Modern Applied Statistics Chap 12: Classification

Yongdai Kim

October 30, 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

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.

Note that B has rank at most $\min(p, g - 1)$.

Discriminant Analysis

Fisher introduced a **linear discriminant analysis** seeking a linear combination xa of the variables that has a maximal ratio of the separation of the class means to the within-class variance.

- Maximizing the ratio $\boldsymbol{a}^T B\boldsymbol{a}/\boldsymbol{a}^T W\boldsymbol{a}$
  - Choose a sphering xS of the variables
  - The problem is to maximize $\boldsymbol{a}^T B\boldsymbol{a}$ subject to $\|\boldsymbol{a}\| = 1$
  - This is solved by taking a to be the eigenvector of B corresponding to the largest eigenvalue.
  - The linear combination a is unique up to a change of sign.

As for principal components, we can take further linear components corresponding to the next largest eigenvalues.

- **Eigenvalues**: the proportions of the between classes variance explained by the linear combinations.
- The corresponding transformed variables are called the **linear discriminants or canonical variates**.
- The linear discriminants are conventionally centred to have mean zero on dataset.

Discriminant Analysis

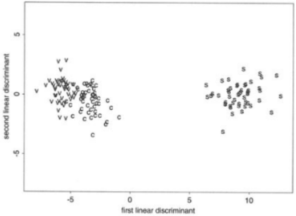

Figure 12.1: The log `iris` data on the first two discriminant axes.

▶ This shows that 99.65% of the between-group variance is on the first discriminant axis.

- Rao (1948) used the unweighted covariance matrix of the group means.
- Our approach uses a covariance matrix **weighted by the prior probabilities of the classes** if these are specified.

Discrimination for normal populations

An alternative approach to discrimination is via probability models.

- Posterior distribution of the classes after observing x is

$$p(c \mid \boldsymbol{x}) = \frac{\pi_c p(\boldsymbol{x} \mid c)}{p(\boldsymbol{x})} \propto \pi_c p(\boldsymbol{x} \mid c)$$

  - π_c: the prior probabilities of the classes
  - $p(\boldsymbol{x} \mid c)$: the densities of distributions of the observations for each class
  - **Bayes rule**: The allocation rule which makes the smallest expected number of errors chooses the class with maximal $p(c \mid x)$

Discriminant Analysis

Suppose the distribution for class c is multivariate normal with mean $\boldsymbol{\mu}_{\mathrm{c}}$ and covariance Σ_c. Then the Bayes rule minimizes

$$
\begin{aligned}
Q_c &= -2 \log p(\boldsymbol{x} \mid c) - 2 \log \pi_c \\
&= (\boldsymbol{x} - \boldsymbol{\mu}_c)\, \Sigma_c^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c)^T + \log |\Sigma_c| - 2 \log \pi_c
\end{aligned}
$$

- The first term is the squared Mahalanobis distance to the class centre.
- The difference between the Q_c for two classes is a quadratic function of x.
  - **Quadratic discriminant analysis.**
- The boundaries of the decision regions are quadratic surfaces in x space.

Discriminant Analysis

Suppose that the classes have a common covariance matrix Σ.

- Differences in the Q_c are then linear functions of $\boldsymbol{x}$

We can maximize $-Q_c/2$ or

$$L_c = \boldsymbol{x}\Sigma^{-1}\boldsymbol{\mu}_c^T - \boldsymbol{\mu}_c\Sigma^{-1}\boldsymbol{\mu}_c^T/2 + \log \pi_c$$

To use Q_c or L_c we have to estimate μ_c and Σ_c or Σ.

- Using obvious estimates, estimate μ_c as the sample mean, Σ_c as covariance matrix, and Σ as W.

Discriminant Analysis

How does this relate to Fisher's linear discrimination?

- L_c gives new variables, the linear discriminants, with unit within-class sample variance.

- On these variables the Mahalanobis distance is

$$\left\|x - \mu_c\right\|^2$$

Only the first r components of the vector depend on c.

$$L_c = \boldsymbol{x}\boldsymbol{\mu}_c^T - \left\|\boldsymbol{\mu}_c\right\|^2/2 + \log \pi_c$$

- We can work in r dimensions.

$$L_2 - L_1 = \boldsymbol{x}\left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\right)^T + \text{ const}$$

- An affine function of the linear discriminant.

Discriminant Analysis

Discriminant Analysis

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