Statistical Analysis of Genetic Data
Math 155, Spring 2008
Jo Hardin
HW #8

Due Friday April 18

1. Explain how SAM calculates FDR. I want the big ideas, not the specific algorithm.

2. For this question, I want you to investigate the triangle inequality for the Pearson distance we discussed in class ($d_P(x, y) = 1 - r_P(x, y)$). I expect you to use R or Excel to do this problem. Please, however, report your correlations and the data values you used in (b).

   (a) Show that the triangle inequality is violated when $x = (1, 1, 0), y = (1, -1, 0), z = (2, 1, 0)$.

   (b) Find three vectors in three space (like in (a) above, except use three different vectors) where the triangle inequality for $d_P(x, y)$ is not violated for any combination of the three variables.

3. What is one main difference between a hierarchical clustering method and a non-hierarchical clustering method?

4. Consider the matrix of distances:

	1	2	3	4
1	0			
2	1	0		
3	11	2	0	
4	5	3	4	0

   Cluster the four items using each procedure:

   (a) Single-linkage hierarchical algorithm

   (b) Complete-linkage hierarchical algorithm

   (c) Average-linkage hierarchical algorithm

   Draw the dendrograms for each, and compare the results in (a), (b), and (c).

5. (We didn't talk about Manhattan distance in class, but the distance function is in your book.) Show that the Euclidean and Manhattan distances do not always result in the same ordering of the distances between pairs of objects. That is, find objects A, B, C, and D such that d(A,B) < d(C,D) for the Euclidean distance and d(A,B) > d(C,D) for the Manhattan distance.

6. Suppose we measure two variables X_1 and X_2 for four items A, B, C, and D. The data are as follows. (Use the Euclidean distance function here.)

	Observations	
Item	x_1	x_2
A	5	4
B	1	-2
C	-1	1
D	3	1

(a) Use the K-means clustering technique to divide the items into k=2 clusters starting with the initial groups (AB) and (CD).

(b) Use the K-means clustering technique to divide the items into k=2 clusters starting with the initial groups (AC) and (BD).

(c) Use the K-means clustering technique to divide the items into k=2 clusters starting with the initial groups (AB) and (CD); this time however, start at the bottom of the list of items and proceed up in the order D, C, B, A.

(d) Compare your 3 solutions. Are they the same? Should they be? Comment.