Cogs 109: Modeling and Data Analysis

Homework 5

Due **Tuesday 11/13**

- **Ridge regression and LASSO for one variable.** (This problem is adapted from ISLR chapter 6, problem 6). In this problem we'll try to understand the effect of regularization on model parameters by considering what happens when n=p=1, i.e. we have just one data point and just one predictor. In this case the equations are simple and by plotting the results we can build intuition.

  a. Consider Eq. (6.12) in the book, the cost function that is minimized in Ridge Regression, with n=p=1:

  $$C_{ridge} = (y - \beta)^2 + \lambda\beta^2$$

  Assuming $y = 2$ and $\lambda = 0$ (i.e. no regularization), make a plot showing C_{ridge} as a function of β for $\beta \in [0, 3]$. What is $\hat{\beta}$, the value of β that minimizes C_{ridge}? (Hint: Recall that to find the maximum or minimum of a function $f(\beta)$, you need to solve the equation $\frac{d}{d\beta} f(\beta) = 0$).

  b. Now plot C_{ridge} as a function of β for $\lambda = 1$ (i.e. increase the regularization). Show that the new $\hat{\beta}$ is given by the formula in (6.14), $\hat{\beta} = y/(1 + \lambda) = y/2$. Describe, in words, how $\hat{\beta}$ changes as λ increases -- this illustrates why ridge regression is a form of *shrinkage*.

  c. Consider Eq. (6.13) in the book, the cost function for LASSO, with p=1:

  $$C_{lasso} = (y - \beta)^2 + \lambda |\beta|$$

  Make plots showing C_{lasso} as a function of β for $\lambda = 0$ and $\lambda = 1$. Comment on the minimum value, $\hat{\beta}$, for each value of λ.

- **Forward stepwise model selection.** Download the Anesthesia daset, anesthesia.csv, from the course website. This dataset contains:

  Time - time in seconds

  F0Hz_1 - EEG power at 0 Hz. (Note the data are the log of the power)

  F1Hz_2 - EEG power at 1 Hz … etc.

  BehaviorResponse - Probability that the subject responded to an auditory stimulus at each time bin

  a. Make a plot showing Time vs. BehaviorResponse. Make sure to label the axes. This shows the timecourse of the study, with the subject starting out awake (BehaviorResponse=1), transitioning into general anesthesia (BehaviorResponse=0), and later emerging from anesthesia (BehaviorResponse=1 again).

  b. Make a scatter plot showing BehaviorResponse vs. EEG power at 0 Hz (F0Hz_1). Make sure to label the axes. Describe, in words, the relationship between these variables.

  c. (1 point) What is the correlation coefficient between BehaviorResponse and EEG power at 0 Hz (F0Hz_1)?

d. Fit a simple linear regression model of the form, BehaviorResponse ~ 1 + F0Hz_1. Is the slope parameter statistically significant?

e. Fit a multiple linear regression that uses all of the EEG power features (i.e. 103 predictors, plus an intercept). What is the p-value of the slope for F0Hz_1? Is it statistically significant? This demonstrates that multiple regression can dramatically change the estimates for specific predictors.

f. Now write a for loop to fit 103 separate linear models of the form, BehaviorResponse ~ 1 + X1, where X1 is one of the EEG power features. For each fit, keep track of the mean squared error. (Note for this step we are not using cross-validation; just use the MSE for the training data.) Finally, make a plot showing MSE vs. feature number. Your code will look something like the pseudo-code below:

```
mse = []; % Initialize an empty array for mse
P = 103; % Number of predictors
for j=1:P
    % Step 1: Fit the model to the data, using predictor j
    model = ...
    % Step 2: Find the predicted values using the current model
    yhat = ...
    % Step 3: Find the MSE for the current model and save
    mse(j) = ...
end
```

Using these results, which single feature gives the best prediction (lowest MSE)?

g. Now write a loop to fit 102 models of the form, BehaviorResponse ~ 1 + X1 + X2, where X1 is the best feature obtained from part (f) and X2 is one of the other features. Which combination of two features gives the best prediction?

h. Using X1 and X2 chosen in part (g), perform 10-fold cross-validation with k=10. What are the training and testing MSE for this model?

i. Extra credit: Write a loop to continue this “forward model selection” process, adding additional features one at a time. Plot the training and testing MSE as a function of the number of model features. Based on these results, how many predictors would you choose to include?

MATLAB:

`plot, readtable, table2array, corr, corrcoef, fitlm, bar`

Python:

`matplotlib.pyplot.plot, pandas.read_csv, scipy.stats.pearsonr, statsmodels.formula.api.ols`