

CBB 200 Problem set 3

Due April 14

- D.1** There are mn patients in a hospital ward and their beds are arranged as n rows by m beds (m, n are even). Suppose one of the patients in bed $m/2$ in row $n/2$ becomes infected and can infect a patient in one of the four neighboring beds. Once a patient is infected they stay contagious for k days and then recover and are cured. For every day of infection the probability that each neighbor gets infected is τ . For any day t what is on average $I(t)$ – number of infected patients, $S(t)$ – the number of patients never infected, and $R(t) = 1 - I(t) - S(t)$ – the number of patients that recovered.

Plot this model for $m = n = 100$ and $\tau = .2$ and $m = n = 10$ and $\tau = .2$.

The patients can now move. The nursing staff sometimes moves patients to other beds. Assume with probability δ a patient will swap with a patient at bed $\lfloor r_2 n + 1 \rfloor, \lfloor r_1 m + 1 \rfloor$ where $r_1, r_2 \sim U[0, 1]$.

Rerun the model with mobility as $\delta = .01$.

We could vaccinate patients. Each day each susceptible individual is vaccinated with probability $\nu = .1$. Rerun the model.

Run the final model 1000 times and plot a histogram. Do this for a variety of ν values.

- D.2** In this problem you have a choice of running either a Support Vector Machine or regularized logistic regression on a gene expression data set. The data set can be found at <http://www.stat.duke.edu/~sayan/homework3/mega.svmfu> each row is a sample and the final column, 16064, is the label. Run a leave-one-out error analysis on this data set.
- D.3** This problem is taken from a CMU machine learning class. For the complete version goto <http://www.cs.cmu.edu/~awm/10701/2004/hw2.pdf>

Goto the webpage

http://www.cs.cmu.edu/~awm/10701/2004/hw2_data.zip

In this problem you will implement a Gaussian mixture model algorithm and will apply it to the problem of clustering gene expression data. The data you will be working with is from yeast, and the measurements were taken to study the cell cycle system in that organism. At the end of each sub-problem where you need to implement a new function we specify the prototype of the function.

1. The file 'alphaVals.txt' contains 18 time points (every 7 minutes from 0 to 119) measuring the log expression ratios of 745 cycling genes. Each row in this file corresponds to one of the genes. The file 'geneNames.txt' which contains the names of these genes. For some of the genes, we are missing some of their values due to problems with the microarray technology. These cases are represented by values greater than 100.
2. Implement an EM algorithm for learning a mixture of five (18-dimensional) Gaussians. It should learn means, covariance matrices and weights for each of the Gaussian. You can assume, however, independence between the different data points, resulting in a diagonal covariance matrix. How can you deal with the missing data? Plot the mean of each of the five classes.

The prototype function is:

$$\text{function}[\mu, s, w] = \text{emcluster}(x, k, \text{ploton});$$

x is input data, where each row is an 18-dimensional sample. Values above 100 represent missing values. k is the number of desired clusters. ploton is either 1 or 0. If 1, then before returning the function plots log-likelihood of the data after each EM iteration (the function will have to store the log-likelihood of the data after each iteration, and then plot these values as a function of iteration number at the end). If 0, the function does not plot anything. The function outputs μ , a matrix with k rows and 18 columns (each row is a center of a cluster), s is also k by 18, with each row being diagonal elements of the corresponding covariance matrix, and w is a column vector of size k , where $w(i)$ is a weight for i th cluster.

D.4 Multidimensional scaling is currently used as an exploratory statistical technique to visualize and characterize ones data. There is much debate over whether one should use MDS in a statistical inference setting—drawing new conclusions based on MDS. This problem will require you to compare clustering techniques from several different data sets and should introduce you to the clustering packages present in R.

R has various packages that can be used for clustering analysis. For more thorough analysis about clustering methods available in R type `help.search("clustering")`. Of importance to you for this assignment will be the hierarchical clustering packages (`library(clustering)`). The file `cluHWDat.zip` contains all the data files as well as a `clustering.R` file which demonstrates the R methods that you will use and can be downloaded from <http://people.genome.duke.edu/~d1m19/cluHWDat.zip>.

We will be comparing the clustering capability of gene expression data from several patients that can be grouped into two classes: patients with lung cancer, and patients without lung cancer. We will be using a small subset of the genes for our cluster analysis— 25. Each group has 16 different expression samples pulled from each class. Instead of clustering based on the presence of a tumor, we are instead interested in how the underlying gene pathways are related to each other— or how the individual genes cluster. We define the similarity between genes as the r between each gene across all samples for each class (where r is the pearson correlation value). We do this for all classes between all genes. We can now create our dissimilarity matrix by subtracting one from our r values. This creates a three dimensional matrix where δ_{ijk} is the dissimilarity between gene i and j for class r (e.g $\delta_{5,8,0}$ is the dissimilarity of gene 5 from 8 for a non cancer patient (where 0 = no cancer, 1 = cancer)).

We can now solve a Threeway MDS problem. We seek to minimize the stress of the function

$$S = \sqrt{\frac{\sum_{ijk} (d_{ijk} - \delta_{ijk})^2}{\sum_{ijk} d_{ijk}^2}}$$

where we define

$$d_{ijk} = \left\{ \sum_{t=1}^p w_{kt} (x_{it} - x_{jt})^2 \right\}^{\frac{1}{2}}$$

. We optimize the weights and coordinates of our stimuli to minimize this function. The solutions are provided for you.

1. Four MDS solutions were solved for 2, 3, 4, and 5 dimensions and are saved as stimuli2.out, stimuli3.out, stimuli4.out, and stimuli5.out respectively. Run Hierarchical clustering on all four of these, plot them. What shape does the dendrogram take as the dimension increases, is it more branched, less branched? What possible explanation does this give for those MDS solutions in lower dimensions?
2. Using only the two dimensional MDS, run a k-means clustering with $k = 3$. Compare it to your hierarchical clustering results, how accurate are they? Calculate silhouette measures on both, which seems to be the better fit? Use clusplot to visualize the partitioning of both k-means and your hierarchical solution.
3. MDS provides neither a robust nor accurate dimension reduction mechanism in this situation. This is in part due to the fact that many of these genes are not correlated with each other at all—leaving high dissimilarity measures between all genes. Suggest another application of bioinformatics where MDS might be more applicable to.