STAT 2270: Data Mining
University of Pittsburgh
Fall 2020

# Homework 3

1. Here we'll look at the relationship between undergraduate GPA and LSAT scores amongst individuals applying to law school.

   (a) Load the `law` data is stored in the `bootstrap` package in R. Does there appear to be a strong relationship? Calculate the correlation.

   (b) Take $B = 1000$ bootstrap replicates of the data to get 1000 bootstrap estimates of correlation and make a histogram of the bootstrap correlations. Insert a red vertical line in the histogram showing the correlation calculated on the original data.

   (c) Calculate the bootstrap percentile confidence interval for correlation and insert blue vertical lines in the histogram from part (b) at the upper and lower limits. Based on this, could we reject the null hypothesis that the true correlation is equal to 0.5?

   (d) Calculate the bootstrap estimate of bias as well as the (standard) bias corrected bootstrap percentile confidence interval. Insert additional green vertical lines in the histogram showing the bounds for this new interval. According to this, could we reject the null hypothesis that the true correlation is equal to 0.5?

   (e) Based on these confidence intervals, you should see strong evidence that the true correlation is not equal to 0. Design and carry out a permutation test to explicitly test this.

2. As we discussed at the beginning of this course, we don't need strong assumptions about linear models in order to derive the optimal parameter estimates with respect to squared error. We do, however, need relatively strong assumptions about the underlying relationships and distributions to provide standard inferential results (confidence intervals and hypothesis tests). Now we'll look at how we could carry out corresponding versions of these with permutation-style-tests. *Note: These tests are not necessarily valid or exact, but they should give you a feel for how you might go about trying to measure familiar characteristics without needing distributional results.*

   (a) Generate 50 observations from the model $Y = X_1 + X_2 + \epsilon$ where $X_1$ and $X_2$ are independent and sampled uniformly at random from $[0, 1]$ and $\epsilon \sim N(0, 0.25^2)$. Generate another test dataset with 30 observations using the same setup. Use the training data to construct a linear model with $X_1$ and $X_2$ – no interaction terms or higher-order polynomial terms. Calculate the MSE on the test set and call this $\text{MSE}_0$.

   (b) Using the test MSE as our statistic, devise a permutation-test-equivalent to the (overall) F-test. Carry out the test with 1000 permutations. Can you reject the null hypothesis that none of the predictors are significant using your test?

(c) As we know, we can examine whether an individual feature is significant by looking at its corresponding t-test. Again using test MSE as the statistic of interest, devise a permutation-test-equivalent to the individual t-test. Carry out the test with 1000 permutations to test whether $X_2$ is significant. Can you reject the null hypothesis that $\beta_2 = 0$ using your test?

(d) In this case, we have only two features so an individual t-test is equivalent to a partial F-test. Let's scale things up a bit. Using the same general procedure and model as above (with all coefficients equal to 1), create a training set with 100 observations on 10 features. Also create a test set with 50 observations.

(e) Using the data from part (d), devise a permutation-test-equivalent to the partial F-test that will evaluate whether any subset of the features is significant. Carry out the test with 1000 permutations on $X_8, X_9$ and $X_{10}$. Can you reject the null hypothesis?

3. This problem will deal with *Stability Selection*. You can find the original paper by Meinshausen and Buhlmann at `https://arxiv.org/pdf/0809.2932.pdf`.

   Consider the (true) model below consisting of 4 sets of predictors:

$$Y = \sum_{i=1}^{3} X_i + \sum_{j=4}^{6} X_j + \sum_{k=7}^{9} X_i + 0 \cdot \sum_{l=10}^{20} X_l + \epsilon$$

   (a) Suppose that each set of features is independent from all others but that within the sets, the correlation is quite strong – let's say 0.9. Also suppose that each feature is normally distributed with mean and variance 1 and that $\epsilon \sim N(0, 1)$ is independent of all of the features. Generate 50 observations from this model.

   (b) Using forward stepwise selection, find the best model of each size and make a 10-fold cross-validation plot of the errors. What size model should you think you ought to pick (knowing the truth)? What size model would you pick based on the cross-validation results?

   (c) Fit a model via lasso using cross-validation to select the value of the tuning parameter.

   (d) Perform stability selection based on 1000 bootstrap samples using lasso to build each model and using the same value of the tuning parameter you chose in part (c). Make a histogram of the selection frequencies for each variable ordered from highest to lowest. What do you observe? Give some intuition for seeing the results you do based on what you know about how lasso behaves.

   (e) Repeat parts (c) and (d) using the group lasso. How does your histogram compare to the one using the standard lasso? Give some intuition for the similarities/differences.

   (f) Depending on the particular sample you drew in part (a), you may see different things happen in parts (d) and (e). Given a (true) model of the form shown above, what model(s) do you think would be reasonable to use in practice if you only wanted a model with 3 terms (covariates). That is, knowing the truth, what do you think would be the best 3-variable models? Is this what was selected in part (d)? In this context, why might

stability selection pick a model different from what you think is best? Play around with the within-group correlation strength and variance of the error term until you find two different settings: one in which stability selection does the "right" thing and one where it does the "wrong" thing.

4. In addition to the kind of permutation tests we discussed in class, there are also *rank*-based tests that rely on permutations and work in a very similar fashion. Suppose that instead of working with the original (raw) observations, we replaced each observation with it's rank. That is, given say two groups of observations, I find my smallest overall observation and replace that value with "1". Then I find the next smallest and replace it with "2" etc. The idea here is that under the null hypothesis, no group should have "ranks" (values) systematically higher than the other. Once I've done the replacing, I could do something like permuting which ranks belong to which group and comparing the differences in sample means (i.e. means of the ranks) to the original difference in means of the ranks.

   (a) What are the advantages to doing something like this as opposed to the standard permutation tests? (*Hint: Suppose you have a small number of samples in each group.*

   (b) What are the disadvantages to doing something like this as opposed to the standard permutation tests? (*Hint: Suppose you have a small number of samples in each group.* Does this actually "solve a problem" or does it just add an extra assumption? Think carefully about this – the answer isn't as obvious as you might first think.)

5. We haven't talked much yet about "variable importance" but we will when we talk about tree-based methods in the coming weeks. Think of this as a primer for the discussion to come.

   (a) Based on the name "variable importance", what do you think we're actually trying to measure? Write down a formal mathematical definition of this. (Note: This doesn't need to be "right." Just based on your intuition, write down a definition that you think would be appropriate and justify it.)

   (b) Suppose I want to define the notion of "importance rank." That is, I just want to order my predictor variables from most important to least important. Given a particular dataset, I propose to do this by something like forward stepwise selection. The variable that goes in first I'll call most important, the next variable to go in would be the second most important etc. Does this seem like a good way to define this? Why or why not – justify your answer.

   (c) Suppose that based on the measure we defined in part (b), you're not actually sure whether the variables are in the right order. To compensate for this, you take $B$ bootstrap samples and do the forward stepwise procedure on each, each time recording the variable order. After that's done, you order the variables by their average rank. Does that solve any issues you may have had with the definition in part (b)?

   (d) Let's try it. Generate $n = 25$ observations from the model $Y = X_1 + X_2 + \epsilon$ where $X_1$ and $X_2$ are independent and sampled uniformly at random from $[0, 1]$ and $\epsilon \sim N(0, 0.25^2)$. Knowing the true model here, is either variable more "important" than the other? Note: This answer probably seems obvious – it's not.

(e) Using the dataset from part (d), carry out the procedure from part (c) with $B = 500$. (Note: with only two variables here and knowing what you do about the model, you don't *actually* need to code a forward selection procedure to accomplish this.) Let $p_1$ denote the percentage of times $X_1$ was selected first and similarly for $p_2 = 1 - p_1$. Define $p = \max\{p_1, p_2\}$.

(f) Repeat part (e) 100 times and make a histogram of $p$. Is this centered where you'd expect?

(g) Repeat (e) and (f) for a few larger values of $n$. Are these histograms centered where you'd expect?

(h) You should be seeing a somewhat counterintuitive result here. Something that a larger $n$ – our usual knight in shining armor as statisticians – doesn't necessarily help fix. Explain what's going on and why you're seeing the results that you are.