Chapter 7

Japanese Automobile Data

- Explanation 2 of Matryoshka Feature-Selection Method (Method 2) -

Shuichi Shinmura

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

Abstract. The Japanese automobile data consists twenty-nine regular cars and fifteen small cars with six independent variables such as the emission rate ($X_1$), the price ($X_2$), the number of seats ($X_3$), CO2($X_4$), Fuel ($X_4$) and Sales ($X_6$). The following reasons are important for this book.

1) The discrimination of LSD
   We can easily recognize this data is linearly separable data (LSD) because $X_1$ and $X_3$ can separate two classes completely by two box-whisker plots.

2) Problem 3
   Forward stepwise procedure chooses $X_1$, $X_2$, $X_3$, $X_4$, $X_5$, and $X_6$, in this order. Although MNM of Revised IP-OLDF and NM of a quadratic discriminant function (QDF) is zero in the 1-variable model ($X_1$), QDF misclassifies all regular cars to the small cars after $X_3$ enters the model because the values of $X_3$ in the small car are four. This data is very suitable for the explanation of Problem 3 because it is easier than exam scores using 100 items.

3) The explanation of the Method 2 by this data
   When we discriminate six microarray data by eight LDFs, only Revised IP-OLDF can make the feature selection naturally and reduce the high-dimension gene space to the small gene subspace that is linearly separable model. We call these subspaces as Matroska. We establish the Matroska feature selection method for the microarray data (Method 2) and the data consist several small Matroska with MNM = 0. Because the discrimination of LSD is not popular now and the Method 2 has several unknown ideas, we explain these ideas by this data in addition to Swiss banknote data in Chapter 6.

   If the data is LSD, full model is the biggest Matroska that contains all smaller Matroska. We had already known the smallest Matroska (the basic gene set, BGS) can describe the structure of Matroska completely by MNM decreases monotonously. On the other hand, the LASSO tries to make the feature selection. If it cannot find BGS in the data, it cannot explain the structure of the data.

Keywords: Linearly Separable Data (LSD), Matroska Feature Selection Method for Microarray Data (Method 2), Problem 3, LASSO, Fisher's Linear Discriminant Function (Fisher's LDF), Quadratic Discriminant Function (QDF), Support Vector Machine (SVM), Minimum Number of Misclassifications (minimum NM, MNM), Revised IP-OLDF, Revised IPLP-OLDF, Revised LP-OLDF.

7.1 Introduction

7.2 Japanese Automobile Data

7.2.1 Data Outlook

7.2.2 Comparison of Nine Discriminant Functions for Non-LSD

7.2.3 Consideration of Statistical Analysis

7.3 100-fold Cross-Validation (Method 1)

7.3.1 Comparison of Best Model

7.3.2 95% CI of Coefficients by Six MP-Based LDFs

7.3.2.1 Revised IP-OLDF Versus H-SVM

7.3.2.2 Revised IPLP-OLDF, Revised LP-OLDF, and other LDFs

7.3.2 95% CI of Coefficients by Fisher's LDF and Logistic Regression
7.4 Matryoshka Feature-Selection Method (Method 2)

7.4.1 Feature-Selection by Revised IP-OLDF

7.4.2 Coefficient of H-SVM and SVM4

7.5 Summary

REFERENCES

Bibliography
