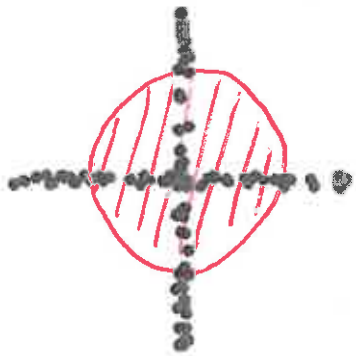
Math 3900 Midterm

March 8, 2018

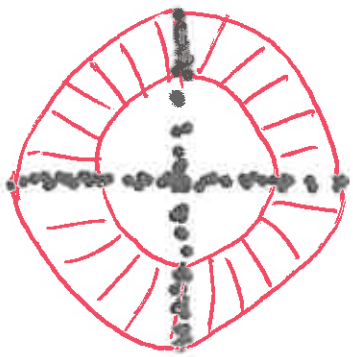
SHOW ALL WORK

1.) Let D be the data set below. Suppose we use eccentricity as the filter function with image [0, 2], i.e., $f : D \to [0, 2]$. Also suppose we cover the image with 3 intervals with 50% overlap,

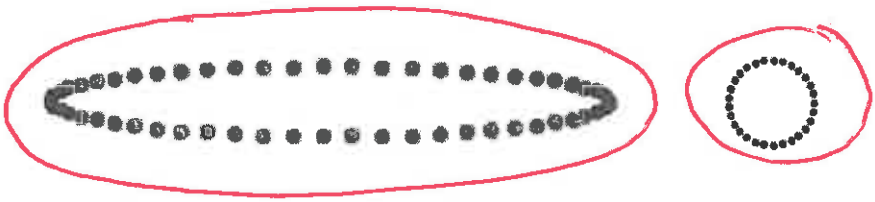
[5] 1a.) Clearly indicate on the graph below what the bin corresponding to $f^{-1}([0, 1])$ could look like (include the use of shading or hatch marks to identify the shape of this bin).



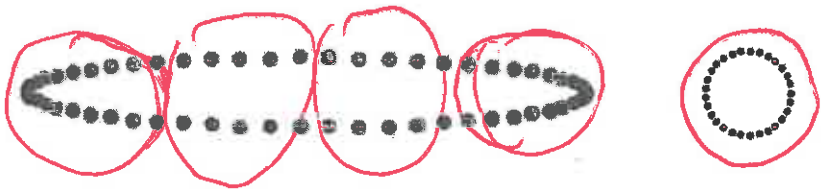
[5] 1b.) Clearly indicate on the graph below what the bin corresponding to $f^{-1}([1, 2])$ could look like (include the use of shading or hatch marks to identify the shape of this bin).



[2] 2a.) In the graph below, circle possible clusters if single linkage clustering is used.



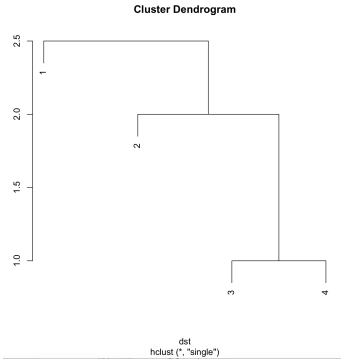
[2] 2b.) In the graph below, circle possible clusters if complete linkage clustering is used.



3.) Given the following distance matrix for a data set with 4 elements,

$$\begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 0 & 2.5 & 3.5 & 4.5 \\ 2.5 & 0 & 2 & 3 \\ 3.5 & 2 & 0 & 1 \\ 4.5 & 3 & 1 & 0 \end{pmatrix} \end{array}$$

[8] 3a.) Create the **single linkage** dendrogram for the data set with distance matrix above.



List the clusters at each merging height:

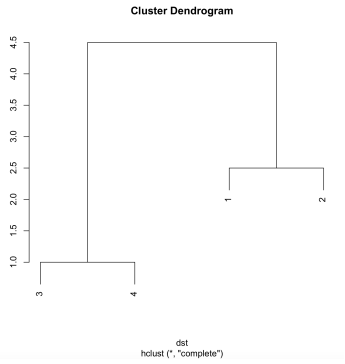
The clusters at time 0 are $\{a\},\ \{b\},\ \{c\},\ \{d\}$

The clusters at merging height $= \underline{1}$ are $\{c, d\},\ \{a\},\ \{b\}$

The clusters at merging height $= \underline{2}$ are $\{b, c, d\},\ \{a\}$

The clusters at merging height $= \underline{2.5}$ are $\{a, b, c, d\}$

[7] 3b.) Create the **complete linkage** dendrogram for the data set with distance matrix above.



List the clusters at each merging height:

The clusters at time 0 are $\{a\},\ \{b\},\ \{c\},\ \{d\}$

The clusters at merging height = $\underline{1}$ are $\underline{\{c, d\},\ \{a\},\ \{b\}}$

The clusters at merging height = $\underline{2.5}$ are $\underline{\{c, d\},\ \{a.b\}}$

The clusters at merging height = $\underline{4.5}$ are $\underline{\{a, b, c, d\}}$

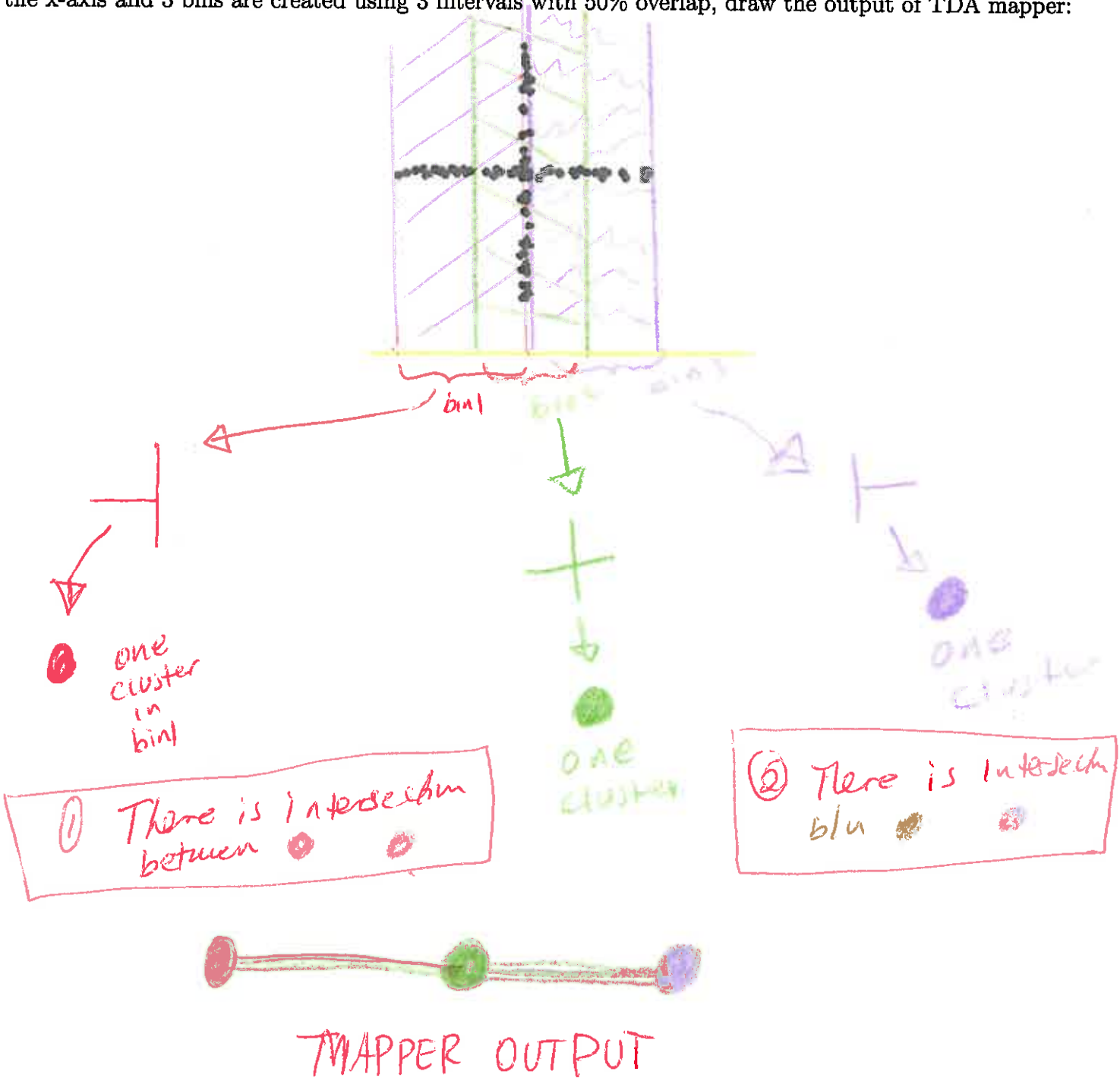
[3] 4.) Given the distance matrix:

$$\begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 0 & 2.5 & 3.5 & 4.5 \\ 2.5 & 0 & 2 & 3 \\ 3.5 & 2 & 0 & 1 \\ 4.5 & 3 & 1 & 0 \end{pmatrix} \end{array}$$

, then $\delta_3(b) =$ _____

Recall δ_k refers to the knn distance and $\delta_1(b) = 0$

[13] 5.) Suppose TDA mapper is used to analyze the following data set where filter function = projection to the x-axis and 3 bins are created using 3 intervals with 50% overlap, draw the output of TDA mapper:



Multiple Choice: **Circle the best answer.**

---

[5] 6.) The key idea(s) of topology that make extracting patterns via TDA mapper possible are

iv.) All of the above.

---

[5] 7.) If two vertices are close in the graph created by TDA mapper, then the data points represented by these vertices are close in the original data set.

ii.) False

---

[5] 8.) Mapper requires you to use a particular clustering algorithm.

ii.) False

---

[5] 9.) Suppose we perform the Kolmogorov-Smirnov two sample test on two 1-dimension data sets. If D is large, than these two data sets likely come from different distributions and thus differ in a significant way. Recall D is the maximum distance between the two empirical distribution functions (ECDF) for the two data sets.

i.) True

---

[5] 10.) Suppose eccentricity is used as a filter function. Let x and y be two points in the data. If x is further away from the center of the data than y is from the center of the data, filter value of x ____ filter value of y?

iii.) $>$

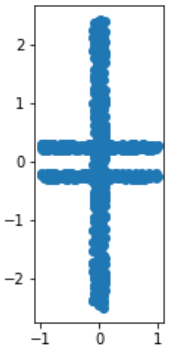
---

[5] 11.) Suppose the TDA mapper algorithm is applied to a group of n students who are interested in math, computer science, or both. We will create 2 bins without using a filter function. Bin 1 consists of all students who are interested in math while Bin 2 consists of all students interested in computer science. Suppose Bin 1 contains 10 students and Bin 2 contains 20 students. If the output of the TDA mapper algorithm is a graph with 2 vertices and 1 edge, then what do we know about n = the number of students in this group? Select the best answer.

v.) $20 \leq n < 30$

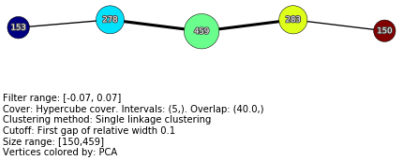
---

[5] 12.) You are given the following dataset to analyze in TDA mapper:



If the filter function is projection onto the first principle component, then which of the following is the most likely output of TDA mapper?

iv.)

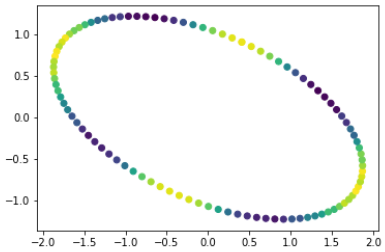


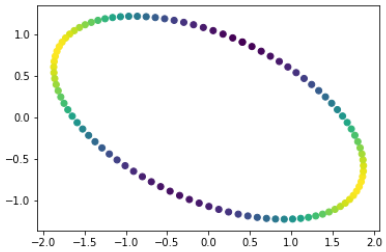
---

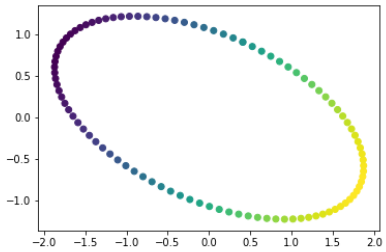
[5] 13.) One **benefit** to the variety of parameters one can choose in TDA mapper is that

ii.) You can probe the data set from a variety of different perpectives.

[10] 14.) The dataset shown below is colored according to various filter functions (eccentricity, knn with k = 45, projection to first principle component). On the line underneath the data set, state the filter function used to color the data set.







$\underline{knn(k = 45)}$ $\underline{eccentricity}$ $\underline{first - principle - component}$

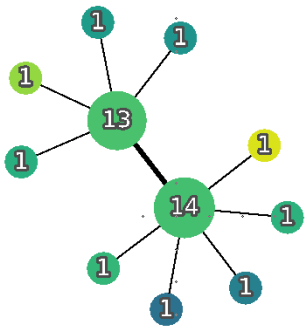
---

[5] 15.) Suppose you have a data set that contains information about the Chicago Cubs baseball team (The Cubs began play in 1870 and played about 80 home games last year). The following statistics are listed for each of the twenty five players on the team: number of hits, number of walks, and number of home runs. If each player represents a data point, how many data points do you have and in what dimension does the data live in?

There are $\underline{25}$ data points living in $\mathbb{R}^n$ where $n = \underline{3}$

---

[5] 16.) The output of python mapper applied to a particular data set is given below. How many data points would you estimate are in this data set? Why?



Note: Ball-13 intersects with four Ball-1s that do not intersect with Ball-14,and Ball-14 intersects with five Ball-1s that do not intersect with Ball-13. Thus the number of points in the intersection of Ball-13 and Ball-14 is 0 < n < 10. If there is 9 points in the intersection,then the number of data points (in the entire dataset) is 18. If the there is only one point in the intersection, then the number of data points (in the entire dataset) is 26. However since the edge between ball-13 and ball-14 isthicker than other edges, there is more than one point in the intersection between these two clusters. If there are only 2 points in the intersection, then the number of data points (in the entire dataset) is 25. Hence the number of data points (in the entire dataset) is between 18 and 25, inclusive.