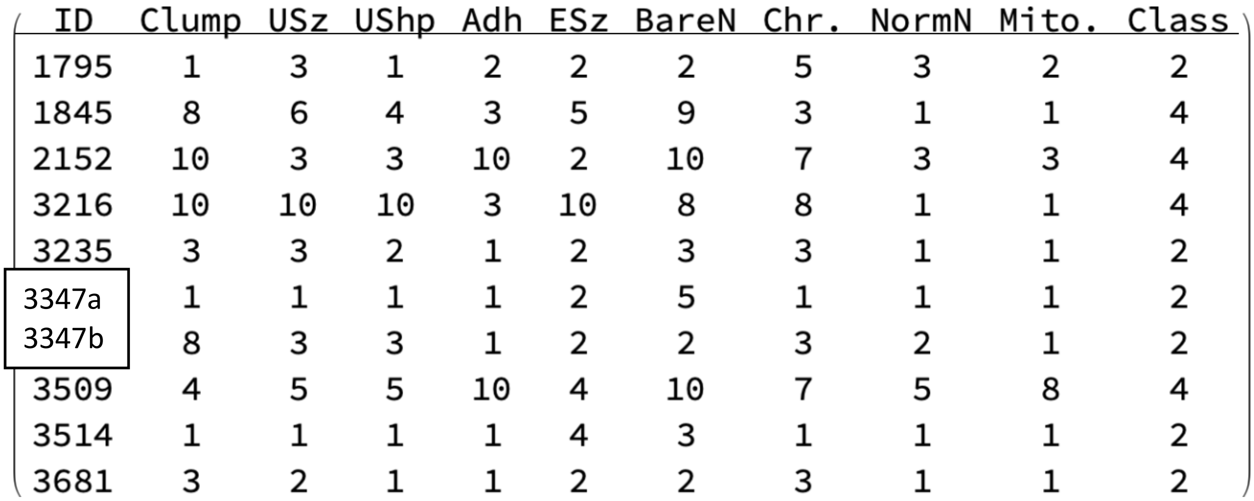
CSE514, Fall 2021, HW 6 Name: Anh Le Student ID: 488493

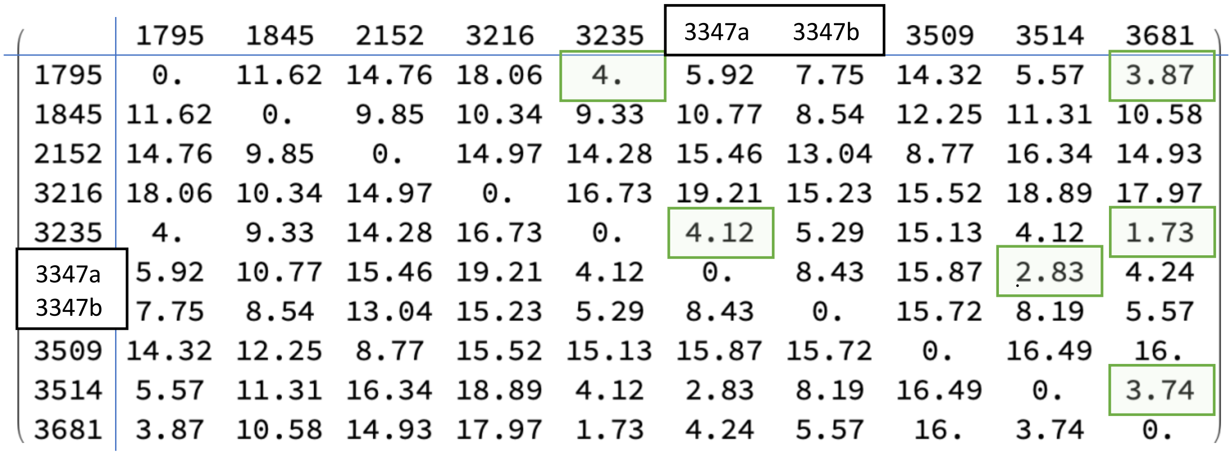
Note: This homework is worth a total of 15 points

Here's a tiny subset of the breast cancer data from your Programming Assignment 2:
(*ID numbers had the first three digits of 117 dropped to save space)



ID	Clump	USz	UShp	Adh	ESz	BareN	Chr.	NormN	Mito.	Class
1795	1	3	1	2	2	2	5	3	2	2
1845	8	6	4	3	5	9	3	1	1	4
2152	10	3	3	10	2	10	7	3	3	4
3216	10	10	10	3	10	8	8	1	1	4
3235	3	3	2	1	2	3	3	1	1	2
3347a	1	1	1	1	2	5	1	1	1	2
3347b	8	3	3	1	2	2	3	2	1	2
3509	4	5	5	10	4	10	7	5	8	4
3514	1	1	1	1	4	3	1	1	1	2
3681	3	2	1	1	2	2	3	1	1	2

Here's a Euclidean distance matrix for this subset, ignoring Class:



	1795	1845	2152	3216	3235	3347a	3347b	3509	3514	3681
1795	0.	11.62	14.76	18.06	4.	5.92	7.75	14.32	5.57	3.87
1845	11.62	0.	9.85	10.34	9.33	10.77	8.54	12.25	11.31	10.58
2152	14.76	9.85	0.	14.97	14.28	15.46	13.04	8.77	16.34	14.93
3216	18.06	10.34	14.97	0.	16.73	19.21	15.23	15.52	18.89	17.97
3235	4.	9.33	14.28	16.73	0.	4.12	5.29	15.13	4.12	1.73
3347a	5.92	10.77	15.46	19.21	4.12	0.	8.43	15.87	2.83	4.24
3347b	7.75	8.54	13.04	15.23	5.29	8.43	0.	15.72	8.19	5.57
3509	14.32	12.25	8.77	15.52	15.13	15.87	15.72	0.	16.49	16.
3514	5.57	11.31	16.34	18.89	4.12	2.83	8.19	16.49	0.	3.74
3681	3.87	10.58	14.93	17.97	1.73	4.24	5.57	16.	3.74	0.

*The six smallest distance values are highlighted green.

Q1 (2pts): Since there are only ten datapoints, each with nine features, this tiny dataset is very sparse. Explain why this would make MeanShift and DBSCAN poor choices for clustering:

- MeanShift is a poor choice because if the datapoints are far from each other, we might not be able to have overlapping windows. If all division of data has density of 1 (its own points), we will not be able to assign clusters. We don't want big windows since it might classify all points into 1 cluster.
- DBSCAN is a poor choice because the dataset is very spare so if the distance ε is relatively small, there will be no neighbors within ε, so all datapoints will be labeled noise. Meanwhile, if ε is large, all datapoints will be labeled into the same cluster.

Q2 (7pts): Agglomerative clustering

Q2a (1pts): What samples are the first to get merged?

- 3235 and 3681 (cluster a)

Q2b (3pts): Remember that single-linkage clustering means that you merge clusters by picking the smallest *minimum* distance between members of each cluster.
Using the single-linkage clustering method, what are the second and third merges?

Second:

- 3347a and 3514 (cluster b)

Third:

- 3235, 3681, 3347a and 3514
- Cluster a and cluster b will be merged since 3.74 is the third smallest distance

Q2c (3pts): Remember that complete-linkage clustering means that you merge clusters by picking the smallest *maximum* distance between members of each cluster.
Using the complete-linkage clustering method, what are the second and third merges?

Second:

- (3347a and 3514) since 2.83 will be the smallest maximum distance

Third:

- Next smallest distance: 3.74 – not the smallest maximum distance between (3347a, 3514) and (3681, 3235) since 3681 to 3347a is bigger (4.24)

- Next smallest distance: 3.87 – not the smallest maximum distance between (3681, 3235) to (1795) since 3235 to 1795 is bigger (4.00)

- Next smallest distance: 4.00 – it is the smallest maximum distance between (3235, 681) and (1795) → merge (3235, 681, and 1795)

Q3 (6pts): k-Means
Let’s say you’ve randomly selected the first (1795) and second (1845) samples to be the starting centroids of your 2-Means clustering.

Q3a (3pts): Which samples are assigned to which cluster?

Cluster 1 w/ (1795):

- 1795
- 3235
- 3347a
- 3347b
- 3514
- 3681

Cluster 2 w/ (1845):

- 1845
- 2152
- 3216
- 3509

Q3b (3pts): What are the new centroid values after this first iteration of clustering?

	Cluster 1	Cluster 2
Clump	2.833	8
USz	2.167	6
UShp	1.5	5.5
Adh	1.667	6.5
ESz	2.333	5.25
BareN	2.833	9.25
Chr.	2.667	6.25
NormN	1.5	2.5
Mito.	1.167	3.25