Computational Molecular Medicine

Assignment 3

Due March 11

Problem 1: The datasets "class_q1" and "features_q1" consist of 5000 features for each of 50 individuals, each labeled by 0 or 1. The features are in reality independent of the labels.

- Intuitively, the best possible error rate corresponds to random guessing. Derive the Bayes error rate e^*.
- Select the 100 features having the largest 100 correlation values with the class label, and then estimate the performance of the nearest neighbor classifier using k-fold cross- validation. (Just take the standard Euclidean metric in 100 dimensions and use any value of k you wish). Report the error rate estimated by cross-validation. Is your result consistent with your answer to the previous question?
- Now estimate the generalization error of the following classifier using k-fold CV: pick the 100 features most highly correlated with the class label and then classify with the nearest-neighbor algorithm. Report the estimated error rate you find with your method. Is it consistent with the bound e^* you computed earlier?

Problem 2: The files "class_q2" and "features_q2" come from a real dataset of 85 Glioma tumors. The class labels in "class_q2" correspond to good prognosis vs bad prognosis and "features_q2" contains mRNA data. The goal is to create a single feature classifier, i.e., a classifier of the form $F_{(k,t)}(x) = \delta(x_k > t)$ for a particular gene k and a threshold t.

- The first problem is that there are too many genes (over 20,000) and possible thresholds to consider every combination exhaustively. Narrow down the possible set of genes by selecting the top 100 most differentially expressed features based on the Wilcoxon rank-sum test. Print your top 10 features.

- Split the data into a training set and a testing set by selecting the first half of each class label (0 or 1) for the training set. Leave the rest for testing. Train a single gene classifier on the entire training set in the following way: Choose any (k, t) pair (with k restricted to the top 100) that gives the best training error when classifying the labels with $F_{(k,t)}$. Notice that you can restrict yourself to the the sample values for the choice of the thresholds.

- Evaluate the test error on the test dataset. What could be a reason for a difference in accuracy of the classifier on the test set and the training set?

- Repeat the parts above with the difference being that you will use only the training set to first derive the top 100 most differentially expressed features. Compare with the previous classifier performance.

Problem 3: Using the dataset of Problem 2:

- Again, narrow down the possible set of genes by selecting the top 100 most differentially expressed features based on the Wilcoxon rank-sum test and using all the samples.

- Split the data into a training set and a testing set by selecting the first half of each class label (0 or 1) for the training set. Pick the most differentially expressed feature among the top 100 using only the training set. Use that feature as a discriminant function to plot an ROC curve for the training set and another ROC curve for the test set. How do you explain the difference between the two ROC curves?

- Repeat the previous question with the difference being that you will use only the training set to derive the most differentially expressed feature. Compare the ROC curve obtained for the test set with the ROC curve on the test set obtained in the previous question. How do you explain the difference?