STAT448 - Advanced Data Analysis

Homework 6

Name: Xinyan Yang

Exercise 1 Solution:

(a).

Multivariate Statistics and F Approximations					
S=2 M=0.5 N=14.5					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.62153127	2.08	8	62	0.0513
Pillai's Trace	0.41001651	2.06	8	64	0.0527
Hotelling-Lawley Trace	0.55817134	2.12	8	42.028	0.0546
Roy's Greatest Root	0.44379952	3.55	4	32	0.0166
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

According to the results of MANOVA tests, the p-values are nearly 0.05, and we can reject the null hypothesis and conclude that this classification model is meaningful.

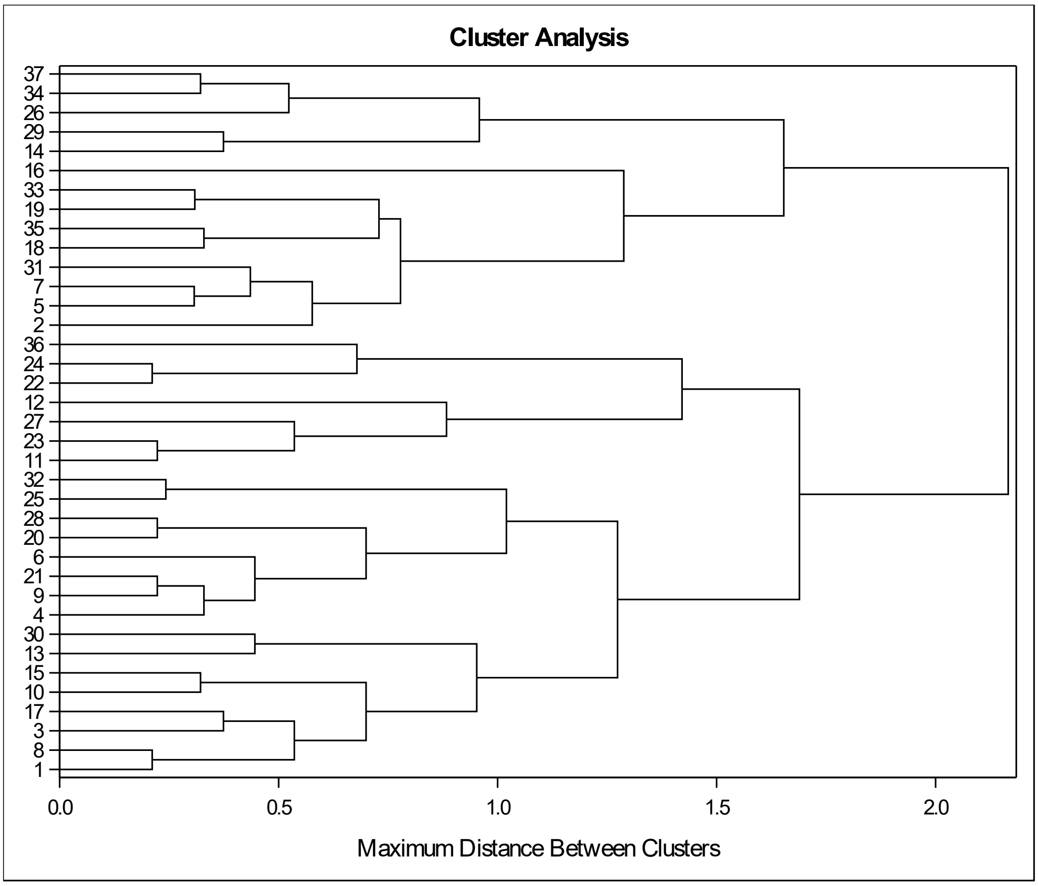
Cross-validation Summary using Linear Discriminant Function

Number of Observations and Percent Classified into proc				
From proc	**aversion therapy**	**immediate stopping**	**tapering**	**Total**
aversion therapy	9 60.00	3 20.00	3 20.00	15 100.00
immediate stopping	6 42.86	7 50.00	1 7.14	14 100.00
tapering	5 62.50	3 37.50	0 0.00	8 100.00
Total	20 54.05	13 35.14	4 10.81	37 100.00
Priors	0.40541	0.37838	0.21622	

Error Count Estimates for proc				
	aversion therapy	**immediate stopping**	**tapering**	**Total**
Rate	0.4000	0.5000	1.0000	0.5676
Priors	0.4054	0.3784	0.2162	

The cross-validation error is 0.5676. It seems like the discrimination doesn't match the proc procedures very well because we can see from the above table that the tapering group has 0 accuracy which means that none of the observations in the tapering group was classified right. Moreover, the aversion therapy and immediate stopping group don't have high accuracy too.

(b).



Based on this dendrogram, I would probably choose 4 clusters because we can see that around maximum distance 1.65-1.70, the observations are split into 4 clusters almost at the same distance.

(c).

Table of CLUSTER by proc				
CLUSTER	**proc**			
Frequency	**aversion therapy**	**immediate stopping**	**tapering**	**Total**
1	6	6	4	16
2	1	6	0	7
3	8	2	4	14
Total	15	14	8	37

We can see that the aversion therapy procedure is prominent in cluster1 and cluster3. And the immediate stopping procedure is more common in cluster1 and 2. The tapering procedure is prominent in cluster 1 and 3.

Therefore the cluster1 has all of three procedures in it. Cluster2 is mainly composed of immediate stopping procedure. And cluster3 is mainly composed of aversion therapy and tapering procedures.

The rating for immediate stopping procedure may be different from the other two procedures because cluster2 has mostly this procedure. Also, the rating for aversion therapy and tapering may have some similarities because these two procedures both appear in cluster1 and cluster3.

Exercise 2 Solution:

(a).

Multivariate Statistics and F Approximations					
S=4 M=-0.5 N=70					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.66358580	3.90	16	434.45	<.0001
Pillai's Trace	0.35330557	3.51	16	580	<.0001
Hotelling-Lawley Trace	0.48181908	4.25	16	278.06	<.0001
Roy's Greatest Root	0.42509538	15.41	4	145	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					

According to the MANOVA test results, we should reject the null hypothesis and conclude that this classification model is meaningful.

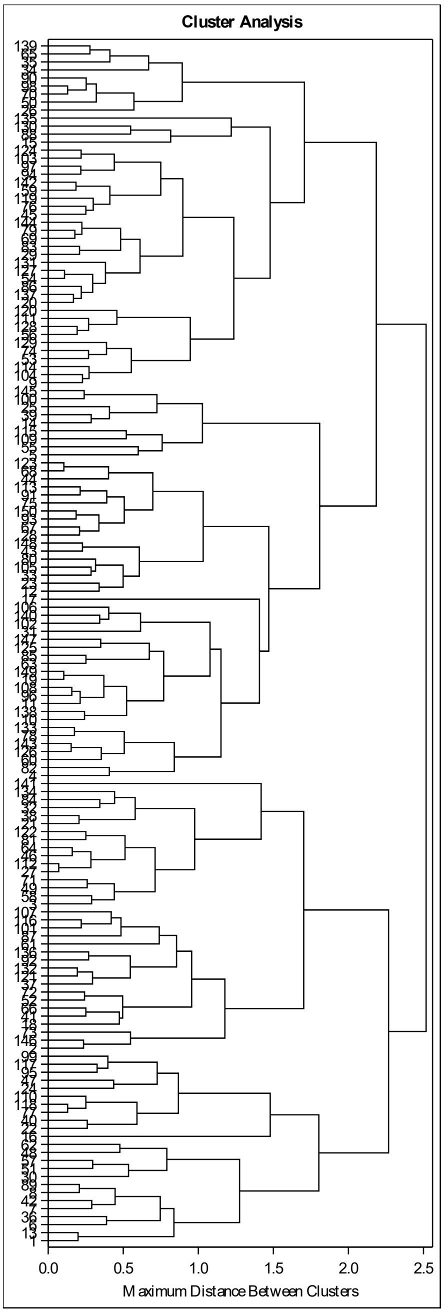
Cross-validation Summary using Linear Discriminant Function

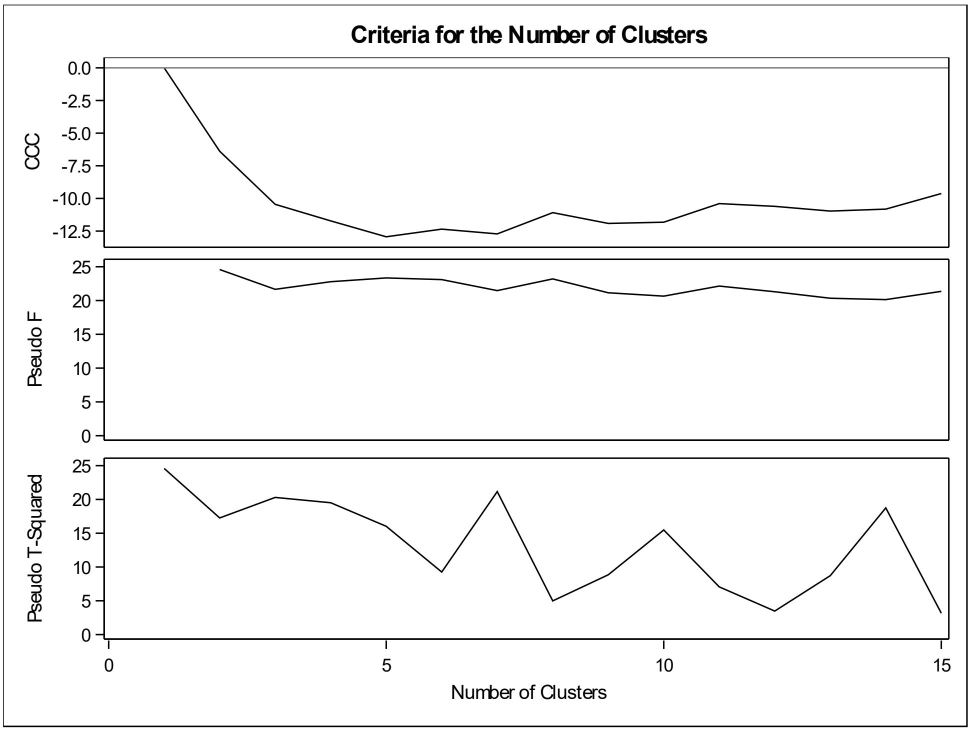
Number of Observations and Percent Classified into epoch						
From epoch	**1**	**2**	**3**	**4**	**5**	**Total**
1	9 30.00	10 33.33	5 16.67	4 13.33	2 6.67	30 100.00
2	11 36.67	7 23.33	5 16.67	4 13.33	3 10.00	30 100.00
3	6 20.00	4 13.33	12 40.00	2 6.67	6 20.00	30 100.00
4	3 10.00	3 10.00	7 23.33	5 16.67	12 40.00	30 100.00
5	2 6.67	4 13.33	4 13.33	10 33.33	10 33.33	30 100.00
Total	31 20.67	28 18.67	33 22.00	25 16.67	33 22.00	150 100.00
Priors	0.2	0.2	0.2	0.2	0.2	

Error Count Estimates for epoch						
	1	2	3	4	5	Total
Rate	0.7000	0.7667	0.6000	0.8333	0.6667	0.7133
Priors	0.2000	0.2000	0.2000	0.2000	0.2000	

The cross-validation error is 0.7133. It seems like the discrimination doesn't match the epochs very well for this high error rate. We can see from the above table that epoch 1 and epoch 2 are misclassified into each other's group. So as epoch 4 and epoch5.

(b).





Based on the dendrogram, I would probably choose 4 or 8 clusters because we can see that around maximum distance 2.0-2.3, the observations are split into 4 clusters almost at the same time. So as the reason for 8 clusters.

Based on the ccc measure, I would choose 6, 8 or 11 clusters. Based on pseudo F, 5, 8 or 11 clusters would be chosen. Based on pseudo T squared, 6, 8 or 12 clusters would be chosen.

To sum up, I would probably choose 8 clusters according to all these measurements.

(c).

Table of CLUSTER by epoch						
CLUSTER	epoch					
Frequency	1	2	3	4	5	Total
1	13	12	5	4	0	34
2	6	6	15	4	9	40
3	6	4	6	16	11	43
4	5	8	4	3	4	24
5	0	0	0	3	6	9
Total	30	30	30	30	30	150

As you can see, epoch 1 and epoch 2 are mostly classified into cluster 1, which means that skull measurements for these too epochs may have some similarities. Also, cluster 3 is mostly composed of epoch 4 and epoch5, which means that these too also have some similarities. We can also notice that cluster2 is mainly composed of epoch 3, so epoch 3 may have some different measures from the other epochs.

Exercise 3 Solution:

Stepwise Selection Summary										
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	bl		0.1864	8.31	<.0001	0.81358903	<.0001	0.04660274	<.0001
2	2	mb		0.1194	4.88	0.0010	0.71644543	<.0001	0.07107651	<.0001

The variables selected are bl and mb.

Then we perform the discriminant analysis using these two variables.

Multivariate Statistics and F Approximations

S=2 M=0.5 N=71

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.71644543	6.53	8	288	<.0001
Pillai's Trace	0.28430603	6.01	8	290	<.0001
Hotelling-Lawley Trace	0.39473084	7.08	8	203.4	<.0001
Roy's Greatest Root	0.39205554	14.21	4	145	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

We can notice that the manova test results become more significant than before.

Number of Observations and Percent Classified into epoch

From epoch	1	2	3	4	5	Total
1	14 46.67	8 26.67	3 10.00	2 6.67	3 10.00	30 100.00
2	12 40.00	6 20.00	9 30.00	2 6.67	1 3.33	30 100.00
3	5 16.67	4 13.33	7 23.33	4 13.33	10 33.33	30 100.00
4	3 10.00	4 13.33	5 16.67	3 10.00	15 50.00	30 100.00
5	3 10.00	3 10.00	6 20.00	4 13.33	14 46.67	30 100.00
Total	37 24.67	25 16.67	30 20.00	15 10.00	43 28.67	150 100.00
Priors	0.2	0.2	0.2	0.2	0.2	

Error Count Estimates for epoch

	1	2	3	4	5	Total
Rate	0.5333	0.8000	0.7667	0.9000	0.5333	0.7067
Priors	0.2000	0.2000	0.2000	0.2000	0.2000	

The cross-validation error rate has decreased a little bit. We can see that the cluster1 is mainly composed of epoch1 and epoch2. And cluster5 is mainly made up of epoch3, 4 and 5. Actually I don’t think this classification result works better than the previous one.

Then we perform cluster analysis using two selected variables.

Table of CLUSTER by epoch						
CLUSTER	**epoch**					
Frequency	**1**	**2**	**3**	**4**	**5**	**Total**
1	10	10	7	10	8	45
2	4	6	13	14	14	51
3	12	11	4	0	1	28
4	2	1	6	5	7	21
5	2	2	0	1	0	5
Total	30	30	30	30	30	150

We can see from the table that epoch1 and epoch2 mainly falls in cluster 1 and cluster3 while epoch4 mainly falls in cluster1 and cluster2. Epoch3 and epoch5 mainly falls in cluster2. Actually I don’t think this clustering model performs better because it still can not split these 5 epochs well.