"Clustering of Pattern by using Discrimination analysis"

Praveen Kumar Pandey, Asst. Professor
Department of Mechanical Engineering, Faculty of Engineering & Technology
Gurukul Kangari University, Haridwar.

1. Introduction

Traditionally grouping of an object in known class was done by various method like cluster analysis, membership-roster concept, common property concept, feature extraction, error estimation, minimum distance method} etc on the basis of similarity of their characteristics. The primary purpose of the discriminant function is to predict the group of unknown objects based on cut off.

Discriminant Analysis is used to distinguish between two or more predefined 'groups'. The analysis identifies those variables that contribute most to the differences between groups, it is also possible to use Discriminant Analysis as a classification technique that can be used to place an unknown case into one of the groups.

Discriminant Analysis works by combining the variables in such a way that the differences between the predefined groups are maximized . Note that group membership must be known before using Discriminant Analysis. The discriminant problem is how do we best predict or assign an object whose population identity we do not know to one of the known populations of interest?

The discriminant function can use several quantities variables, each of which makes an independent contribution to the overall discrimination. Taking into consideration the effect of all variables this discriminant function produces the statistical decision for guessing.

Discriminant function analysis or DA is used to classify cases into the values of a categorical dependent, usually a dichotomy. If discriminant function analysis is effective for a set of data, the classification table of correct and incorrect estimates will yield a high percentage correct. Multiple discriminant function analysis is used when the dependent has three or more categories.

Discriminant function analysis is used to determine which variables discriminate between two or more naturally occurring groups. For example, an educational researcher may want to investigate which variables discriminate between high school graduates who decide (1) to go to college, {2) to attend a trade or professional school, or (3) to seek no further training or education. For that purpose the researcher could collect data on numerous variables prior to students graduation. After graduation, most students will naturally fall into one of the three categories. Discriminant Analysis could then be used to determine which variable are the best predictors of student's subsequent educational choice.

A medical researcher may record different variables relating to patients backgrounds in order to learn which variables best predict whether a patient is likely to recover completely (group 1), partially (group 2), or not at all (group 3). A biologist could record different characteristics of similar types (groups) of flowers, and then perform a discriminant function analysis to determine the set of characteristics that allows for the best discrimination between the types.

There are several purposes for DA:

. To classify cases into groups using a discriminant prediction equation.
. To investigate independent variable mean differences between groups formed by the dependent variable.
. To determine the percent of variance in the dependent variable explained by the independents.
. To determine the percent of variance in the dependent variable explained by the independents over and above the variance accounted for by control variables, using sequential discriminant analysis.
. To assess the relative importance of the independent variables in classifying the dependent variable.
. To discard variables which are little related to group distinctions.
. To test theory by observing whether cases are classified as predicted.

1.1 BASIC ELEMENTS OF DISCRIMINANT FUNCTION ANALYSIS

1.1.1 DISCRIMINATING VARIABLES: These are the independent variables, also called predictors.

1.1.2 THE CRITERION VARIABLE : This is the dependent variable, also called the grouping variable. It is the object of classification efforts.

1.1.3 DISCRIMINANT FUNCTION

A discriminant function, also called a canonical root, Is a latent variable which is created as a linear combination of discriminating (independent) variables, such that $L = b_1x_1 + b_2x_2 + \ldots\ldots + b_nx_n + c$, Where the b's are discriminant coefficients, the x's are discriminating variables, and c is a constant. This is analogous to multiple regressions, but the b's are discriminant coefficients, which maximize the distance between the means of the criterion (dependent) variable. Note that the foregoing assumes the discriminant function is estimated using ordinary least squares, the traditional method.

1.1.4 NUMBER OF DISCRIMINANT FUNCTIONS

There is one discriminant function for 2-group discriminant analysis, but for higher order DA, the number of functions (each with its own cut-off value) is the lesser of (g - 1), where g is the number of categories in the grouping variable, each discriminant function is orthogonal to the others. A dimension is simply one of the discriminant functions when there is more than one, in multiple discriminant analysis.

1.1.5 EIGEN VALUE

The Eigen value of each discriminant function reflects the ratio of importance of the dimensions, which classify cases of the dependent variable. If there is more than one discriminant function, the first will be the largest and most important, the second next most important in explanatory power, and so on. The Eigen values assess relative importance because they reflect the percent's of variance explained in the dependent variable, cumulating to 100% for all functions.

1.1.6 THE DISCRIMINANT SCORE

The discriminant score also called the DA score, is the value resulting from applying a discriminant function formula to the data for a given case.

1.1.7 CUTOFF

If the discriminant score of the function is less than or equal to the cutoff, the case is classed as 0, or if above it is classed as 1. When group sizes are equal, the cutoff is the mean of the two centroids (for two-group DA). If the groups are unequal, the cutoff is the weighted mean.

1.2 ASSUMPTION

The minimum set of conditions necessary to conduct linear discriminant is:

1. There should be two or more than two priori groups or classifications of the population of entities.
2. A sample of entities known to belong to each group exists.
3. Each entity can be described by a set of quantitative variables.
4. The variance and relationship among each of the variables are the same for each group.

1.3 ANALYSIS OF TWO-GROUP DISCRIMINANT FUNCTION

In the two-group case, discriminant function analysis can also be thought of as (and is analogous to) multiple regression (see Multiple Regression; the two-group discriminant analysis is also called Fisher linear discriminant analysis after Fisher, 1936; computationally all of these approaches are analogous). If we code the two groups in the analysis as 1 and 2, and use that variable as the dependent variable in a multiple regression analysis, then we would get results that are analogous to those we would obtain via Discriminant Analysis. In general, in the two-group case we fit a linear equation of the type:

$$Group = a + b_1 * x_1 + b_2 * x_2 + \ldots \ldots + b_m * x_m$$

Where a is a constant and b_1 through bm are regression coefficients. The interpretation of the results of a two-group problem is straight forward and closely follows the logic of multiple regressions: Those variables with the largest (standardized) regression coefficients are the ones that contribute most to the prediction of group membership.

1.4 DISCRIMINANT FUNCTIONS FOR MULTIPLE GROUPS

When there are more than two groups, then we can estimate more than one discriminant function like the one presented above. For example, when there are three groups, we could estimate (1) a function for discriminating between group 1 and groups 2 and 3 combined, and (2) another function for discriminating between group 2 and group 3. For example, we could have one function that discriminates between those high school graduates that go to college and those who do not (but rather get a job or go to a professional or trade school), and a second function to discriminate between those graduates that go to a professional or trade school versus those who get a job. The b coefficients in those discriminant functions could then be interpreted as before.

MATHEMATICAL EXAMPLE

In this example a simple data set will be used. The data are from 10 males and 10 females. Three variables were recorded: height (inches!), weight (pounds!) and age (years).

Data Summary

Variable	Male	Female	Difference
Height	70.3	66.2	4.1
Weight	165.9	130.7	35.2
Age	42.3	33.9	8.4

We need a method, which will maximize the group differnces displayed by these three discriminating variable, when they are combined into a single discriminating variable. This will be achieved by calculating a Discriminant Function of the type:

score = $w_1$height + $w_2$weight + $w_3$age

Thus our problem is finding suitable values for wi.

First calculate the variance-covariance matrix A

$$\text{Let } A= \begin{array}{llll} & \text{Height} & \text{Weight} & \text{Age} \\ \text{Height} & 5.21 & 24.49 & 10.68 \\ \text{Weight} & 24.49 & 207.72 & 65.56 \\ \text{Age} & 10.68 & 64.56 & 155.17 \end{array}$$

Let w be a vector containing the unknown weights:

$w=[w_1\ \ w_2\ w_3\,]$

and d be a vector of the group differences (as shown above):

d=[4.1 35.2 8.4]

It can be shown that A.w = d

This is a relatively simple matrix algebra calculation since the equation be rewritten as w = A^{-1}.d and we need only find A^{-1} , the inverse of matrix A, to solve the equation.

Using standardized variables we find that

$$W= \begin{array}{l} 0.0029 \text{ height} \\ 1.028 \text{ weight} \\ -0.096 \text{ age} \end{array}$$

Hence;

Discriminant score = 0.0029 height + 1.028 weight - 0.096 age

The group centroids (mean scores) are females -1.25 males +1.23

Consequently a positive score (>0) indicates a male, a negative score (<0) indicates a female. The means would not be symmetrical if the group sizes differed.

CONCLUSION

1. The model is capable of grouping them simultaneously for given data.
2. It has been observed that for input data no need to do calculation manually, computer program is capable to classify the data in various groups.

3. Method presented is very useful to classify the object in various field like banking, share market, traffic control, etc.

5.3 SCOPE FOR FUTURE WORK

The procedure and algorithm developed are based on certain assumptions. However, according to our work on pattern recognition system, the works that can be carried out in the future are here under:

This procedure is very use full for classification so it can be use in many fields to classify the object in different groups. For example

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.
- Land use: Identification of areas of similar land use in an earth observation database.
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost:

References

1. Choulakian V. and Almhana J., ‘An Algorithm for nonmetric discriminant analysis’, Computational Statistics& Data Analysis, 35, 253 264, 2001.
2. Dai D.Q. and Pong C. Y., 'Rgularized Discriminant analysis and its application’, Journal of the Pattern Recognition Society, 36, 845-847, 2003.
3. Gavin C. C. and Nicola L. C. Talbot, ‘Efficient leave-one-out cross- validation of Kernel Fisher Discriminant analysis’, Journal of the Pattern Recognition Society, 36, 2585-2592, 2003,
4. Gonzalez R. C, and Thomason M.G., ‘Syntactic Pattern Recognition an Introduction’, Addison-Wesley Publishing Company, 1978.
5. Gupta 3.C. and Kapoor V.K., ‘Fundamentals of Mathematical Statistics’, Sultan Chand & Sons, 1999.
6. Lotikar R. and Kothari R., ‘Adaptive linear dimensionality reduction for classification society’, Journal of the Pattern Recognition Society, 33, 185-189, 2000.
7. Ordowski M. and Gerard G.L. Meyer, ‘Geometric linear
8. Discriminant analysis for pattern recognition’, Journal of the Pattern Recognition Society, 37, 421-428, 2004.