Homework 9
BIOS-7659/CPBS-7659
Due 12/10 9AM on Canvas

1. Classification

- Download the data provided on Canvas (dataHW9-breastcancer.Rdata). This file contains the "breastcancer" object, which is a dataset for 46 breast tumor samples where 23 are positive for an estrogen receptor (ER+) and 23 were negative (ER-) (West et al., PNAS 2001 98:11462-11467). This data set contains processed expression levels for 7129 genes (in variable `x`) by 46 samples (in variables `y`).
- Install the `class` package from CRAN.
- Use the following code to read in the data

```
library(class) #package with the knn() function
load("dataHW9-breastcancer.Rdata")
train = breastcancer[[1]] #training expression data
trainclass = breastcancer[[2]] #training classes
test = newpatients #new expression data
testclass = trueclasses #new classes
#To run knn(), take the transpose of "train" (or "test")
```

(a) Describe the k-nearest neighbor algorithm and how it classifies observations. Using the function `knn()` from the `class` package, run k-nearest neighbors with $k = 3$ to train and test on the same training data set (`train`). What percentage of subjects were correctly classified? It is not good practice to train and test on the same data set, why not?

(b) Repeat part a) but for multiple values of k. What values and range of k are suitable for k-nearest neighbor? Plot k versus error rate. What value of k would you select based on this plot and why?

(c) Using k selected in part b), predict the tumor class for three new subjects in `newpatients` using their k nearest-neighbors in the training data. The correct classes are in `trueclasses`. How well did you do?

(d) Extra credit: Perform 5-fold cross validation to determine your error rate (perform on only one random partitioning). Since 46/5 is not an integer, create 4 folds of 10 samples, and one fold with 6 samples. Average over the 10-sample folds (so only 4 folds). Plot your results for different values of k. How does this compare to the error rates from part a) above.

(e) Extra Credit: Write your own code for k-nearest neighbor classification using one minus the absolute correlation as a distance. Apply your code to the samples and plot the error rate for different values of k. What value for k do you select based on the training data? How does that compare with the results using the Euclidean distance from part a).

2. Clustering

- Download the data provided on Canvas (dataHW9-cellcycle.txt) for yeast gene expression over two cell cycles (Cho *et al.*, Molecular Cell 1998, 2:65-73). In this experiment, mRNA was extracted from yeast cells at 10 minute intervals after reinitiation of the cell cycle. The data entries are fluorescence intensities for a single dye at each time point. The first column is the yeast gene ID and the second column is the gene name. HSP genes encode heat shock proteins, RPS genes encode ribosomal proteins and MCM genes encode for members of the MCM (mini-chromosome maintenance) complex.
- Read sections 10.4 and 11.5.1 provided on Canvas (Computational Genome Analysis, Deonier, Tavare & Waterman, 2005).
- The `stats` package contains both the `hclust()` and `kmeans()` functions used in this problem.

(a) Using the R code in section 11.5.1 as an example, run the k-means algorithm to cluster the data (use the standardized data - see Step 2). Try different values for $k = 2, 3, \ldots$ and create a plot of "Within Sum of Squares" versus k. What is your best choice of k? For the final clusters, what genes are grouped together?

(b) Using the R code in section 10.4 as an example, apply the hierarchical clustering algorithm on the data. Calculate the Euclidean distance (using `dist()` then `as.dist()`) with standardized data. How does the cluster membership obtained with hierarchical clustering compare with the results from k-means in part a)? Describe the "method" option in `hclust` discussed in class. Try different values for "method". What happens when you change this option? Repeat this analysis with the Manhattan (L_1) distance instead. How do the results change?