MAE 345
Robotics and Intelligent Systems

Assignment #6
due: December 15, 2011

The goal of this assignment is to use a neural network to classify genetic data as belonging to one of three breast cancer types. Using cDNA microarrays, gene expression levels have been measured for 22 tissue samples[1], each of which is characterized by over 3,200 genes. Seven samples were drawn from tumors that had mutated BRCA1 genes, eight were drawn from tumors with mutated BRCA2 genes, and seven were drawn from sporadic tumors, i.e., tumors in which neither gene is mutated.

The expression levels – at this point, numbers on a spreadsheet – form the basis for classification. The first step is to find those genes that are most likely to be over- or under-expressed in each tumor class. The paper by Hedenfalk *et al* used 51 genes to perform the classification. The second step is to train neural networks that classify tumor samples on the basis of the genetic data alone. You are encouraged to use the MATLAB Neural Network Toolbox to do this assignment. The manual for this toolbox is at *http://www.mathworks.com/access/helpdesk/help/pdf_doc/nnet/nnet.pdf.*

An Excel spreadsheet of this dataset can be found on the *MAE 345 Blackboard* web page. Each row of the spreadsheet contains the microarray measurements of a single gene type for all of the samples, as suggested by the table below. Each column contains all of the gene measurements for a single sample. The first 7 samples were taken from tumors with mutation of the BRCA1 gene, the second 7 samples were taken from tumors with mutations of the BRCA2 gene, and the last 7 samples were taken from sporadic tumors.[2]

	mutant BRCA1 sample #1	mutant BRCA1 sample #2	...	mutant BRCA2 sample #8	mutant BRCA2 sample #9	...
Gene #1	138	68	...	14	-51	...
Gene #2	3	23	...	33	29	...
Gene #3	...	...	...	...	...	...

[1] Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O.-P., Wilfond, B., Borg, Å., and Trent, J., "Gene-Expression Profiles in Hereditary Breast Cancer, *The New England Journal of Medicine*, Vol. 344, No. 8, Feb. 22, 2001, pp. 539-548. (Full text at http://www.nhgri.nih.gov/DIR/Microarray/NEJM_Article.pdf. Supplemental information at http://research.nhgri.nih.gov/microarray/NEJM_Supplement/)

[2] Two mutant-BRCA2 samples were taken from the same patient. Only the first of these, identified as Sample 10a, is included in the training set. Sample 10b is included on the spreadsheet; it can be used as a novel sample for validation.

The mean values and standard deviations for gene expression levels of each class also are presented on the spreadsheet. A few additional statistics have been calculated. We'll describe these for the mutant-BRCA1 case, and those for mutant-BRCA2 and sporadic cases are defined accordingly. The remainder mean value for each gene is

$$Remainder\ Mean\ Value = Average(BRCA2\ levels\ and\ Sporadic\ levels)$$

The remainder standard deviation for each gene is

$$Remainder\ Standard\ Deviation = Standard\ Deviation(BRCA2\ levels\ and\ Sporadic\ levels)$$

The ratios of mutant-BRCA1 to remainder means and standard deviations are presented, as are the *t* values for each gene in each class comparison.

1) For each of the three cases, choose 6 gene sets (i.e., rows) that you feel best characterize the cases. You might, for example, choose the 6 that have highest or lowest *t* value for mutant-BRCA1 tumors (and then for mutant-BRCA2 and sporadic tumors), highest or lowest mean ratio, highest or lowest standard deviation ratio, or some combination of these. It may be easiest to do this step in Excel, using the sorting operation. As an alternative, you might choose 18 of the 51 genes used in the Hedenfalk paper. Be sure to make a backup copy of the spreadsheet if you choose to modify it. List the 18 genes chosen and the reason for choosing them.

Your neural network will have three outputs, each of which should be zero or one, depending on the prediction made by the network. The first output is one if "mutant-BRCA1" is predicted and zero otherwise; the second is one if "mutant-BRCA2" is predicted and zero otherwise; the third is one if "sporadic" is predicted and zero otherwise. These values establish the target matrix for supervised training of the networks. The target matrix contains (3 x 21) elements, with ones and zeros where appropriate. The training information is contained in an (18 x 21) matrix of gene expression levels (i.e., the input contains 18 rows with 21 elements each, with ordering corresponding to the target vector ordering) and the target vector (the intended output).

You can cut-and-paste the training data from the spreadsheet into your MATLAB m-file. In operation following training, the neural network would accept a single gene set (i.e., a new column of data) and place a "1" or "0" in the appropriate output as a prediction that the gene sample represents a mutant BRCA1, mutant BRCA2, or sporadic tumor. It is likely that the outputs would not be precisely "1" or "0", so results should be rounded to the nearest integer.

2) Design a two-layer sigmoid network with 18 gene inputs and 3 outputs to classify the entire data set (21 samples) into three categories: mutant BRCA1, mutant BRCA2, or sporadic tumor. The first layer contains sigmoid nodal activation functions, and the second layer contains linear nodes. Vary the number of sigmoid nodes (e.g., 3, 6, and 9) to assess the effect on network training. Show a

plot of the network training convergence, and comment on the accuracy of the classification as functions of the number of sigmoid nodes, the learning rate or algorithm, and the number of epochs used for training. Indicate the number of incorrect classifications in your results.

3) Use leave-one-out validation to assess the effect of the number of sigmoid nodes in the first layer, and comment on your results.

4) Use your trained neural network to classify Sample 10b. What does the network predict?