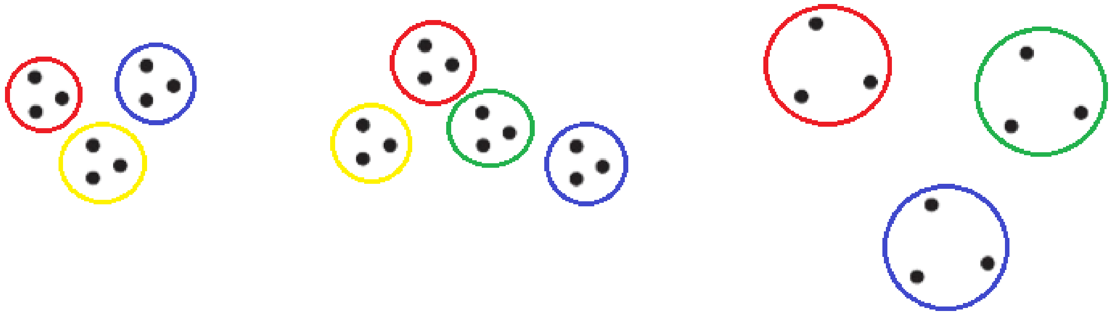
CS 422: Data Mining

Department of Computer Science
Illinois Institute of Technology
Jithin Joyson

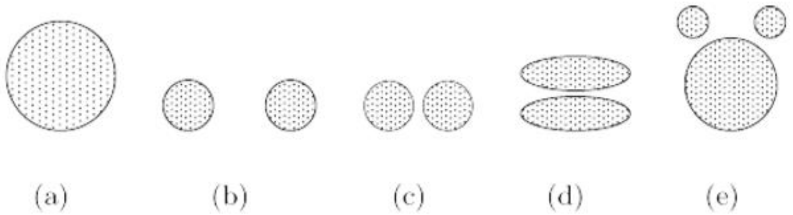
Spring 2019: Homework 2

1 Exercises (2, 6,11,12,16)

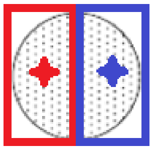
1.1 2.



6.

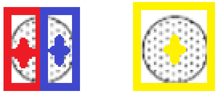

(a) (b) (c) (d) (e)

a.

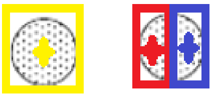


Since the clustering is on a circle and K =2, there is an infinite amount of ways to split the circle into two clusters where both clusters will have the same area and the circle is bisected into 2 at any angle. Since the clusters are the same area and shape, the centroids will be placed in the exact symmetric positions.

b.

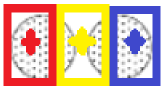


or



Since the circles are separated by a distance greater than the radii of the circle, you cannot divide the whole area of the circles into 3 portions equally. The middle cluster will overlap and take away from the cluster 1 and 2. Therefore, two clusters must share a circle and the last cluster will be the left alone cluster. The shared circle is the same as k =2 from (a) and it can be divided at any angle. The second circle will be an easy capture of the whole circle.

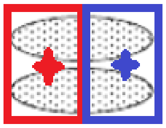
c.



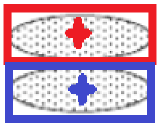
K = 3. The distance between the edges of the circles is much less than the radii of the circles.

Since the circles are separated by less than the radii, introducing a centroid in the middle of the circles yield in a cluster. Thus, the full area of the circles (including the between distances) will be divided as shown.

d.

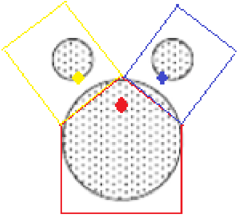


or



Since the circles are ovals, the area of the points is divided equally at 90 degrees or 0 degrees. Furthermore, both create ideal clusters with low SSE but the first division is less optimal since it divides through the separated ovals.

e.



Since k =3 and the proportions of the bottom circle is different from the left and right circles, the top two clusters have to mimic each other in ideal clustering (lowest SSE). Thus, the bottom cluster's centroid moves toward he top to minimize the distances between the centroid and the data. Furthermore, the bottom clusters hold more data than the top two clusters (top two clusters hold the same amount of points) and have to take most of the bottom circle's are to lower the within ss in each cluster.

11. **If the SSE for one variable is low for all clusters**

Since this variable is present in all clusters and there is no variability (low for all clusters), it is meaningless and doesn’t provide much help when creating the clusters.

if the SSE for one variable is low for just one cluster

Since this variable is present in all clusters and the SSE is low for cluster A, then this variable helped when creating cluster A.

if the SSE for one variable is high for all clusters

Since this variable is present in all clusters and there is no variability (high for all clusters), it is meaningless and doesn’t provide much help when creating the clusters. This variable can also be classified as noise since the error is high.

if the SSE for one variable is high for just one cluster

Since this variable is present in all clusters and the SSE is high for cluster A, then this variable doesn’t provide much help when creating cluster A. Other variables with lower SSE relative to other clusters may contribute to creating the cluster though.

How could you use the per variable SSE information to improve your clustering?

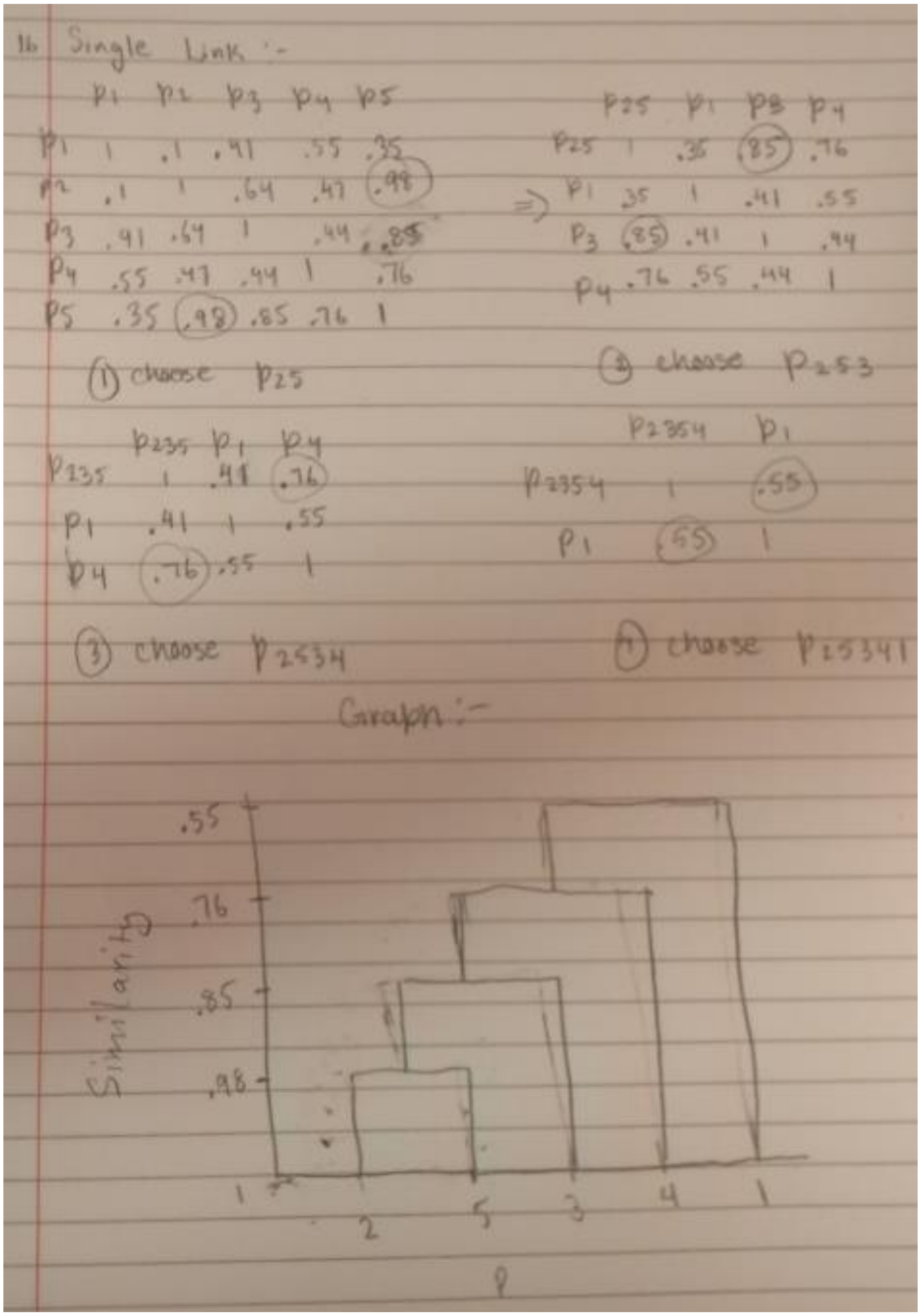
As mentioned before, having a variable with high SSE for all clusters or low SSE for all clusters does not help when creating the clusters. Having high SSE in a cluster for a variable relative to other clusters is good but introduce noise. Finally, having a low SSE in a cluster for a variable relative to other clusters helps the most since it helps create that cluster without worrying about the noise.

12.

a. The K-means algorithm tries to clean up it’s centroids by running several iterations. Compared to the leader algorithm, the k-means takes a longer time solidify the clusters. This is also a disadvantage since it means the leader algorithm is prone to error (SSE) and k-means have better clusters.
Another disadvantage is that the leader algorithm always creates the same clusters and the number of clusters cannot be directly forced as k-means where the k-means algorithm will different based on centroid placements and clusters are dependent on the centroids.
b. Since the biggest disadvantage is that the algorithm does only one pass, we need to address this to have the leader algorithm improved. Thus, the

thresholds and placement of the “leaders” needs to be well thought out. To do this, the distribution of the data needs to be analyzed (by random sampling) and then good leaders need to choose with the closeness of other points to the leader (distances).

16.



16 Single Link :-

	p_1	p_2	p_3	p_4	p_5
p_1	1	.1	.41	.55	.35
p_2	.1	1	.64	.47	.98
p_3	.41	.64	1	.44	.85
p_4	.55	.47	.44	1	.76
p_5	.35	.98	.85	.76	1

(1) choose p_{25}

=>

	p_{25}	p_1	p_3	p_4
p_{25}	1	.35	.85	.76
p_1	.35	1	.41	.55
p_3	.85	.41	1	.44
p_4	.76	.55	.44	1

(2) choose p_{253}

	p_{235}	p_1	p_4
p_{235}	1	.41	.76
p_1	.41	1	.55
p_4	.76	.55	1

(3) choose p_{2534}

	p_{2354}	p_1
p_{2354}	1	.55
p_1	.55	1

(4) choose p_{25341}

Graph :-

Complete Link: (minimize links)

	p1	p2	p3	p4	p5
p1	1	.1	.41	.55	.35
p2	.1	1	.64	.47	.98
p3	.41	.64	1	.44	.85
p4	.55	.47	.44	1	.76
p5	.35	.98	.85	.76	1

⇒

	p25	p1	p3	p4
p25	1	.1	.64	.47
p1	.1	1	.41	.55
p3	.64	.41	1	.44
p4	.47	.55	.44	1

⇒

① Choose p25

② Choose p253

	p253	p1	p4
p253	1	.1	.44
p1	.1	1	.55
p4	.44	.55	1

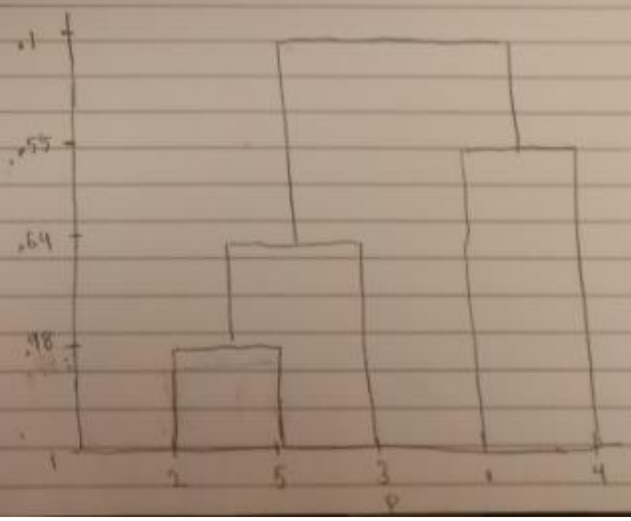
⇒

	p253	p14
p253	1	.1
p14	.1	1

③ Choose p14

④ choose p25314

Graph:


.1
.55
.64
.98
1
2
5
3
1
4
p