Homework 3

DSO 530: Applied Modern Statistical Learning Methods

Spring 2022

Deadline. Thursday, April 7th, 5 pm, Los Angeles time.

Submission instruction. To submit your homework, please go to the **assessments** folder on Blackboard and find HW3. The submitted document should be in **pdf** format. Submission of other formats will incur a 20% penalty. Before you submit, please double check if your document is readable. You can submit your homework multiple times before the deadline, and the TA will grade the last submission.

Late submission policy. All late submission should be sent to xint@marshall.usc.edu. Late submission will incur 20% penalty within 24 hours after the deadline, and 40% penalty between 24 and 48 hours. **No submission after 48 hours of the deadline will be accepted**. As we have a large class this year, I wish that everybody respects this policy.

Additional instructions for HW3. When the path to an answer involves coding, please show the **Python codes and proper output**. Answering questions clearly and concisely is better than writing page-long convoluted responses. Discussion among students is strongly encouraged. But everyone should write up their own solution. **If one copies others' homework or lets others to copy the homework, they will receive 0 for this homework and face further penalty in the final grade**.

1. (i) Did you review up to and including lecture 6 and tutorial 9? If you haven't, please do so first. This homework, like every other homework, only covers a small part of the course contents. (ii) Did your group meet for the first time on the DSO530 group project?

2. Are the following statements correct? (1) Ridge regression tends to give sparser models compared to LASSO. (2) Backward stepwise selection can be applied to situations in which the number of predictors is more than the sample size. (You just need to answer Yes or No)

3. (i) When the number of totaly predictors is 19 (i.e., $p = 19$), how many models will best subset selection go through? (ii) If we run forward stepwise selection up to (and including) k^* ($1 \leq k^* < p$) number of predictors, how many models will be investigated?

4. Using the `Auto` dataset which you can find together with Tutorial 8, compare two linear regression models: (1) `mpg` as the response variable, `displacement` and `horsepower` as the predictors; and (2) `mpg` as the response variable, `acceleration` and `weight` as the predictors. Pleases compare via 10-fold CV and start with the code: `kfolds_regresssion = KFold(n_splits = 10, random_state = 2, shuffle = True)` and use the `cross_val_score` function with R squared as the CV criterion.

5. Write down C_p and AIC criteria for linear regression. Explain why they give the same ranking of models.