BMI/CS 576 Fall 2015
Homework #5

Prof. Colin Dewey

Due Monday, December 7th, 2015 by 11:59pm

The goal of this assignment is to become more familiar with clustering methods.

To turn in your assignment, copy all relevant files to the directory:
`/u/medinfo/handin/bmi576/hw5/USERNAME`
where `USERNAME` is your account name for the BMI network. Please note the homework policies posted at `http://www.biostat.wisc.edu/bmi576/hw.html`

1. (20 points) Run the k-means algorithm on the following set of one-dimensional points: $X = (x_1, x_2, x_3, x_4, x_5) = (2, 4, 5, 8, 10)$. Let $k = 2$ and the initial cluster centers be $f_1 = 2$ and $f_2 = 5$. After each iteration, show (i) the assignment of points to clusters and (ii) the updated cluster centers.

2. (20 points) Run the EM algorithm for Gaussian mixture model-based clustering on the set of points in Problem 1 for three iterations. Let $k = 2$, the initial cluster means be $\mu_1 = 2$ and $\mu_2 = 5$, the initial cluster prior probabilities be $P_1 = P_2 = 0.5$, and the variances be $\sigma_1^2 = \sigma_2^2 = 1$. After each iteration, show (i) the probabilities of each point being assigned to each cluster, (ii) the updated cluster means, and (iii) the updated cluster prior probabilities. You should treat the variances as fixed parameters that are not updated during EM.

3. (60 points) Write a program, `cluster`, that implements the agglomerative hierarchical clustering discussed in class. Your program should perform a full hierarchical clustering, but output a description of a specified number of partitional clusters (the top k clusters represented in the hierarchy) using the method discussed in class. The program will be designed to take as input a small gene expression data set. Your algorithm should use Euclidean distance as the distance between two expression profiles, and be able to do single-link, average-link, or complete-link when determining the distances between clusters.

   Your program should be run via a simple shell script `cluster.sh` that calls your program in the appropriate manner. A template script is available from:

`http://www.biostat.wisc.edu/bmi576/hw/hw5/cluster.sh`
The script will take three command-line arguments:

(a) the name of a file containing the expression data

(b) a single character indicating the type of hierarchical clustering to perform ('S' for single-link, 'C' for complete-link, and 'A' for average link),

(c) an integer value, k, indicating the number of clusters to return.

The expression data file will contain a table in tab-delimited format. Each line of the file describes a single gene, and tabs are used to delimit the individual columns of a line. The first column lists an identifier for each gene. The second column lists a common name for the gene and a description of its function. The remaining columns list expression values for the gene under various conditions. These values are log expression ratios. Your program should not use the gene descriptions in doing the clustering.

Your program should not output a description of the complete hierarchy induced by the clustering method. Instead you should print a description of each of the k highest clusters represented by the hierarchy. In particular, your program should list the identifiers and descriptions of the genes that belong to each cluster. You should list these genes, one per line, on consecutive lines. Additionally, after listing the genes that belong to a given cluster, you should list the average expression values for the cluster on a separate line. There should be a blank line between each of these cluster descriptions. You should print the clusters and the genes within each cluster in order as follows. The clusters should be ordered by the average expression ratio, across the various measurements, of the genes they contain (from smallest to largest). Similarly, the genes within a cluster should be ordered by their average expression ratio (from smallest to largest).

Sample input and output files, are provided at: `http://www.biostat.wisc.edu/bmi576/hw/hw5/` For example, using the tiny example input, below is a command we should be able to run with its expected output:

```
$ sh cluster.sh tiny-yeast-set S 2
YPL090C RPS6A  PROTEIN SYNTHESIS        RIBOSOMAL PROTEIN S6A -0.468
YOR182C RPS30B PROTEIN SYNTHESIS        RIBOSOMAL PROTEIN S30B -0.298
YOR369C RPS12  PROTEIN SYNTHESIS        RIBOSOMAL PROTEIN S12 -0.295
-0.354

YPR001W CIT3   TCA CYCLE                CITRATE SYNTHASE -0.140
YLR038C COX12  OXIDATIVE PHOSPHORYLATIO CYTOCHROME-C OXIDASE, SUBUNIT VIB -0.106
YGR270W YTA7   PROTEIN DEGRADATION      26S PROTEASOME SUBUNIT; ATPASE -0.034
YLR395C COX8   OXIDATIVE PHOSPHORYLATIO CYTOCHROME-C OXIDASE CHAIN VIII -0.022
YKL145W RPT1   PROTEIN DEGRADATION, UBI 26S PROTEASOME SUBUNIT -0.018
```

```
YGL048C RPT6   PROTEIN DEGRADATION      26S PROTEASOME REGULATORY SUBUNIT 0.029
YDL066W IDP1   TCA CYCLE                ISOCITRATE DEHYDROGENASE (NADP+) 0.043
YFL018C LPD1   TCA CYCLE                DIHYDROLIPOAMIDE DEHYDROGENASE 0.074
YGR183C QCR9   OXIDATIVE PHOSPHORYLATIO UBIQUINOL CYTOCHROME-C REDUCTASE SUBUNIT 9 0.092
-0.009
```

(Note that the output ends in a blank line).