Advanced Statistical Inference
Clustering

1 Aims

- To implement K-means and use it to cluster some data.

This lab exercise is a bit more open ended than previous ones. You should by now be familiar with Matlab – you'll need some of the techniques you've used in previous labs.

1. Download `kmeansdata.mat` from the course webpage.
2. Implement the K-means algorithm covered in the lecture (not the kernelised version). Hints:
    - The distance between all of the data points in `X` and a mean vector `mu(k,:)` can be computed by:

      ```
      di(:,k) = sum((X - repmat(mu(k,:),N,1)).^2,2)
      ```
    - Alternatively, you could write a loop over the N data points.
    - Your code needs to alternate between assigning points to the cluster that they're closest to and recalculating the means by taking the average of these points.
    - You may need to add something to your code to deal with the problem of no points being assigned to a particular cluster – set the mean randomly.
    - You'll find it easiest to maintain a variable that indicates which cluster each point is assigned to. Be careful not to assign a point to more than one cluster.
    - Your algorithm should converge in fewer than 20 steps (or thereabouts). If it takes 100, something isn't right!
3. Run your algorithm for $K = 2, 3, 4, 5$ and, in each case, plot the data using a different symbol for each cluster. Show the results