Chapter 1

New Theory of Discriminant Analysis

Shuichi Shinmura

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

Abstract. In this book, we explain a new theory of discriminant analysis after R. Fisher. There are five serious problems of the discriminant analysis. We solve these problems by four mathematical programming based linear discriminant functions (MP-based LDFs) completely. At first, we develop an optimal LDF using integer programming (IP-OLDF) based on a minimum number of misclassifications (minimum NM, MNM) criterion. We consider discriminating the data with n-cases by p-variables. The case $x_i = (x_{i1} + \ldots + x_{ip})$ is p-vector ($i = 1, \ldots, n$). Because we formulate it on the p-dimensional discriminant coefficient space $b$, n-linear hyperplanes ($x^*b + 1 = 0$) divide the coefficient space into finite convex polyhedron (CP). LDF corresponding to an interior point of CP misclassify same k-cases, and it reveals the relation of NM and discriminant coefficient clearly. Because there are finite CPs on the discriminant coefficient space, we had better chosen the interior point of CP with MNM. We call this CP as an optimal CP (OCP). MNM decreases monotonously (MNM_{k} >= MNM_{k+1}). Therefore, if MNM_{k} = 0, all MNMs of models including these k-variables are zero. If data is a general position, IP-OLDF looks for the vertex of true OCP. However, if data is not general position such as student data, IP-OLDF may not look for the vertex of true OCP. Therefore, we develop Revised IP-OLDF that looks for the interior point of true OCP directly. If LDF corresponds to the vertex or edge of CP, there are over p-cases on the discriminant hyperplane and LDF cannot discriminate these cases correctly (Problem 1). This fact means NM may not be true. Only Revised IP-OLDF is free from Problem 1. When we discriminate Swiss banknote data having six variables, MNM of the two-variables model (X4, X6) is zero. Therefore, sixteen models including (X4, X6) are zero, and forty-seven models are not linearly separable. Although a hard-margin SVM (H-SVM) tells us a linearly separable data (LSD) clearly, there are few types of research about the discrimination of LSD. Most statisticians misunderstand the purpose of the discrimination is to discriminate the overlapping data, not LSD. All LDFs except for H-SVM and Revised IP-OLDF may not discriminate LSD correctly (Problem 2). Moreover, those can not decide whether the data is overlap or LSD because MNM = 0 means LSD and MNM > 0 means overlapping. We show Fisher’s LDF and a quadratic discriminant function (QDF) cannot judge the pass/fail determination using exam scores and eighteen error rates of both LDFs are very high. We explain the defect of the generalized inverse matrices technique and QDF misclassify all cases of class1 to class2 for a particular case (Problem 3). Fisher never formulated the equation of standard error (SE) of error rate and discriminant coefficient (Problem 4). We solve Problem 4 by the k-fold cross-validation for small sample method (Method 1). It offers the mean of error rates, M1 and M2, in the training and validation samples in addition to the 95% confidence interval (CI) of error rate and coefficient. We propose the simple and powerful model selection procedure to choose the best model with minimum M2 instead of the leave-one-out (LOO) procedure. The best models of Revised IP-OLDF are better than another seven LDFs. For more than 10 years, many researchers have been struggling to analyze microarray data that is LSD (Problem 5). Only Revised IP-OLDF can make feature selection naturally. We develop the Matroska feature selection method (Method 2). It finds the surprising structure of microarray data that is the union of several Matroska those are linearly separable models. From now on, we can analyze microarray data very quickly.

Keywords: Fisher’s LDF, logistic regression, RDA, QDA, the minimum number of misclassifications (MNM), Revised IP-OLDF, Revised LP-OLDF, Revised IPLP-OLDF, three SVMs, k-fold cross-validation for small sample method (Method 1), Matroska feature selection method for microarray data (Method 2), LASSO.

1.1 Introduction

1.1.1 Theme of Theory

1.1.2 Five Problems

1.1.2.1 Problem 1
1.1.2.2  Problem 2

1.1.2.3  Problem 3

1.1.2.4  Problem 4

1.1.2.5  Problem 5

1.1.2.6  Summary

1.2  Motivation for Our Research

1.2.1  Contribution by Fisher

1.2.2  The defect of Fisher’s assumption for Medical Diagnosis

1.2.3  Research Outlook

1.2.4  Method 1 and Problem 4

1.3  Discriminant Functions

1.3.1  Statistical Discriminant Functions

1.3.2  Before and After SVM

1.3.3  IP-OLDF and Four New Facts of Discriminant Analysis

1.3.4  Revised IP-OLDF, Revised LP-OLDF and Revised IPLP-OLDF

1.4  The Unresolved Problem (Problem 1)

1.4.1  Perception Gap of Problem 1

1.4.2  Student data

1.5  The Discrimination of LSD (Problem 2)

1.5.1  Importance of this problem

1.5.2  Pass/Fail Determination

1.5.3  Discrimination by Four Testlets

1.6  Generalized Inverse Matrix (Problem 3)

1.7  K-Fold Cross-validation (Problem 4)
1.7.1 100-Fold Cross-validation

1.7.2 LOO and K-Fold Cross-validation

1.8 Matroska Feature-Selection Method (Problem 5)

1.9 Summary

Reference


[65] Shinmura S (2015j) Comparison Fisher’s LDF by JMP and Revised IP-OLDF by LINGO for Microarray Data (6). Research Gate (6), Nov.11, 2015: 1-10


