

BINF 733 CSI739 – Spring 2004 Weller/Solka
Dimensionality Reduction HW - 1

Obtain the Golub data set and class files from our 2003 class website <http://www.binf.gmu.edu/%7Ejsolka/s2003/csi739/csi73903.htm>. Please try to have the problems completed by 2/26/04. Email me your solutions, jsolka@gmu.edu, as a pdf file if possible.

#1 (20 pts) View the data as 72 observations in 7129 dimensions. Plot the AML, ALL T-cell and ALL B-cell observations in the first two principal components. Use a different symbol and color for the AML, ALL T-cell and ALL B-cell observations. Do the work in both MATLAB and R.

#2 (20 pts) Provide a “scree plot” for the data set of problem number 1. Do the work in both MATLAB and R.

#3 (20 pts) Let’s consider the Golub data again. Suppose one would like to be able to distinguish between the ALL observations and the AML observations. Write an R and MATLAB function that will take compute the ratio of S_B/S_W for the 7129 Golub features. Provide a sorted list of the feature numbers from the feature with the largest S_B/S_W ratio to the feature with the smallest S_B/S_W ratio.

#4 (20 pts) Plot the AML, ALL T-cell and ALL B-cell observations as a scatter plot in the first 3 coordinates (gene expression levels) that were obtained in problem 3. Do the work in both MATLAB and R.

#5 (20 pts) Download and install the R tree package by Brian Ripley from <http://lib.stat.cmu.edu/R/CRAN/>. Using this as a starting point attempt to implement code to perform the random forests gene selection discussed in class. Apply your code to the Golub data. How do your results compare to those discussed in class and in the paper on the author’s website. Implement all code in R.

#BONUS (10 pts) Download the Golub Leukemia dataset from

http://www-genome.wi.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43

and parse the dataset using either a MATLAB or R parser.