STATS 415 Homework 6

Due at 2:30pm, November 14, 2019

1. This exercise investigates properties of the adjusted R^2. (15 points each question)

   (a) We know that R^2 is guaranteed to be between 0 and 1. Are both bounds (≥ 0 and ≤ 1) true for R_a^2? For each bound, either prove it is true or give a counterexample.

   (b) Suppose you have p=500 predictors and 501 observations in your dataset, and you fit a linear regression model. Predictors 1-50 are correlated with the response, and when a linear model with just these 50 predictors is fit, we get $R^2 = 0.5$. The remaining 450 predictors have 0 correlation with the response, so adding any of them to the model does not change the R^2. How many of these extra "uninformative" predictors added to the model will make the *adjusted* R^2 exactly 0?

2. (20 points) Consider the following linear model with fixed design:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}^* + \epsilon_i, \quad i \in [n],$$

   where $\{\mathbf{x}_i\}_{i\in[n]}$ are p-dimensional, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $\{\epsilon_i\}_{i=1}^n$ are independent. Derive the AIC of a candidate model $\mathcal{S} \subset [p]$. We assume that σ^2 is known in advance.

3. In this exercise, we will predict the acceptance rate of a college (number of applications accepted / number of applications received) using the `College` dataset from the ISLR package. (10 points each question)

   (a) Split the data set into a training set and a test set. Fix the random seed to the value 234, choose 30% (rounded down to the nearest integer) of the data at random for testing, and use the rest for training. Define a new response variable `Accept/Apps`. Plot this variable against every variable in the dataset (make sure you use the appropriate type of plot for each predictor). Comment on which variables appear to be most predictive.

(b) Fit a linear model using least squares on the training set, and report the training and test error obtained, with `Accept/Apps` as the response variable and all other variables as predictors.

(c) Perform forward and backward selection on the full model with the threshold $\alpha = 0.05$ to select a potentially smaller model. Report which model each method chose, and the training and test errors for their chosen models.

(d) Use AIC, BIC, and adjusted R^2 to select a potentially smaller model instead, from the set of all possible predictors used in 3b. Report which model each method chose, and the training and test errors for their chosen model(s).

(e) Use 5-fold cross-validation to estimate the test error from the training data, for the candidate smaller model(s) you found so far, and for the full model from 3b. Compare the training, CV, and test errors and comment on the results.

Please limit your answer to Q3 to 8 pages, organized into a coherent typed data analysis report. Answers to Q1 and Q2 may be either typed or handwritten. Please staple everything together and clearly write your name, your UMID, and your GSI/lab number on the homework.