Cogs 109: Modeling and Data Analysis

Homework 4

Due **Friday 10/27 in class**

1. ISLR chapter 5, exercise 2 (page 197-198)
2. In this problem we will continue to explore the data set used in HW3 (hw3_divseq_data.csv). Recall that the data set contains gene expression levels for 2 genes (*Malat1* and *Lars2*) for each of 817 cells. Some of the cells are mature (mature=1) and others are immature neurons (mature=0).
    a. In HW3, problem 4 part (j), you fit a logistic regression and used it to classify the cells as mature or immature using both genes' expression levels. Make a confusion matrix, i.e. a 2x2 table showing the number of cells correctly and incorrectly classified in each category (mature and immature). Be sure to label the rows and columns of your matrix. What is the fraction of all mature cells that are correctly classified? What fraction of immature cells are correctly classified? Was is the fraction of all cells (both mature and immature) that are correctly classified?
    b. Recall that our classifier used a Bayesian criterion, $P(mature \mid Malat1,\ Lars2) > Threshold$, where the threshold was set at $Threshold = 0.5$. However, in this data set there are many more immature (80%) compared to mature (20%) neurons. Thus, the Bayesian classifier may give greater importance to immature cells since they make up the majority of the data set. Suppose we want to give greater importance to correctly classifying mature neurons -- i.e. we want to reduce the errors in classifying mature neurons, at the expense of potentially increasing errors for immature neurons. To do this, should we increase the threshold to $Threshold = 0.8$ or decrease it to $Threshold = 0.2$? Justify your answer.
    c. Based on your answer in part (b), implement the revised classifier and report the new confusion matrix. Report the same three performance measures as in part (a) and comment on the new values.
    d. Different researchers may want to choose different threshold values to achieve a particular false positive rate. Make a plot showing the false positive rate as a function of $Threshold$. The x-axis and y-axis should both range from 0 to 1.
    e. Make a plot showing the true positive rate as a function of $Threshold$.
    f. Make a ROC plot, i.e. show the true positive rate vs. the false positive rate for all $Threshold$ values. Use a plot symbol (e.g. circle or cross) to show the point on the ROC curve that corresponds to the Bayesian classifier ($Threshold = 0.5$), and another symbol to show the point that corresponds to your modified classifier in part (c).
    g. Is it possible to find a threshold value that gives false positive rate ≤ 10% and true positive rate ≥ 95%? Justify your answer.

3. In this problem we will use cross-validation and the bootstrap to estimate the reliability of our classifier.

    a. We will first use 10-fold cross-validation to estimate the test set error rate of our classifier. Divide the 817 observations into 10 (roughly) equal subsets, or “folds”. NOTE: You must randomize the order of the observations before splitting them up, otherwise some folds may end up with all of the mature cells and others with none. In MATLAB you can use the function `randperm` to randomly permute the order of your observations, e.g.:

    ```
    data = readtable('hw3_divseq_data.csv');
    data = data(randperm(size(data,1)),:);
    ```

    Write a for-loop that selects one fold at a time to be held-out as a test set for validation. Use the remaining 9 folds to fit a logistic regression. Compute the total error rate (i.e. the error for all cells, whether mature or immature) for both the training and test sets. Finally, print a table showing the training and test error for each of the 10 folds.

    b. What is the mean training error and what is the mean test set error? Comment on their relative values.

    c. Now use bootstrap resampling to fit the logistic regression using N=100 randomly resampled data sets. Make sure that you resample *with replacement*. Hint: The matlab function `randi` may be useful. For a data table `data`, you can use:

    ```
    nsamples = size(data,1);
    data_resample = data(randi(nsamples,nsamples,1),:);
    ```

    For each sample, keep track of the value of the slope parameter (coefficient) for *Malat1*. Report the mean and standard deviation of the coefficient across all 100 bootstrap samples.

    d. What is a 95% confidence interval for the coefficient *Malat1*?