the Karhunen-Loeve expansion to feature selection and ordering. IEEE Trans. C-19, 311-318 (1970).

7. J. Kittler. Mathematical methods of feature selection in pattern recognition. Int. J. Man-Machine Studies, 7, 609-637, (1975).

8. J. Kittler and P. C. Young. A new approach to feature selection based on the Karhunen-Loeve expansion, Pattern Recognition, Vol. 5, 335-352, (1973).