Homework 4 - Hierarchical Regression and Bayesian Model Averaging

DS6040 Bayesian Machine Learning
(Summer 2022, Instructor: Taylor Brown)

2022-07-19

1 The Problem

This assignment has two distinct parts. The first will ask you to use hierarchical regression to examine how the personality characteristics of store managers in a regional chain of coffee shops impact sales of coffee and food. The second part of this assignment will ask you to apply BMA with logistic regression to revisit classifying wine quality using the white wine dataset we had been working with previously.

What you will need:

- `sales-ds6040.csv` - Found in on the Collab site. This dataset has the following variables.
  - sales - Standardized number of sales
  - food - 0/1 indicator, were these sales for food or for coffee.
  - con - Standardized conscientiousness rating for the store manager.
  - neur - Standardized neuroticism rating for the store manager.
  - store - 0-indexed store ID.

  There are 20 stores (each store having the same manager for the year), and 12 months of observations for food and coffee sales (so, 24 observations per store.)
- `hw4Companion-ds6040.ipynb` - Jupyter notebook containing an example hierarchical model and an code/example for the BMA component.
- `whitewine-training-ds6040.csv` and `whitewine-testing-ds6040.csv` - Our white wine datasets.

Prepare your own Jupyter Notebook for submission (you may submit a HTML or PDF). You may discuss this assignment with other students in the class, but you must submit your own answers to the questions below. **Include an honor pledge with your submission.**

2 Part 1: Bayesian Hierarchical Modelling (60 points)

You have been hired by a regional chain of coffee shops to help improve sales and to examine how the personality characteristics of individual store managers might impact the sales numbers of both coffee and food. The client as the following questions they need answered:

- How does conscientiousness and neuroticism impact the sales of coffee and food, and are coffee and food impacted differently?
- Once you control for the personality characteristics of the store managers, what stores should be performing well? (i.e. the rest of the employees might be great, but the store manager might be bringing sales down)

IMPORTANT: This part of the assignment is meant to simulate the experience of a working data scientist. For this section, prepare your response as though you are preparing a report for your client. This means that it needs to be readable, well formatted, and detail your reasoning. To make these requirements a bit more explicit, this is a minimal list of what I want to see:

1. Problem Statement - What problem are you tackling?

2. Approach - Describe the model you are using. Present it both in equation form, as well as a written description of the model. Importantly, you are not writing this for another data scientist, you are writing this for someone who is capable of understanding what a regression is, and what these sorts of models can provide, but has never worked with data analyses or statistics before (so smart, but without the same knowledge base you have.)
3. Prior Rationale - List your prior choices and why those were chosen.
4. Findings - This is where you present your findings.
5. Summary - This is where you summarize and interpret what your analyses uncovered. Again, this is for the client, so it needs to be usable information for them.
6. Diagnostics - This is where you put information/plots as to how the estimator performed. This is for technical reference (I like to always have these in the reports I create, but this information is not really for a client per say, more for another data scientist to validate your work.) This doesn't need to be long.

I (Taylor) am intentionally leaving the directions to this assignment vague, as I want to see how you approach a "real"data science problem. I do have some hints to help you out though:

- The unit of clustering is the store, and the store variable provided is a 0-indexed variable. This means it can be used analogously to the "county"variable in the example model.
- Yes, you do need to justify your prior *hyperparameter* choices, but to make it a bit simpler, you can use Normal priors for the regression coefficients, and retain the Half Cauchys for the variance distributions (as is already in the radon example).
- There are several ways of examining how the personality variables impact the sales in each category separately, but a simple way of specifying is via *multiplicative interaction terms*. For example. Let β be the effect of conscientiousness for coffee, and β_f be the difference in that effect for food. The correct way of using these coefficients is $\beta * \text{Con} + \beta_f * \text{Con} * \text{Food}$. This setup avoids the need to subset the data into food and coffee sales.
- If you use the previous method for specifying interactions, take care in your interpretation. In the previous example, β is the effect of conscientiousness on coffee sales, but β_f is not the effect of conscientiousness on food sales, rather it is the offset from the coffee effect. $\beta + \beta_f$ is the effect of conscientiousness on food sales.
- You should only have 2 hierarchical effects in your model, and these are analogous to the hierarchical effects in the radon example. You will have 4 other effects, but these won't differ by store. Think carefully about which effects are the hierarchical ones. If you choose incorrectly, this should be immediately apparent when examining the posterior distributions (they will exhibit very weird behavior.)
- I leave it up to you to determine what are the appropriate plots/tables to provide. Focus on providing information to the client that fully answers their questions while providing the all important "quantification of uncertainty."

3 Part 2: Bayesian Model Averaging with Logistic Regression (40 points)

Continuing our adventures in classifying wine, in this section you will be applying (pseudo)-Bayesian Model Averaging with logistic regression.

1. First, revisit your HW2 and calculate the misclassification rate and the cross tabs for 3 variable models that used **flat priors** that performed best on the **testing data**. You will have 1 model for LDA and 1 model for QDA.
2. Next, use the provided `BMA_Wine` class to fit a Bayesian Model Averaged logistic regression using the **training data**. Output the variable inclusion probabilities using the `summary()` function and interpret.
3. Finally, obtain the miss-classification rates and cross tabs for the BMA model applied to the training data and the testing data. Compare the performance of the BMA models to the performance of the best LDA and QDA models.

4 Extra Credit (10 points)

- (5 points) - Compare and contrast the LDA/QDA approach to a multinomial logistic regression. How are these methods different from one another? (I am looking for the technical details here, so a full credit answer will display understanding of precisely what each of these methods is doing statistically.)
- (5 points) - Explain the relation between the BIC and Bayes Factor. (Again, I am looking for technical details here. Simply saying that the BIC is an approximation to the Bayes Factor is insufficient. This will require some reading/investigating on your end.)