Breiman Questions

Alyssa Andrichik

11/13/2019

Two Cultures

Reading Questions

1. What are the two cultures outlined by Breiman?

The data modeling culture, which includes regression models and goodness-of-fit tests, and the algorithmic modeling culture, which is mostly decision trees and neutral nets that focus on predictive accuracy.

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 used the supervised learning method to create a regression model. The model had a lot of bias since it gave too much power to one variable leading the model to be under fit, so more variables should have been added to the model to decrease the bias of the one predictor that caused many problems.

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

It is a linear regression model.

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?

Too many parameters in a model causes it to take into account the random error. The large amount of variance causes the model to be less bias to a singular variable, but the over-fitting leads to a false perception of accuracy and thus a low bias overall.

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

The Rashomon Effect is when there is a multitude of different equations of f(x) in a class of functions giving about the same minimum error rate, meaning many different models seem like they all would work just as well at predicting. We haven't worked on number 5 yet because we haven't gone over it in class.

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

I feel like decision trees are able to make a simple model that is relatively accurate if you prune well. Prunning removes sections of a decision tree that do not add significant predictive power to the overall model and thus simplifies a the tree model.

Discussion Questions

7. The most illuminating point for me in this paper was. . .

his discussion on the large amount of misleading conclusions that have been published claiming a connection between 2 or more things that may not even be accurate at all, but it passed goodness-of-fit tests and residual checks. These publications could be quite damaging while not even being true.

8. The most confusing point for me in this paper was. . .

the concept of black boxes.

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

I found Cox's response to be the most incisive since he focused on the Breiman's focus on models with good predictability. Cox did not agree with the idea of using a simple standard model to predict and pretend that what the model produces to be accurate. He broke down why that was clearly.

10. Which do you think is the strongest single criticism of Breiman's paper that is levelled by the commentators?

The criticism I saw the most was about Breiman's focus on predicting. Predicting is not necessarily the most important part of statistics, rather it is about trying to interpret and understand what factors led to a specific outcome. Predicting can also cause lots of problems if wrong.

11. 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?

Algorithmic model. I feel like it accounts for diverse opinions in political science a bit more than data models.