Discriminant Analysis

Introduction

Cluster analysis: find groups among data.

Discriminant analysis: *given* groups, find out how data differ. Use information in variables to get (as near as possible) separation into correct groups.

Echoes of regression: explain dependent variable (group membership) in terms of independent (other) variables.

Two methods (Fisher/Mahalanobis), look different, come out same.

Potential applications

- Exploratory investigation of eg. marketing data to determine differences between heavy/light users of a product
- Learning how to classify patients as having/not having disease on basis of symptoms
- Categorizing people by risk level for loans, credit, insurance etc. using demographic information

Two-group discriminant analysis

```
            Plot of Y*X.  Symbol is value of GROUP.

  Y |
1.0 +          2     2                      22
    |                                     2
    |                        2      2    2
    |                    2    2 2                         1
    |                              2      1           1
    |                                             1
0.5 +         2 2    2       2         1
    |         2
    |                   2         1
    |        2  2              1                         1
    |                2            1   1                     1
    |  2     2               1     1   1
0.0 +    2                                11
    ---+----------+----------+----------+----------+----------+--
      0.0        0.2        0.4        0.6        0.8        1.0
```

Easier to visualize with 2 independent variables (can draw picture). Picture: variables X, Y, groups 1, 2.

Points in bottom right are 1's, those top left 2's.

Within each group, positive correlation between X and Y.

Idea: *draw line* to best separate groups. Horizontal or vertical line not best; need line at angle. What line?

Mahalanobis approach

Let x denote (column) vector representing a point; let $\bar{x}_1$ be (vector) mean of points in group 1, $\bar{x}_2$ be mean of points in group 2.

Assume equal var/cov matrices for two groups. How to estimate this from data?

Recall two-sample t with equal variances: use *pooled variance*

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Weighted average of group variances.

Same idea here, with matrices: let S_1, S_2 be sample var/cov matrices for each group separately, based on n_1, n_2 observations.

Then pooled var/cov matrix is

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}.$$

For any point x, calculate Mahalanobis distance to each group mean, using pooled var/cov matrix:

$$\begin{aligned} D_1^2 &= (x - \bar{x}_1)' S_p^{-1} (x - \bar{x}_1) \\ D_2^2 &= (x - \bar{x}_2)' S_p^{-1} (x - \bar{x}_2) \end{aligned}$$

Using Mahalanobis distance allows for covariance among variables.

Idea: each observation supposed to be in group whose mean closer (in Mahalanobis distance). So draw line between groups by finding where $D_1^2 = D_2^2$ (locus of points).

Example: suppose

$$S_p^{-1} = \frac{1}{3} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$$

and that $\bar{x}_1 = (0, 0)'$, $\bar{x}_2 = (1, 0)'$. Then, for any point $x = (x_1, x_2)'$,

$$\begin{aligned} D_1^2 &= \frac{1}{3}(2x_1^2 - 2x_1 x_2 + 2x_2^2) \\ D_2^2 &= \frac{1}{3}(2(x_1 - 1)^2 - 2(x_1 - 1)x_2 + 2x_2^2) \end{aligned}$$

Quadratic forms, but places where equal form line: set difference $D_1^2 - D_2^2 = 0$:

$$0 = 2(2x_1 - 1 - x_2)$$

since all else cancels; line is $x_2 = 2x_1 - 1$. Anything on one side of line in group 1, anything on other in group 2.

Mahalanobis in general

In general, can find separation between groups by setting $D_1^2 = D_2^2$ and finding the x points solving this.

D_1^2 and D_2^2 quadratic forms, but locus of points where equal always linear combination of elements of x. Proof: write $D_1^2 - D_2^2$, expand out, collect terms to get

$$2\left(\frac{\bar{x}_1 + \bar{x}_2}{2} - x\right)' S_p^{-1}(\bar{x}_1 - \bar{x}_2) = 0.$$

As function of x, linear; can write as $(x - \bar{x})'k = 0$ where $k = S_p^{-1}(\bar{x}_1 - \bar{x}_2)$ and $\bar{x}$ is halfway between the two group means.

(Not mean of all data together unless groups same size.)

Revisit example:

$$k = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} -1 \\ 0 \end{pmatrix} = \frac{1}{9}\begin{pmatrix} -2 \\ 1 \end{pmatrix}$$

and $\bar{x} = (\frac{1}{2}, 0)$, so

$$0 = \frac{1}{9}\begin{pmatrix} x_1 - \frac{1}{2} & x_2 \end{pmatrix}\begin{pmatrix} -2 \\ 1 \end{pmatrix} = \frac{1}{9}(-2x_1 + 1 + x_2),$$

giving same line as before.

Discriminant scores, hit/miss table

Since $(x - \bar{x})'k = 0$ for points x on the line equidistant from the group means, and $\bar{x}$ is a constant for any particular data set, the value $x'k$ for any observation determines which group it should belong to: large for one group, small for the other.

Thus the vector Xk gives a **discriminant score** for each observation.

Using discriminant score, can estimate which group observation would have come from (had we not known). Like fitted values in regression.

Then make 2-way table of actual group vs. fitted group. Shows good, bad predictions of group membership.

Example: books by mail

Text example: new art book, "Art History of Florence", offered by book club. Try to understand who buys it. Send out mailing about book to subscribers. For each subscriber, note:

- how many months since last book purchase
- number of art books purchased
- whether subscriber purchased this book (1), or not (0).

Guess that subscribers who often buy books, or who buy many art books, more likely to buy this book.

Data in "books.dat". SAS code:

```
options ls=65;

data books;
  infile "books.dat";
  input id sincelst art florence;

proc discrim;
  var sincelst art;
  class florence;
```

"ID" is customer number, ignored in analysis.
Results (edited):

FLORENCE	Frequency	Weight	Proportion	Prior Probability
0	917	917.0000	0.917000	0.500000
1	83	83.0000	0.083000	0.500000

	FLORENCE 0	1
CONSTANT	-1.38666	-1.80584
SINCELST	0.19952	0.14291
ART	0.69853	2.26698

First: 83 people bought the new book, 8.3% of total.

Lower table: ignore Constant line; other values are $S_p^{-1}\bar{x}_1$ and $S_p^{-1}\bar{x}_2$, so *difference* (2nd minus 1st) gives $k = (-0.056, 1.577)'$.

Number of Observations and Percent Classified into FLORENCE:

From FLORENCE	0	1	Total
0	702	215	917
	76.55	23.45	100.00
1	35	48	83
	42.17	57.83	100.00
Total	737	263	1000
Percent	73.70	26.30	100.00

Of 917 non-buyers, 702 (77%) correctly classified as non-buyers; of 83 buyers, only 48 (58%) correctly classified. **Hit-miss table.**

Easier to predict that someone will not buy the book.

Testing for equality of var/cov matrices

Assumed that two groups have same var/cov matrix. Should test whether this is true.

Box's Test given on p. 443 of text, also in SAS. H_0 : var/cov matrices same; H_a: different.

Based on determinants of pooled and unpooled var/cov matrices. Idea: if pooling does not work, det of pooled matrix bigger than average det of unpooled matrices.

To get in SAS, add `pool=test` to PROC DISCRIM line. Output for book data:

```
Test Chi-Square Value =    77.119054
with      3 DF      Prob > Chi-Sq = 0.0001
```

Reject H_0, conclude that var/cov matrices different, should not have pooled. But sample size very large: very small difference in matrices could be significant.

Test also works with more than 2 groups.

What if var/cov matrices not equal?

Can still do Mahalanobis distances, but using separate var/cov matrices for each group. Locus of points equidistant from each group mean now no longer line, but general quadratic.

Analysis based on not pooling therefore called **quadratic discriminant analysis**.

SAS can do this: to prevent pooling, use `pool=no` on PROC DISCRIM line; to test first, use `pool=test` as above. Then SAS chooses linear/quadratic based on test result.

Quadratic analysis doesn't give discriminant coefficients, but still gives hit-miss table.

For book data:

```
Number of Observations and Percent Classified into FLORENCE:

From FLORENCE              0              1           Total

            0            743            174             917
                       81.03          18.97          100.00

            1             39             44              83
                       46.99          53.01          100.00

        Total            782            218            1000
      Percent          78.20          21.80          100.00
```

Correctly predicts more non-buyers (81% vs. 77%), but fewer buyers (53% vs. 58%).

Testing for significant differences between groups

Always possible that apparent group differences actually chance. How to test that group means significantly different?

Use two-sample version of Hotelling's T^2 (akin to two-sample t test). Let vector d be difference between sample means, let n_1, n_2 be sample sizes, p be number of variables, S_p be pooled var/cov matrix. Then test statistic is

$$T^2 = \frac{(n_1 + n_2 - p - 1)n_1 n_2}{p(n_1 + n_2 - 2)(n_1 + n_2)} d' S_p^{-1} d$$

and P-value from F_{p, n_1+n_2-p-1}.

Books example: $n_1 = 917, n_2 = 83, d = (-3.3, 0.67)$,

$$S_p^{-1} = \begin{pmatrix} 0.016 & -0.006 \\ -0.006 & 2.324 \end{pmatrix};$$

$$T^2 = \frac{(997)(917)(83)}{(1996)(1000)} \begin{pmatrix} -3.3 & 0.67 \end{pmatrix} S_p^{-1} \begin{pmatrix} -3.3 \\ 0.67 \end{pmatrix} = 47.23$$

with 2 and 997 df. P-value is very small, so differences between group means are real and not chance.

SAS does variations on this test; to get, add `manova` to the PROC DISCRIM line. For book data:

```
         Multivariate Statistics and Exact F Statistics

                      S=1    M=0    N=497.5

Statistic                Value         F    Num DF   Den DF  Pr > F

Wilks' Lambda           0.914041   46.881       2      997  0.0001
Pillai's Trace          0.085959   46.881       2      997  0.0001
Hotelling-Lawley Trace  0.094043   46.881       2      997  0.0001
Roy's Greatest Root     0.094043   46.881       2      997  0.0001
```

which all have same conclusion.

Bias in hit rate

In estimating hits and misses, using same data as used to estimate discriminant function in first place. Expect to *over-estimate* how well future observations would be classified.

Fix: use different data for estimation and classifying.

Could split data into two parts, but wasteful of data. Another idea: "jackknife" or "crossvalidation": build discriminant function from $n - 1$ observations, classify n-th, repeat for all observations.

In SAS, add `crossvalidate` to PROC DISCRIM line. Output is revised, more honest hit-miss table. Though for book data, no change (because n so large.)

Multiple-group discriminant analysis

With more than two groups, how to do discriminant analysis?

Fisher & Mahalanobis approaches now different; latter simpler.

Mahalanobis: calculate pooled var/cov matrix S_p, calculate distance of observation x from mean $\bar{x}_g$ of group g as

$$D_g^2 = (x - \bar{x}_g)' S_p^{-1} (x - \bar{x}_g).$$

No longer worry about finding lines distinguishing groups; simply classify each observation into group for which D_g^2 smallest.

Example: iris data

Looked at Fisher's iris data in assignment; tried to distinguish groups using principal components.

Use 4 variables (sepal & petal length & width) to explain grouping.

Test for equal group var/cov matrices:

```
Test Chi-Square Value =   140.943050
with     20 DF      Prob > Chi-Sq = 0.0001

Since the chi-square value is significant at the  0.1 level,
the within covariance matrices will be used in
the discriminant function.
```

So reject equality, do quadratic discrimination.

Test hypothesis of equal group means:

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.023439	199.15	8	288	0.0001
Pillai's Trace	1.191899	53.466	8	290	0.0001
Hotelling-Lawley Trace	32.47732	580.53	8	286	0.0001
Roy's Greatest Root	32.19193	1167	4	145	0.0001

Means are definitely not all same.

Hit-miss table, using crossvalidation:

From SPECIES	setosa	versicol	virginic	Total
setosa	50	0	0	50
	100.00	0.00	0.00	100.00
versicol	0	47	3	50
	0.00	94.00	6.00	100.00
virginic	0	1	49	50
	0.00	2.00	98.00	100.00
Total	50	48	52	150
Percent	33.33	32.00	34.67	100.00

Only 4 of the 150 iris misclassified: 3 versicolor as virginica, 1 virginica as versicolor. All setosas classified correctly.

(Similar results without crossvalidation: only 2 versicolor classified as virginica.)

Predicting new observations

Important use of discriminant analysis is to be able to classify new observations (of unknown groups) into groups.

Mahalanobis distance: if observation x much closer to group 1 than group 2, $D_1^2 << D_2^2$. x almost certainly in group 1. But if x almost equidistant, $D_1^2 \simeq D_2^2$, x could almost equally be in either group.

Assume: data distributed as multivariate normal in each group, var/cov matrices same. Try to estimate probability that new observation belongs in each group.

Result: $P(\text{group } i|x) \propto \exp(-D_i^2/2)$.

Small example:

$$S_p^{-1} = \frac{1}{3} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix},$$

$\bar{x}_1 = (0,0)'$, $\bar{x}_2 = (1,0)'$. Classify $x = (2,1)$ and $x = (2,0)$:

x	D_1^2	D_2^2	$\exp(-D_1^2/2)$	$\exp(-D_1^2/2)$	$P(1\|x)$
$(2,1)$	1	$\frac{2}{3}$	0.607	0.717	0.459
$(2,0)$	$\frac{8}{3}$	$\frac{2}{3}$	0.264	0.717	0.269

1st obs. could be from either group, but 2nd most likely group 2.

Also in SAS. Example: classify these iris measurements (sepal length, width; petal length, width):

```
7.8 3.9 4.7 1.7
7.2 3.0 1.3 2.0
5.6 3.0 6.7 0.3
```

Save in `irispred.dat`. Code now as below. Note variable names must match:

```
data irisnew;
  infile "irispred.dat";
  input sepallen sepalwid petallen petalwid;

proc discrim data=iris testdata=irisnew testlist;
  class species;
```


Results include classification probabilities for each new obs:

```
                  Posterior Probability of Membership in SPECIES:
     Obs     Classified
             into SPECIES       setosa     versicol     virginic

       1     versicol           0.0000       0.9999       0.0001
       2     setosa             1.0000       0.0000       0.0000
       3     versicol           0.0000       0.9924       0.0076
```

Clear-cut because groups well separated.