Problem 1: K-means Clustering

Part a

1) The coordinates of the 3 cluster centers are:

[[3.94151515 4.03818182]

 [2.94094444 -4.96888889]

 [0.86250714 2.02731837]]

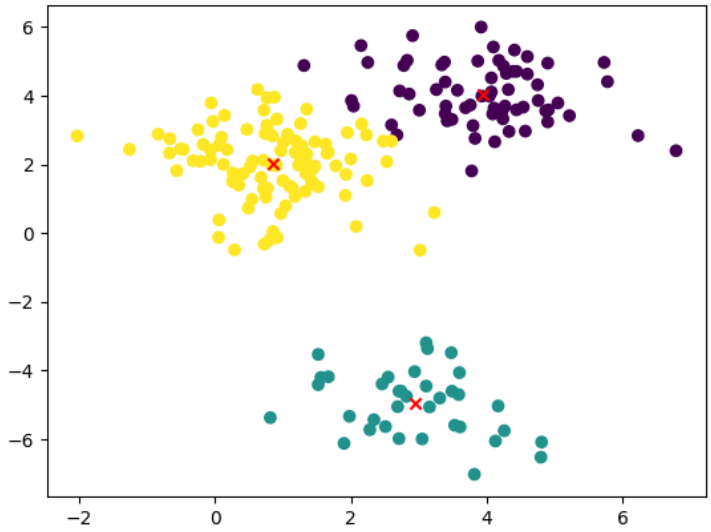
2) The sizes of the clusters are:

cluster size of label 0: 66

cluster size of label 1: 36

cluster size of label 2: 98

3) graph (red ‘x’s are the centers)



Part b

1) The coordinates of the 3 cluster centers are:

[[1.18686585 1.09676098]

 [2.94094444 -4.96888889]

 [4.04412698 4.03301587]

 [0.68707   2.76916667]]

2) The sizes of the clusters are:

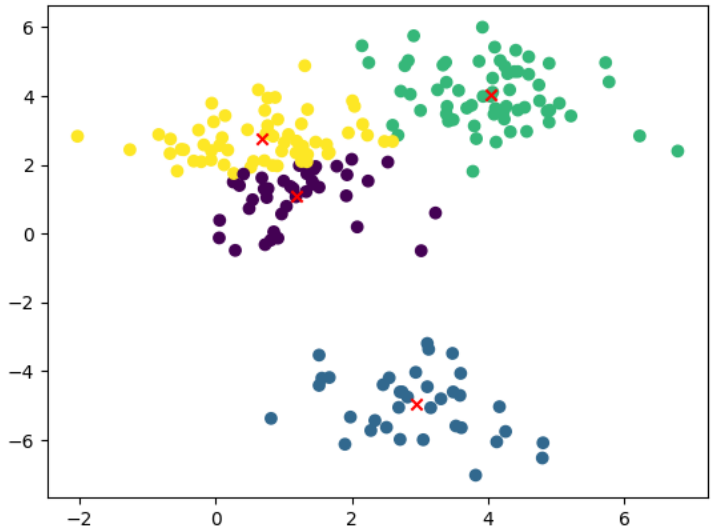
cluster size of label 0: 40

cluster size of label 1: 36

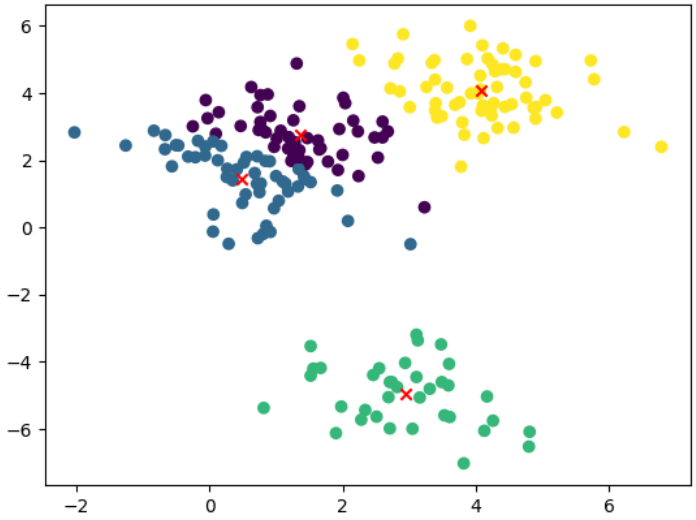
cluster size of label 2: 63

cluster size of label 3: 61

3) graph (red ‘x’s are the centers)



Part c



Center coordinates:

[[1.36657843 2.77435294]

[0.48981154 1.45452308]

[2.94094444 -4.96888889]

[4.09016393 4.06672131]]

Part d

K-means minimizes the sum of squared center-point distances for all clusters. Thus, the best clustering is the one with the least sum of squared center-point distances.

$$\min \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - u_i||^2,$$ u_i = center of cluster S_i

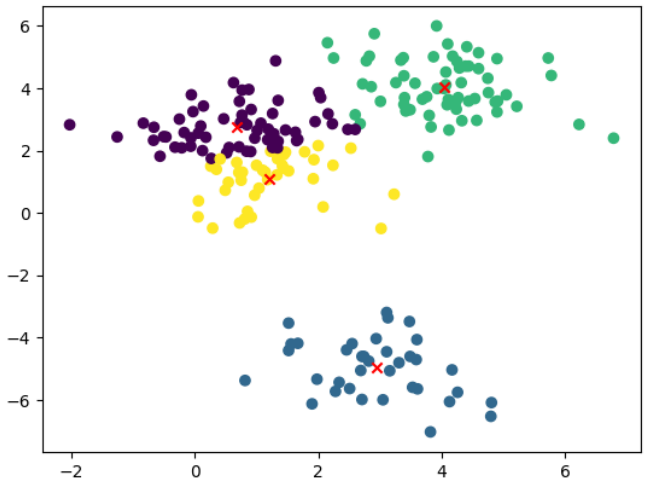
Part e

By this criterion, we run the program for 30 times and compute the sum of squared center-point distances:

Run #	Cluster size	distances
1	69 32 36 63	283.0469478430381
2	65 60 36 39	284.2225709390983
3	51 50 36 63	282.40704795229055
4	63 36 38 63	281.9067366627193
5	64 49 51 36	282.51663762771375
6	52 36 64 48	282.5346825524412
7	36 63 58 43	282.1831751313205
8	49 63 36 52	282.4018184933067
9	63 40 61 36	282.020825142376
10	61 36 63 40	281.8922416624156
11	64 36 61 39	281.9922054529733
12	63 39 36 62	281.90362966647837

13	40 60 64 36	281.99468140152777
14	36 63 61 40	282.020825142376
15	36 53 63 48	282.8117499119038
16	49 63 36 52	282.54418498854176
17	64 55 36 45	282.25284708353536
18	61 36 52 51	292.95868768611325
19	36 63 46 55	282.89638487205684
20	43 36 56 65	284.48588551495334
21	50 63 36 51	282.4310280558003
22	41 64 36 59	282.12189750676936
23	47 61 56 36	292.4693468258907
24	42 62 60 36	292.9463794340613
25	40 36 61 63	281.9440439777117
26	39 62 36 63	281.90362966647837
27	63 49 52 36	282.4018184933067
28	36 61 40 63	281.9440439777117
29	50 51 36 63	282.40704795229055
30	36 63 38 63	281.9067366627193

Among these 3 runs, the sum of center-point distances has a smallest value in the 10th run, with the value of 281.8922416624156 and cluster size 61 36 63 40.



Problem 2: Hierarchical clustering

Part a

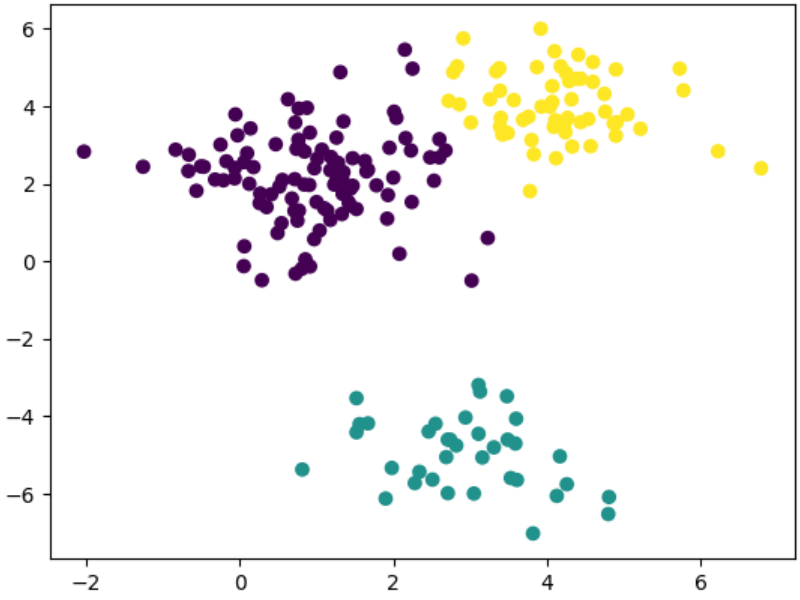
1) The sizes of the clusters are:

cluster size of label 0: 105

cluster size of label 1: 36

cluster size of label 2: 59

2) graph



The clusters are a little different than the K-means in problem 1, part a. The cluster split of data instances in the upper part of the graph differs a little (105 vs 98, 59 vs 66).

Part b

1) The sizes of the clusters are:

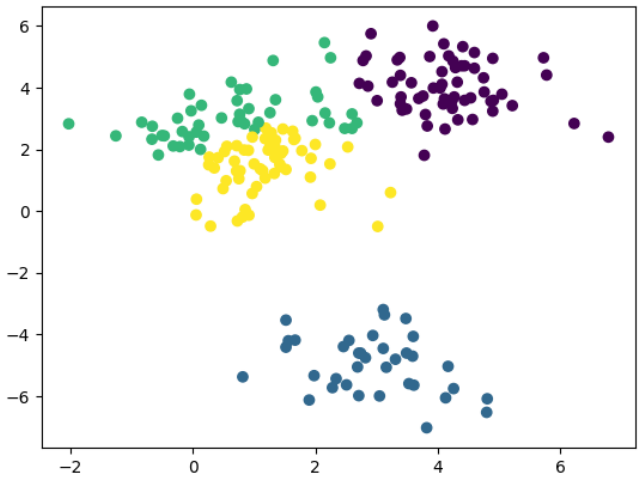
cluster size of label 0: 59

cluster size of label 1: 36

cluster size of label 2: 47

cluster size of label 3: 58

2) graph



The clustering is a little different from problem1, part b. Still, it's the cluster split of data instances in the upper part of the graph differs a little.

Part c

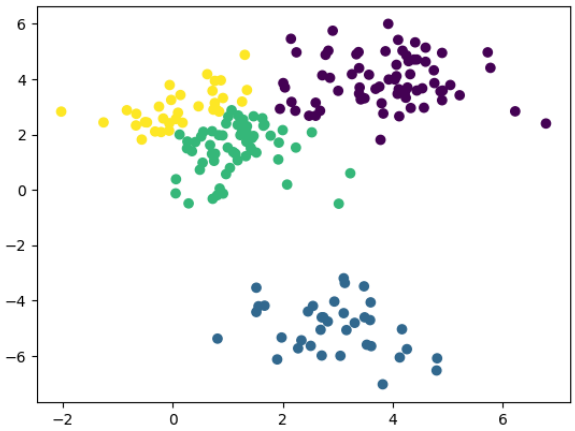
1) Ward

cluster size of label 0: 70

cluster size of label 1: 36

cluster size of label 2: 61

cluster size of label 3: 33



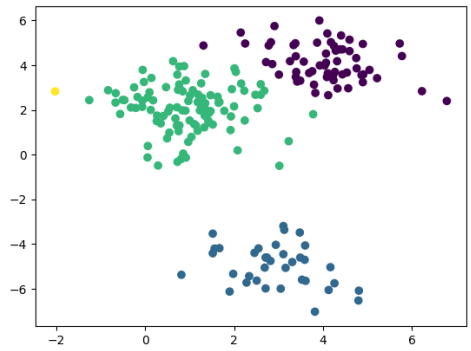
2) Average

cluster size of label 0: 61

cluster size of label 1: 36

cluster size of label 2: 102

cluster size of label 3: 1



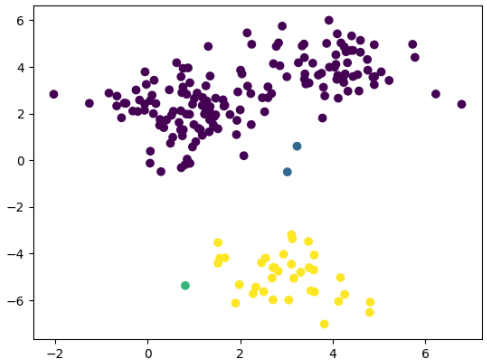
3) Single

cluster size of label 0: 162

cluster size of label 1: 2

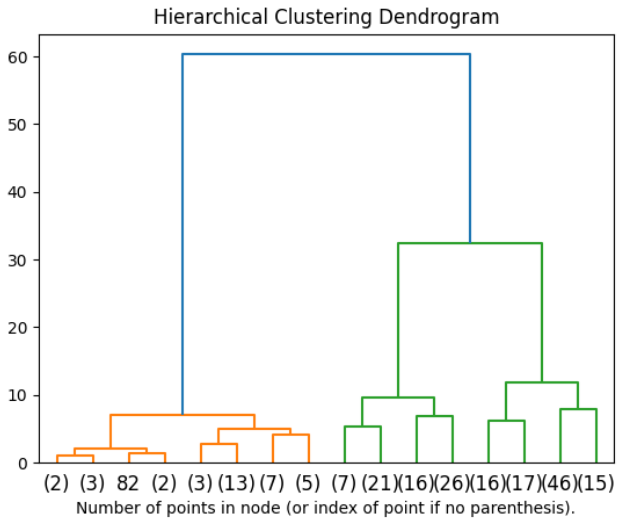
cluster size of label 2: 1

cluster size of label 3: 35



Compare the 4 types of linkages, complete linkage divides the clusters most evenly, and ward linkage is less evenly but still remaining similar sizes. Average and single linkages make the clusters very imbalanced, resulting in some clusters only have 1 or 2 instances. Between these 2, single linkage is more imbalanced.

Extra Credit



Problem 3: Feature/Input ranking

Part a

Rank	label	fisher score
1	48	0.32296908826105625
2	25	0.21718861805514905
3	21	0.19340370867423126
4	70	0.19116545711426614
5	65	0.1706993849366968
6	40	0.1683883407191055
7	29	0.1675001501102931
8	19	0.1418240847823315
9	57	0.1263666836103104
10	20	0.12213582023664572
11	24	0.10065683677897679
12	30	0.09626396302415782
13	12	0.0864726786128256
14	47	0.08567079579782078
15	61	0.06157126237847705
16	10	0.05878478730683284
17	34	0.053084461773596194
18	27	0.04683667668196662
19	39	0.046569437697087115
20	41	0.04242582045035129

Part b

Rank	label	AUROC score
1	25	0.7339474502487562
2	29	0.6840796019900497
3	11	0.670164800995025
4	47	0.666122512437811
5	19	0.6313355099502488
6	34	0.6173818407960199
7	32	0.6021066542288558
8	30	0.6019900497512438
9	9	0.5998911691542289
10	56	0.5970149253731343
11	27	0.5953047263681592
12	60	0.5929726368159203
13	51	0.5883473258706468
14	26	0.5869869402985075
15	53	0.5843827736318409
16	7	0.5796797263681592
17	10	0.5708955223880597
18	61	0.5687189054726367
19	43	0.556708644278607
20	44	0.5423274253731343

The 2 order lists are not the same but similar: there are 9 out of 20 labels appear in both lists. Both the fisher score and the auroc score measure the predictability of individual features, so I expect the order lists to be similar if not exactly the same.