Assignment #4 STA437H1S/2005H1S

due Friday April 8, 2016

Instructions: All students do both problems.

1. Consider the single and complete linkage clustering methods. Given disjoint clusters A, B, and C, we can (for each of the two methods) define a distance measure $d(A, B)$ between clusters A and B.
(a) For single linkage clustering, show that

$$d(A, B \cup C) = \min\{d(A, B), d(A, C)\} = \frac{1}{2}d(A, B) + \frac{1}{2}d(A, C) - \frac{1}{2}|d(A, B) - d(A, C)|$$

(b) For complete linkage clustering, show that

$$d(A, B \cup C) = \max\{d(A, B), d(A, C)\} = \frac{1}{2}d(A, B) + \frac{1}{2}d(A, C) + \frac{1}{2}|d(A, B) - d(A, C)|$$

(c) Suppose that for single component clusters a, b, c, the triangle inequality holds, that is,

$$d(a, c) \leq d(a, b) + d(b, c).$$

Show that for complete linkage clustering, the triangle inequality $d(A, C) \leq d(A, B)+d(B, C)$ holds for disjoint clusters A, B, and C.

2. [**One last crack at the crabs data!**] In this problem, we will assume that we do not have at our disposal the sex/species identifiers for the 200 crabs and use normal mixture models to try to identify clusters in the data. To this end, you will use the function `EM` (available on Blackboard as `EM.txt`) to estimate the parameters of a two component multivariate normal mixture model. (`EM` is very slow so it would be wise to have other activities planned while it runs!)
As before, the data can be read into R as follows:

```
> x <- scan("crabs.txt",skip=1,what=list("c","c",0,0,0,0,0,0))
> FL <- x[[4]]
> RW <- x[[5]]
> CL <- x[[6]]
> CW <- x[[7]]
> BD <- x[[8]]
> y <- cbind(FL,RW,CL,CW,BD)
```

(a) Start by doing 30 iterations of the EM algorithm:

```
> r30 <- EM(y,k=2,em.iter=30)
```

The component `r$cluster` will contain the estimated cluster (either 1 or 2) for each observation. These can be seen on the pairwise scatterplot as follows:

```
> colour <- rep("blue",200)
> colour[r30$cluster==2] <- "red"
> pairs(y,col=colour)
```

Do the clusters estimated after 30 iterations seem reasonable?
(b) Repeat the procedure in part (a) now using 100 iterations of EM:

```
> r100 <- EM(y,k=2,em.iter=100)
```

Comment on the difference the estimated clusters here and those from part (a).
(c) [Optional but recommended] Repeat the procedure above estimating four clusters; you will probably need to significantly more than 100 iterations of the EM algorithm. How do the clusters compare to the sex/species groupings?