Comprehensive Exam: *Computational Learning*

Spring 2020

Professor: Carlotta Domeniconi

Date: April 2, 2020
Time: 2 hours
Maximum Points: 100

This test is closed-book and may *not* be circulated. No notes are allowed. This test is governed by the GMU Honor Code. The paper you turn in must be your sole work. Help may be obtained from the instructor to understand the description of the problem, but the solution must be the student's own work. Any deviation from this is considered a Honor Code violation.

Good Luck!

Student Name: Tommy Jones

Mason ID: ____________________

Question 1	
Question 2	
Question 3	
Question 4	

1. **Support Vector Machines** [25 points]

Consider a linear Support Vector Machine and the following training data from two categories:

$$x_1 = (1,1)^T, x_2 = (2,2)^T, x_3 = (2,0)^T \text{ with class labels } y_1 = y_2 = y_3 = -1$$

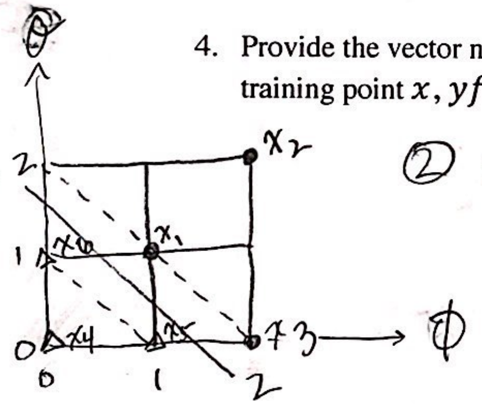
$$x_4 = (0,0)^T, x_5 = (1,0)^T, x_6 = (0,1)^T \text{ with class labels } y_4 = y_5 = y_6 = +1$$

1. Plot these six training points, and plot by inspection the optimal hyperplane.
2. Provide the equation of the optimal hyperplane.
3. Write the equation of the optimal hyperplane in the form:

$$f(x) = w^T x + b = 0.$$

4. Provide the vector normal to the hyperplane. (Hint: keep in mind that for each training point $x, yf(x) > 0$, where y is the class label of x.)

①



② $\theta = 1.5 - 1 \cdot \phi$

I can see the support vectors are x_1 x_3 x_5 x_6
In these cases $\vec{x}_i \cdot \vec{w} + b \in \{-1, 1\}$ exactly. Solving this, I get $b = 3$ and $\vec{w} = (-2, -2)^T$

Scanned with CamScanner

2. Clustering: *k*-means vs. *k*-medoids [25 points]

Consider the following five two-dimensional points:

$$a = (1,0), b = (2,0), c = (-1,0), d = (-2,0), e = (-8,0)$$

1. Cluster the given five points in two groups using the *k*-means algorithm (k=2) with Euclidean distance. The initial centroids are (-1,0) and (1,0). For each iteration of the algorithm, provide the resulting partition and centroids.
2. Cluster the given five points in two groups using the *k*-medoids algorithm (k=2) with Euclidean distance. The initial medoids are (-1,0) and (1,0). Ties are broken in favor of the medoid to the left. For each iteration of the algorithm, provide the resulting partition and medoids.
3. Comment on the differences between the two obtained clusterings in view of the properties (pros and cons) of *k*-means and *k*-medoids.

①

Iter	Center$_1$	Pts$_1$	Center$_2$	Pts$_2$	
0	(-1,0)	{c,d,e}	(1,0)	{a,b}	
1	$(-3.\overline{6},0)$	{c,d,e}	(1.5,0)	{a,b}	
2	$(-3.\overline{6},0)$	{c,d,e}	(1.5,0)	{a,b}	← Stop

②

Iter	Center 1	Pts 1	Center 2	Pts 2	
0	{c}	{c,d,e}	{a}	{a,b}	
1	{d}	{c,d,e}	{a}	{a,b}	
2	{d}	{c,d,e}	{a}	{a,b}	← Stop

③ In this example, the obtained clusterings were the same. However, the cluster centers were different. For k-means, the cluster center was pulled more heavily by outliers (i.e. {e}). Is this good? If we suspect that {e} is not an outlier and we'll see more points like it in the future, maybe it's ok.

the difference in densities between clusters was higher for k-means.

Kmeans density 1 $= \frac{1}{3}(-3.67+8)^{-2}+(-3.67+1)^{-2}+(-3.67+2)^{-2} \approx 0.18$

KMeans density 2 $= \frac{1}{2}(1.5-1)^{-2}+(1.5-2)^{-2} \approx 8$

Kmedoids density 1 $= 1/3\,(-2+8)^{-2}+(-2+1)^{-2} \approx 0.34$

K medoids density 2 $= 1/2\,(1-2)^{-2} = 0.5$

3. Clustering: agglomerative hierarchical clustering [25 points]

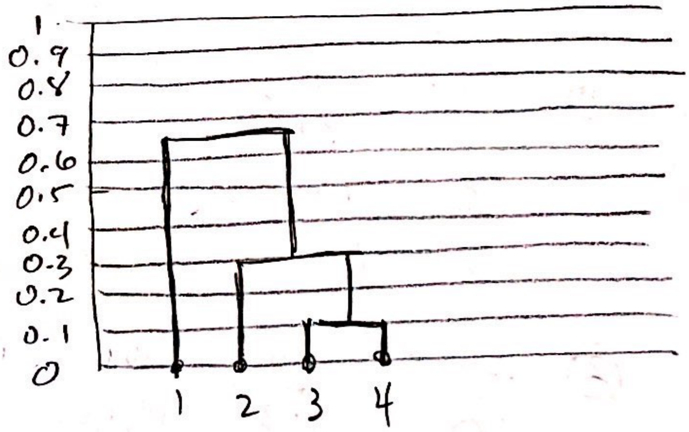
Use the following *similarity* matrix to perform single (MIN) and complete (MAX) link hierarchical clustering (agglomerative). Show your results by drawing a dendrogram. The dendrogram must clearly show the order in which the points are merged, and must show the distance values of the pair of clusters being merged at each step (on the vertical axis).

Point	P1	P2	P3	P4
P1	1	0.68	0.6	0.55
P2	0.65	1	0.7	0.6
P3	0.6	0.7	1	0.9
P4	0.55	0.6	0.9	1

2, 3, 4
3, 1, 4
4, 2 1
3, 2, 1

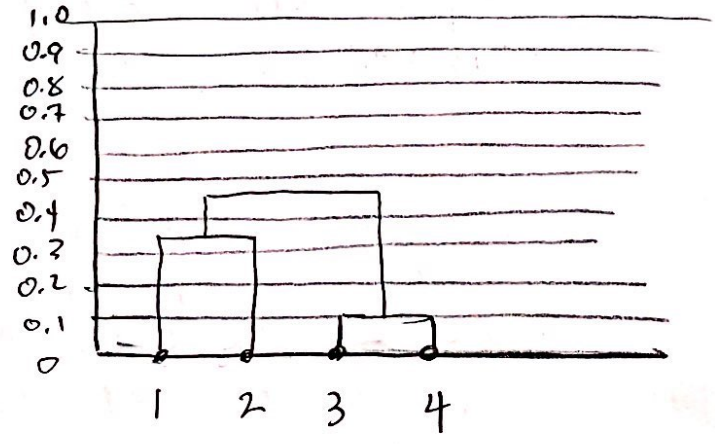
Single Link

Iteration	Clusters
0	{1} {2} {3} {4}
1	{1} {2} {3,4}
2	{1} {2, 3, 4}
3	{1 , 2, 3, 4}



Complete Link

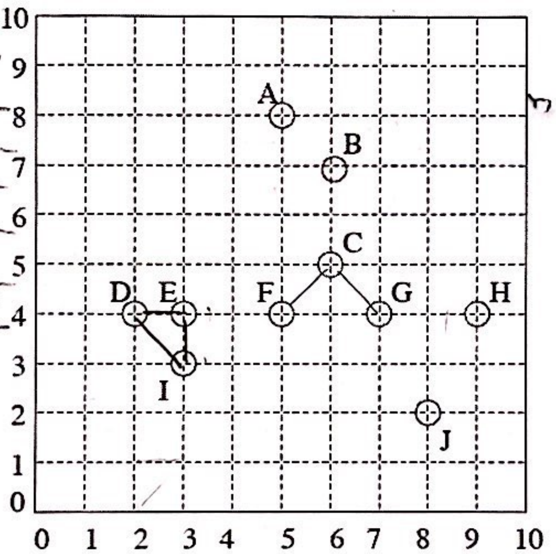
Iteration	Clusters
0	{1}{2} {3} {4}
1	{1} {2} {3,4}
2	{1, 2} {3, 4}
3	{1, 2, 3, 4}



4. **Clustering: DBSCAN** [25 points]

Consider the set of points given in the Figure below. Assume that eps =$\sqrt{2}$ and minpts = 3 (including the center point). Using the Euclidean distance and DBSCAN, find all the density-based clusters. List the final clusters (with the points in lexicographic order, i.e., from A to J) and the outliers. Show your work, i.e. indicate which are the core, the border, and the noise points, and the final clustering.

Point	# Pts w/i Eps	Pts w/i Eps	Class
A	2	+ B	Noise
B	2	+ A	Noise
C	3	+ F, G	Core
D	3	+ E, I	Core
E	3	+ D, I	Core
F	2	+ C	Border
G	2	+ C	Border
H	1	∅	Noise
I	3	+ D, E	Core
J	1	∅	Noise



Final Clusters

1 = {D, E, I}

2 = {C, F, G}

Noise Points = {A, B, H, J}