WEEK 5 ASSIGNMENT

Concepts of Statistics 2 – DATA-51200 | Spring 2 2020

Christina Morgenstern

---

1. In your own words, discuss (in less than one page) the differences between Multiple Regression Analysis and Multiple Discriminant Analysis.

Multiple Regression Analysis (MRA) and Multiple Discriminant Analysis (MDA) are both multivariate statistical techniques examining dependence relationships among the implicated variables. The major distinction between the two approaches is the nature of the single dependent variable which in the case of MRA is metric and nonmetric for MDA. Thus, MDA allows to predict a categorical dependent variable based on a set of independent metric variables while MRA deals with a continuous dependent variable.

Another distinction between MRA and MDA is the number of variates in a single analysis. While MRA is limited to a single variate, MDA can deal with several variates through calculating a discriminant function for each one.

When assessing model fit, the classification matrix and hit ratio value replace the R^2 value in MDA. The error in MRA is calculated as the residual between actual value and predicted value. Errors in MDA are evident from the classification matrix as misclassified examples.

2. For the data set associated with this homework (HBAT and HBAT_Test (you may use any software and programming language you feel comfortable dealing with. Make sure to include your codes, diagrams and results). Using X4 as the non-metric variable and (X6 up to X18) as the metric variables:

a. What does each variable represent? (go back to chapter 1)

The dependent nonmetric variable X4 in the HBAT survey represents the region i.e. the location of the customer which can be either 0 for being a USA/North America resident or 1 for customers located outside North America.

The metric independent variables X6 to X18 encode different HBAT business functions which were rated by customers on a metric scale from 0 (poor) to 10 (excellent):

Variable X6 denotes the perceived level of quality of HBAT´s paper products.
The overall image and user-friendliness of the HBAT´s website is stored in variable X7.
Variable X8 assesses the extent to which technical support is offered to help solve product and service issues.
The extent to which any complaints are resolved in a timely and complete manner is measured by variable X9 complaint resolution.
Variable X10 measures the perceptions of HBAT´s advertising campaigns in all types of media.
The depth and breadth of HBAT´s product line is stored in variable X11.
Variable X12 stands for the overall image of HBAT´s salesforce.
Competitive pricing is represented by variable X13 and assesses the extent to which HBAT offers competitive prices.
Warranty and claims issues are stored in variable X14.
The extent to which HBAT develops and sells new products is perceived with variable X15.
Efficiency in ordering and billing is measured using variable X16.

Variable X17 denotes the perceived willingness of HBAT sales representatives to negotiate price on purchases of paper products.
Variable X18 measures the amount of time it takes to deliver the paper products once an order has been confirmed.

b. How many groups does X4 has?

The variable X4 has two groups: 0 and 1, for customers located in USA/North America and outside North America, respectively.

c. Apply linear discriminant analysis to the data (HBAT) and find:

Using SAS studio, I applied Linear Discriminant Analysis using X4 as the dependent categorical variable and 13 independent metric variables X6 to X18.

The following SAS code was generated:

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '18/04/2020 21:16'
 * Generated by 'sasdemo'
 * Generated on server 'LOCALHOST'
 * Generated on SAS platform 'Linux LIN X64 2.6.32-754.6.3.el6.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 *  Generated on browser 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6)
AppleWebKit/605.1.15 (KHTML, like Gecko) Version/13.1 Safari/605.1.15'
 *               Generated                on              web              client
'http://localhost:10080/SASStudio/38/main?locale=en_GB&zone=GMT%252B02%253A00&http%
3A%2F%2Flocalhost%3A10080%2FSASStudio%2F38%2F='
 *
 */

ods noproctitle;

proc discrim data=WORK.IMPORT pool=yes crossvalidate crosslisterr distance
             posterr list listerr;
      class x4;
      var x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 x16 x17 x18;
      priors prop;
run;
```

• The linear discriminant function for X4.

The discriminant function for the variable X4 and the two classes, 0 (outside inside the USA) and 1 (outside the USA) is shown in Table 1. For every independent variable, discriminant weights are calculated which represent the contribution of each variable. The strongest effect on the discriminant function exhibit variables X11, X14, X17 and X18. Thus, the product line, the salesforce image, price flexibility and delivery speed seem to strongly impact group distinction.

Linear Discriminant Function for x4			
Variable	Label	0	1
Constant		-191.92135	-194.33409
x6	x6	8.32797	7.65248
x7	x7	4.20485	1.20101
x8	x8	-2.06370	-2.09852
x9	x9	-3.62428	-3.60295
x10	x10	-1.62571	-2.03642
x11	x11	58.68681	58.34999
x12	x12	1.69711	4.70052
x13	x13	3.64274	4.22872
x14	x14	13.54926	13.38761
x15	x15	0.00591	0.28730
x16	x16	-2.84573	-2.25264
x17	x17	62.42839	64.51335
x18	x18	-101.40047	-103.57123

Table 1. Linear Discriminant Function for X4.

• By applying the LDF to the training data (HBAT): How many observations were misclassified? What are they? Find the confusion matrix and the probability of (error) misclassification.

Table 2 shows the classification matrix listing the number of correct and wrong classifications. Out of 100 observations, 5 observations were misclassified. Although the true class of these observations (observations 3, 22, 38, 60, 94) is 1 (outside the USA/North America), the linear discriminant analysis (LDA) classified the observations as class 0 (outside USA/North America).

Classification Results for Calibration Data: WORK.IMPORT
Resubstitution Results using Linear Discriminant Function

Number of Observations and Average Posterior Probabilities Classified into x4		
From x4	0	1
0	39 0.8790	0 .
1	5 0.8463	56 0.9453
Total	44 0.8753	56 0.9453
Priors	0.39	0.61

Table 2. Classification matrix for LDA of HBAT training data set.

The rate of making this mistake, a Type I error, is with 0.0820 (approx. 8%) relatively low. See Table 3 for the error count estimates of variable x4 and the two classes.

Error Count Estimates for x4			
	0	1	Total
Rate	0.0000	0.0820	0.0500
Priors	0.3900	0.6100	

Table 3. Probability of misclassification for LDA on HBAT training data set.

• By applying the LDF to the test data (HBAT_Test): How many observations were misclassified? What are they? Find the confusion matrix and the probability of (error) misclassification.

Load the HBAT_Test data into SAS (WORK.IMPORT1) and perform LDA using this test set.

Generated SAS code:

```
ods noproctitle;

proc discrim data=WORK.IMPORT testdata=WORK.IMPORT1 pool=yes crossvalidate
             crosslisterr distance posterr list listerr testlist testlisterr;
      class x4;
      var x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 x16 x17 x18;
      priors prop;
run;
```

When using the previously trained LDA model on the unseen test data set, the number of misclassifications increased to 13 with 11 observations classified as class 0 although being class 1 (Type I error) and 2 observations classified as class 1 although being class 0 (Type II error). See Table 4 for the results of the classification and the error for misclassification. The rate of committing a Type I or Type II error is 0.1803 and 0.0513, respectively. Thus, the model does well on classifying customers in the USA/North America market, whereas it makes mistakes in classifying customers outside the USA/North America.

Number of Observations and Percent Classified into x4			
From x4	0	1	Total
0	37 94.87	2 5.13	39 100.00
1	11 18.03	50 81.97	61 100.00
Total	48 48.00	52 52.00	100 100.00
Priors	0.39	0.61	

Error Count Estimates for x4			
	0	1	Total
Rate	0.0513	0.1803	0.1300
Priors	0.3900	0.6100	

Table 4. Classification matrix and probability of misclassification for LDA on HBAT test data set.

Total Sample Size	100	DF Total	99
Variables	13	DF Within Classes	98
Classes	2	DF Between Classes	1

Number of Observations Read	100
Number of Observations Used	100

Class Level Information

x4	Variable Name	Frequency	Weight	Proportion	Prior Probability
0	_0	39	39.0000	0.390000	0.390000
1	_1	61	61.0000	0.610000	0.610000

Pooled Covariance Matrix Information

Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
13	-8.01842

Squared Distance to x4

From x4	0	1
0	0	7.45466
1	7.45466	0

F Statistics, NDF=13, DDF=86 for Squared Distance to x4

From x4	0	1
0	0	11.97157
1	11.97157	0

Prob > Mahalanobis Distance for Squared Distance to x4

From x4	0	1
0	1.0000	<.0001
1	<.0001	1.0000

Generalized Squared Distance to x4

From x4	0	1
0	1.88322	8.44325
1	9.33787	0.98859

Linear Discriminant Function for x4

Variable	Label	0	1
Constant		-191.92135	-194.33409
x6	x6	8.32797	7.65248
x7	x7	4.20485	1.20101
x8	x8	-2.06370	-2.09852
x9	x9	-3.62428	-3.60295
x10	x10	-1.62571	-2.03642
x11	x11	58.68681	58.34999
x12	x12	1.69711	4.70052
x13	x13	3.64274	4.22872
x14	x14	13.54926	13.38761
x15	x15	0.00591	0.28730
x16	x16	-2.84573	-2.25264
x17	x17	62.42839	64.51335
x18	x18	-101.40047	-103.57123

Classification Results for Calibration Data: WORK.IMPORT
Resubstitution Results using Linear Discriminant Function

Posterior Probability of Membership in x4

Obs	From x4	Classified into x4		0	1
1	1	1		0.0103	0.9897
2	0	0		0.9989	0.0011
3	1	0	*	0.8181	0.1819
4	1	1		0.0566	0.9434
5	0	0		0.8269	0.1731
6	1	1		0.0123	0.9877
7	1	1		0.0008	0.9992
8	1	1		0.0004	0.9996
9	1	1		0.0004	0.9996
10	1	1		0.0298	0.9702
11	0	0		0.7504	0.2496
12	1	1		0.0050	0.9950
13	0	0		0.9220	0.0780
14	0	0		0.9990	0.0010
15	1	1		0.0274	0.9726
16	0	0		0.9980	0.0020
17	1	1		0.0309	0.9691
18	1	1		0.0072	0.9928
19	1	1		0.0001	0.9999
20	1	1		0.0009	0.9991
21	1	1		0.0103	0.9897
22	1	0	*	0.9920	0.0080
23	0	0		0.9929	0.0071
24	1	1		0.1817	0.8183
25	1	1		0.0237	0.9763
26	1	1		0.0003	0.9997
27	0	0		0.9952	0.0048
28	1	1		0.0171	0.9829
29	0	0		0.9868	0.0132
30	1	1		0.0130	0.9870
31	0	0		0.9818	0.0182
32	1	1		0.4376	0.5624
33	1	1		0.0843	0.9157
34	1	1		0.0059	0.9941
35	1	1		0.0011	0.9989
36	0	0		0.9980	0.0020
37	0	0		0.9397	0.0603
38	1	0	*	0.8503	0.1497
39	1	1		0.0023	0.9977
40	1	1		0.0017	0.9983
41	1	1		0.0023	0.9977
42	0	0		0.6015	0.3985
43	0	0		0.8872	0.1128
44	1	1		0.0002	0.9998
45	0	0		0.9975	0.0025
46	1	1		0.0365	0.9635
47	0	0		0.9970	0.0030
48	1	1		0.0001	0.9999
49	1	1		0.1019	0.8981
50	0	0		0.8055	0.1945
51	1	1		0.1036	0.8964
52	0	0		0.9971	0.0029
53	1	1		0.1996	0.8004
54	0	0		0.9598	0.0402
55	1	1		0.0015	0.9985
56	0	0		0.6783	0.3217
57	1	1		0.0046	0.9954
58	0	0		0.7772	0.2228
59	0	0		0.9990	0.0010
60	1	0	*	0.7572	0.2428
61	0	0		0.9972	0.0028
62	1	1		0.0409	0.9591
63	0	0		0.5946	0.4054
64	1	1		0.4852	0.5148
65	1	1		0.3120	0.6880
66	1	1		0.0003	0.9997
67	1	1		0.0136	0.9864
68	1	1		0.0560	0.9440
69	1	1		0.0143	0.9857
70	1	1		0.0002	0.9998
71	1	1		0.0014	0.9986
72	0	0		0.9811	0.0189
73	1	1		0.0165	0.9835
74	1	1		0.4355	0.5645
75	1	1		0.0249	0.9751
76	0	0		0.8383	0.1617
77	1	1		0.0149	0.9851
78	0	0		0.8987	0.1013
79	0	0		0.8987	0.1013
80	1	1		0.0265	0.9735
81	0	0		0.7325	0.2675
82	0	0		0.8172	0.1828
83	0	0		0.9498	0.0502
84	1	1		0.0005	0.9995
85	0	0		0.6598	0.3402
86	1	1		0.0474	0.9526
87	1	1		0.0020	0.9980
88	0	0		0.5442	0.4558
89	0	0		0.7779	0.2221
90	1	1		0.0003	0.9997
91	0	0		0.8146	0.1854
92	1	1		0.1368	0.8632
93	0	0		0.9986	0.0014
94	1	0	*	0.8139	0.1861
95	0	0		0.8117	0.1883
96	0	0		0.8776	0.1224
97	1	1		0.0062	0.9938
98	0	0		0.9984	0.0016
99	1	1		0.0112	0.9888
100	1	1		0.0029	0.9971

* Misclassified observation

Classification Summary for Calibration Data: WORK.IMPORT
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into x4

From x4	0	1	Total
0	39 100.00	0 0.00	39 100.00
1	5 8.20	56 91.80	61 100.00
Total	44 44.00	56 56.00	100 100.00
Priors	0.39	0.61	

Error Count Estimates for x4

	0	1	Total
Rate	0.0000	0.0820	0.0500
Priors	0.3900	0.6100	

Classification Results for Calibration Data: WORK.IMPORT
Resubstitution Results using Linear Discriminant Function

Number of Observations and Average Posterior Probabilities Classified into x4

From x4	0	1
0	39 0.8790	0 .
1	5 0.8463	56 0.9453
Total	44 0.8753	56 0.9453
Priors	0.39	0.61

Posterior Probability Error Rate Estimates for x4

Estimate	0	1	Total
Stratified	0.0125	0.1321	0.0855
Unstratified	0.0125	0.1321	0.0855
Priors	0.3900	0.6100	

Classification Results for Calibration Data: WORK.IMPORT
Cross-validation Results using Linear Discriminant Function

Posterior Probability of Membership in x4

Obs	From x4	Classified into x4		0	1
3	1	0	*	0.9388	0.0612
22	1	0	*	0.9999	0.0001
24	1	0	*	0.6602	0.3398
32	1	0	*	0.5773	0.4227
38	1	0	*	0.9394	0.0606
42	0	1	*	0.4490	0.5510
60	1	0	*	0.9396	0.0604
63	0	1	*	0.3361	0.6639
64	1	0	*	0.6576	0.3424
74	1	0	*	0.7029	0.2971
81	0	1	*	0.4834	0.5166
88	0	1	*	0.3170	0.6830
94	1	0	*	0.9345	0.0655

* Misclassified observation

Classification Summary for Calibration Data: WORK.IMPORT
Cross-validation Summary using Linear Discriminant Function

Number of Observations and Percent Classified into x4

From x4	0	1	Total
0	35 89.74	4 10.26	39 100.00
1	9 14.75	52 85.25	61 100.00
Total	44 44.00	56 56.00	100 100.00
Priors	0.39	0.61	

Error Count Estimates for x4

	0	1	Total
Rate	0.1026	0.1475	0.1300
Priors	0.3900	0.6100	

Classification Results for Calibration Data: WORK.IMPORT
Cross-validation Results using Linear Discriminant Function

Number of Observations and Average Posterior Probabilities Classified into x4

From x4	0	1
0	35 0.8689	4 0.6036
1	9 0.8167	52 0.9556
Total	44 0.8582	56 0.9305
Priors	0.39	0.61

Posterior Probability Error Rate Estimates for x4

Estimate	0	1	Total
Stratified	0.0317	0.1458	0.1013
Unstratified	0.0317	0.1458	0.1013
Priors	0.3900	0.6100	

Classification Results for Test Data: WORK.IMPORT1
Classification Results using Linear Discriminant Function

Posterior Probability of Membership in x4

Obs	From x4	Classified into x4		0	1
1	1	1		0.0102	0.9898
2	0	0		0.9990	0.0010
3	1	0	*	0.8181	0.1819
4	1	1		0.0566	0.9434
5	0	1	*	0.4568	0.5432
6	1	1		0.0116	0.9884
7	1	1		0.0006	0.9994
8	1	1		0.0005	0.9995
9	1	1		0.0006	0.9994
10	1	1		0.0313	0.9687
11	0	1	*	0.3942	0.6058
12	1	1		0.0050	0.9950
13	0	0		0.9268	0.0732
14	0	0		0.9991	0.0009
15	1	1		0.0274	0.9726
16	0	0		0.9996	0.0004
17	1	1		0.2404	0.7596
18	1	1		0.0271	0.9729
19	1	1		0.0014	0.9986
20	1	1		0.0118	0.9882
21	1	0	*	0.6756	0.3244
22	1	0	*	0.9986	0.0014
23	0	0		0.9995	0.0005
24	1	0	*	0.5897	0.4103
25	1	1		0.4453	0.5547
26	1	1		0.0049	0.9951
27	0	0		0.9991	0.0009
28	1	1		0.3116	0.6884
29	0	0		0.9976	0.0024
30	1	1		0.3453	0.6547
31	0	0		0.9995	0.0005
32	1	0	*	0.9750	0.0250
33	1	1		0.0588	0.9412
34	1	1		0.0078	0.9922
35	1	1		0.0004	0.9996
36	0	0		0.9961	0.0039
37	0	0		0.9229	0.0771
38	1	0	*	0.9082	0.0918
39	1	1		0.0035	0.9965
40	1	1		0.0023	0.9977
41	1	1		0.0023	0.9977
42	0	0		0.6015	0.3985
43	0	0		0.8872	0.1128
44	1	1		0.0002	0.9998
45	0	0		0.9975	0.0025
46	1	1		0.0365	0.9635
47	0	0		0.9970	0.0030
48	1	1		0.0001	0.9999
49	1	1		0.1019	0.8981
50	0	0		0.8055	0.1945
51	1	1		0.1036	0.8964
52	0	0		0.9971	0.0029
53	1	1		0.1996	0.8004
54	0	0		0.9598	0.0402
55	1	1		0.0015	0.9985
56	0	0		0.6783	0.3217
57	1	1		0.0046	0.9954
58	0	0		0.7772	0.2228
59	0	0		0.9990	0.0010
60	1	0	*	0.7572	0.2428
61	0	0		0.9972	0.0028
62	1	1		0.0409	0.9591
63	0	0		0.5946	0.4054
64	1	1		0.4852	0.5148
65	1	1		0.3120	0.6880
66	1	1		0.0003	0.9997
67	1	1		0.0136	0.9864
68	1	1		0.0560	0.9440
69	1	1		0.0926	0.9074
70	1	1		0.0005	0.9995
71	1	1		0.2064	0.7936
72	0	0		0.9845	0.0155
73	1	1		0.0670	0.9330
74	1	0	*	0.9919	0.0081
75	1	1		0.3673	0.6327
76	0	0		0.8883	0.1117
77	1	1		0.0251	0.9749
78	0	0		0.5896	0.4104
79	0	0		0.8533	0.1467
80	1	1		0.0196	0.9804
81	0	0		0.9848	0.0152
82	0	0		0.9482	0.0518
83	0	0		0.9282	0.0718
84	1	1		0.0029	0.9971
85	0	0		0.9156	0.0844
86	1	0	*	0.8741	0.1259
87	1	1		0.2922	0.7078
88	0	0		0.8636	0.1364
89	0	0		0.9432	0.0568
90	1	0	*	0.7074	0.2926
91	0	0		0.6306	0.3694
92	1	1		0.1920	0.8080
93	0	0		0.9986	0.0014
94	1	0	*	0.8139	0.1861
95	0	0		0.8620	0.1380
96	0	0		0.9224	0.0776
97	1	1		0.0175	0.9825
98	0	0		0.9990	0.0010
99	1	1		0.0233	0.9767
100	1	1		0.0029	0.9971

* Misclassified observation

Classification Summary for Test Data: WORK.IMPORT1
Classification Summary using Linear Discriminant Function

Observation Profile for Test Data

Number of Observations Read	100
Number of Observations Used	100

Number of Observations and Percent Classified into x4

From x4	0	1	Total
0	37 94.87	2 5.13	39 100.00
1	11 18.03	50 81.97	61 100.00
Total	48 48.00	52 52.00	100 100.00
Priors	0.39	0.61	

Error Count Estimates for x4

	0	1	Total
Rate	0.0513	0.1803	0.1300
Priors	0.3900	0.6100	

Classification Results for Test Data: WORK.IMPORT1
Classification Results using Linear Discriminant Function

Number of Observations and Average Posterior Probabilities Classified into x4

From x4	0	1
0	37 0.9039	2 0.5745
1	11 0.8282	50 0.9146
Total	48 0.8865	52 0.9015
Priors	0.39	0.61

Posterior Probability Error Rate Estimates for x4

Estimate	0	1	Total
Stratified	-0.0911	0.2315	0.1057
Unstratified	-0.0911	0.2315	0.1057
Priors	0.3900	0.6100	