ML4Bio 2012

Assignment #2

Due Date: February 10th, 2012 at the beginning of class

Email your completed assignments to ml4bio@gmail.com

Please include any R code you used for answering the questions, as well as a short description of what you did to answer those questions and any figures that you generated. We can read PS, PDF or Word documents.

Question #1: Download the `ass2-simpleclust.txt` data file from the course website and read the file into R.

(a) Plot the data in 2-D.

(b) Calculate a distance matrix and cluster it using `hclust()`, plot the results.

(c) Plot the centres you get from a K-means with the data. How many centers did you use?

(d) Apply mclust to these data, what is the best model? Do you agree? Why or why not?

Question #2: Download the `ass2-hardclust.txt` data file from the course website and read the file into R.

(a) Plot the data in 2-D

(b) Calculate a distance matrix and cluster it using `hclust()`, plot the results using a heatmap – what do you see?

(c) Try K-means with two centres, four centres, and ten centres. Plot the data with the centres, what do you see?

(d) Apply `mclust` to these data (you'll need to install the package), what is the best model? Plot the model with the centres (and possibly the standard deviations). Do you think that `mclust` found the right model? Why or why not?

Question #3: Download the yeast microarray data file `ass2-phodata.txt` from the course website and read it into R. In this file, each row represents a gene and each column represents a microarray experiment. You will be clustering genes not experiments.

(a) One important practical consideration when choosing clustering algorithms is how their performance depends on dataset size. Using the microarray data you downloaded, time the performance of hierarchical clustering and Gaussian mixture model-based clustering for subsets of the data containing ~200, 500, 1000, 2000, and 5000 points. Plot the time each algorithm takes as a function of the dataset size.

How do the algorithms scale with the number of datapoints? (hints: you'll need to convert the data to a 'matrix' in order to apply the clustering algorithms. Once you have the matrix, you'll also need to remove all missing datapoints: if X is the matrix, `X<-X[rowSums(is.na(X))==0,]` will keep only the rows that have no missing data. Finally, you can use the `system.time()` command to tell you how long it takes for any R command to run on your computer.)

(b) Repeat the experiment in (a), but now keep the number of datapoints constant and increase the number of microarray experiments from 2,4 to 8. How do the algorithms scale with the number of dimensions?

(c) In this yeast dataset, a small subset of genes shows consistent increased expression. Use K-means clustering to find these genes. What are the genes? Use gene set enrichment analysis to find out what functions they are involved in.