CSC 411, Fall 2006 — Assignment #4

*Due at **start** of lecture on December 6. Note that this assignment is to be done by each student individually. You may discuss it in general terms with other students, but the work you hand in should be your own. Handing in work that is not your own is a serious academic offense. Fabricating results, such handing in fake output that was not actually produced by your program, is also an academic offense.*

For this assignment, you will try using PCA to reduce the dimensionality of gene expression data obtained using DNA microarrays, and will also see how such a reduction in dimensionality affects performance of a k-nearest-neighbor classifier.

The data (available from the course web page) gives the expression levels of 2000 genes in tissue samples from 62 people. It was used in the following paper:

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999) "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", *Proceedings of the National Academy of Sciences (USA)*, vol. 96, pp. 6745-6750.

40 of the tissue samples are from people with colon cancer; 20 are from people without cancer. A second data file contains these indicators.

In this assignment, you will look this data in two ways — as data on 62 tissue samples (with 2000 variables for each sample), and as data on 2000 genes (with 62 variables for each gene). In both cases, you should reduce dimensionality using PCA. You should subtract the sample mean of each variable when doing this, but not divide by the standard deviation. (The variables are all measurements of the same type, so the original scaling may be meaningful.)

You should find the principal component directions and the projections of the points on these directions by applying matrix operations in functions you write yourself, not using some existing package for PCA. You should use the method discussed in the lecture slides for finding principal components when the number of variables is larger than the number of points.

You should reduced the dimensionality of the data on the 2000 genes to the first 10 principal components. You should then look at scatterplots of pairs of principal components, and comment on whether anything that seems of interest can be found.

You should reduce the dimensionality of the data on the 62 tissue samples to the first 25 principal components. You should then consider the performance of a k-nearest-neighbor classifier (with Euclidean distance) for whether the tissue is cancerous or not, with $k = 5$. You should try this using the original data (2000 variables), using just the first principal component, using just the first two principal components, etc., up to using the first 25 principal components. You should evaluate performance using leave-one-out cross validation, in terms of error rate (plus any other measure you find interesting).

You may use the k nearest neighbor classifier (in R) on the course web page, or any other software for doing that. You should write your own function for doing the leave-one-out cross validation, however.

You should hand in listings of the functions you wrote, a discussion of the results, and suitable plots or other output that illustrate and justify your conclusions.