Four Problems of the Discriminant Analysis

Shuichi Shinmura*
Seikei Univ. Faculty of Economics, Tokyo, Japan – shinmura@econ.seikei.ac.jp

Abstract

There are four problems of the discriminant analysis. In this study, the fourth problem is discussed. Fisher formulated the linear discriminant function (Fisher's LDF) but never formulated the SEs of error rate and discriminant coefficient by the traditional approach based on the normal distribution. Both 95% confidence intervals (C.I.) are obtained by the "k-fold cross-validation for small sample" method. Revised IP-OLDF based on the minimum number of misclassifications (minimum NM) compares with another seven LDFs by this method. The mean of error rates of Revised IP-OLDF is the best among eight LDFs in the training and validation samples.

Keywords: k-fold cross-validation for small sample; 95 % confidence interval (C.I.); error rate; linear discriminant coefficient; Revised IP-OLDF; SVM; logistic regression; Fisher's LDF.

1. Introduction

There are four problems of the discriminant analysis. We address the discrimination of two classes having p-independent variables (p-variables) and n cases. Let's f(**x**) be a linear discriminant function (LDF), and y_i be 1 for the cases belong to class1 and -1 for the cases belong to class2. The discriminant rule is very simple as follows: If y_i*f ($\mathbf{x_i}$) > 0, $\mathbf{x_i}$ is classified to class1/class2 correctly. If y_i*f($\mathbf{x_i}$) < 0, $\mathbf{x_i}$ is misclassified. This simplistic scenario may hide four problems. Revised IP-OLDF based on the minimum number of misclassifications (minimum NM) already resolves three problems (Shinmura, 2014a). In this paper, we discuss the discriminant analysis is not the inferential statistics the same as the regression analysis. Fisher (1936) formulated Fisher's LDF but never formulated the standard errors (SEs) of error rate and discriminant coefficient by the traditional approach based on the normal distribution. Both 95% confidence intervals (C.I.) are obtained by the "k-fold cross-validation for small sample" method (Shinmura, 2011b; 2013). This method is the combination of the resampling (Efron, 1979) and k-fold cross-validation. This method is very simple and helpful for the researchers who wish to validate the small samples.

LINGO (Schrage, 2006) codes six mathematical programming based (MP-based) LDFs. Six MP-based LDFs are as follows: Revised IP-OLDF can find true MNM(Shinmura, 2007). Revised IPLP-OLDF finds the approximation of MNM(Shinmura, 2014b). Revised LP-OLDF is one of L1-norm LDF. Vapnik (1995) proposed a hard margin SVM and a soft-margin SVM (SVM1 for penalty c=1 and SVM4 for penalty c=10000). JMP (Sall et al., 2004) codes two statistical LDFs such Fisher's LDF and a logistic regression. We compare the means of error rates of eight LDFs by four data sets. Fisher evaluated his LDF by the iris data. The Swiss bank note data (Flury & Rieduyl, 1988) is a linearly separable data and shows us the second problem. The CPD data has multicollinearity. The student data (Shinmura, 2010) shows us the first problem. Revised IP-OLDF were the best among eight LDFs in the training and validation samples (Shinmura, 2014c). In this study, both C.I. of the Swiss bank note data are examined by this method. This data consists 100 genuine and 100 counterfeit bills having six-variables.

2. Three Problems

Three problems are explained by adding new facts using the Swiss bank note data.

2.1 First Problem

We cannot discriminate between cases where $\mathbf{x_i}$ lies on the discriminant hyperplane (f($\mathbf{x_i}$)=0) correctly. If some case rests on the discriminant hyperplane, the numbers of misclassification (**NM**s) or error rates may not be correct. Many researchers treat these cases belong to class1 without valid reason.

Some statisticians believe it is decided by dice because statistics is the study of probability. Both treatments are not logical. All LDFs except for Revised IP-OLDF formulated in the equation (1) cannot avoid this problem.

$$\text{MIN} = \Sigma e_i\ ;\quad y_i{*}\ ({}^t\mathbf{x_i}\ \mathbf{b}+ b_0) >= 1 - M{*}\ e_i\ ;\qquad i=1,\ldots, n. \qquad (1)$$

e_i: 0/1 integer variable corresponding to $\mathbf{x_i}$. y_i: variable having 1/-1 corresponding to class1/class2. $\mathbf{x_i}$: p-elements vector of ith-case. $\mathbf{b}$: p- discriminant coefficient vector. b_0: free decision variable. M: a big M constant such 10000.

$$\text{MIN} = \Sigma e_i\ ;\ \ y_i{*}\ (\mathbf{x_i}'\mathbf{b}+1) >= - M{*}\ e_i\ ; \qquad (2)$$

IP-OLDF in the equation (2) found the first problem when tested on the student data that was not general position (Shinmura, 1998; 2004). In addition, it finds new facts about the discriminant analysis. First fact is that we can understand the relation of NMs and LDFs on the p-discriminant coefficients space (Shinmura, 2000). N cases correspond to n-linear hyperplanes those divide p-discriminant space into finite convex polyhedron. There are the optimal convex polyhedron (OCP), NMs of which are the MNMs. Only Revised IP-OLDF can find the interior point of OCP and avoids the first problem. If another LDFs look for the vertex or edge of the convex polyhedron, these LDFs may not avoid the first problem. IP-OLDF look for the vertex of OCP if data is general position but cannot avoid the first problem if data is not general position. Second fact is the monotonous decrease of MNM. Let MNM_p be the MNM of p-variables and $\text{MNM}_{(p+1)}$ be the MNM of (p+1)-variables to add one variable to existed p-variables model. If MNM_p=0, all MNN_q including these p-variables are zero. IP-OLDF shows us the Swiss bank note data is linearly separable because MNM_2=0 including 2-variables (X4, X6) (Shinmura, 2007). There are 63 discriminant models. Sixteen models including (X4, X6) are linearly separable models. Another 47 models are not linearly separable models. We had better focused on the linearly separable models if data is linearly separable.

Liittschwager & Wang (1978) proposed LDF based on MNM criterion in equation (3). Let's $f_1 \cdot f_2$ and g_1, g_2 be 1. The object function minimizes the error rate instead of NM in equation (1). If we consider 'b' is the constant and 'C' is a big M constant, two constraints are almost the same as the constraints in equation (1). However Revised IP-OLDF insert '1' in the right hand constant of constraints. In addition, two constraints (3.3) and (3.4) are added.

$$\text{Min}=f_1g_1M^{-1}\Sigma_{(i=1,\ldots,M)}P_i + f_2g_2N^{-1}\Sigma_{(j=1,\ldots,N)}Q_j \qquad (3)$$

$$a_1x_{i1}+a_2x_{i2}+\ldots+a_kx_{ik} \leqq b+CP_i \quad (i=1, 2,\ldots, M) \qquad (3.1)$$

$$a_1y_{j1}+a_2y_{j2}+\ldots+a_ky_{jk} \geqq b\text{-}CQ_i, \quad (j=1, 2,\ldots, N) \qquad (3.2)$$

$$-1+2D_r \leqq a_r \leqq 1\text{-}2E_r, \quad (r=1, 2,\ldots, k) \qquad (3.3)$$

$$\Sigma_{(r=1,\ldots k)}D_r+\Sigma_{(r=1,\ldots,k)}E_r=1 \qquad (3.4)$$

$f_1 \cdot f_2$: risk. g_1, g_2 : prior probability. P_i, Q_j : 0/1 integers for each e_i.

M, N : number of cases in both class. a_i: discriminant coefficients.

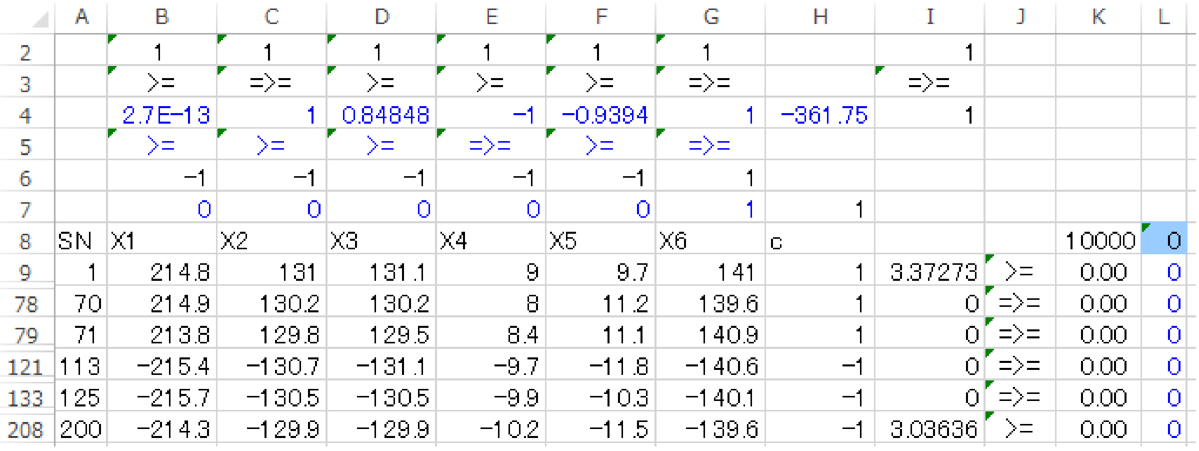
b : discriminant hyperplane. C: positive constant. D_r, E_r : 0/1 integers.

Fig. 1 is the result of Swiss bank note data solved by What's Best!. That is an Excel add-in solver developed by LINDO Systems Inc. Cell ranges of 'B1: G1' and 'B7: G7' correspond to E_r and Q_j , respectively. Cell range of 'B4: G4' corresponds to the discriminant coefficients. Cell H4 is a constant. Cell range 'I2: I4' corresponds to the equation (3.4). Cell ranges of 'B9:H108' and 'B109: H208' correspond to cases of class1 and class2, respectively. Cell ranges of 'I9:I208', 'K9: K208' and 'L9: L208' correspond to the discriminant scores, the values of CP_i/CQ_i and 0/1 integer variables P_i/Q_i, respectively. All 200 P_i/Q_i become zero. Therefore, the object cell 'L8' is zero. However, there are four bills (case No. 70, 71, 113 and 125) on the discriminant hyperplane. Two bills are the genuine bills, and another two bills are the counterfeit bills. We cannot discriminate four bills to the genuine or counterfeit bills. Although this full model is linearly separable, two constraints of the discriminant coefficients summon the first problem. We can conclude about the first problem as follows:

1) If data is not general position, IP-OLDF may not find true MNM.
2) All LDFs except for Revised IP-OLDF cannot avoid the first problem because there is no theoretical guarantee to find the interior point of the convex polyhedron defined by IP-OLDF. This

result means NM of these LDFs may not be correct. Therefore, all statistical software must output the number of cases on the discriminant hyperplane.

3) Because Liittschwager & Wang model restricts the discriminant coefficients, it causes the first problem. If we discriminate the categorical data, the possibility of the first problem may increase.



	A	B	C	D	E	F	G	H	I	J	K	L
2		1	1	1	1	1	1		1			
3		>=	=>=	>=	>=	>=	=>=		=>=			
4		2.7E-13	1	0.84848	-1	-0.9394	1	-361.75	1			
5		>=	>=	>=	=>=	>=	=>=					
6		-1	-1	-1	-1	-1	1					
7		0	0	0	0	0	1	1				
8	SN	X1	X2	X3	X4	X5	X6	c			10000	0
9	1	214.8	131	131.1	9	9.7	141	1	3.37273	>=	0.00	0
78	70	214.9	130.2	130.2	8	11.2	139.6	1	0	=>=	0.00	0
79	71	213.8	129.8	129.5	8.4	11.1	140.9	1	0	=>=	0.00	0
121	113	-215.4	-130.7	-131.1	-9.7	-11.8	-140.6	-1	0	=>=	0.00	0
133	125	-215.7	-130.5	-130.5	-9.9	-10.3	-140.1	-1	0	=>=	0.00	0
208	200	-214.3	-129.9	-129.9	-10.2	-11.5	-139.6	-1	3.03636	>=	0.00	0

Fig. 1 The result of the Swiss bank note data solved by What's Best!

2.2 Second Problem

H-SVM tells us the discrimination of the linearly separable data clearly. However, there is no study about it because of two reasons. H-SVM can recognize the linearly separable data, but it can apply only for the linearly separable data, and S-SVM sometimes cannot recognize the linearly separable data. Therefore, there is few linearly separable data. The pass/fail determination by exam scores has a trivial LDF, MNM of which is zero. The ranges of Fisher's LDF and a quadratic discriminant function (QDF) are [2.2%, 16.7%] and [0.8%, 10.8%], respectively (Shinmura, 2011a).

Some statisticians believe that the purpose of discriminant analysis is to discriminate overlapping cases successfully. However, the definition of overlap is uniquely defined by the condition of MNM > 0 in the world of LDF. All LDFs except H-SVM and Revised IP-OLDF cannot define the overlap theoretically because they cannot discriminate the linearly separable data correctly.

2.3 Third Problem

We assume that cases vary in statistics. If some variable is constant, we cannot compute the inverse matrices of the variance-covariance matrices. Therefore, most statistical packages exclude these variables from the discriminant analysis. However, JMP adopts the generalized inverse matrices and can compute the inverse matrix. However, we found a defect, when the value of variable belonging to only one class is constant, and the value of another class varies. QDF and a regularized discriminant analysis (RDA; Friedman, 1989) misclassify all cases belonging to class2 to class1. After the end of Dec. in 2012, JMP released modified RDA, but QDF enhanced by the generalized inverse matrices cannot discriminate this particular cases correctly until Dec. in 2014. If we add small random noise to the variable with a constant value, the third problem is resolved. There is no need to exclude the constant variable in the discriminant analysis. If the variable X4 of the genuine bills is constant because those are well controlled by the Swiss government, we can quickly confirm this defect if JMP does not resolve this problem.

3. The fourth Problem

In this study, we consider the fourth problem that the discriminant analysis is not the inferential statistics. In section 3.1, we focus on 16 linearly separable models and compare the means of error rates in the training samples (M1s) and validation samples (M2s) of eight LDFs. We propose a new model selection method to find the best model. In section 3.2, we discuss the 95% C.I. of error rates and discriminant coefficients.

3.1. The Mean of Error Rates

Swiss bank note data consists of two kinds of bills: 100 genuine and 100 counterfeit bills. We investigate a total of 63 (=2^6-1) models by six variables. We had better considered about two types of discriminations: 16 linearly separable models and other 47 models. **Table 1** shows only 16 linear separable models. There are mistakes in the paper (Shinmura, 2014c). The “M1 and M2” are the means of error rates in the training and validation samples by the “100-fold cross-validation” method. Only four M2s of Revised IP-OLDF are zero in the third and fourth columns. In this case, we had better chosen the minimum number of variables such (1, 4, 5, 6) by the principle of parsimony. We compare Revised IP-OLDF with another seven LDFs in this best model. Results of H-SVM are the same as SVM4. First rows of each LDFs in sixth and seventh columns are the ranges of 16 models, and second rows are the results of the best model. All models of seven LDFs have the minimum M2. Those values are 0.38 (SVM4), 0.38 (H-SVM), 0.52, 0.27, 0.41, 0.38 and 0.47%, respectively. All M2s are within 0.52%. However, SVM1 and Fisher’s LDF cannot recognize linear separable data because all M1s of both LDFs are not zero.

Table 1. The M1s and M2s of eight LDFs (Shinmura, 2014c)

RIP	Model	M1	M2		M1	M2
1	1-6	0	0	SVM4	0	[0.38,0.71]
2	2-6	0	0.24	/HSVM	0	0.38
3	1,3-6	0	0	SVM1	[0.24,0.57]	[0.52,0.87]
4	1,2,4-6	0	0		0.27	0.52
5	1-4,6	0	0.1	LP	0	[0.27,0.77]
8	3-6	0	0.21		0	0.27
9	2,4-6	0	0.16	IPLP	0	[0.41,0.85]
10	1,4-6	0	0		0	0.41
11	2-4,6	0	0.03	Logistic	0	[0.38,0.73]
12	1,3,4,6	0	0.08		0	0.38
13	1,2,4,6	0	0.1	LDF	[0.44,0.95]	[0.47,1.03]
23	4-6	0	0.15		0.44	0.47
24	3,4,6	0	0.02			
25	1,4,6	0	0.02			
26	2,4,6	0	0.03			
41	4,6	0	0.01			

3.2. The 95% Confidence Intervals of Error Rates and Discriminant Coefficients

We judge the model (X1, X4, X5, X6) as the best model among 16 linearly separable models and compare the 95% C.I. of error rates by this model. **Table 2** shows the four percentile such 0% (minimum), 2.5%, 97.5% and 100% (maximum). The result of Revised IP-OLDF is the best because both 100 error rates of the training and validation samples are zero. The 100 percentile of Fisher’s LDF is less than another six LDFs. Although SVM4, LP, IPLP, and logistic can recognize linear separable models, the results of these LDFs are worse in the validation samples.

Table 2. The 95% C.I. of four linearly separable models.

	Training				Validation			
	0%	2.5%	97.5%	100%	0%	2.5%	97.5%	100%
RIP	0	0	0	0	0	0	0	0
HSVM	0	0	2	2.5	0.58	0.58	1.65	2.98
SVM4	0	0	0	0	0	0	1.87	2.38
SVM1	0	0	1	1	0	0	1.63	2.38

LP	0	0	0	0	0	0	1.86	2.40
IPLP	0	0	0	0	0	0	1.91	2.40
logistic	0	0	0	0	0	0	1.74	2.50
	0	0	1.5	2	0	0	0.5	0.50

Table 3 shows the 95% C.I. of discriminant coefficients by six MP-based LDFs. Fisher's LDF and logistic regression by JMP script do not output 100 discriminant coefficients. If the 95% C.I. includes zero, we judge the pseudo-population discriminant coefficient is zero. The coefficients of X1 and constant, those of SVM4, SVM1, and Revised IPLP-OLDF, are zero by this standard.

Table 3. The 95% C.I. of six LDFs.

		0%	2.5%	50%	97.5%	100%		0%	2.5%	50%	97.5%	100%
RIP	1	0.78	0.91	1.02	1.02	1.02	IPLP	-2.01	-1.34	1.04	2.89	2.89
	4	1.6	2.32	2.97	2.97	2.97		0.14	0.35	1.89	3.71	3.97
	5	2.31	2.67	3	3	3		0.55	0.55	1.81	3.9	4.43
	6	-1.99	-1.99	-1.99	-1.75	-1.48		-2.76	-2.43	-1.72	-0.61	-0.44
	c	0	0	0	0	0		-385.78	-359.05	0	438.13	539.72
HSVM	1	0.07	0.23	1.04	1.43	1.61	LP	0.07	0.23	1.04	1.43	1.61
	4	0.46	0.88	2.18	3.12	3.27		0.46	0.88	2.19	3.12	3.27
	5	0.31	0.51	2.28	4.23	5.12		0.31	0.51	2.28	4.07	4.43
	6	-2.7	-4.3	-2.06	-0.61	-0.44		-2.7	-2.43	-1.84	-0.61	-0.44
	c	0	0	0	0	165.5		0	0	0	0	0
SVM4	1	-0.68	-0.64	1.19	2.18	2.18	SVM1	-0.62	-0.34	0.18	0.86	0.9
	4	0.31	0.42	1.97	2.74	2.81		0.31	0.41	1.03	1.58	1.73
	5	0.6	0.63	1.76	2.96	3.09		0.5	0.6	0.9	1.46	1.57
	6	-2.54	-2.32	-1.73	-0.69	-0.49		-1.8	-1.77	-1.29	-0.75	-0.62
	c	-397	-257	-42.5	281.3	303.4		-72.69	-43.3	115.37	231.73	303.98

If we treat y_i as the object variable of the regression analysis, the obtained regression coefficients in equation (4) are proportional to the discriminant coefficient by the plug-in rule. The numbers in parentheses are SEs. The coefficient of X1 of Fisher's LDF is zero. JMP's logistic regression offers the 95% C.I. of logistic regression coefficients in equation (5). Those are obtained in numerical calculation by the maximum likelihood estimation. The numbers in parentheses are SE calculated by Hessian matrix. The values of SEs are enormous, and all confidence intervals include zero. Therefore, JMP outputs a warning message such the "estimation is unstable" for the linearly separable model (Firth, 1993). However, if we find "NM=0" on the ROC by JMP and "MNM=0" by Revised IP-OLDF, we judge it is the linearly separable model. In general, an exact logistic regression supported by SAS is recommended in order to avoid this complex work.

$$\mathrm{LDF}=-0.02(0.05)X1-0.31(0.02)X4-0.33(0.03)X5+0.41(0.03)X6-46.32(12.18). \qquad (4)$$

$$\mathrm{Logi}= 28.03(19036)*X1+ 37.3(6895)*X4+49.16(10354)*X5-29.49(82)*X6-2771(3938259). \qquad (5)$$

4. Conclusions

In this study, we discuss the fourth problem of discriminant analysis. Fisher never formulated the SEs of error rate and discriminant coefficient. However, some statisticians believe the discriminant analysis is the inferential statistics similar to the regression analysis because Fisher's LDF assumes the Fisher's assumption based on the normal distribution. This claim is not logical. Statistical software reflects the common knowledge obtained by statistical study. A statistical user can infer that the discriminant analysis is not the same as traditional inferential statistics. There is a study about the error rate by the bootstrap method (Konishi & Honda, 1992) by the computer-intensive approach. In this study, we propose the "k-fold cross-validation for small sample" method. This method can resolves the fourth problem and evaluates eight LDFs by the mean error rates in the training and validation samples. The procedure of this method is very simple as follows. We generate pseudo-population or resampling sample from the research data. The sub-samples are used as training samples and pseudo-population is used as a validation sample. There are several merits as follows:

1) It is easy to generate the resampling sample by statistical software.
2) It reflects the relation of pseudo-population and samples. The training sample should be a sub-set of pseudo-population.
3) This method shows good results as explained in this paper.

We can conclude that at least this method is better than LOO method (Lachenbruch & Mickey, 1968).

References

Efron, B. (1979). Bootstrap Methods -Another Look at the Jackknife-. The Annals of Statics, **7**/1, 1–26.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. Biometrika, **80**, 27-39.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, **7**, 179–188.

Flury, B., & Rieduyl, H. (1988). Multivariate Statistics: A Practical Approach. Cambridge University Press.

Friedman, J. H. (1989). Regularized Discriminant Analysis. Journal of the American Statistical Association, **84**/405, 165-175.

Konishi, S., & Honda, M. (1992). Bootstrap Methods for Error Rate Estimation in Discriminant Analysis. Japanese Society of Applied Statistics, **21**/2, 67-100.

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. Technometrics **10**, 1-11.

Liittschwager, J.M., & Wang, C. (1978). Integer programming solution of a classification problem. Management Science, **24**/14, 1515-1525.

Sall, J. P., Creighton, L., & Lehman, A. (2004). JMP Start Statistics, Third Edition. SAS Institute Inc.

Schrage, L. (2006). Optimization Modeling with LINGO. LINDO Systems Inc.

Shinmura, S. (1998). Optimal Linear Discrimrnant Functions using Mathematical Programming. Journal of the Japanese Society of Computer Statistics, **11** / 2, 89-101.

Shinmura, S. (2000). A new algorithm of the linear discriminant function using integer programming. New Trends in Probability and Statistics, **5**, 133-142.

Shinmura, S. (2004). New Algorithm of Discriminant Analysis using Integer Programming. IPSI 2004 Pescara VIP Conference CD-ROM, 1-18.

Shinmura, S. (2007). Overviews of Discriminant Function by Mathematical Programming. Journal of the Japanese Society of Computer Statistics, **20**/1-2, 59-94.

Shinmura, S. (2010). The optimal linear discriminant function. Union of Japanese Scientist and Engineer Publishing.

Shinmura, S. (2011a). Problems of Discriminant Analysis by Mark Sense Test Data. Japanese Society of Applied Statistics, **40**/3, 157-172.

Shinmura, S. (2011b). Beyond Fisher's Linear Discriminant Analysis - New World of Discriminant Analysis -. ISI CD-ROM, 1-6.

Shinmura, S. (2013). Evaluation of Optimal Linear Discriminant Function by 100-fold cross-validation. 2013 ISI CD-ROM, 1-6.

Shinmura, S. (2014a). End of Discriminant Functions based on Variance Covariance Matrices. ICORES, 5-14, 2014.

Shinmura, S. (2014b). Improvement of CPU time of Linear Discriminant Functions based on MNM criterion by IP. Statistics, Optimization and Information Computing, vol. **2**, 14-129.

Shinmura, S., (2014c). Comparison of Linear Discriminant Function by K-fold Cross-validation. Data Analytic 2014, 1-6.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag.