Hans Trautlein Problem Set 7, Andrew Bray

# Two Cultures

# Reading Questions

1. *What are the two cultures outlined by Breiman?*

   **Data Modeling Culture** Assume stochastic data model originally.

   **Algorithmic Modeling Culture** Assume that the function is explaining the outcome from the input is complex and unknown.

2. *Is the Ozone Project supervised or unsupervised? Classification or Regression? Which methods that we've seen could be used to tackle this problem?*

   The Ozone project is supervised. There are elements of both classification and regression here and it is a little unclear which it is entirely. While it sounds like regression's were run to find what would be a continuous value, I'm pretty sure in the end they were trying to put it into specific classes, like "high ozone day, don't drive" or "low ozone day, pick the kids up in a Hummer." It looks like when they talk about $y$ they are really just focusing on the actual ozone number, which would be a regression, but if they then sorted that number into one of those groups it'd be classification. So if I had to pick one it'd be regression, because I bet the consultants were asked to do the numbers while the bureaucracy could decide then what to do with them. We could use the method they describe here for example, linear regression, or we could try more complicated methods, like a tree based method (maybe based on values like "was it sunny yesterday" and "was traffic higher than normal yesterday").

3. *What is the name of the model/method that is discussed in equation R of section 5.1?*

   Simple Linear Regression.

4. *In section 5.4 he states, "If the model has too many parameters, then it may overfit the data and give a biased estimate of accuracy." Where would this model be in terms of the bias-variance tradeoff?*

   It would be flexible but have a high variance, meaning that the bias would likely be low. The language he uses in this paragraph is a little confusing though and suggests the opposite perhaps, using phrases like "But there are ways to remove the bias."[1]

5. *What is the Rashoman effect? Did you run into this effect is question 5 from the last lab?*

   The Rashoman effect is when you have multiple different models that can explain a single outcome. They might come from different points of view so if you were to interpret them you would be a little confused about maybe a disjoint narrative, but their predictions seems to end up in around the same place.

   This occurred during the fifth question from the last lab, where we were asked to take different models and compare their MSE's, one way to evaluate the error in a model. The bagged version and the random forest both had decent explanatory power despite being models that used different predictors and weights on those predictors to get to their outcome. Our bagged MSE versus our random forest MSE ended up being less than 1% different.

6. *Explain how one of the techniques that we've covered could be seen to invoke Occam's Razor.*

   I often think of trees as looking like Occam's Razor at times, as they are often following the easiest path through the predictors to arrive at a prediction. When you do things like block the first node

[1] Who knows, I could very very likely be wrong!

from reappearing I think you solve some of this issue, but otherwise Occam's Razor seems to apply well to the use of trees in predicting data. What Breiman suggests instead is to "grow forests instead of trees," or use a random forest method instead.

# 1 Discussion Questions

1. *The most illuminating point for me in this paper was?*

   I loved the idea of the Rashoman effect. I have never heard this term used to describe this sot of behavior, but I absolutely love it. It accurately talks about what I've come across in my thesis analysis at times, so now I have another fancy word to use during my orals maybe.

2. *The most confusing point for me in this paper was?*

   I think I was a little confused initially by the difference between the algorithmic culture and the data modeling culture. Luckily Breiman expands later on, but I think he might have been more clear the beginning to help a young and less well-read student of his paper like me.

3. *Which of the responses (Cox, Efron, Hoadley, Parzen) do you find the most incisive? Why?*

   Not necessarily incisive in an inflamatory way, but I thought that Hoadley had a particularily potent two point critique of Breiman's paper. I thought the "[n]ew methods always look better than old ones" comment is often true, as well as the comment "[c]omplicated methods are harder to criticize than simple ones." Both of these seem to be powerful to me (with the second being my answer to the next question). I think they both emphasize the need for time perhaps, and more research and thought to be done around the algorithmic modeling culture, something that Breiman perhaps overlooks a little.

4. *Which do you think is the strongest single criticism of Breiman's paper that is leveled by the commentators?*

   I think that the comment that "[c]omplicated methods are harder to criticize than simple ones" is going to be the strongest piece of criticism that I have found within the responses against Breiman's paper. It's a recognition that this is actually a call for more theory in particular, something that might be lacking in the relatively newer and more foreign culture of algorithmic modeling. I have found his above comment to be true over and over when it comes to my thesis. I could perhaps use a more complicated model, but I would lose an explanatory ability that comes with a simpler regression-based model, and also I might have less true critics as only those familiar with the method would be able to criticize. I think the ability of others to recognize where you could have gone wrong can be as important as getting it all right.

5. *The big ticket question: in your area of study, if you had to use methods from only one of Breiman's cultures for the rest of your life, which would it be: Data Model or Algorithmic Model?*

   I think I would honestly end up using the data model culture, probably because I won't be using a huge amount of statistics in my future (sadly :( - I'm trying to find jobs that might change that trajectory) and I probably won't have the time to be a part of the algorithmic modeling culture, which honestly has felt more difficult and I have been less sure of myself throughout the entire process when we do delve into this specific culture. But if I had all the time in the world I think I'd use algorithmic modeling more often!