Stat 435
Arnav Palkhiwala

1. Summary:

This week, we focused on model selection and regularization in building a model and picking the best one to represent the data and predict on new data points. We started out with model subset selection using tools such as the step function and forward and backward selection. We used metrics like AIC, BIC, and Mallow's CP to evaluate a model. Then we went onto L2 (ridge regression) and L1 (lasso) norms to determine the best model.

2. Concepts:

Why do we need alternatives to least squares:

- Prediction accuracy - OLS may not be exact
- Interpretability if there are a lot of features

We have 3 classes of models:

- Subset selection
    - Identify a subset of p-predictors relevant for prediction
    - Fit OLS to this model
- Shrinkage
    - Fit model on all predictors, encourage some of coefficients to be small or zero (regularization)
- Diversion Reduction
    - Project p-predictors into a row-dimensional subspace
    - Fit a prediction model based on projected data P(A)

Subset Selection:

- Predictors X, response Y, data (x, y)
- Approach
    - M_0, model with 0 predictors, uses mean of response to predict
    - For k = 1, 2,...,p
        - Fit all models that contain exactly k predictors using p choose k
        - Pick best model based on smallest RSS (R squared)
    - Best models for each number of predictors M_1 to M_k
    - Select the best model based on AIC, BIC, cross-validation, Mallows CP
- Pros
    - Find best model for each complexity
- Cons

    - Computation - 2 to the power of p possible models, good results may occur by chance and not perform well on other data sets
    - Models are not nested
- Forward stepwise
    - Start with 0 predictors and for each step, include predictors with greatest additional improvement in fit
    - Computations are p^2/2 but the benefit of this is that there is nested behavior
- Choosing the best model
    - Adjustment to training error, CP, AIC, BIC, etc.
    - Cross Validation

Regularization:

- Used to control variance
- Ridge regression adds a regularization term/shrinkage penalty (also called L2 norm)
    - p = 1, with no intercept, add additional term to RSS of lambda times the sub of beta bar squared
    - Take derivative with respect to beta bar
    - Penalty term is small when Beta close to 0
    - Tuning parameter lambda control the impact of parameter
        - If lambda is small priority is fit data and if big priority is shrinkage
    - Coefficients can be very small and not zero, used when all predictors relevant
- Lasso (also called L1 norm) - identify small set of predictors gives good model
- Uses all predictors, some coefficients will be zero
- Penalty term is lambda times sum of absolute value of beta bar
- All coefficients equal 0 as lambda goes to infinity and in general coefs can = 0
- Because it is a convex problem, no closed form solution

3. Uncertainties:

Most of my uncertainties are around the actual derivation of L1 and L2 norms because mathematically they were complex and there were a few things I missed during lecture in terms of proving that L1 norm can set coefficients to 0, but L2 norm cannot. I need to look into the math for this as I understand it partially that the absolute value function is non-differentiable which leads to this. I am also just curious about use-cases for all of these as model subset selection just seems too intense computationally, so I wonder when this is actually used in practice? Also, when you do not know how related the predictors are, which is the best or safer option to pick? Is the L2 norm better because there could be related predictors so just to be safe you should account for this or is this a different way to deal with it?