Parametric Models for Discrimination & Classification

John Maindonald

August 7, 2007

1 Linear Methods for Discrimination

See Ripley (1996); Venables and Ripley (2002); Maindonald & Braun (2007, Section 12.2).

The methods discussed here may be contrasted with the strongly non-parametric tree-based methods that are discussed in Maindonald & Braun (2007, Section 11.7, with a brief overview of Sections 11.1–11.6).

Notation

Observations are rows of a matrix $\mathbf{X}$ with p columns. The vector $\mathbf{x}$, is a row of $\mathbf{X}$, but in column vector form. The outcome is categorical, one of g classes. The matrix $\mathbf{W}$ estiamtes the within class variance-covariance matrix, while $\mathbf{B}$ estimates the between class variance-covariance matrix. Details of the estimators used are not immediately important. Note however that they may differ somewhat between computer programs.

Methods discussed here will all work with linear functions of the columns of $\mathbf{X}$. By allowing columns that are non-linear functions of the initial variables, additive non-linear effects can be accommodated.

1.1 Canonical discriminant analysis

Fisher's linear disciminant analysis was a version of canonical discriminant analysis that used a single discriminant axis. The more general case, where there can be as many as $r = \min(g-1, p)$ discriminant functions, is described here.

In this context it is convenient to use a version of $\mathbf{X}$ that does not have an initial column of ones. The discussion is simplified if a linear transformation is applied to the data so that the estimate of the within class variance-covariance matrix becomes the identify matrix. This can be achieved by replacing $\mathbf{x}$ by

$$\mathbf{z} = \mathbf{U'}^{-1}\mathbf{x}$$

where $\mathbf{U}$ is an upper triangular matrix such that $\mathbf{U'U} = \mathbf{W}$. [The usual estimate of the variance-covariance matrices is positive definite, providing that the same observations are used in calculating all elements in the variance-covariance matrix and no variable is redundant.]

The between classes variance-covariance matrix becomes

$$\tilde{\mathbf{B}} = \mathbf{U'}^{-1}\mathbf{B}\mathbf{U}^{-1}$$

The ratio of between to within class variance of the linear combination $\alpha'\mathbf{z}$ is then

$$\alpha'\tilde{\mathbf{B}}\alpha/\tilde{\alpha}'\tilde{\alpha}$$

The matrix $\tilde{\mathbf{B}}$ admits the principal components decomposition

$$\tilde{\mathbf{B}} = \lambda_1\mathbf{u_1}\mathbf{u_1'} + \lambda_2\mathbf{u_2}\mathbf{u_2'} + \ldots + \lambda_r\mathbf{u_r}\mathbf{u_r'}$$

The choice $\alpha = \mathbf{u_1}$ maximizes the ratio of the between to the within group variance, a fraction λ_1 of the total. The choice $\alpha = \mathbf{u_2}$ accounts for the next largest proportion λ_2, and so on.

The vectors $\mathbf{v}_1, \ldots \mathbf{v}_r$ are known as "linear discriminants" or "canonical variates". Scores, which are conveniently centered about the mean over the data as a whole, are available on each observation for each discriminant. These locate the observations in r-dimensional space, where r is at most $\min(g-1, p)$. A simple rule is to assign observations to the group to which they are nearest, i..e., the distance d_c is smallest in a Euclidean distance sense.

Fewer than the maximum number of discriminants may be used, depending on the effect on discriminatory power. If there are more than three groups, it is pertinent to inquire whether the two-dimensional representation that uses the first two discriminants misses non-trivial information.

Variables have been scaled so that within group variance-covariance matrix is the identity. Hence the variance should be the same in every direction. An equal scaled plot should therefore be used to plot the scores.

1.2 `lda()` and `qda()`

The functions `lda()` and `qda()` in the *MASS* package implement a Bayesian decision theory approach.

- A prior probability π_c is assigned to the cth class ($i = 1, \ldots g$).
- The density $p(\mathbf{x}|c)$ of $\mathbf{x}$, conditional on the class c, is assumed multivariate normal, i.e., rows of $\mathbf{X}$ are sampled independently from a multivariate normal distribution.
- For linear discrimination, classes are assumed to have a common covariance matrix Σ. For quadratic discrimination, different $p(\mathbf{x}|c)$ are allowed for different classes.
- Use Bayes' rule to derive $p(c|\mathbf{x})$. The allocation rule that gives the largest expected accuracy chooses the class with maximal $p(c|\mathbf{x})$; this is the Bayes' rule.
- More generally, assign cost L_{ij} to allocating a case of class i to class j, and choose c to minimize $\sum_i L_{ic} p(i|\mathbf{x})$.
- The Bayes rule requires knowledge of $p(c|\mathbf{x})$. These are unknown; hence a parametric family $p(c|\mathbf{x}; \theta)$ is assumed. For $g = 2$ classes, a logistic model is assumed, while for $g > 2$ a multinomial loglinear model is assumed.
- For estimation of the posterior probabilities, the simplest approach is to replace $p(c|\mathbf{x}; \theta)$ by $p(c|\mathbf{x}; \hat{\theta})$ for calculation of posterior probabilities (the 'plug-in' rule). The functions `predict.lda()` and `predict.qda()` offer the alternative estimate `method="predictive"`, which takes account of uncertainty in $p(c|\mathbf{x}; \hat{\theta})$. Note also `method="debiased"`, which may be a reasonable compromise between `method="plugin"` and `method="predictive"`

Note that `lda()` and `qda()` use the prior weights, if specified, as weights in combining the within class variance-covariance matrices.

1.2.1 Connection with Fisherian linear discriminant analysis

The theory underlying `lda()` assigns `x` to the class that maximizes the likelihood. This is equivalent to choosing the class c that minimizes $d_c + \log(\pi_c)$, where if the same estimates are used for $\mathbf{W}$ are $\mathbf{B}$, d_c is the distance as defined for Fisherian linear discriminant analysis. Recall that π_c is the prior probability of class c.

The output from `lda()` includes the list element `scaling`, which is a matrix with one row for each column of $\mathbf{X}$ and one column for each discriminant function that is calculated. This gives the discriminant(s) as functions of the values in the matrix $\mathbf{X}$.

1.2.2 Calculations using `lda()` from R's *MASS* package

The data frame `fgl` in the *MASS* gives 10 measured physical characteristics for each of 214 glass fragments that are classified into 6 different types. The following may help make sense of the information in the list element `scaling`.

```
library(MASS}
fgl.lda <- lda(type ~ ., data=fgl)
scores <- predict(fgl.lda, dimen=5)$x  # Default is dimen=2
## Now calculate scores from other output information
checkscores <- as.matrix(fgl[, -10])%*%fgl.lda$scaling
## Center columns about mean
checkscores <- scale(checkscores, center=TRUE, scale=FALSE)
plot(scores[,1], checkscores[,1])  # Repeat for remaining columns
## Check other output information
fgl.lda
```

93% of the information, as measured by the trace, is in the first two discriminants.

1.3 Logistic Regression

This may be handled using R's function `glm()`. Logistic regression is a special case of a Generalized Linear Model (GLM). The approach is to model $p(c|\mathbf{x};\hat{\theta})$ using a parametric model that may be the same logistic model as for linear and quadratic discriminant analysis.

In this context it is convenient to change notation slightly, and give $\mathbf{X}$ an initial column of ones. In the linear model and generalized linear model contexts, $\mathbf{X}$ has the name "model matrix".

The vector $mathbfx$ is a row of $\mathbf{X}$, but in column vector form. Then if π is the probability of membershipin the second group, the model assumes that

$$\log(\pi/(1-\pi) = \beta'\mathbf{x}$$

where d is a constant.

Compare logistic regression with linear discriminant analysis:

- Inference is conditional on the observed $\mathbf{x}$. A model for $p(\mathbf{x}|c)$ is not required. Results are therefore more robust against the distribution $p(\mathbf{x}|c)$.
- Parametric models with "links" other than the logit $f(\pi) = \log(\pi/(1-\pi)$ are available. Where there are sufficient data to check whether one of these other links may be more appropriate, this should be done. Or there may be previous experience with comparable data that suggests use of a link other than the logit.
- Observations can be given prior weights.
- There is no provision to adjust predictions to take account of prior probabilities.
- The fitting procedure minimizes the deviance, which is twice the difference between the loglikelihood for the model that is fitted and the loglikelihood for a 'saturated' model in which predicted values from the model equal observed values. This does not necessarily maximize predictive accuracy.
- Standard errors and Wald statistics (roughly comparable to t-statistics) are provided for parameter estimates. These are based on approximations that may fail if predicted proportions are close to 0 or 1 and/or the sample size is small.

1.4 Model choice, and comparison with highly non-parametric approaches

The linearity assumptions are restrictive, even allowing for the use of regression spline terms to model non-linear effects. It is not obvious how to choose the appropriate degree for each of a number of terms. The attempt to investigate and allow for interaction effects adds further complications. In order to make progress with the analysis, it may be expedient to rule out any but the most obvious interaction effects. These issues affect regression methods (including GLMs) as well as discriminant methods.

On a scale in which highly parametric methods lie at one end and highly non-parametric methods at the other, linear discriminant methods lie at the parametric end, and tree-based methods and random forests at the non-parametric extreme. An attraction of tree-based methods and random forests is that model choice can be pretty much automated.

1.5 Visualization

In linear discriminant analysis, discriminant scores in as many dimensions as seem necessary are used to classify the points, and thus emerge directly from the analysis. Each pair of dimensions gives a two-dimensional projection of the data. If there are three groups and at least two explanatory variables, the two-dimensional plot is a complete summary of the analysis. Even where higher numbers of dimensions are required, it may capture most of the information. This can be checked.

With most other methods, a low-dimensional representation does not arise so directly from the analysis. The following approach, which can be used directly with random forests, can be adapted for use with other methods. The proportion of trees in which any pair of points appear together at the same node may be used as a measure of the "proximity" between that pair of points. Then, using 1-proximity as a measure of distance, an ordination method can be used to find a representation of those points in a low-dimensional space.

1.6 Reference

References

Maindonald, J. H. and Braun, W.J. 2007. *Data Analysis and Graphics Using R – An Example-Based Approach.* 2^{nd} edition, Cambridge University Press.

Ripley, B. D., 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press.

Venables, W. N. and Ripley, B. D., 2002. *Modern Applied Statistics with* S. Springer-Verlag, 4 edition. See also R Complements to Modern Applied Statistics with S.
`http://www.stats.ox.ac.uk/pub/MASS4/`
[Note especially pp.331–341 (lda and qda) and pp.187–198 (logistic and other GLMs). In the third edition, see pap.344-354 and pp.211-226]