Week 04 Problems: Cross Validation, Bootstrap.

Submit Assignment

Due Friday by 11:59pm **Points** 12
Submitting a text entry box, a website url, or a file upload
Available Feb 26 at 12am - Mar 5 at 11:59pm 8 days

1. Cross Validation done wrong.

- Pick 50 training points uniformly at random from a space of 1,000 predictors, each with values between -1 and 1.
- Assign a class of 0 or 1 at random to each training point.
- Choose the 50 predictors with highest correlation (positive or negative) to the class labels.
- Use those 50 predictors to fit a multivariate logistic model to the data, using 5-fold CV.
- What is the estimated test error? What will the actual test error be?
- Discuss what's wrong and how to fix it.

2. LOOCV. This problem is modeled on chapter 5, problem 8, but instead of $y = x - 2x^2 + \epsilon$ I will ask you to use the function and models from the Week 02 assignment. You may start with your solution or start with my solution instead.

A. Modify the Week 02 assignment notebook to generate just two training set of 15 points each instead of 10 training sets.
B. Create a scatterplot of Y versus X. for each training set. Comment on what you find.
C. Fit each of the five polynomial models from the Week 02 assignment to both training sets. Compute the LOOCV estimate of test error each time. That's 10 estimates.
D. Compare the five estimates from the first training set to the five estimates from the second training set and comment on what you see.
E. Compare the five models. Which had the smallest LOOCV estimate for the first training set? For the second training set? Is this what you expected? Explain.
F. Print a summary of each model fit. Consider the statistical significance of each variable in each model fit. Do these reported statistical significances agree with the conclusions you drew from the LOOCV estimates in part E?

3. Bootstrap. Chapter 5, problem 9.