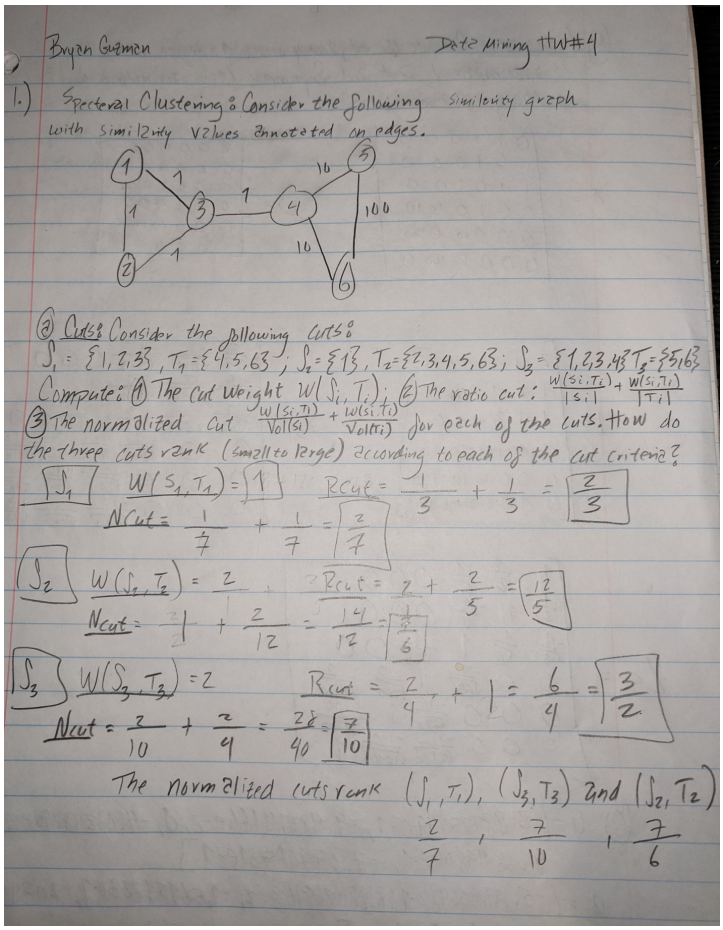
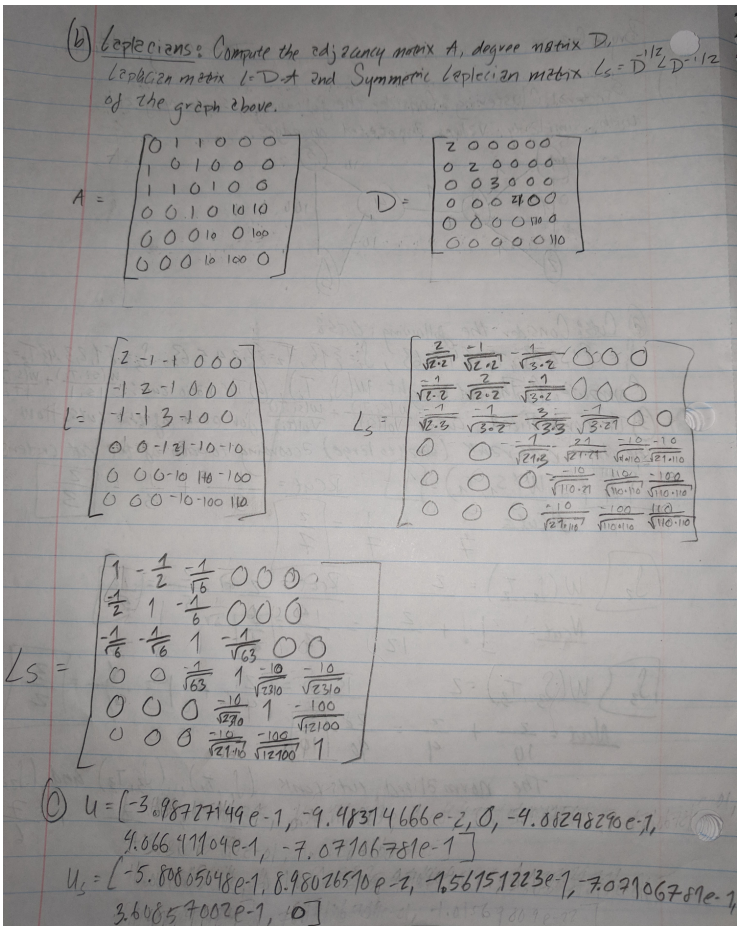
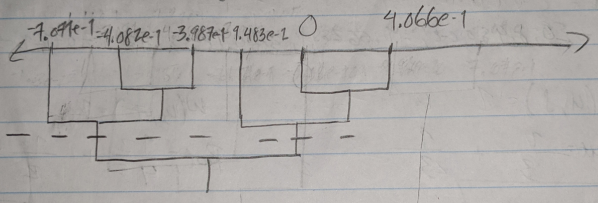
Bryan Guzman
ICSI-435: Data Mining
Homework #4
001265918

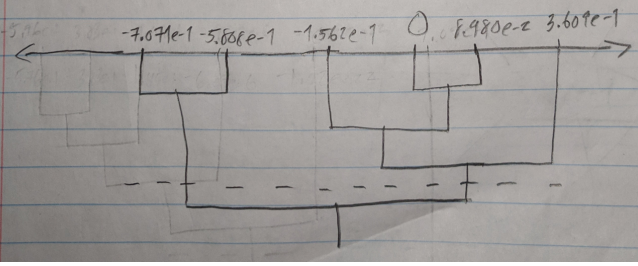
1.)



(d) Hierarchical Cluster from eigenvectors: Here we will do agglomerative single link clustering of the nodes represented by their $u$ values and by their $u_5$ values computed in the previous part. Note that we will get two dendograms, one for $u_5$ and one for $u_5$ and in each the nodes will be one-dimensional points. Draw the two dendograms.




(e) Hierarchical Clusters from eigenvectors: Cut the dendograms above to obtain two clusters. Show the resulting partitions in your solution file. What are their cut weights, normalized cuts and ratio cuts? Discuss why we obtain these partitions by relating the Laplacians to the specific cut measures.

$u = \{0, 4.066e\text{-}1, 9.483e\text{-}2\}$ $S = \{-4.082e\text{-}1, -3.987e\text{-}1, -7.071e\text{-}1\}$

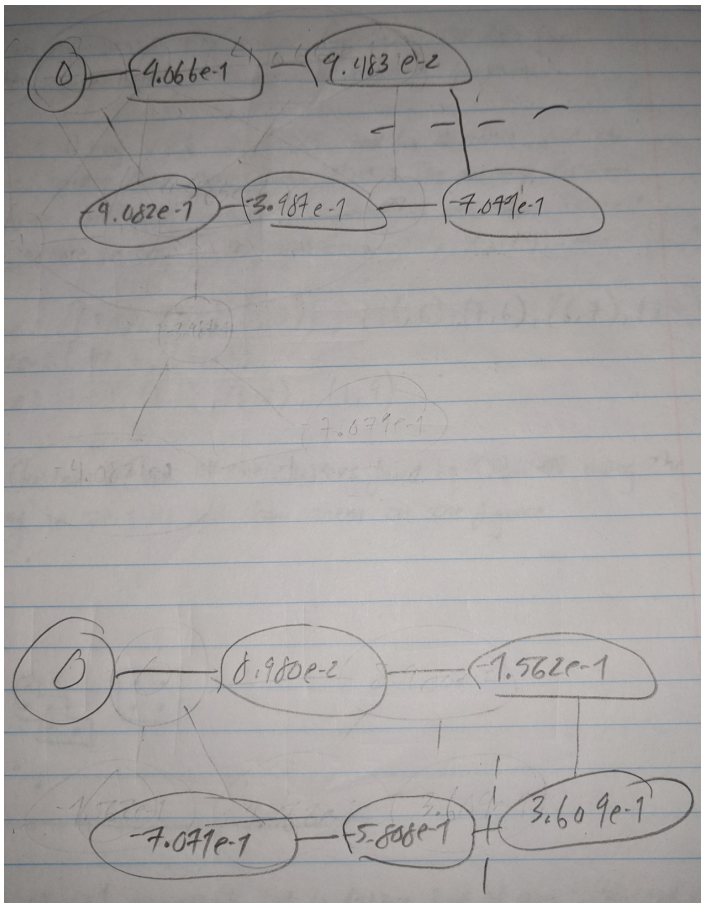
$u_5 = \{0, 8.980e\text{-}2, -1.562e\text{-}1, 3.609e\text{-}1\}$ $T = \{-7.071e\text{-}1, -5.808e\text{-}1\}$

$w(u,s) = 1$ $\quad w(u_5, t) = 1$

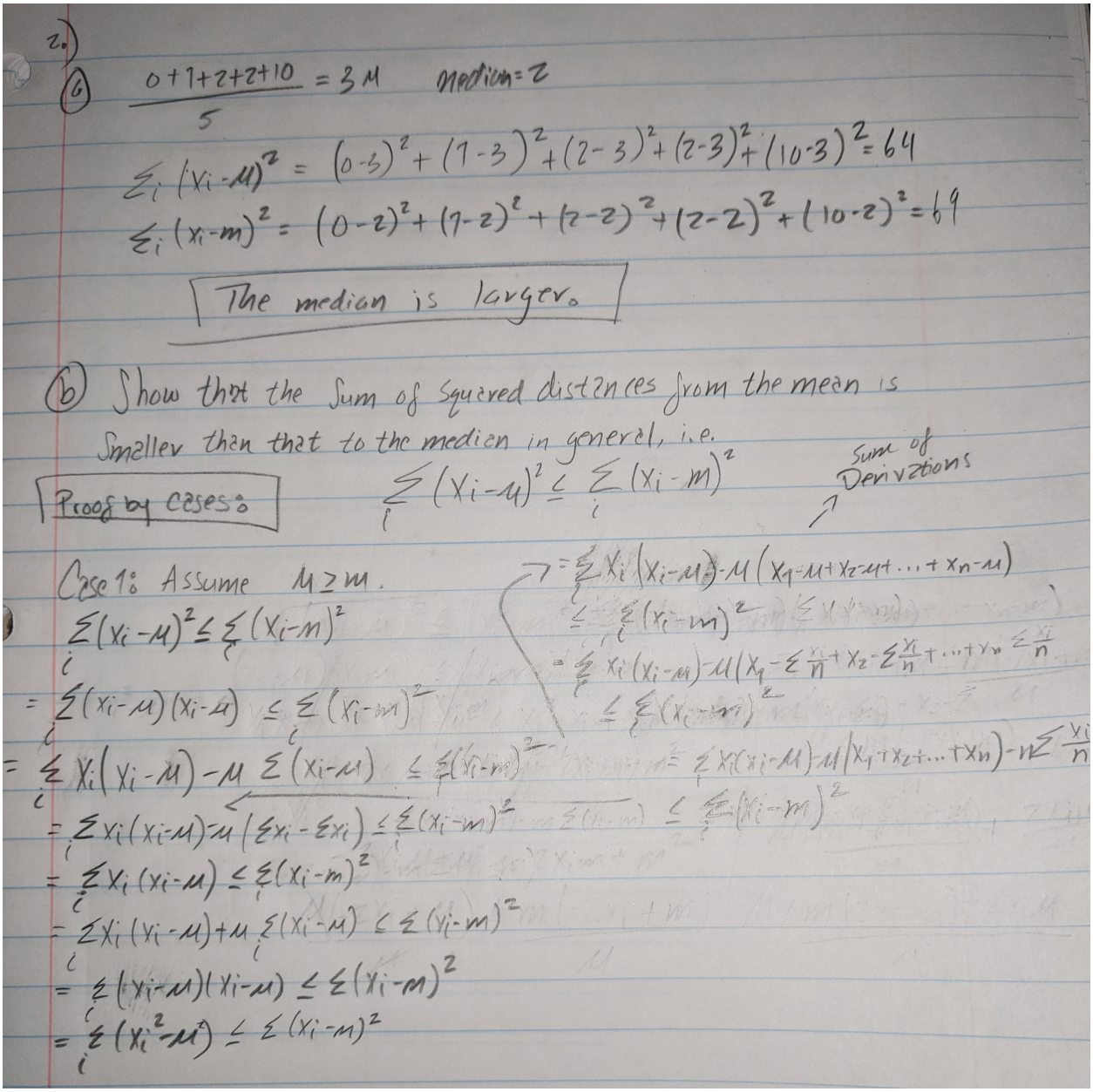
$Rcut = \frac{1}{3} + \frac{1}{3}$ $\quad Rcut = \frac{1}{4} + \frac{1}{2} = \boxed{3/4}$

$\boxed{Rcut = 2/3}$

$Ncut = \frac{1}{5} + \frac{1}{5} = \boxed{2/5}$ $\quad Ncut = \frac{1}{7} + \frac{1}{3} = \boxed{\frac{10}{21}}$



2.)



2.)

(a) $\frac{0+7+2+2+10}{5} = 3 \mu$ median = 2

$\sum_i (x_i-\mu)^2 = (0-3)^2+(7-3)^2+(2-3)^2+(2-3)^2+(10-3)^2 = 64$

$\sum_i (x_i-m)^2 = (0-2)^2+(7-2)^2+(2-2)^2+(2-2)^2+(10-2)^2 = 69$

The median is larger.

(b) Show that the sum of squared distances from the mean is smaller than that to the median in general, i.e.

$\sum_i (x_i-\mu)^2 \le \sum_i (x_i-m)^2$

Proof by cases:

Case 1: Assume $\mu \ge m$.

$\sum_i (x_i-\mu)^2 \le \sum_i (x_i-m)^2$

$= \sum_i (x_i-\mu)(x_i-\mu) \le \sum_i (x_i-m)^2$

$= \sum_i x_i(x_i-\mu) - \mu \sum_i (x_i-\mu) \le \sum_i (x_i-m)^2$

$= \sum_i x_i(x_i-\mu) - \mu(\sum x_i - \sum x_i) \le \sum_i (x_i-m)^2$

$= \sum_i x_i(x_i-\mu) \le \sum_i (x_i-m)^2$

$= \sum_i x_i(x_i-\mu) + \mu \sum_i (x_i-\mu) \le \sum_i (x_i-m)^2$

$= \sum_i (x_i-\mu)(x_i-\mu) \le \sum_i (x_i-m)^2$

$= \sum_i (x_i^2-\mu^2) \le \sum_i (x_i-m)^2$

Sum of Derivations

$= \sum_i x_i(x_i-\mu) - \mu(x_1-\mu+x_2-\mu+\dots+x_n-\mu)$

$\le \sum_i (x_i-m)^2$

$= \sum_i x_i(x_i-\mu) - \mu(x_1-\sum \frac{x_i}{n} + x_2 - \sum \frac{x_i}{n} + \dots + x_n \le \frac{x_i}{n}$

$\le \sum_i (x_i-m)^2$

$= \sum_i x_i(x_i-\mu) - \mu(x_1+x_2+\dots+x_n) - n\sum \frac{x_i}{n}$

$\le \sum_i (x_i-m)^2$

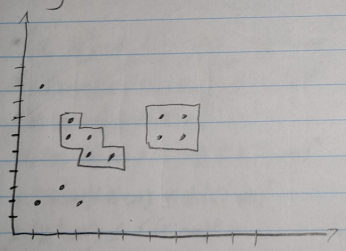
3.)

3.) Density-based Clusterings: Considering the data points in the figure below, answer the following questions.

(a) $L_\infty$: Using $\epsilon = 1$, minpts = 3, and $L_\infty$ distance, find all core, border and noise points. Hints: The distance between points according to $L_\infty$ is defined as maximum absolute difference in any of the dimensions, i.e. $d_\infty(x,y) = \max_{i=1..d} |x_i - y_i|$

Core: $\{(2,6), (3,6), (3,5)\}$, $\{(6,6), (7,6), (6,7), (7,7)\}$
Border: $\{(2,7), (4,5)\}$
Noise: $(1,2), (2,3), (3,2), (1,9)$

(b) $L_\infty$ Clusters: Find all the clusters found by DBSCAN using the setting in part(a) and show them on the figure.



(c) $L_2$: Using $\epsilon = 1$, minpts = 3, and $L_2$ distance, find all core, border and noise points. Compare your results with part (a). What difference do you see? Why?

Core: $\{(2,6), (3,6), (3,5)\}$, $\{(6,6), (7,6), (6,7), (7,7)\}$
Border: $\{(2,7), (4,5)\}$
Noise: $(1,2), (2,3), (3,2), (1,9)$

I see no difference with any of the core, border and noise points. Nothing changed since when using Euclidean distance of 1, it introduce no new points into the clusters. The noise points and already present clusters had 2 distance greater than 1.

(d) Find all the clusters found by DBSCAN using the setting in part c and show them on the figure. Compare your results with part(b). What difference do you see? Why?

* See question 3b for the figure.

There is no difference in figure 3d when compared to figure 3b. They are exactly the same due to what I discussed in question 3c.