Assignment 5: Classification Analysis

DA 410

Marjorie Blanco

Problem 9.7 (a)(b)(c)

Do a classification analysis on the beetle data of Table 5.5 as follows:

(a) Find the classification function and and the cutoff point

Let G_1 = Haltica oleracea and G_2 = Haltica carduorum

Mean vectors

```
y_bar1 <- colMeans(oleracea)
y_bar2 <- colMeans(carduorum)
```

$\overline{y_1}$ =

y1	y2	y3	y4
194.4737	267.0526	137.3684	185.9474

$\overline{y_2}$ =

y1	y2	y3	y4
179.55	290.8	157.2	209.25

Covariance matrices

```
S1 <- cov(oleracea)
S2 <- cov(carduorum)
```

S1 =

	y1	y2	y3	y4
y1	187.59649	176.86257	48.37134	113.58187
y2	176.86257	345.38596	75.97953	118.78070
y3	48.37134	75.97953	66.35673	16.24269
y4	113.58187	118.78070	16.24269	239.94152

S2 =

	y1	y2	y3	y4
y1	101.83947	128.06316	36.98947	32.59211
y2	128.06316	389.01053	165.35789	94.36842
y3	36.98947	165.35789	167.53684	66.52632
y4	32.59211	94.36842	66.52632	177.88158

Pooled covariance matrix

```
Spl <- (1/ (nrow(oleracea) + nrow(carduorum) - 2)) *
  ((nrow(oleracea) - 1) * S1 + (nrow(carduorum) - 1) * S2)
```

Spl =

	y1	y2	y3	y4
y1	143.55910	151.8034	42.52660	71.99253
y2	151.80341	367.7878	121.87653	106.24467
y3	42.52660	121.8765	118.31408	42.06401
y4	71.99253	106.2447	42.06401	208.07290

$a' = (y_1 - y_2)' S_{pl}^{-1} y$

```
a_prime <- t(y_bar1 - y_bar2) %*% inv(Spl)
```

a' =

0.345249	-0.1303878	-0.1064337	-0.1433534

```
z_bar1 <- a_prime %*% y_bar1
```

$\overline{z_1}$ =

-8.955386

```
z_bar2 <- a_prime %*% y_bar2
```

$\overline{z_2}$ =

-22.6554

```
z <- (z_bar1 + z_bar2) /2
```

z =

-15.80539

Assign y to G_1 if z $\geq$ -15.8053945

Assign y to G_2 if z $<$ -15.8053945

Find the classification table using the linear classification function in part (a).

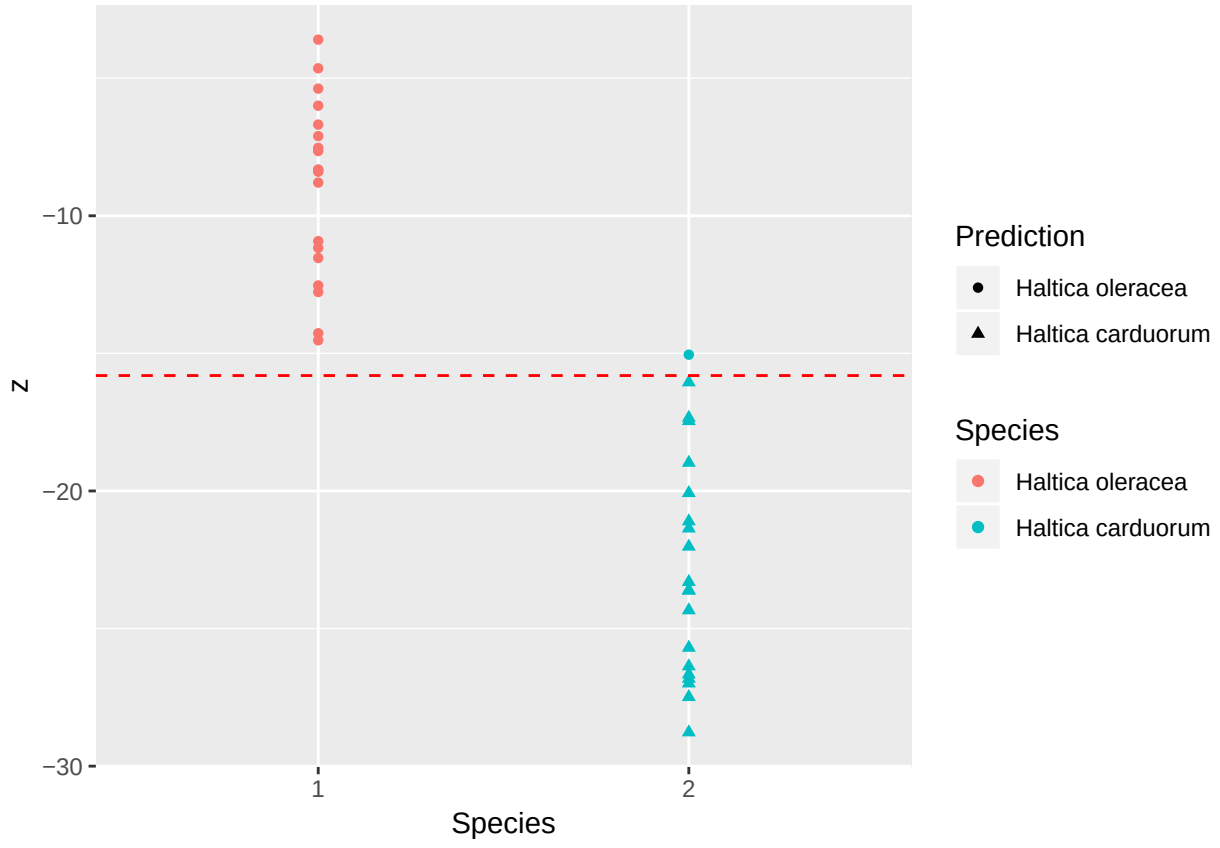
Group	y1	y2	y3	y4	z	prediction
1	189	245	137	163	-4.640979	1
1	192	260	132	217	-12.769966	1
1	217	276	141	192	-3.599013	1
1	221	299	142	213	-8.333793	1
1	171	239	128	158	-8.398465	1
1	192	262	147	173	-8.319697	1
1	213	278	136	201	-5.998797	1
1	192	255	128	185	-7.104982	1
1	170	244	128	192	-14.269669	1
1	201	276	146	186	-8.795046	1
1	195	242	128	192	-5.377667	1
1	205	263	147	192	-6.685562	1
1	180	252	121	167	-7.531410	1
1	192	283	138	183	-11.533472	1
1	200	294	138	188	-10.922513	1
1	192	277	150	177	-11.168229	1
1	200	287	136	173	-7.646630	1
1	181	255	146	183	-12.531822	1
1	192	287	141	198	-14.524626	1
2	181	305	184	209	-26.822884	2
2	158	237	133	188	-17.458697	2
2	184	300	166	231	-26.373166	2
2	171	273	162	213	-24.334835	2
2	181	297	163	224	-25.694975	2
2	181	308	160	223	-26.666586	2
2	177	301	166	221	-27.486763	2
2	198	308	141	197	-15.047923	1
2	180	286	146	214	-21.363050	2
2	177	299	171	192	-23.600907	2
2	176	317	166	213	-28.771390	2
2	192	312	166	209	-22.022053	2
2	176	285	141	200	-20.074542	2
2	169	287	162	214	-26.994117	2
2	164	265	147	192	-21.101548	2
2	181	308	157	204	-23.623570	2
2	192	276	154	209	-16.050886	2
2	181	278	149	235	-23.304421	2
2	175	271	140	192	-17.341100	2
2	197	303	170	205	-18.974638	2

(b) Find the classification table using the linear classification function in part (a).

Classification Table for the Beetle Data of Table 5.5

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  2
##          1 19  0
##          2  1 19
##
```

```
##                Accuracy : 0.9744
##                  95% CI : (0.8652, 0.9994)
##     No Information Rate : 0.5128
##     P-Value [Acc > NIR] : 1.858e-10
##
##                   Kappa : 0.9488
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9500
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 0.9500
##              Prevalence : 0.5128
##          Detection Rate : 0.4872
##    Detection Prevalence : 0.4872
##       Balanced Accuracy : 0.9750
##
##        'Positive' Class : 1
##
```



(c) Find the classification table using the nearest neighbor method. Assign 3rd point from two datasets into group1 or group2.

We compare y_3 to each y_i, i=1, 2,... ,k, by the distance function

$D_i^2(y_3) = (y_3 - \overline{y}_i)' S_{pl}^{-1} (y_3 - \overline{y}_i)$

and assign y to the group for which $D_i^2(y_3)$ is smallest.

```
for (i in 1:nrow(bettle_matrix))
{
   if (i != 3)
     bettle_df[i,"Distance"] <- t((bettle_matrix[3,] - bettle_matrix[i,])
                                  %*% solve(Spl)
                                  %*% (bettle_matrix[3,]- bettle_matrix[i,]))
}
```

Let $y_3 =$

y1	y2	y3	y4
217	276	141	192

This table shows the distance between vector 3 from G_1 and all other vectors.

ID	Group	y1	y2	y3	y4	z	prediction	Distance
1	1	189	245	137	163	-4.640979	1	7.186748
2	1	192	260	132	217	-12.769966	1	13.359185
4	1	221	299	142	213	-8.333793	1	4.241619
5	1	171	239	128	158	-8.398465	1	16.459603
6	1	192	262	147	173	-8.319697	1	6.627530
7	1	213	278	136	201	-5.998797	1	1.493747
8	1	192	255	128	185	-7.104982	1	5.441329
9	1	170	244	128	192	-14.269669	1	20.140582
10	1	201	276	146	186	-8.795046	1	3.324057
11	1	195	242	128	192	-5.377667	1	5.008894
12	1	205	263	147	192	-6.685562	1	2.357370
13	1	180	252	121	167	-7.531410	1	13.701430
14	1	192	283	138	183	-11.533472	1	10.577422
15	1	200	294	138	188	-10.922513	1	9.918989
16	1	192	277	150	177	-11.168229	1	8.800756
17	1	200	287	136	173	-7.646630	1	8.624116
18	1	181	255	146	183	-12.531822	1	11.737611
19	1	192	287	141	198	-14.524626	1	12.821197
20	2	181	305	184	209	-26.822884	2	44.300984
21	2	158	237	133	188	-17.458697	2	29.827634
22	2	184	300	166	231	-26.373166	2	38.865076
23	2	171	273	162	213	-24.334835	2	36.309813
24	2	181	297	163	224	-25.694975	2	37.039060
25	2	181	308	160	223	-26.666586	2	41.757627
26	2	177	301	166	221	-27.486763	2	43.414263
27	2	198	308	141	197	-15.047923	1	18.488708
28	2	180	286	146	214	-21.363050	2	28.069688
29	2	177	299	171	192	-23.600907	2	36.161963
30	2	176	317	166	213	-28.771390	2	51.391438
31	2	192	312	166	209	-22.022053	2	26.510273
32	2	176	285	141	200	-20.074542	2	28.491238
33	2	169	287	162	214	-26.994117	2	44.015818
34	2	164	265	147	192	-21.101548	2	31.943316
35	2	181	308	157	204	-23.623570	2	35.073377
36	2	192	276	154	209	-16.050886	2	12.752907
37	2	181	278	149	235	-23.304421	2	34.980377
38	2	175	271	140	192	-17.341100	2	21.399913
39	2	197	303	170	205	-18.974638	2	18.906154

Using k = 2 the group assign to vector 3 to G_1.

Let k = 2 for our k-Nearest Neighbor algorithm. This means we classify y_3 according to the two point in the training set it is closet to. In this case, y_3 is closest to y_7 and y_{12}, and therefore we classify y_3 to G_1.

ID	Group	y1	y2	y3	y4	Distance
7	1	213	278	136	201	1.493747
12	1	205	263	147	192	2.357370

Using k = 3 the group assign to vector 3 to G_1.

Let k = 3 for our k-Nearest Neighbor algorithm. This means we classify y_3 according to the three point in

the training set it is closet to. In this case, y_3 is closest to y_7, y_{10} and y_{12}, and therefore we classify y_3 to G_1.

ID	Group	y1	y2	y3	y4	Distance
7	1	213	278	136	201	1.493747
12	1	205	263	147	192	2.357370
10	1	201	276	146	186	3.324057

Using k = 4 the group assign to vector 3 to G_1.

Let k = 4 for our k-Nearest Neighbor algorithm. This means we classify y_3 according to the four point in the training set it is closet to. In this case, y_3 is closest to y_7, y_{10}, y_{12} and y_4, and therefore we classify y_3 to G_1.

ID	Group	y1	y2	y3	y4	Distance
7	1	213	278	136	201	1.493747
12	1	205	263	147	192	2.357370
10	1	201	276	146	186	3.324057
4	1	221	299	142	213	4.241619

Using k = 5 the group assign to vector 3 to G_1.

Let k = 5 for our k-Nearest Neighbor algorithm. This means we classify y_3 according to the five point in the training set it is closet to. In this case, y_3 is closest to y_7, y_{10}, y_{12}, y_4 and y_{11}, and therefore we classify y_3 to G_1.

ID	Group	y1	y2	y3	y4	Distance
7	1	213	278	136	201	1.493747
12	1	205	263	147	192	2.357370
10	1	201	276	146	186	3.324057
4	1	221	299	142	213	4.241619
11	1	195	242	128	192	5.008894

We compare y_{22} to each y_i, i=1, 2,. . . ,k, by the distance function

$D_i^2(y_{22}) = (y_{22} - \overline{y}_i)' S_{pl}^{-1} (y_{22} - \overline{y}_i)$

and assign y to the group for which $D_i^2(y_{22})$ is smallest.

```
for (i in 1:nrow(bettle_df))
{
   if (i != 22)
     bettle_df[i,"Distance"] <- t((bettle_matrix[22,] - bettle_matrix[i,])
                                  %*% solve(Spl)
                                  %*% (bettle_matrix[22,]- bettle_matrix[i,]))
}
```

Let y_{22} =

y1	y2	y3	y4
184	300	166	231

This table shows the distance between vector 3 from G_2 and all other vectors.

ID	Group	y1	y2	y3	y4	z	prediction	Distance
1	1	189	245	137	163	-4.640979	1	40.7341707
2	1	192	260	132	217	-12.769966	1	16.6109996
3	1	217	276	141	192	-3.599013	1	38.8650757
4	1	221	299	142	213	-8.333793	1	27.3155849
5	1	171	239	128	158	-8.398465	1	35.4340213
6	1	192	262	147	173	-8.319697	1	29.0935742
7	1	213	278	136	201	-5.998797	1	31.1068119
8	1	192	255	128	185	-7.104982	1	29.1682912
9	1	170	244	128	192	-14.269669	1	17.3931606
10	1	201	276	146	186	-8.795046	1	24.5301446
11	1	195	242	128	192	-5.377667	1	33.6595782
12	1	205	263	147	192	-6.685562	1	29.9901650
13	1	180	252	121	167	-7.531410	1	35.5914825
14	1	192	283	138	183	-11.533472	1	21.8066227
15	1	200	294	138	188	-10.922513	1	23.9299938
16	1	192	277	150	177	-11.168229	1	22.4349862
17	1	200	287	136	173	-7.646630	1	34.0815413
18	1	181	255	146	183	-12.531822	1	18.2726217
19	1	192	287	141	198	-14.524626	1	13.5062845
20	2	181	305	184	209	-26.822884	2	7.0030128
21	2	158	237	133	188	-17.458697	2	15.9581388
23	2	171	273	162	213	-24.334835	2	2.9379156
24	2	181	297	163	224	-25.694975	2	0.2898419
25	2	181	308	160	223	-26.666586	2	2.0672898
26	2	177	301	166	221	-27.486763	2	1.0414160
27	2	198	308	141	197	-15.047923	1	18.8806698
28	2	180	286	146	214	-21.363050	2	4.3830684
29	2	177	299	171	192	-23.600907	2	9.0526574
30	2	176	317	166	213	-28.771390	2	5.9202830
31	2	192	312	166	209	-22.022053	2	4.7662794
32	2	176	285	141	200	-20.074542	2	9.2692307
33	2	169	287	162	214	-26.994117	2	2.1908846
34	2	164	265	147	192	-21.101548	2	8.5822387
35	2	181	308	157	204	-23.623570	2	6.5498073
36	2	192	276	154	209	-16.050886	2	8.2511501
37	2	181	278	149	235	-23.304421	2	3.3795586
38	2	175	271	140	192	-17.341100	2	11.0426743
39	2	197	303	170	205	-18.974638	2	7.9075827

Using k = 2 the group assign to vector 3 to G_2.

Let k = 2 for our k-Nearest Neighbor algorithm. This means we classify y_{22} according to the two point in the training set it is closet to. In this case, y_{22} is closest to y_{24} and y_{26}, and therefore we classify y_{22} to G_2.

ID	Group	y1	y2	y3	y4	Distance
24	2	181	297	163	224	0.2898419
26	2	177	301	166	221	1.0414160

Using k = 3 the group assign to vector 3 to G_2.

Let k = 3 for our k-Nearest Neighbor algorithm. This means we classify y_{22} according to the three point in

the training set it is closet to. In this case, y_{22} is closest to y_{24}, y_{26}, and y_{25}, and therefore we classify y_{22} to G_2.

ID	Group	y1	y2	y3	y4	Distance
24	2	181	297	163	224	0.2898419
26	2	177	301	166	221	1.0414160
25	2	181	308	160	223	2.0672898

Using k = 4 the group assign to vector 3 to G_2.

Let k = 4 for our k-Nearest Neighbor algorithm. This means we classify y_{22} according to the four point in the training set it is closet to. In this case, y_{22} is closest to y_{24}, y_{26}, y_{25}, and y_{33}, and therefore we classify y_{22} to G_2.

ID	Group	y1	y2	y3	y4	Distance
24	2	181	297	163	224	0.2898419
26	2	177	301	166	221	1.0414160
25	2	181	308	160	223	2.0672898
33	2	169	287	162	214	2.1908846

Using k = 5 the group assign to vector 3 to G_2.

Let k = 5 for our k-Nearest Neighbor algorithm. This means we classify y_{22} according to the five point in the training set it is closet to. In this case, y_{22} is closest to y_{24}, y_{26}, y_{25}, y_{33} and y_{23}, and therefore we classify y_{22} to G_2.

ID	Group	y1	y2	y3	y4	Distance
24	2	181	297	163	224	0.2898419
26	2	177	301	166	221	1.0414160
25	2	181	308	160	223	2.0672898
33	2	169	287	162	214	2.1908846
23	2	171	273	162	213	2.9379156