B365 Homework 8

1. Suppose our data set is given by $\{(0,1)^t, (1,1)^t, (0,2)^t, (1,2)^t, (0,3)^t, (1,3)^t\}$ and we begin with $m_1 = (0,1)^t$ and $m_2 = (0,2)^t$ as our initial configuration.

   (a) Construct the initial partition of the data constructed by the K-means algorithm and give the mean for each set in the partition.

   (b) Working from the mean values you constructed in the previous part, compute the partition and mean values for the 2nd iteration of the algorithm.

   (c) Carry the algorithm through to completion giving the final resulting clusters and mean values.

   (d) How can you tell that you have reached the terminal point of the algorithm?

2. Let $x_1, \ldots, x_n$ be a collection of numbers and consider $H(m) = \sum_{i=1}^{n}(x_i - m)^2$ for an unknown number m. We can minimize this function by differentiating with respect to m and setting the derivative equal to 0. Do this to show that the minimizing value is $m = \frac{1}{n}\sum_{i=1}^{n} x_i$.

3. Consider the K-means algorithm in one dimension (clustering a collection of numbers, $x_1, \ldots, x_n$), using the usual interpretation of our clustering function $c(i)$. That is, $c(i)$ is the cluster that x_i currently belongs to.

   (a) Show that the objective function
   $$H = \sum_{k=1}^{K} \sum_{i:c(i)=k} (x_i - m_k)^2$$
   decreases at each iteration of the algorithm.

   (b) Argue the the algorithm cannot cycle (repeat an earlier configuration of the clustering function), thus must terminate.

4. For the data points in problem 1, find a "stable configuration" of m_1 and m_2 that is not "globally optimal." A stable configuration is one that doesn't change after an iteration of the K-means algorithm. Globally optimal giving the lowest value of the objective criterion from the previous problem when compared with all possible choices of m_1 and m_2.