Cogs 109: Modeling and Data Analysis

Homework 3

Due **Tuesday 10/23**

1. ISLR chapter 4, exercise 4 (page 168-169)
2. ISLR chapter 4, exercise 6
3. ISLR chapter 4, exercise 7
4. Most neurons in the brain develop before you are born and remain with you throughout your life. A small but important part of the brain called the dentate gyrus of the hippocampus continues to create new neurons past birth and into adulthood. These "adult newborn neurons" are thought to be important for creating distinct memories of similar events. In this problem, we will use a recently published data set containing gene expression measurements from single neurons to classify cells by their age. The study by Habib et al. is titled "*Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons*" (Science, 2016: https://www.ncbi.nlm.nih.gov/pubmed/27471252).

   Download the data set hw3_divseq_data.csv. There are 3 variables; the first 2 rows are shown here:

Lars2	*Malat1*	mature
9.95	6.69	1
10.54	8.53	1

- *Lars2* and *Malat1* are the expression levels[1] of two genes, both of which become highly expressed as neurons mature.
- Mature is a binary variable coding whether a cell is mature (1, corresponding to cell_age=14) or immature (0, corresponding to cell_age<14).
  a. Create a box plot showing the expression level of *Lars2* for immature and mature neurons. Do the same for *Malat1*.
  b. Based on these plots, comment on whether you expect that a classifier could perfectly predict a neuron's maturity based on *Lars2* expression alone.
  c. Fit a logistic regression to predict mature based on *Lars2* alone; do not use *Malat1*. What is the p-value for coefficient (slope) of *Lars2*? What can you infer, i.e. what conclusion can you draw?
  d. Using your model, calculate the predicted probability that each neuron is mature, i.e. $p = P(mature \mid Lars2)$. Make a plot showing *Lars2* on the x-axis vs. p on the y-axis. The plot should have a sigmoid shape. Based on this plot, what prediction would you make for the maturity of a cell with *Lars2* = 8?

[1] Gene expression is measured in units called "log TPM", or log(transcripts per million).

e. Use a Bayesian classification criterion to predict, for each cell, whether or not it is mature. Recall that a Bayesian classifier chooses the most likely category; in this case, that means that it should predict "mature" whenever $P(mature \mid Lars2) > 0.5$. Using these predictions, compute the sensitivity of your classifier, i.e. the fraction of mature cells that are correctly classified as mature.

f. Compute the specificity of your classifier, i.e. the fraction of immature cells that are correctly classified as immature.

g. Try predicting the maturity level for each cell with a threshold of 20%, i.e. predict mature whenever $P(mature \mid Lars2) > 0.2$. What are the sensitivity and specificity? Explain why the sensitivity is increased, while the specificity is decreased. In what circumstance might you prefer to use this classification threshold (20%) instead of the Bayesian threshold (50%)?

h. Now we will incorporate data from both genes to try to improve our prediction. First, make a scatter plot showing *Lars2* expression (x-axis) vs. *Malat1* expression (y-axis). Use a different color and/or plot symbol for cells that are immature and mature. Make sure to label the axes of the plot and include a legend explaining which color/symbol corresponds to which condition.

i. Fit a logistic regression using both *Lars2* and *Malat1* as predictors. Print the regression summary table showing the coefficients, SE, t-statistic and p-value for each term. Which predictors have a significant effect?

j. Use your new model to predict whether each neuron is mature, using a Bayesian decision threshold, i.e. $P(mature \mid Malat1,\ Lars2) > 0.5$. What are the sensitivity and specificity for this new prediction? Compare these values to the sensitivity and specificity you calculated in part (e).