Northeastern University
College of Science

Homework 13

1. (70 points) Analysis of the ALL data set.
(a) Define an indicator variable IsB such that IsB=TRUE for B-cell patients and IsB=FALSE for T-cell patients.

(b) Use two genes "39317_at" and "38018_g_at" to fit a classification tree for IsB. Print out the confusion matrix. Plot ROC curve for the tree.

(c) Find its empirical misclassification rate (mcr), false negative rate (fnr) and specificity. Find the area under curve (AUC) for the ROC curve.

(d) Use 10-fold cross-validation to estimate its real false negative rate (fnr). What is your estimated fnr?

(e) Do a logistic regression, using genes "39317_at" and "38018_g_at" to predict IsB. Find an 80% confidence interval for the coefficient of gene "39317_at".

(f) Use n-fold cross-validation to estimate misclassification rate (mcr) of the logistic regression classifier. What is your estimated mcr?

(g) Conduct a PCA on the scaled variables of the whole ALL data set (NOT just the two genes used above). We do this to reduce the dimension in term of genes (so this PCA should be done on the transpose of the matrix of expression values). To simply our future analysis, we use only the first K principal components (PC) to represent the data. How many PCs should be used? Explain how you arrived at your conclusion. Provide graphs or other R outputs to support your choice.

(h) Do a SVM classifier of IsB using only the first five PCs. (The number K=5 is fixed so that we all use the same classifier. You do not need to choose this number in the previous part (g).) What is the sensitivity of this classifier?

(i) Use leave-one-out cross-validation to estimate misclassification rate (mcr) of the SVM classifier. Report your estimate.

(j) If you had to choose between classifiers in part (e) and in part (h), which one would you choose? Why?

You should put answers to the questions in the PDF file. That means, for (a), provide the R command; for (b), provide the printout and the plot; for (c) provide the numerical answers; et al. Remember to answer each question directly. The grader should not have to pick out the numerical answers from the R outputs. The R commands that you used to get those printout/plots et al. should be submitted in the separate R script file.

2. (30 points) Choosing Classifiers and Number of Principal Components for PCA reduced iris data set.

In the last example of this module, we compared three classifiers on the iris data by working on the first three principal components. We choose the best classifiers based on cross-validated misclassification rate. We can also choose the number of principal components to use by cross-validation, instead of fixing it at K=3.

Use the leave-one-out cross-validation to choose the number of principal components together with the classifier. Please report the empirical misclassification rates (on whole data set) and the leave-one-out cross-validation misclassification rates for each value of K=1, 2, 3, 4 principal components and for each of the three classifiers: logistic regression, support vector machine and classification tree. Based on those rates, what is your choice?

Note: when you fit the logistic regression with K=1 principal component, then the PC1 becomes a vector instead of a matrix. You will need to modify the code for logistic regression for K=1 differently from the other values of K=2, 3, 4.