ECE 532: Homework 1

Due on Tuesday, Sept. 9

Robert Nowak 11:00 am

Elijah Bernstein-Cooper

September 9, 2014

Problem 1

Design an algorithm to find the genes that are most predictive of a disease or phenotype. DNA microarrays can be used to measure the amount of protein produced by each gene (the so-called "gene expression level"). Suppose we measure the gene expression levels of a number of people with and without a particular disease. Assume m genes are measured. What is your data matrix A? How would you decide which genes are most important to or involved in the disease process?

Our data matrix, A, will be a $k \times m$ matrix, with k people and m genes. To determine which genes are most likely to cause disease we first assign a label matrix L, a column vector of length k where the i^{th} element corresponds to the label for the i^{th} person in our matrix A. Diseased people will be labeled 1, healthy people will be labeled 0.

Our problem will then be reduced to solving the linear equation

$$Ax = L \tag{1}$$

where x will be a column vector of length m. We then solve for x. x will describe the relative contributions of genes to disease or health.

Problem 2

Now suppose we have gene expression data for n different strains of yeast. We also measure phenotypic similarities between the different strains (e.g., similarity measures based on the observable characteristics or traits of the strains like shaper or color). In other words, we have an n x n matrix of similarity values, say ranging continuously between 0 (dissimilar) to 1 (identical). How would you determine which genes are most important for predicting phenotypic similarities?

We have an $n \times n$ matrix S with similarity values between the i^{th} yeast strain and the j^{th} yeast strain. The matrix will be symmetric about the diagonal. We can compute a singular value decomposition of S, whereby we can rank the singular values to rank the phenotypical similarities.