Homework 5 – Model selection and validation sets

CS584, Fall, 2017

Due October 26, 3pm, Blackboard

The learning objective is to practice performing model selection using

- Best subset,
- Forward
- Backward
- Validation
- K-fold cross validation

In general, I would like you to repeat the "lab" in ISLR for the above techniques (lab 6.5 in ISLR) using the Auto data set. This is how I would like you to treat the Auto data set:

- Predict mpg
- Do not use the "name" variable
- Treat origin as a categorical variable
- Use adjusted R2 as your metric if we need a metric (note, in practice, you can also use the other metrics, like BIC, but I use adjusted R2 because I believe it's more commonly implemented and easier to understand)

Answer the following questions and show your work / code.

1. What is the best model using best subset selection as a function of adjusted R2?
    a. What is the best adjusted R2?
    b. Show the coefficients.
    c. Show a plot of the adjusted R2 as a function of number of variables.
2. Repeat the above with forward selection. In addition to the questions above, answer d) Is the best result the same / better / worse than the best subset? e) does this technique pick any different subset of k features ($k < p$) than does best subset for the same k? If so, for which k and indicate the feature differences.
3. Repeat the above but with backward stepwise selection.

Now, we consider a validation set. In this case, please use the technique used in the text to create the test and training sets in the text. Set your seed to 1 and then create a vector indicating the rows that should be in the test and training set. If you are using Python, do whatever is analogous to this so that grading is easier and more consistent.

4. Using the single test set approach on the models generated by best-subset, which subset had the lowest validation error. (The book uses mean squared error.)
    a. What is the best validation error?
    b. Plot the validation error as a function of k, the number of features.
    c. Show the coefficients.
    d. Is this result different than the one you generated in question 1 for best subsets?

    e. Now that you figured out the features to use, retrain your model on the full data set. Show the coefficients you yield.
5. Repeat question 4 but with 3-fold cross validation. I use 3 instead of 10 here because the data set is so small…

Comments:

- It is very important to understand that you may get different "recommendations" depending on the technique you use in Machine Learning – we see this in model selection here, but we have seen it in other places! Because of this ambiguity, you should have a "process" in place to help you make your choice before you actually see the results. Your specific process can be developed with experience. I personally like people to use K-fold cross validation for most work.
    - As a side note, you may see combinations of features selected by these algorithms that don't really make sense, but have the best "performance." In practice, people sometimes pick features that make more sense if the relative performances are "close." Again, what those features are and what "close" means is up to the you! You should be able to defend your choice!
- When you use a test set to perform model selection, remember to rebuild your model with the full data set after you figure out the set of features to use. Using more data in model building should make your model more accurate and robust.
- When you use a test set for tuning your model, you might "overfit" the test set, especially if you do a lot of tuning on the same test set. This may make your error estimation using the test set overly optimistic. To handle this, sometimes people use a third set, not used at all for model tuning or selection.