Modern Applied Statistics Chap 12: Classification

Yongdai Kim

November 2, 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.

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

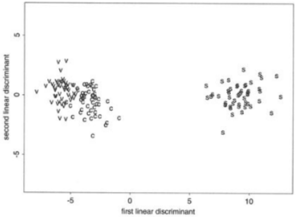

Figure 12.1: The log `iris` data on the first two discriminant axes.

▶ 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 $\boldsymbol{\mu}_c$ 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.

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.

Crabs dataset

Construct a rule to predict the sex of a future Leptograpsus crab of unknown colour form.

- Linear discriminant analysis, for what are highly non-normal populations, finds a variable that is essentially $\mathrm{CL}^3\mathrm{RW}^{-2}\mathrm{CW}^{-1}$, a dimensionally neutral quantity.
- Six errors are made, all for the blue form

Crabs dataset

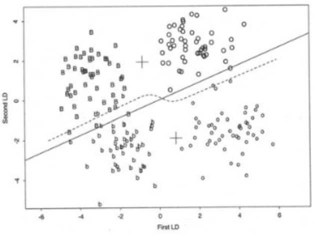

Figure 12.2: Linear discriminants for the `crabs` data. Males are coded as capitals, females as lower case, colours as the initial letter of blue or orange. The crosses are the group means for a linear discriminant for sex (solid line) and the dashed line is the decision boundary for sex based on four groups.

- The first two linear discriminants dominate the between-group variation.
- Using the first two linear discriminants as the data will provide a very good approximation

Robust estimation of multivariate location and scale

- Apply a robust location estimator to each component of a multivariate mean.
- Consider the estimation of mean and variance simultaneously.
- Two methods for robust covariance estimation
  - Our function cov.rob [1]
  - The S-PLUS functions `cov.mve` and `cov.mcd` and covRobin in library section robust.

    - MVE(minimum volume ellipsoid method): seeks an ellipsoid containing $h = \lfloor (n+p+1)/2 \rfloor$ points that is of minimum volume.

    - MCD(minimum covariance determinant method): eeks h points whose covariance has minimum determinant.

Robust estimation of multivariate location and scale

- The search for an MVE or MCD
  - provides h points whose mean and variance matrix give an initial estimate.
  - This is refined by selecting those points whose Mahalanobis distance from the initial mean using the initial covariance is not too large
  - and returning their mean and variance matrix.
- Alternative approach
  - Fitting a multivariate t_ν distribution for a small number ν of degrees of freedom.
  - This is implemented in our function cov.trob
- Normally `cov.trob` is faster than `cov.   rob`, but it lacks the latter's extreme resistance.

Classification Theory

Classification Theory

- In the terminology of pattern recognition
  - Training set: The given examples together with their classifications.
  - Test set: Future cases form.
  - **Our primary measure of success is the error rate.**
- The type of errors
  - A confusion matrix gives the number of cases with true class i classified as of class j.
  - We assign costs L_{ij} to allocating a case of class i to class j.

Classification Theory

The average error cost is minimized by the Bayes rule, which is to allocate to the class c minimizing $\sum_i L_{ic} p(i \mid \boldsymbol{x})$.

- $p(i \mid \boldsymbol{x})$ is the posterior distribution of the classes after observing $\boldsymbol{x}$.
- If the costs of all errors are the same, this rule amounts to choosing the class c with the largest posterior probability $p(c \mid \boldsymbol{x})$.

The minimum average cost is known as the Bayes risk.

- Estimate a lower bound for it by the method of Ripley (1996, pp. 196-7).

In Section 12.1,

- How $p(c \mid \boldsymbol{x})$ can be computed for normal populations.
- How estimating the Bayes rule with equal error costs leads to linear and quadratic discriminant analysis.

As our functions `predict.lda` and `predict.qda` return posterior probabilities.

- ▶ They can be used for classification with error costs.

Classification Theory

The posterior probabilities $p(c \mid \boldsymbol{x})$ may also be estimated directly.

- For two classes: we can model $p(1 \mid \boldsymbol{x})$ using a logistic regression, fitted by glm.
- For more than two classes: we need a multiple logistic model.
  - ▶ Using a surrogate log-linear Poisson GLM model.
  - ▶ Using the multinom function in library section `nnet`.

Classification trees model the $p(c \mid \boldsymbol{x})$ directly.

▶ Since the posterior probabilities are given by the `predict` method it is easy to estimate the Bayes rule for unequal error costs.

Predictive and 'plug-in' rules

To find the Bayes rule we need to know the posterior probabilities $p(c \mid \boldsymbol{x})$.

- Since these are unknown, we use an parametric family $p(c \mid \boldsymbol{x}; \theta)$.
- We act as if $p(c \mid x; \hat{\theta})$ were the actual posterior probabilities.
  - $\hat{\theta}$ is an estimate computed from the training set $\mathcal{T}$.
  - **'plug-in' rule**

The 'correct' estimate of $p(c \mid \boldsymbol{x})$ is (Ripley, 1996, §2.4) to use the predictive estimates

$$\tilde{p}(c \mid \boldsymbol{x}) = P(c = c \mid \boldsymbol{X} = \boldsymbol{x}, \mathcal{T}) = \int p(c \mid \boldsymbol{x}; \theta) p(\theta \mid \mathcal{T}) \mathrm{d}\theta$$

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