Evaluation of Optimal Linear Discriminant Function by 100-fold cross validation

Shuichi Shinmura
Faculty of Economics, Seikei Univ.
3-3-1 Kichijoji-kitamachi, Musashino-shi, Tokyo 180-8633, Japan
E-mail: Shinmura@econ.seikei.ac.jp

Abstract

I develop new linear discriminant function called 'Revised IP-OLDF' based on MNM criterion. It is compared with Fisher's linear discriminant function, quadratic discriminant function, logistic regression and soft margin SVM by 100 fold cross-validation. One hundred re-sampling data sets are generated from four kinds of original data such as: Fisher's Iris data (15 models), Swiss bank note data (16 models), CPD (Cephalo Pelvis Disproportion) data (19 models) and student data (31 models). The mean of error rates of 81 models of these methods are computed by LINGO and JMP. It is concluded that Revised IP-OLDF is better than other methods. In addition to these results, LDF and QDF never recognize linear separable data.

Keywords: Linear Discriminant Function, Logistic Regression, Minimum Number of Misclassifications, Quadratic Discriminant Function, Revised IP-OLDF, Soft margin SVM

1. Introduction

In this paper, Optimal Linear Discriminant Functions based on **MNM** (Minimum Number of Misclassifications) are proposed. **IP-OLDF** looks for the vertex of Optimal Convex defined on the discriminant coefficient space if data is in general position. The number of misclassifications (NM) of the interior point of Optimal Convex equals to MNM. It may not find the vertex of Optimal Convex if data isn't in general position. **Revised IP-OLDF** looks for the interior point of true Optimal Convex directly. Linear discriminant functions corresponding to the interior point discriminates or misclassifies same cases, and there is no cases on the discriminant hyper-plane $f(\mathbf{x}_i)=0$. On the other hand, all discriminant functions except for Revised IP-OLDF can't count NM correctly, because there may be the cases on $f(\mathbf{x}_i)=0$. Shinmura (2010) shows that IP-OLDF has solved many problems of the discriminant analysis, and gives us new knowledge such as "monotonous decrease of MNM". And Revised IPLP-OLDF is compared with Fisher's linear discriminant function (LDF) and logistic regression by 100 fold cross validation (Shinmura, 2011).

The aim of this research is to show the following three points.

1) Revised IP-OLDF is compared with four methods (LDF, QDF, the logistic regression and S-SVM) by four kinds of real data such as Fisher's iris data, CPD (Cephalo Pelvis Disproportion) data,

Swiss bank note data and student test data.

2) Next, 100 bootstrap samples are generated from above real data. Revised IP-OLDF and three methods (LDF, the logistic regression and S-SVM) are compared with the mean error rates of 81 different discriminant models by 100 fold cross validations. The mean error rates of Revised IP-OLDF are almost less than those of three methods.
3) Revised IP-OLDF and four methods are compared with the pass/fail determination of six examinations, which are trivial linear separable data such as Swiss bank note data. It is concluded that LDF and QDF and S-SVM (if penalty c is small such as c=1) can't recognize the linear separable data. S-SVM (if penalty c is large number such as c=10000) and logistic regression can almost recognize the linear separable data.

2. Many Problems of discriminant analysis

Discriminant Rule is very simple as follows: If $y_i * f(\boldsymbol{x_i}) > 0$, $\boldsymbol{x_i}$ is classified to class1/class2 correctly. If $y_i * f(\boldsymbol{x_i}) < 0$, $\boldsymbol{x_i}$ is miss-classified. There are many unresolved problems hidden in this simplicity. These problems are resolved by IP-OLDF and Revised IP-OLDF.

1) LDF and QDF based on variance-covariance matrices can't recognize the linear separable data (MNM=0), therefore these methods should not be used in pattern recognition, medical diagnosis, genome diagnosis, and the pass/fail of exams that is trivial linear separable data. In addition to this, these methods of SPSS and JMP can't work properly for the data if some variables belonging to one class /both classes have constant values.
2) All methods except for Revised IP-OLDF can't count error rates correctly, because these can't prevent the case $\boldsymbol{x_i}$ on the discriminant hyper-plane ($f(\boldsymbol{x_i})=0$) .
3) IP-OLDF explains the relation of the number of misclassifications (NM) and the discriminant coefficients. And MNM decreases monotonously.

3. Discriminant Functions

After 1997, several new methods are developed such as **IP-OLDF** and **Revised IP-OLDF**, **LP-OLDF** and **Revised IPLP-OLDF**. In this paper, Revised IP-OLDF is compared with S-SVM, LDF, QDF and logistic regression by four kinds of real data as training data. Next, Revised IP-OLDF is compared with S-SVM, LDF, and logistic regression by 100 fold cross validation. Revised IP-OLDF and S-SVM are solved by LINGO that is mathematical programming solver (Schrage, 2003), Shinmura(2010)). LDF, QDF and logistic regression are solved by JMP (Sall et.al. (2004)).

3.1 Optimal Linear Discriminant Functions

IP-OLDF is defined in formula (3.1). If $\boldsymbol{x_i}$ is classified correctly, $e_i=0$ and $y_i * f_i(\boldsymbol{b}) = y_i * (\boldsymbol{x_i}' \boldsymbol{b}+1) >= 0$. If $\boldsymbol{x_i}$ is misclassified, $e_i = 1$ and $y_i * f_i(\boldsymbol{b}) >= -999999$. This means that IP-OLDF choose the discriminant hyper-plane $f_i(\boldsymbol{b})=0$ for classified cases, and $f_i(\boldsymbol{b}) = -999999$ for misclassified cases by 0/1 decision variable. If e_i are non-negative real variables, it is changed to LP-OLDF that is one of

L_1-norm linear discriminant functions. Its computational time is faster than IP-OLDF.

$$MIN = \Sigma e_i ; \quad y_i * (\boldsymbol{x_i}' \boldsymbol{b} + 1) >= - M * e_i ; \tag{3.1}$$

xi : p-independent variables, b : p-discriminant coefficients,yi = 1 / -1 for xi ∈ class1/class2,
ei : 0/1 decision variable , M: 1000,000 (Big constant)

Revise IP-OLDF in formula (3.2) can find true MNM. This means there is no cases on $f_i(\boldsymbol{b})=0$. If e_i are non-negative real variables, it is changed to Revised LP-OLDF.

$$MIN = \Sigma e_i ; \quad y_i * (\boldsymbol{x_i}' \boldsymbol{b} + b_0) >= 1 - M * e_i ; \tag{3.2}$$

b_0: free decision variables

3.2 H-SVM and S-SVM

S-SVM is defined in formula (3.3) . There is no rule to decide c properly, and different c gives us different results.

$$MIN = ||\boldsymbol{b}||^2/2 + c * \Sigma e_i ; \quad y_i * (\boldsymbol{x_i}' \boldsymbol{b} + b_0) >= 1 - e_i ; \tag{3.3}$$

c : penalty c to combine two objectives

3.3 Statistical discriminant functions

Fisher defines LDF to maximize the ratio of (between classes/within class) in formula (3.4). It is solved by non-linear programming (NLP). If within class variance is fixed to 1, it is solved by quadratic programming (QP) in (3.5). NLP can obtain the local solution before 2000, and can obtain the global solution (MIN/MAX) after 2000. H-SVM and S-SVM are defined by QP instead of NLP, also.

$$\text{MIN}=\boldsymbol{b}'(\mathbf{x}_{m1}-\mathbf{x}_{m2})\,(\mathbf{x}_{m1}-\mathbf{x}_{m2})'\boldsymbol{b}/\boldsymbol{b}'\textstyle\sum\boldsymbol{b}; \tag{3.4}$$

$$\text{MIN}=\boldsymbol{b}'(\mathbf{x}_{m1}-\mathbf{x}_{m2})\,(\mathbf{x}_{m1}-\mathbf{x}_{m2})'\boldsymbol{b}; \quad \boldsymbol{b}'\textstyle\sum\boldsymbol{b}=1; \tag{3.5}$$

If we accept Fisher's assumption, the same formula of LDF is obtained in (3.6). This formula define the formula of LDF explicitly, nevertheless formula (3.5) define the formula of LDF implicitly. Therefore, statistical software packages adopt this formula. Discriminant analysis is independent of inferential statistics. Therefore, the leave one out (**LOO**) method is proposed to decide the proper discriminant model. In addition to LOO method, we can use model selection methods of regression analysis, if objective values of two groups are 1/-1 dummy variables.

$$\text{LDF}: f(\mathbf{x})=\{\mathbf{x}-(\mathbf{m}_1+\mathbf{m}_2)/2\}'\textstyle\sum^{-1}(\mathbf{m}_1-\mathbf{m}_2) \tag{3.6}$$

Most of real data doesn't satisfy Fisher's assumption. If most of real data doesn't satisfy variance covariance of two classes are not same ($\sum_1 \neq \sum_2$)**.** In this case, QDF is formulated in (3.7).

$$f(\mathbf{x})=\mathbf{x}'(\textstyle\sum_2^{-1}-\sum_1^{-1})\,\mathbf{x}/2+(\mathbf{m}_1'\sum_1^{-1}-\mathbf{m}_2\sum_2^{-1})\mathbf{x}+c \tag{3.7}$$

Mahalanobis distance in (3.8) is used for the discrimination of many classes and MT (Mahalanobis Taguchi) method in quality control.

$$D=\text{SQRT}\,((\mathbf{x}-\mathbf{m})'\textstyle\sum^{-1}(\mathbf{x}-\mathbf{m})) \tag{3.8}$$

These discriminant functions are applied for many areas such as pattern recognition, medical

diagnosis, genome diagnosis, and the pass/fail of exams that is trivial linear separable data. But these discriminant functions are not calculated if some independent variables are constant. There are three cases. First, some variables that belong in both classes are the same constant. Second, some variables that belong in both classes are the different constant. Third, some variable that belong in one class is constant. SPSS excludes all variables in three cases. QDF of JMP outputs wrong discriminant result in case three, because it doesn't assume this case.

4. Experimental Study by the original data and bootstrap data sets

In this study, four kinds of real data are used for evaluation. In first stage, these data are used to examine the validity of new methods. These methods are compared with S-SVM, LDF, QDF and logistic regression. In second stage, 100 bootstrap samples are generated from real data. And 81 different discriminant models of Revised IP-OLDF, S-SVM, LDF and logistic regression are evaluated by 100 fold cross validation.

4.1 Four kinds of Real data

Iris data (Edgar,1935) consists of 100 cases having four independent variables. Object variable consists of two species such as 50 versicolor and 50 virginica. All combinations of independent variables ($15= 2^4-1$) are investigated. This data is used for the evaluation of LDF. Swiss bank notes data (Flury & Rieduyl, 1988) consists of 200 cases having six independent variables. Object variable consists of two kinds of bills such as 100 genuine and 100 counterfeit bills. There are 16 different models for experimental sturdy. Student data consists of 40 students having five independent variables. Object variable consists of two groups such as 25 students who pass the examination and 15 students who don't pass. All combinations of independent variables ($31= 2^5-1$) are investigated. CPD data consists of 240 patients having 19 independent variables. Object variable consists of two groups such as 180 pregnant women whose babies are born by the natural delivery and 60 pregnant women whose babies are born by Caesarian operation. Nineteen models selected by forward stepwise method are analyzed, because we can't examine ($2^{19}-1$) models by all combinations of independent variables. There are three multi-collinearities in this data.

4.2 100 re-sampling data sets and 100 fold cross validation

From above real data sets, 100 re-sampling data sets are generated as follows. 1) JMP copies 100 data sets and adds the variable that have the uniform random number, 2) sort it by object variable (1/-1) and divides it into 100 data sets, 3) 100 fold cross validation are done using these 100 data sets.

5. Results

5.1 Original Data

Table 1 shows the result of Iris data. "VAR" is independent variables. "p" is the number of independent variables. There are 15 models in all combination of independent variables. AIC, BIC and Cp statistics are obtained by regression analysis. These statistics recommend full mode. LOO is the

NM of LOO method by SPSS. This recommend model 1 and model 3. pLDF, pQDF and pLogi are NM of LDF, QDF and logistic regression. MNM is MNM by Revised IP-OLDF. It recommend full model. SVM100 is NM of S-SVM in the case C=100. LDF, QDF logistic regression and S-SVM can't exclude the cases on $f(\mathbf{x}_i)=0$. If we can count the number of the case on $f(\mathbf{x}_i)=0$ and this number is not zero, we had better estimate NM as (NM in the table + this number). Error rates of LDF, QDF, logistic regression and S-SVM are higher 2%, 2%, 1% and 2%, because sample size are 100.

Results of other three real data show that Revised IP-OLDF is better than four methods.

Table 1. Result of Iris data

Model	p	VAR	AIC	BIC	Cp	LOO	pLDF	pQDF	pLogi	MNM	SVM100
1	4	X1-X4	143	158	5	3	3	3	2	1	3
2	3	X2-X4	149	161	10	4	4	4	2	2	3
3	3	X1,X3,X4	152	164	14	3	3	3	2	2	3
4	3	X1,X2,X4	164	176	27	5	5	6	4	4	7
5	3	X1-X3	174	187	40	7	7	8	4	2	5

5.2 100 fold cross validation

One hundred bootstrap samples are generated from real data. These data sets have the same size (number of cases and variables) of real data. Eighty one different discriminant models of LDF, logistic regression, S-SVM and Revised IP-OLDF are compared by 100 fold cross-validations. There are 100-NM and discriminant functions for 81 different models. Mean error rates, and 95% confidence intervals of error rates and discriminant coefficients are calculated. **Table2** shows the difference between mean of error rates in three methods and Revised IP-OLDF.

Table2. The difference between mean of error rates in three methods and Revised IP-OLDF

	LDF - MNM		Logi - MNM		SVM - MNM	
	Training	Evaluation	Training	Evaluation	Training	Evaluation
	Min/Max	Min/Max	Min /Max	Min/Max	Min /Max	Min/Max
Iris(15)	0.9/12.94	-0.0(1)/5.56	0.7/10.1	0.4/8.2	0.9/12.94	-0.0(1)/5.56
Bank(16)	0.5/0.98	-1.3(2)/1.03	0/0	-1.3(2)/0.68	0.6/1.49	-1.1(1)/1.37
Student(31)	1.2/8.61	-1.6(3)/6.77	-2.4 (2)/6.63	-3.1(7)/5.66	-0.6 (1)/15.4	-5.6(12)/2.22
CPD(19)	3.1/7.31	1.9/6.05	0.12/2.90	0/1.63	-0.9 (2)/2.57	-0.4(4)/1.68

First two columns are the comparison of the difference between mean of error rates in LDF and Revised IP-OLDF for the training and evaluation samples. Minimum value of Iris bootstrap samples (15 different discriminant models) is 0.9%. This means that the mean error rate of LDF is 0.9% greater than those of Revised IP-OLDF. Therefore, all 15 discriminant models of LDF are inferior to those of Revised IP-OLDF. Maximum value is 12.94%, nevertheless Iris data is considered to satisfy Fisher's

hypothesis. Swiss bank note data is linear separable data for the two variable model including (X4, X6). If MNM_p shows MNM for p variables model, $MNM_{(P+1)}$ to add one variable in the model is always small ($MNM_p >= MNM_{(P+1)}$). Therefore, all 16 models including (X4, X6) are linear separable. Minimum value of Swiss bank note bootstrap samples (16 linear separable models) is 0.5%. Maximum value is 0.98%. The difference between LDF and Revised IP-OLDF is very small, because these data are linear separable. Minimum/maximum values of student bootstrap samples (31 models) are 1.2/8.61%. Minimum/ maximum values of CPD bootstrap samples (19 models selected by forward stepwise) are 3.1/7.31%. The difference is big, because there are multi- collinearities in CPD data. All 81 models of LDF are inferior to those of Revised IP-OLDF. On the other hand, only 6 models of LDF are superior to those of Revised IP-OLDF in the case of evaluation samples. Only 2 and 9 models of logistic regression are superior to those of Revised IP-OLDF in the case of training and evaluation samples. Only 3 and 18 models of S-SVM (C=10000) are superior to those of Revised IP-OLDF in the case of training and evaluation samples. P

6. Conclusion

Revised IP-OLDF can recognize MNM=0 data correctly, and can avoid the cases on $f(\mathbf{x}_i)=0$. MNM is the lower limit of NM for the training bootstrap samples. Two models of (Logi-MNM) and three models of (SVM-MNM) in Table2 are negative. This result shows logistic regression and S-SVM can't avoid the cases on $f(\mathbf{x}_i)=0$. The mean error rates of Revised IP-OLDF are better than LDF, logistic regression and S-SVM.

REFERENCES (RÉFERENCES)

Edgar, A. (1935). The irises of the Gaspé Peninsula. Bulletin of the American Iris Society, **59**, 2–5.

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

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

Shinmura A,S. (2011). Beyond Fisher's Linear Discriminant Analysis- New World of Discriminant Analysis -. ISI proceedings, 1-10.

Sall, J. P., Creighton, L. and Lehman, A (2004): JMP Start Statistics, Third Edition. SAS Institute Inc. (Japanese version is edited by Shinmura, S.)

Schrage, L.(2003). Optimizer Modeling with LINGO. LINDO SYSTEMS Inc., Chicago.

Shinmura, S. (2010). Optimal Linear Discriminant Function. Union of Japanese Scientists and Engineers Publishing, Tokyo.