Modern Applied Statistics Chap 12: Classification

Yongdai Kim

October 27, 2022

Seoul National University

Outline

Introduction

Introduction

In the statistical literature the word is used in two distinct senses.

- The sense of cluster analysis discussed in Section 11.2
- The other meaning (Ripley, 1997) of allocating future cases to one of g prespecified classes

It is sometimes helpful to distinguish discriminant analysis in the sense of describing the differences between the g groups from classification, allocating new observations to the groups.

- The first provides some measure of explanation
- The second can be a 'black box' that makes a decision without any explanation.

Discriminant Analysis

Discriminant Analysis

Suppose that we have a set of g classes, and for each case we know the class. We can then use the class information to help reveal the structure of the data.

The sample covariance matrices

$$W = \frac{(X - GM)^T (X - GM)}{n - g}, \quad B = \frac{(GM - 1\bar{x})^T (GM - 1\bar{x})}{g - 1}$$

- W : the within-class covariance matrix
- B : the between-classes covariance matrix
- M : the $g \times p$ matrix of class means
- G : the $n \times g$ matrix of class indicator variables

  - Then the predictions are GM
- $\bar{x}$: the means of the variables over the whole sample.

Discriminant Analysis

In pattern-recognition terminology the distinction is between supervised and unsupervised methods.

- **Iris data**
  Iris data has 150 cases, which are stated to be 50 of each of the three species *Iris setosa, virginica and versicolor*. Each case has four measurements on the length and width of its petals and sepals.

A *priori* this seems a supervised problem, and the obvious questions are to use measurements on a future case to classify it, and perhaps to ask how the variables vary among the species. However, the classification of species is uncertain, and similar data have been used to identify species by grouping the cases.

Discriminant Analysis

Krzanowski (1988) and Mardia, Kent and Bibby (1979) are two general references on multivariate analysis. For pattern recognition we follow Ripley (1996), which also has a computationally-informed account of multivariate analysis. Most of the emphasis in the

literature and in this chapter is on continuous measurements, but we do look briefty at multi-way discrete data in Section 11.4. Colour can be used very effective1y to differentiate groups in the

plots of this chapter, on screen if not on paper. The code given here uses both colours and symbols, but you may prefer to use only one of these to differentiate groups.

Classification Theory

Classification Theory

Non-parametric Rules

Non-parametric Rules

Neural Networks

Neural Networks

Support Vector Machine

Support Vector Machine

Forensic Glass Example

Forensic Glass Example

Calibration Plots

Calibration Plots