HW 9

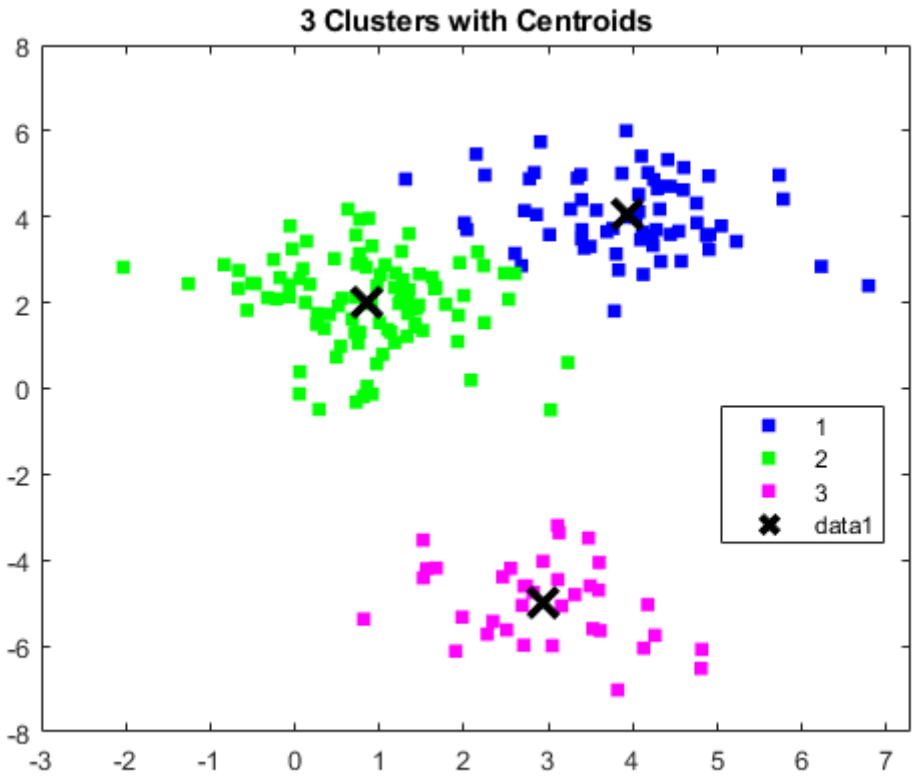
Adam Karl

April 8, 2021

1 K-means clustering

a.

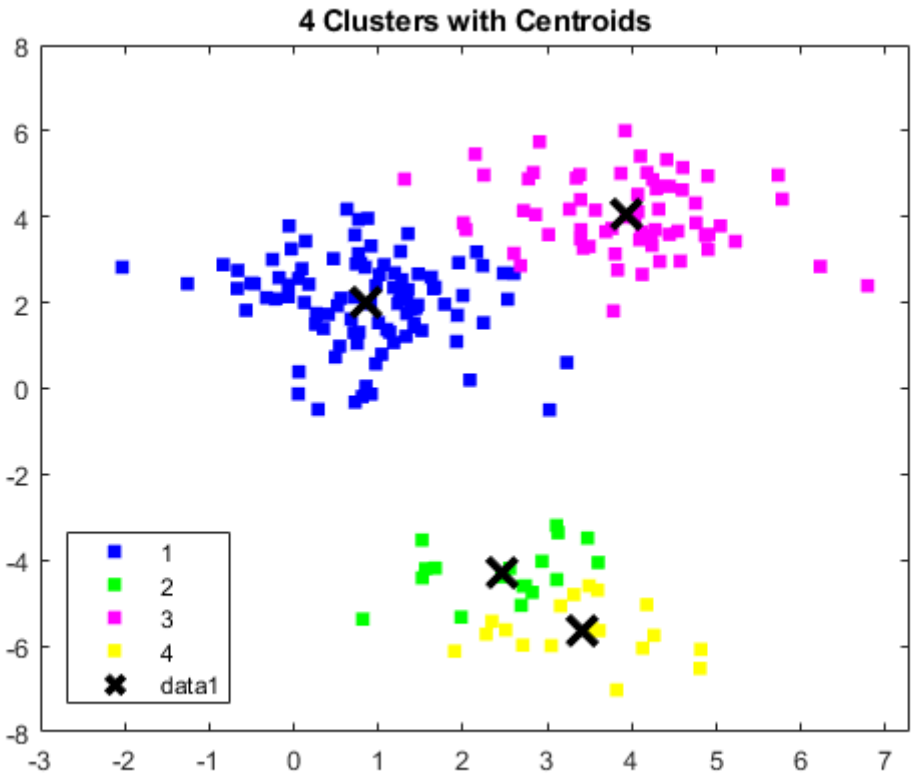
- cluster 1
  - 66 elements
  - center: (3.94, 4.04)
- cluster 2
  - 98 elements
  - center: (0.86, 2.03)
- cluster 3
  - 36 elements
  - center: (2.94. -4.97)



b.

- cluster 1

  - 99 elements
  - center: (0.86, 2.03)
- cluster 2
  - 18 elements
  - center: (2.47, -4.29)
- cluster 3
  - 66 elements
  - center: (3.94, 4.04)
- cluster 4
  - 18 elements
  - center: (3.42, -5.65)

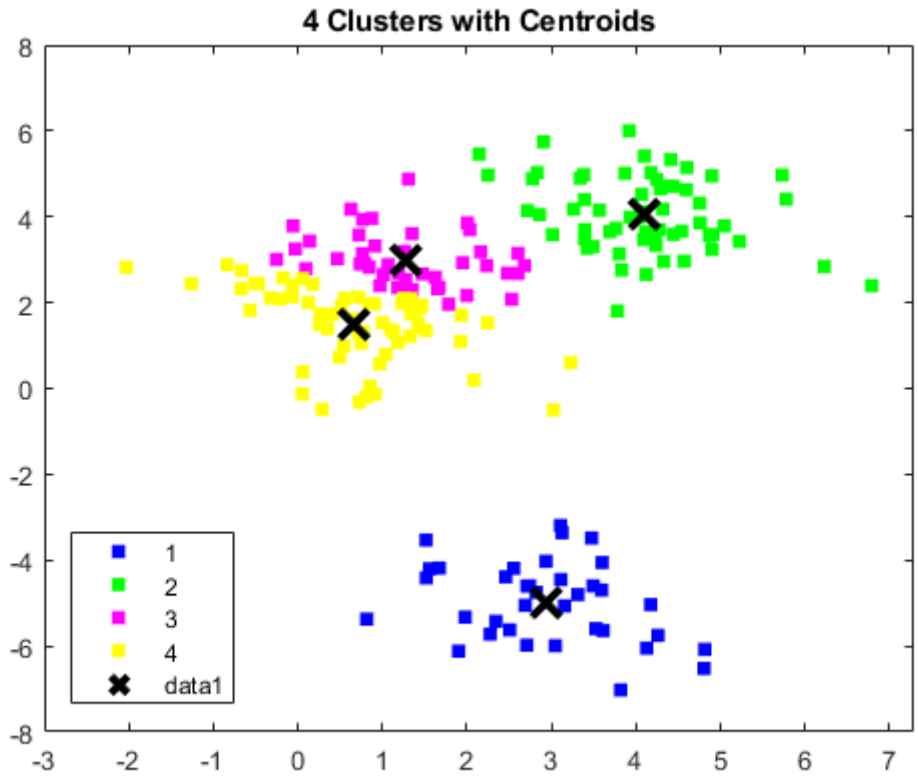


c.

- cluster 1
  - 36 elements
  - (2.94, -4.97)
- cluster 2
  - 61 elements
  - (4.09, 4.07)
- cluster 3
  - 42 elements

  - (1.29, 3.00)

- cluster 4
  - 61 elements
  - (0.67, 1.50)



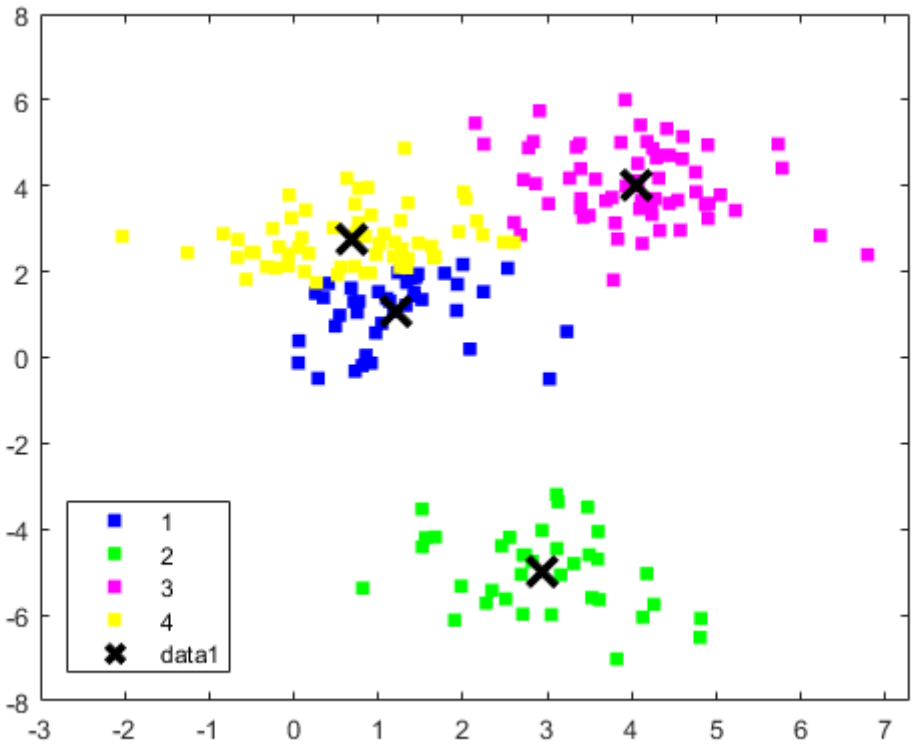
d. The k-means algorithm seeks to minimize the sum of the squared distances from every data point to its respective center. Thus, we can compare the kmeans models by looking for the one with the lower value of:

$$\sum_{i=1}^{k} \sum_{x_j \epsilon S_i} ||x_j - u_i||^2$$

Where u_i is the respective center of cluster S_i.
e. Here are the cluster sizes for each of the 30 trials:

	1	2	3	4
1	66	98	26	10
2	95	36	31	38
3	13	98	66	23
4	98	10	26	66
5	46	36	29	89
6	63	36	69	32
7	40	36	96	28
8	29	82	36	53
9	66	18	98	18
10	98	66	26	10
11	98	66	18	18
12	92	33	36	39
13	96	36	40	28
14	66	13	98	23
15	98	18	18	66
16	61	36	61	42
17	66	18	98	18
18	66	98	10	26
19	56	47	36	61
20	63	36	69	32
21	40	36	63	61
22	36	47	91	26
23	60	65	39	36
24	61	56	36	47
25	63	32	36	69
26	36	60	50	54
27	29	36	89	46
28	52	61	51	36
29	98	13	66	23
30	96	36	40	28

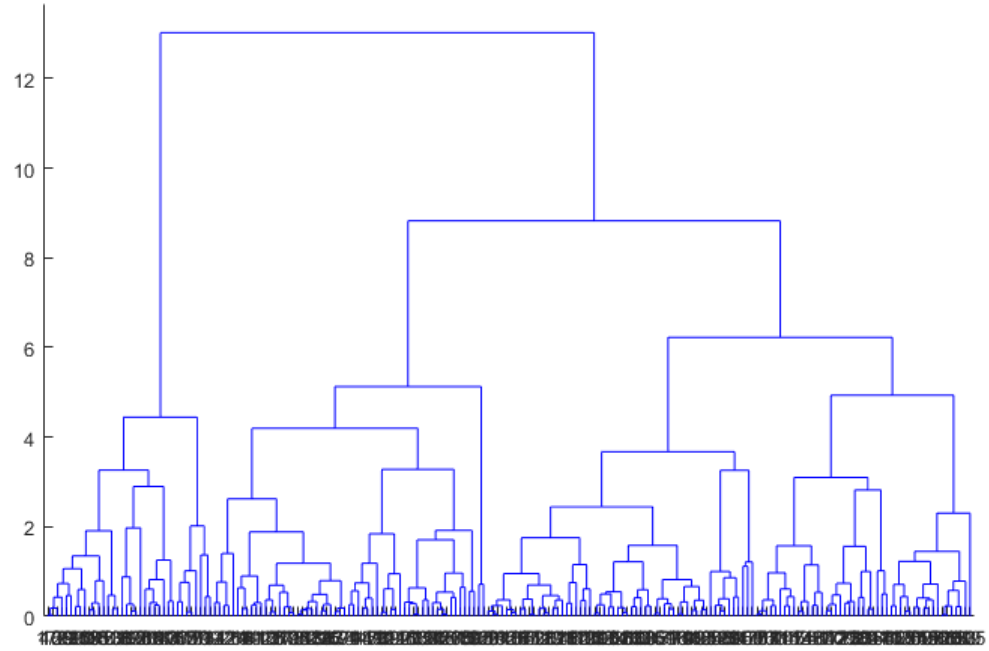
The best model resulted in a total squared distance of 281.9, and clusters of sizes 40, 36, 63, and 61.



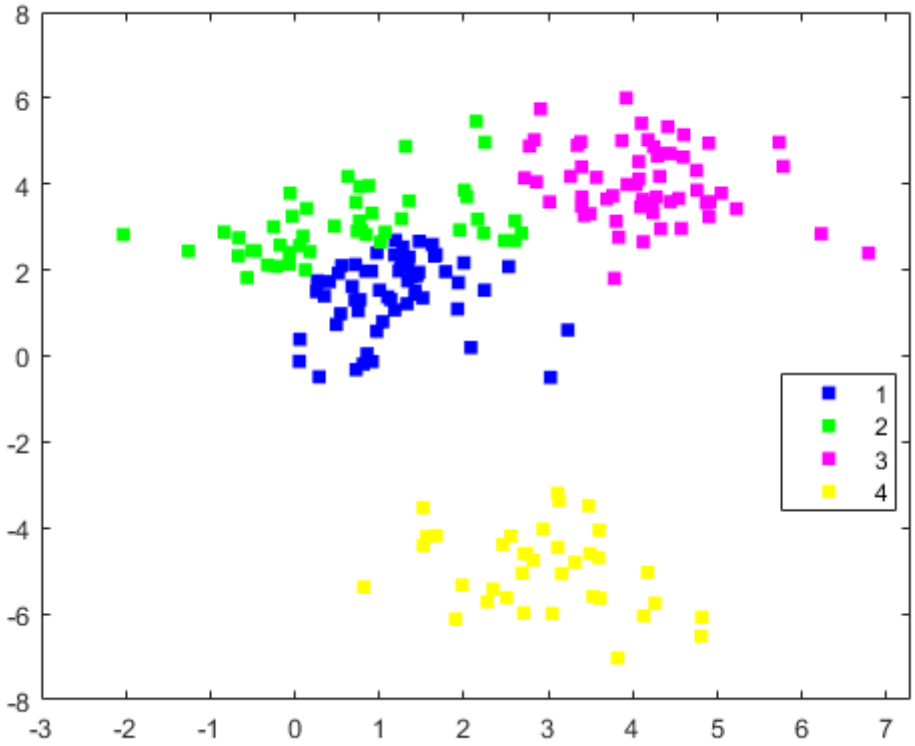
Best model

2 Hierarchical clustering

a.



b. Using the hierarchical model to split the data into 4 groups, the clusters look like:



This appears to be practically identical to the best model using the k-means approach.

3 Feature/Input Ranking

a. The top 20 dimensions according to their Fisher scores are as follows:

	1	2
1	48	0.3192
2	25	0.2140
3	21	0.1910
4	70	0.1892
5	65	0.1693
6	40	0.1673
7	29	0.1650
8	19	0.1402
9	57	0.1255
10	20	0.1212
11	24	0.0995
12	30	0.0950
13	12	0.0858
14	47	0.0846
15	61	0.0607
16	10	0.0579
17	34	0.0527
18	27	0.0462
19	39	0.0461
20	41	0.0422

b. The top 20 dimensions according to their AUROC scores are as follows:

	1	2
1	25	0.7340
2	29	0.6837
3	11	0.6695
4	47	0.6661
5	19	0.6315
6	34	0.6174
7	32	0.6021
8	30	0.6021
9	9	0.6000
10	56	0.5971
11	27	0.5953
12	60	0.5929
13	51	0.5881
14	26	0.5874
15	53	0.5845
16	7	0.5797
17	10	0.5709
18	61	0.5686
19	43	0.5567
20	44	0.5422

Although there are some similarities (for instance dimensions 25 and 29 rank highly for both), the AUROC rankings are drastically different from the Fisher score rankings. Generally this is as expected: some similarities for the most/least important factors in classification, but different since the Fisher scores and AUROC scores go about the ranking in wildly different manners.