CSCI 6350/7350
HOMEWORK 7
VIJAY KOJU

Problem 8.3
$P_1(22, 1, 42, 10)$, and $P_2(20, 0, 36, 8)$.
Euclidean distance $= [(22-20)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2]^{\frac{1}{2}} = 6.71$
Manhatten distance $= |22-20| + |1-0| + |42-36| + |10-8| = 11$
Minkowski distance (q = 3) $= [(22-20)^3 + (1-0)^3 + (42-36)^3 + (10-8)^3]^{\frac{1}{3}} = 6.15$

Problem 8.4
a) Contingency matrices:

		Caroline		
		1	0	sum
	1	2	0	2
Kevan	0	0	2	2
	sum	2	2	4

		Eric		
		1	0	sum
	1	0	2	2
Kevan	0	2	0	2
	sum	2	2	4

		Eric		
		1	0	sum
	1	0	2	2
Caroline	0	2	0	2
	sum	2	2	4

b) Simple matching coefficient
d(Kevan,Caroline) $= \frac{4-4}{4} = 0$
d(Kevan,Eric) $= \frac{4-0}{4} = 1$
d(Caroline,Eric) $= \frac{4-0}{4} = 1$

c) Jaccard coefficient
d(Kevan,Caroline) $= \frac{0+0}{2+0+0} = 0$
d(Kevan,Eric) $= \frac{2+2}{0+2+2} = 1$
d(Caroline,Eric) $= \frac{2+2}{0+2+2} = 1$

d) According to the Jaccard coefficient, Kevan and Caroline would make the best pair of penpals and Kevan and Eric, and Caroline and Eric would be the least compatible.

e) Based on the Jaccard coefficient, including the symmetric variable gender in our analysis would not change the result because it would have an equal effect on all the cases. Hence, Kevan and Caroline would still be the most compatible pair.

Problem 4.6
$A_1(2,10), A_2(2,5), A_3(8,4), B_1(5,8), B_2(7,5), B_3(6,4), C_1(1,2), C_2(4,9)$.
center 1 (Cluster 1) $\rightarrow m_1 = A_1$
center 2 (Cluster 2) $\rightarrow m_2 = B_1$
center 3 (Cluster 3) $\rightarrow m_3 = C_1$

a) Find new cluster centers after the first round execution?

First round execution:

Points	Euclidean distance			Cluster
A_1	$d(A_1, m_1) = 0.0$,	$d(A_1, m_2) = 3.6$,	$d(A_1, m_3) = 8.0$	Cluster 1
A_2	$d(A_2, m_1) = 5.0$,	$d(A_2, m_2) = 4.2$,	$d(A_2, m_3) = 3.2$	Cluster 3
A_3	$d(A_3, m_1) = 8.5$,	$d(A_3, m_2) = 5.0$,	$d(A_3, m_3) = 7.3$	Cluster 2
B_1	$d(B_1, m_1) = 3.6$,	$d(B_1, m_2) = 7.1$,	$d(B_1, m_3) = 7.2$	Cluster 2
B_2	$d(B_2, m_1) = 0.0$,	$d(B_2, m_2) = 3.6$,	$d(B_2, m_3) = 4.1$	Cluster 2
B_3	$d(B_3, m_1) = 7.2$,	$d(B_3, m_2) = 6.7$,	$d(B_3, m_3) = 5.4$	Cluster 2
C_1	$d(C_1, m_1) = 8.1$,	$d(C_1, m_2) = 7.2$,	$d(C_1, m_3) = 0.0$	Cluster 3
C_2	$d(C_2, m_1) = 2.2$,	$d(C_2, m_2) = 1.4$,	$d(C_2, m_3) = 7.6$	Cluster 2

Therefore, new cluster centers after the first round execution are:
$m_1' = (2, 10)$
$m_2' = (\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}) = (6, 6)$
$m_3' = (\frac{2+1}{2}, \frac{5+2}{2}) = (1.5, 3.5)$.

b) Find the final three clusters?

Second round execution:

Points	Euclidean distance			Cluster
A_1	$d(A_1, m_1') = 0.0$,	$d(A_1, m_2') = 5.6$,	$d(A_1, m_3') = 6.5$	Cluster 1
A_2	$d(A_2, m_1') = 5.0$,	$d(A_2, m_2') = 4.2$,	$d(A_2, m_3') = 1.6$	Cluster 3
A_3	$d(A_3, m_1') = 8.5$,	$d(A_3, m_2') = 2.8$,	$d(A_3, m_3') = 6.5$	Cluster 2
B_1	$d(B_1, m_1') = 3.6$,	$d(B_1, m_2') = 2.2$,	$d(B_1, m_3') = 5.7$	Cluster 2
B_2	$d(B_2, m_1') = 7.1$,	$d(B_2, m_2') = 1.4$,	$d(B_2, m_3') = 5.7$	Cluster 2
B_3	$d(B_3, m_1') = 7.2$,	$d(B_3, m_2') = 2.0$,	$d(B_3, m_3') = 4.5$	Cluster 2
C_1	$d(C_1, m_1') = 8.1$,	$d(C_1, m_2') = 6.4$,	$d(C_1, m_3') = 1.6$	Cluster 3
C_2	$d(C_2, m_1') = 2.2$,	$d(C_2, m_2') = 3.6$,	$d(C_2, m_3') = 6.1$	Cluster 1

Therefore, new cluster centers after the first round execution are:
$m_1'' = (\frac{2+4}{2}, \frac{10+9}{2}) = (3, 9.5)$
$m_2'' = (\frac{8+5+7+6}{4}, \frac{4+8+5+4}{4}) = (6.5, 5.25)$
$m_3'' = (\frac{2+1}{2}, \frac{5+2}{2}) = (1.5, 3.5)$.

Third round execution:

Points	Euclidean distance			Cluster
A_1	$d(A_1, m_1'') = 1.1,$	$d(A_1, m_2'') = 6.5,$	$d(A_1, m_3'') = 6.5$	Cluster 1
A_2	$d(A_2, m_1'') = 4.6,$	$d(A_2, m_2'') = 4.5,$	$d(A_2, m_3'') = 1.6$	Cluster 3
A_3	$d(A_3, m_1'') = 7.4,$	$d(A_3, m_2'') = 1.9,$	$d(A_3, m_3'') = 6.5$	Cluster 2
B_1	$d(B_1, m_1'') = 2.5,$	$d(B_1, m_2'') = 3.1,$	$d(B_1, m_3'') = 5.7$	Cluster 1
B_2	$d(B_2, m_1'') = 6.0,$	$d(B_2, m_2'') = 0.6,$	$d(B_2, m_3'') = 5.7$	Cluster 2
B_3	$d(B_3, m_1'') = 6.3,$	$d(B_3, m_2'') = 1.3,$	$d(B_3, m_3'') = 4.5$	Cluster 2
C_1	$d(C_1, m_1'') = 7.8,$	$d(C_1, m_2'') = 6.4,$	$d(C_1, m_3'') = 1.6$	Cluster 3
C_2	$d(C_2, m_1'') = 1.1,$	$d(C_2, m_2'') = 4.5,$	$d(C_2, m_3'') = 6.1$	Cluster 1

Therefore, new cluster centers after the first round execution are:
$m_1''' = (\frac{2+5+4}{3}, \frac{10+8+9}{3}) = (3.67, 9)$
$m_2''' = (\frac{8+7+6}{3}, \frac{4+5+4}{3}) = (7, 4.33)$
$m_3''' = (\frac{2+1}{2}, \frac{5+2}{2}) = (1.5, 3.5)$.

Fourth round execution:

Points	Euclidean distance			Cluster
A_1	$d(A_1, m_1''') = 1.9,$	$d(A_1, m_2''') = 7.5,$	$d(A_1, m_3''') = 6.5$	Cluster 1
A_2	$d(A_2, m_1''') = 4.3,$	$d(A_2, m_2''') = 5.1,$	$d(A_2, m_3''') = 1.6$	Cluster 3
A_3	$d(A_3, m_1''') = 6.6,$	$d(A_3, m_2''') = 1.1,$	$d(A_3, m_3''') = 6.5$	Cluster 2
B_1	$d(B_1, m_1''') = 1.7,$	$d(B_1, m_2''') = 2.0,$	$d(B_1, m_3''') = 5.7$	Cluster 1
B_2	$d(B_2, m_1''') = 5.2,$	$d(B_2, m_2''') = 0.7,$	$d(B_2, m_3''') = 5.7$	Cluster 2
B_3	$d(B_3, m_1''') = 5.5,$	$d(B_3, m_2''') = 1.1,$	$d(B_3, m_3''') = 4.5$	Cluster 2
C_1	$d(C_1, m_1''') = 7.5,$	$d(C_1, m_2''') = 6.4,$	$d(C_1, m_3''') = 1.6$	Cluster 3
C_2	$d(C_2, m_1''') = 0.3,$	$d(C_2, m_2''') = 5.5,$	$d(C_2, m_3''') = 6.1$	Cluster 1

k-means algorithm converges at this point.
Thus, the final three clusters are:
Cluster 1 $\rightarrow (A_1, B_1, C_2)$,
Cluster 2 $\rightarrow (A_3, B_2, B_3)$, and
Cluster 3 $\rightarrow (A_2, C_2)$.