

## Discriminant Analysis

Discriminant Analysis is a technique concerned with separating distinct sets of objects or observations into groups and allocating new objects into previously defined groups.

Goals of Discriminant Analysis:

1. To **describe** the distinguishing features of observations from several known populations. That is, we try to find “discriminants” that separate the groups as much as possible.
2. To sort observations into two or more labeled classes and derive a “rule” that can be used to **assign** new objects into the labeled class.

### Examples

Separate good and poor credit risks	by income, age, family size, # of credit cards
Separate successful and unsuccessful college students	standardized test score, GPA, high school activities
Define authorship of Federalist Papers (James Madison or Alexander Hamilton)	frequencies of certain words, lengths of sentences

You can think of Discriminant Analysis like a two sample tests or MANOVA in reverse.

- In MANOVA, we want to see if there is a difference among the groups on certain factors.
- In discriminant analysis, we want to find what factors discriminate among the groups. That is, we usually have samples from two (or more) populations and we compare aspects of the two groups. In this case, we seek to discriminate or separate the two groups based on certain aspects. The emphasis is on defining a discriminating rule for future observations.

Suppose we wish to separate purchasers of a new product from non-purchasers. We use income, education, and family size in attempt to “discriminate” between purchasers and non purchasers.

Clearly Income is the “best” discriminator.

Graphically, we might observe the following incomes from a sample of purchasers and non purchasers.

We seek to define the point of separation

In 2-D we might observe

Again, we seek to find the separation point (line).

Let  $\Pi_1$  and  $\Pi_2$  be the two “classes” or **populations** we discriminate between.

The discrimination will be based on  $n$  observations of  $p$  random variables.

Let  $f_1(x)$  and  $f_2(x)$  be the probability density functions associated with the  $p \times 1$  vector of observations,  $X$ , for the respective populations,  $\Pi_1$  and  $\Pi_2$ .

Each observation must be classified as coming from either  $\Pi_1$  or  $\Pi_2$ .

Based on our sample, we define a portion as coming from  $\Pi_1$  and a portion from  $\Pi_2$ . Let  $R_1$  be the portion of our sample classified as originating from  $\Pi_1$ , and  $R_2$  is the portion of our sample that is classified as originating from  $\Pi_2$ .

Clearly we will misclassify some objects.

Our goal is to minimize this occurrence of misclassifications.

Let  $P(2|1)$  be the probability of classifying an object as  $\Pi_2$  when it is really from  $\Pi_1$ .

Likewise,

In one dimension

If, for example, we knew a priori what proportion of observations originates from each group, we would certainly want to take this into account.

Let  $p_1$  be the probability of an observation coming from  $\Pi_1$  and let  $p_2$  be the probability of an observation coming from  $\Pi_2$  ( $p_1 + p_2 = 1$ ).

P [ observation is correctly classified as  $\Pi_1$  ] =

P [ observation is misclassified as  $\Pi_1$  ] =

P [ observation is correctly classified as  $\pi_2$  ] =

P [ observation is incorrectly classified as  $\pi_2$  ] =

In discriminant analysis, there are two types of error that might occur.

- 1) one might classify an object in group 2 when it is really from group 1.  
Let  $C(2|1)$  be the cost of this misclassification.
  
- 2) one might classify an object in group 1 when it is really from group 2.  
Let  $C(1|2)$  be the cost of this misclassification.

The Expected Cost of Misclassification (ECM) is given by:

It can be shown that the discrimination rule which minimizes ECM is given by:

Assign to  $\Pi_1$  if

Assign to  $\Pi_2$  if

Suppose prior probabilities are equal. This rule simplifies to:

a)

Suppose costs of misclassification are equal:

b)

Suppose costs and prior probabilities are equal:

c)

When misclassification costs and prior probabilities are unknown they are often assumed to be equal. Thus, rule C is used arbitrarily in practice.

In order to actually use this “rule”, we must know  $f_1(x)$  and  $f_2(x)$ .

Suppose  $f_1(x)$  and  $f_2(x)$  are multivariate normal densities given by:

$\mu_1$  is the 1<sup>st</sup> mean vector and

$\Sigma_1 = \Sigma_2 = \Sigma$  is the covavariance matrix.

Simplifying the ratios given above, we have the following classification rules:

In practice, we estimate the mean vectors  $\mu_1$  and  $\mu_2$  with  $\bar{X}_1$  and  $\bar{X}_2$ . We estimate  $\Sigma$  with  $S_{\text{pooled}}$ ,

Thus our allocation rule becomes,

Allocate the observation  $X_0$  to  $\Pi_1$  if

Suppose we have the cases of equal cost and prior probabilities,

Thus our minimum ECM rule gives us

In essence, we have reduced the dimensions from  $p$  to 1 by taking appropriate linear combinations of the original variables.

In computing applications, the linear discriminant function

is split into two linear functions, one for each group.

That is, we rewrite the linear discriminant function as follows:

Most software packages print the equations separately, and the difference in these equations gives the linear discriminant rule.

Example: A man had been arrested in Kansas for stealing his neighbors turkeys. He claimed that the meat in his freezer actually came from wild turkeys. Bone measurements could be made on the confiscated turkeys. Using this information, the state was able to prove (in a court trial) that the turkeys in the man's freezer were actually domestic and not wild turkeys. How? Discriminant Analysis.

Of the 158 original turkeys studied, 33 cases were available for the analysis. 19 of these cases were domestic turkeys and 14 were wild.

### Discriminant Analysis

Linear Method for Response: TYPE  
 Predictors: HUM RAD ULN FEMUR TIN CAR D3P COR SCA

Group	DOMESTIC	WILD
Count	19	14

33 cases used 49 cases contain missing values

#### Summary of Classification

Put into	....True Group....	
Group	DOMESTIC	WILD
DOMESTIC	18	1
WILD	1	13
Total N	19	14
N Correct	18	13
Proportion	0.947	0.929

N = 33 N Correct = 31 Proportion Correct = 0.939

#### Summary of Classification with Cross-validation

Put into	....True Group....	
Group	DOMESTIC	WILD
DOMESTIC	17	2
WILD	2	12
Total N	19	14
N Correct	17	12
Proportion	0.895	0.857

N = 33 N Correct = 29 Proportion Correct = 0.879

Squared Distance Between Groups

	DOMESTIC	WILD
DOMESTIC	0.0000	13.1804
WILD	13.1804	0.0000

Linear Discriminant Function for Group

	DOMESTIC	WILD
Constant	-1071.5	-1086.0
HUM	2.5	2.4
RAD	0.6	0.5
ULN	5.2	4.6
FEMUR	2.5	2.3
TIN	-0.2	0.7
CAR	0.9	0.9
D3P	0.8	0.7
COR	-7.7	-6.9
SCA	3.4	3.4

The discriminant function is given by,

$$14.5709 + 0.1048 * HUM + 0.08 * RAD + 0.6195 * ULN + 0.2074 * FEM \\ - 0.9126 * TIN + 0.0325 * CAR + 0.1032 * D3P - 0.8216 * COR + 0.012 * SCA$$

this is  $\hat{a} x_0 - \hat{m}$

We compare this to  $k = 0$  for each observation.

if  $\hat{a} x - \hat{m} \geq 0$  we classify the turkey as domestic

otherwise we classify the turkey as wild.

Consider turkey # 13. Plugging into the discriminant function, we obtain

Thus turkey #13 is classified as domestic.

## How do we evaluate the “goodness” of our discriminant function?

- One way is the percent incorrect classifications. APER, or apparent error rate is frequently reported. This is the % misclassified.
- If, however, we measure this based on the existing data set, the percentage correctly classified will be inflated. A model, always works best when retrospectively applied to the data the model was derived from.
- If a large data set is available, one might split the data set and develop a discriminant function based on part of the data. The percent incorrectly classified will be determined from the remainder. The idea is to get a more realistic idea of the “goodness” of the model. This is known as the holdout method.

More often than not, large data sets are not available. In this case, we can use cross validation.

Cross Validation works like this:

1. remove observation 1 from data.
2. develop discriminant function based on the remaining  $(n-1)$  observations.
3. determine classification for observation 1.
4. repeat steps 1-3 for remaining  $n-1$  observations.

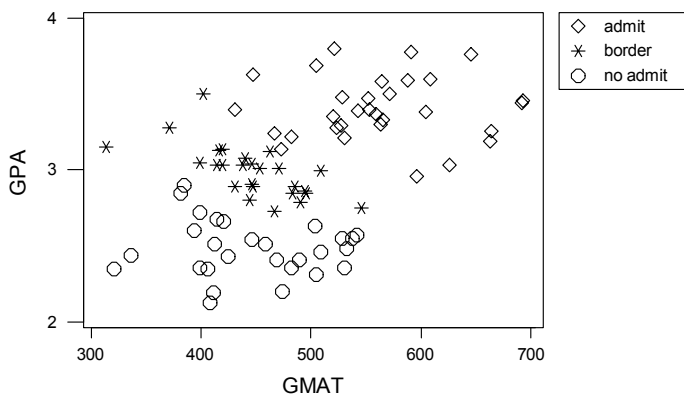
This is also known as jackknifing.

Example: The admissions officer of a business school has an ‘index’ of undergraduate GPA and GMAT scores to help decide which applicants should be admitted to the graduate program. Recent applicants are categorized as,

$$\Pi_1 = \text{admit}, \Pi_2 = \text{borderline}, \Pi_3 = \text{do not admit}$$

according to

$$X_1 = \text{GPA and } X_2 = \text{GMAT}$$



### Discriminant Analysis

Linear Method for Response: STATUS  
 Predictors: GPA GMAT

Group	admit	border	no admit
Count	31	26	28

#### Summary of Classification

Put into	....True Group....		
	admit	border	no admit
admit	27	1	0
border	4	25	2
no admit	0	0	26
Total N	31	26	28
N Correct	27	25	26
Proportion	0.871	0.962	0.929

N = 85      N Correct = 78      Proportion Correct = 0.918

Summary of Classification with Cross-validation

Put into Group	....True Group....		
	admit	border	no admit
admit	26	1	0
border	5	24	2
no admit	0	1	26
Total N	31	26	28
N Correct	26	24	26
Proportion	0.839	0.923	0.929

N = 85      N Correct = 76      Proportion Correct = 0.894

Squared Distance Between Groups

	admit	border	no admit
admit	0.0000	10.0634	31.2888
border	10.0634	0.0000	7.4336
no admit	31.2888	7.4336	0.0000

Linear Discriminant Function for Group

	admit	border	no admit
Constant	-240.37	-177.32	-133.90
GPA	106.25	92.67	78.09
GMAT	0.21	0.17	0.17

How would we classify a student who had a 2.5 GPA and a 500 GMAT?

Plug these values into each discriminant equation. The one with the largest value receives the classification.

Summary of Misclassified Observations

Observation	True Group	Pred Group	X-val Group	Group	Squared Distance		Probability	
					Pred	X-val	Pred	X-val
1 **	admit	admit	border	admit	5.490	6.221	0.59	0.49
				border	6.207	6.140	0.41	0.51
				no admit	15.030	15.235	0.00	0.01
2 **	admit	border	border	admit	4.9246	5.5379	0.12	0.09
				border	0.9477	0.9944	0.88	0.90
				no admit	13.0965	13.5209	0.00	0.00
3 **	admit	border	border	admit	3.199	3.516	0.37	0.33
				border	2.096	2.144	0.63	0.66
				no admit	16.719	16.935	0.00	0.00
24 **	admit	border	border	admit	4.803	5.392	0.48	0.40
				border	4.616	4.561	0.52	0.60
				no admit	23.691	23.428	0.00	0.00
31 **	admit	border	border	admit	3.762	4.166	0.30	0.26
				border	2.034	2.069	0.70	0.74
				no admit	17.005	17.160	0.00	0.00
58 **	no admit	border	border	admit	21.099	20.955	0.00	0.00
				border	2.078	2.056	0.75	0.80
				no admit	4.329	4.865	0.24	0.20
59 **	no admit	border	border	admit	18.950	18.886	0.00	0.00
				border	1.521	1.504	0.87	0.90
				no admit	5.278	6.008	0.13	0.10
66 **	border	admit	admit	admit	6.937	6.909	0.53	0.69
				border	7.206	8.474	0.47	0.31
				no admit	28.744	31.197	0.00	0.00
75 **	border	border	no admit	admit	17.504	17.572	0.00	0.00
				border	1.924	2.107	0.51	0.49
				no admit	2.020	2.035	0.49	0.51

Consider measurements taken on three varieties of Iris.

$\Pi_1 = \text{Iris Setosa}$ ,  $\Pi_2 = \text{Iris Versicolor}$ ,  $\Pi_3 = \text{Iris Virginica}$

And we have four variables,

X1 = Sepal Length      X2 = Sepal Width

X3 = Petal Length      X4 = Petal Width

Is there a difference between the groups? One might answer this using :

MANOVA:

**Analysis of Variance (Balanced Designs)**

Factor	Type	Levels	Values
Iris	fixed	3	1    2    3

Analysis of Variance for sep leng

Source	DF	SS	MS	F	P
Iris	2	63.212	31.606	119.26	0.000
Error	147	38.956	0.265		
Total	149	102.168			

Analysis of Variance for sep widt

Source	DF	SS	MS	F	P
Iris	2	11.3449	5.6725	49.16	0.000
Error	147	16.9620	0.1154		
Total	149	28.3069			

Analysis of Variance for pet leng

Source	DF	SS	MS	F	P
Iris	2	437.10	218.55	1180.16	0.000
Error	147	27.22	0.19		
Total	149	464.33			

Analysis of Variance for pet widt



Classification Function Created Using Variable(s)	Proportion Correctly Classified (using cross-validation method)
X1	0.747
X2	0.520
X3	0.933
X4	0.960
X1, X2	0.793
X1, X3	0.960
X1, X4	0.953
X2, X3	0.953
X2, X4	0.960
X3, X4	0.960
X1, X2, X3	0.960
X1, X2, X4	0.947
X1, X3, X4	0.973
X2, X3, X4	0.967
X1, X2, X3, X4	0.980

Choose the most parsimonious model.

## Discriminant Analysis

Linear Method for Response: Iris  
Predictors: pet widt

Group	1	2	3
Count	50	50	50

### Summary of Classification

Put into	....True Group....		
Group	1	2	3
1	50	0	0
2	0	48	4
3	0	2	46
Total N	50	50	50
N Correct	50	48	46
Proportion	1.000	0.960	0.920

N = 150      N Correct = 144      Proportion Correct = 0.960

### Summary of Classification with Cross-validation

Put into	....True Group....		
Group	1	2	3
1	50	0	0
2	0	48	4
3	0	2	46
Total N	50	50	50
N Correct	50	48	46
Proportion	1.000	0.960	0.920

N = 150      N Correct = 144      Proportion Correct = 0.960

### Squared Distance Between Groups

	1	2	3
1	0.0000	27.8499	75.6513
2	27.8499	0.0000	11.6996
3	75.6513	11.6996	0.0000

### Linear Discriminant Function for Group

	1	2	3
Constant	-0.722	-20.991	-49.003
pet widt	5.874	31.661	48.374

Summary of Misclassified Observations

Observation	True Group	Pred Group	X-val Group	Group	Squared Distance		Probability	
					Pred	X-val	Pred	X-val
71 **	2	3	3	1	57.661	59.483	0.00	0.00
				2	5.365	5.762	0.11	0.10
				3	1.220	1.258	0.89	0.90
78 **	2	3	3	1	50.478	51.325	0.00	0.00
				2	3.340	3.536	0.40	0.38
				3	2.538	2.580	0.60	0.62
120 **	3	2	2	1	37.5467	39.0835	0.00	0.00
				2	0.7229	0.7525	0.95	0.96
				3	6.6061	7.1601	0.05	0.04
130 **	3	2	2	1	43.774	44.824	0.00	0.00
				2	1.793	1.836	0.78	0.80
				3	4.333	4.620	0.22	0.20
134 **	3	2	2	1	37.5467	39.0835	0.00	0.00
				2	0.7229	0.7525	0.95	0.96
				3	6.6061	7.1601	0.05	0.04
135 **	3	2	2	1	31.7971	33.7745	0.00	0.00
				2	0.1307	0.1389	0.99	0.99
				3	9.3568	10.3484	0.01	0.01