

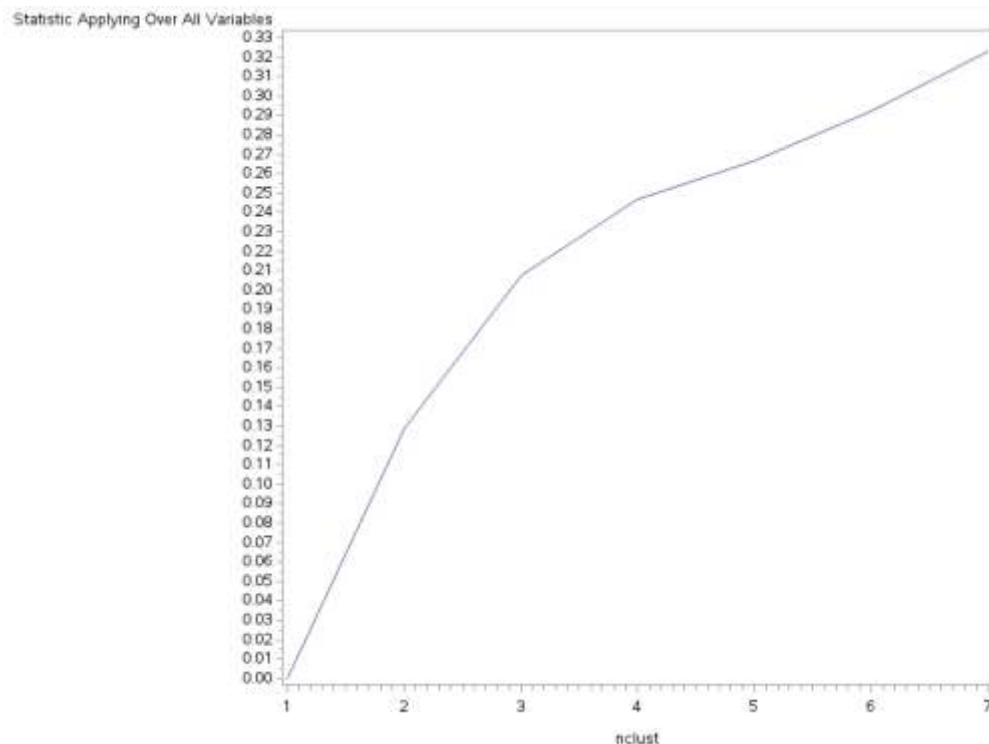
Machine learning for data analysis

Week 4 assignment

Running a k-means cluster analysis

I chose addhealth as my dataset. 12 variables (H1GH1 BIO_SEX H1GH2 H1GH6 H1GH59 H1GH60 H1FS11 H1WP9 H1WP13 H1DA1 H1DA2 H1DA3) were analyzed. BIO_SEX is a binary variable. H1GH59 and H1GH60 are quantitative variables. Others are ordered categorical variables. Each of them has more than 3 levels, so that they can be treated as quantitative variables.

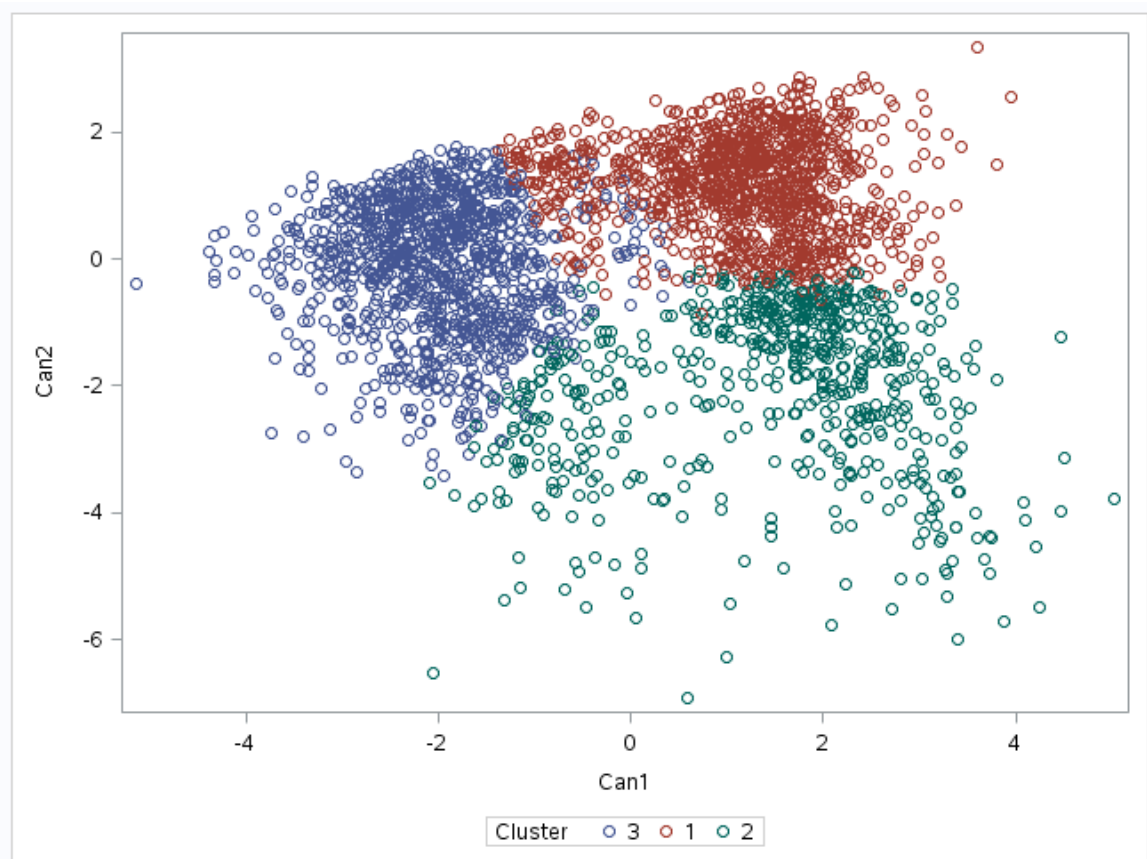
K=7. There is a bend at nclust equal to 2, 3 and 4. So I further evaluated these cluster solutions.



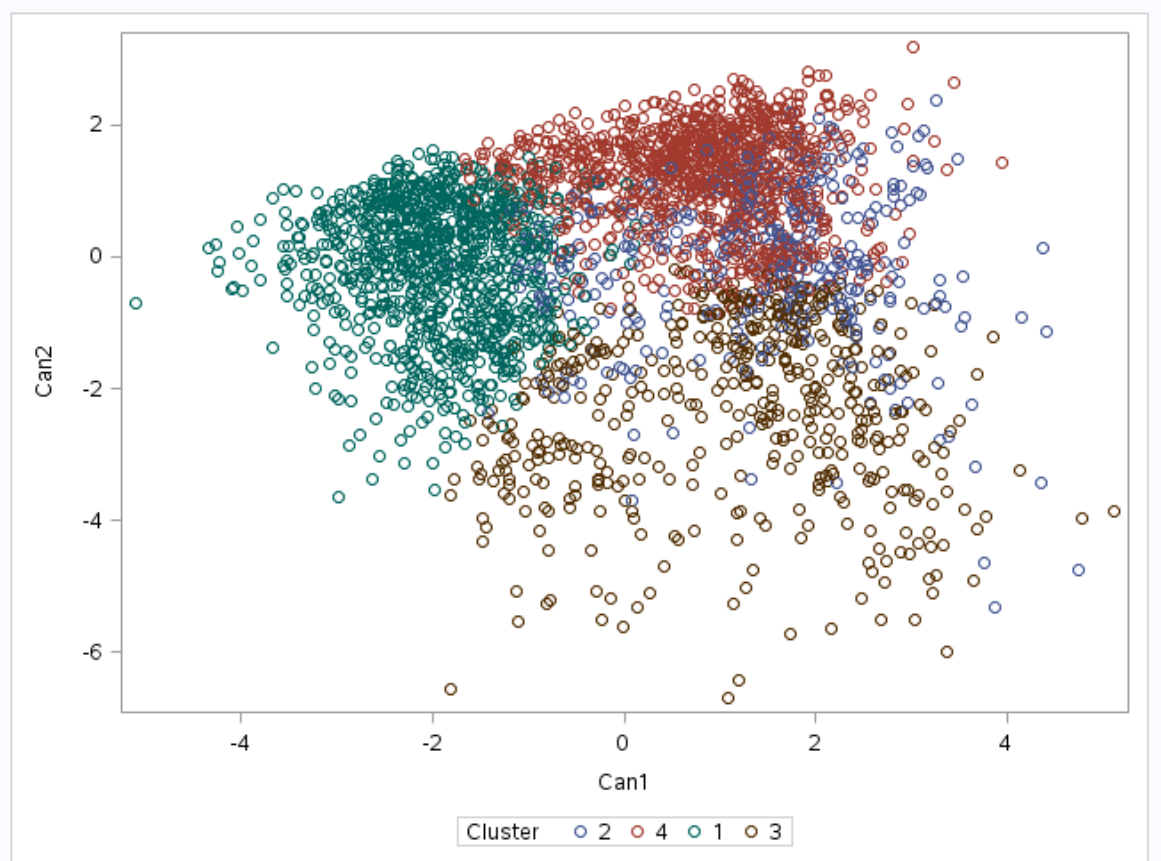
For nclust=2, can2 is invalid.

Class Means on Canonical Variables	
CLUSTER	Can1
1	1.771294998
2	-2.234977624

For nclust=3, the cluster 2 is more spread out than the other two clusters, meaning is variance within cluster is much higher. There is almost no overlap between each of the clusters.



For $n_{\text{clust}}=4$, the most part of cluster 2, cluster 4 and cluster 3 are overlapped, indicating the variance between cluster is not high enough. This should not be the best cluster solution.



Overall, $n_{\text{clust}}=3$ may be the best cluster solution.

My code:

```
1 *Load data;
2 LIBNAME mydata "/courses/d1406ae5ba27fe300" access=readonly;
3 data new; set mydata.addhealth_pds;
4
5 if H1GH1=6 then delete; if H1GH1=8 then delete;
6 if H1GH2=6 then delete; if H1GH2=8 then delete;
7 if H1GH6=6 then delete; if H1GH6=8 then delete;
8 if H1GH59A=96 then delete; if H1GH59A=98 then delete; if H1GH59A=99 then delete;
9 if H1GH59B=96 then delete; if H1GH59B=98 then delete; if H1GH59B=99 then delete;
10 if H1GH60=996 then delete; if H1GH60=998 then delete; if H1GH60=999 then delete;
11 if H1FS11=6 then delete; if H1FS11=8 then delete;
12 if H1WP9=6 then delete; if H1WP9=7 then delete;
13 if H1WP9=8 then delete; if H1WP9=9 then delete;
14 if H1WP13=6 then delete; if H1WP13=7 then delete;
15 if H1WP13=8 then delete; if H1WP13=9 then delete;
16 if H1DA1 = 6 then delete; if H1DA1 = 8 then delete;
17 if H1DA2 = 6 then delete; if H1DA2 = 8 then delete;
18 if H1DA3 = 6 then delete; if H1DA3 = 8 then delete;
19
20 H1GH59=H1GH59A * 12 + H1GH59B;
21
22 keep AID H1GH1 BIO_SEX H1GH2 H1GH6 H1GH59 H1GH60 H1FS11 H1WP9 H1WP13 H1DA1 H1DA2 H1DA3;
23
24 run;
25
26 * split the data set into training and test data;
27 proc surveyselect data=new out=train test seed=100 samprate=0.7 method=srs outall;
28 run;
29
30 data new_train; set train test;
31 if selected = 1;
32 run;
33
34 data new_test; set train test;
35 if selected = 0;
36 run;
37
38 *standardized the clustering variable;
39 proc standard data=new_train out=stdtrain mean=0 std=1;
40 var H1GH1 BIO_SEX H1GH2 H1GH6 H1GH59 H1GH60 H1FS11 H1WP9 H1WP13 H1DA1 H1DA2 H1DA3;
41 run;
42
43 %macro kmean(K);
44 proc fastclus data=stdtrain out=outdata&K. outstat=clustat&K.
45     maxclusters=&K. maxiter=300;
46 var H1GH1 BIO_SEX H1GH2 H1GH6 H1GH59 H1GH60 H1FS11 H1WP9 H1WP13 H1DA1 H1DA2 H1DA3;
47 run;
48 %mend;
49
```

```
50 %kmean(1);
51 %kmean(2);
52 %kmean(3);
53 %kmean(4);
54 %kmean(5);
55 %kmean(6);
56 %kmean(7);
57
58 data clus1; set clustat1;
59 nclust = 1;
60 if _type_ = 'RSQ';
61 keep nclust over_all;
62 run;
63
64 data clus2; set clustat2;
65 nclust = 2;
66 if _type_ = 'RSQ';
67 keep nclust over_all;
68 run;
69
70 data clus3; set clustat3;
71 nclust = 3;
72
72 if _type_ = 'RSQ';
73 keep nclust over_all;
74 run;
75
76 data clus4; set clustat4;
77 nclust = 4;
78 if _type_ = 'RSQ';
79 keep nclust over_all;
80 run;
81
82 data clus5; set clustat5;
83 nclust = 5;
84 if _type_ = 'RSQ';
85 keep nclust over_all;
86 run;
87
88 data clus6; set clustat6;
89 nclust = 6;
90 if _type_ = 'RSQ';
91 keep nclust over_all;
92 run;
93
94 data clus7; set clustat7;
```

```

95 nclust = 7;
96 if _type_ = 'RSQ';
97 keep nclust over_all;
98 run;
99
100 data cluster_r2; set clus1 clus2 clus3 clus4 clus5 clus6 clus7;
101 run;
102
103 symbol interpol=join;
104 proc gplot data=cluster_r2;
105 plot over_all * nclust;
106 run;
107
108
109 *****further evaluation*****;
110 * nclust = 2;
111 proc candisc data=outdata2 out=clustercan2;
112 class cluster;
113 var H1GH1 BIO_SEX H1GH2 H1GH6 H1GH59 H1GH60 H1FS11 H1WP9 H1WP13 H1DA1 H1DA2 H1DA3;
114 run;
115
116
117 proc sgplot data=clustercan2;
118 scatter y = can2 x = can1 / group=cluster;
119 run;
120
121 *nclust = 3;
122 proc candisc data=outdata3 out=clustercan3;
123 class cluster;
124 var H1GH1 BIO_SEX H1GH2 H1GH6 H1GH59 H1GH60 H1FS11 H1WP9 H1WP13 H1DA1 H1DA2 H1DA3;
125 run;
126 proc sgplot data=clustercan3;
127 scatter y = can2 x = can1 / group=cluster;
128 run;
129
130 *nclust = 4;
131 proc candisc data=outdata4 out=clustercan4;
132 class cluster;
133 var H1GH1 BIO_SEX H1GH2 H1GH6 H1GH59 H1GH60 H1FS11 H1WP9 H1WP13 H1DA1 H1DA2 H1DA3;
134 run;
135 proc sgplot data=clustercan4;
136 scatter y = can2 x = can1 / group=cluster;
137 run;

```