E401/M518: Empirical Challenge

Model Selection and Regularization

Fall 2023

October 2023

Please work on this challenge with a partner. All challenges are based on real-world data and are similar to what you might encounter in your future career. The main purpose of this challenge is not on the application, but on getting you (and your class mates) more familiar with applying the techniques that we discussed in the lecture. Nevertheless, whatever you present should make economic sense and if it doesn't you should think about what could be going wrong with the data and/or your analysis. Please keep in mind that these are (potentially dirty) real-world data and I haven't checked every detail of it. Therefore, you are likely to run into a lot of problems. I strongly encourage you to come to my office hour to discuss any issues as well as your overall plan for your presentation a few days before the respective class. There is always a risk that there is not much interesting in your data set. As long as you are able to clearly document what you tried and have some conjecture/explanation for why you get the results you get, this is totally fine. It's very likely that you will be in similar situations regularly when taking a job as a data scientist. I designed this challenge to be pretty open-ended on purpose. When diving into the data you may find aspects that are totally different from what I had in mind. This is totally fine and another likely outcome in data science projects.

You are expected to give a presentation of roughly 25-30 minutes in class. Think of this presentation as one you would give to your boss or at a board meeting of a company or policy institution that hired you as a data scientist. Other students should think of themselves as board members who attend your presentation and are strongly encouraged to ask critical questions about your analysis and you should be prepared to answer them. Your presentation should contain the following elements: (1) a brief discussion of the data, i.e., where is it coming from, what are the most important variables, what is the unit of observation, what concerns do you have about the quality of the data etc., (2) the big picture business or policy question that you are trying to address with these data (other students have not necessarily read the questions in advance), (3) overview of the methodology that you used to answer the question, (4) your empirical results, (5) discussion of the results, policy implications, and potential caveats and suggestions for further steps. Lastly, this is not a presentation class, so don't invest in fancy PowerPoint slides! Having prepared a RScript in RStudio that generates all your results as we click through it is totally fine! However, I ask you to only work with code scripts. Avoid manual manipulation or loading of the data from a graphical interface at *all costs!*

Main Techniques

In this challenge I will ask you to work mostly with linear model selection and regularization techniques, such as ridge regression and LASSO.

Data

In this challenge you will use several data sets on the browsing and online shopping behavior of 10,000 consumers. The file `browser-domains.csv` is the core data and contains the following columns:

1. `id`: machine id that identifies a unique consumer
2. `site`: name of the website visited
3. `visits`: time spent on the website (in minutes)

The file `browser-sites.txt` contains a list of all the websites that are tracked in the data. The file `browser-totalspend.csv` contains a list of how much each consumer spent online during the previous year.

Before you run any analysis, make sure you familiarize yourself with the data on players and player attributes and examine the data quality. Briefly mention in your presentation, if some features look dubious to you.

Business question

You are working for a marketing firm. A group of online retailers has approached you to develop a model to identify profitable high-spenders based on their browsing behavior. This would allow the retailers to predict how much a consumer is likely to spend online based on consumers' cookie data, and so target marketing activities more effectively. Your client has several fairly specific questions and instructions for you.

1. To get a better idea of the overall browsing and purchase behavior, please provide a few descriptive statistics of the data.
2. Based on the data provided, which variables would you include in your model? Does it make sense to estimate your model via OLS? Please carefully explain your answer and point out potential problems of a linear regression model.
3. Your clients insist that a regularization-based model, such as ridge regression or LASSO, would definitely be a good choice. Do you agree? How would such a model differ from a linear regression model?
4. Please explain how we need to prepare the data in order to estimate a LASSO model that predicts online spending with browsing data. What dimensions do the inputs for your model have? When we created the predictor matrices for some of our previous

models, many of our computers ran out of memory. Can you come up with a solution to this problem? Can you briefly walk us through the relevant code?

5. Estimate both a ridge regression and a LASSO model. What kind of decisions do you have to make when setting up a ridge regression/LASSO model? Are your spending predictions sensitive to the model specification used? If so, which one would you prefer? Explain clearly how you arrived at your conclusions and what some of its limitations are. What would be the most important thing you need to obtain or do in order to overcome some of the limitations?
6. Finally, one of the retailers has identified consumers with IDs 1 to 10 as particularly important for their business. Could you let us know which websites these 10 consumers visited and how much time they spent on each site?