Problem Set 3

Econ 4676: Big Data and Machine Learning for Applied Economics

Due Date: There are two deadlines for this problem set. The usual deadline for the document is November 19 at 11:00 am. The deadline for predictions is November 18 at 8:00 pm.

1 Theory Exercises

1. *LDA vs QDA.* I showed in class where the term "linear" comes from LDA. However, I skipped steps in Lecture 19 slide 18.

   (a) Show the complete steps from odds ratio and equal variance assumption that allows reaching a linear function for the odds ratio.

   (b) Lift the equal variance assumption, and show that you reach a quadratic function.

   (c) Generate some simulated data and plot the decision boundaries for both classifiers.

2. *Binary Response Online Updating.* The problem is simple but yet complicated. Online updating is essential because it breaks the storage barrier and helps with computation. Assume that a new observation arrives, and instead of refitting the entire model, you want to update your binary model estimates. Show that the contribution made by the observation i to the likelihood function is

$$l(y, \beta) = \sum_{i=1}^{N} (y_i log F(x_i'\beta) + (1 - y_i) log(1 - F(x_i'\beta))) \tag{1}$$

   is globally concave with respect to β if the function F is such that $F(-x) = 1 - F(x)$, and if it's derivative f, and it's second derivative f' satisfy the condition

$$f'(x)F(x) - f^2(x) < 0 \tag{2}$$

   for all real finite x. Show that this condition is satisfied by the logistic function $\Lambda(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}$

3. *Fake it until you make it.* Suppose that you have the following model

$$y_i^* = \beta_0 + \beta_1 x_i + u_i \tag{3}$$
$$u_i \sim N(0,1) \tag{4}$$
$$x_i \sim N(0,1) \tag{5}$$

but you observe

$$y_i = \begin{cases} 1 & if\ y_i^* > 0 \\ 0 & if\ y_i^* \leq 0 \end{cases} \tag{6}$$

(a) Generate 500 samples of 20 observations of (x_i, y_i):

i. 100 assuming that $\beta_0 = 0$ and $\beta_1 = 1$
ii. 100 assuming that $\beta_0 = 1$ and $\beta_1 = 1$
iii. 100 assuming that $\beta_0 = -1$ and $\beta_1 = 1$
iv. 100 assuming that $\beta_0 = 0$ and $\beta_1 = 2$
v. 100 assuming that $\beta_0 = 0$ and $\beta_1 = 3$

(b) For each of the 500 samples, estimate a Probit model. You'll note that the estimation fails because of perfect classifiers. What proportion of the time does this failure happen?

(c) Why there are more failures in some cases than in others?

(d) Repeat the exercise but now increase the sample size to 40. Is there a change? Can you say something about the effect of the sample size?

(e) Don't forget to set a seed so your results can be replicated.

4. *Piano piano va lontano.* Suppose you want to minimize the following function

$$f(\beta_1, \beta_2) = \frac{1}{2}(\beta_1^2 - \beta_2)^2 + \frac{1}{2}(\beta_1 - 1)^2 \tag{7}$$

(a) Compute the gradients $\frac{\partial f}{\partial \beta_1}$ and $\frac{\partial f}{\partial \beta_2}$

(b) Using gradient descent for the win. Write the following function

i. Give initial values β_1 and β_2
ii. Until $f(\beta_1^i, \beta_2^i)$ *"does not change much do"*
- $\beta_1^{i+1} = \beta_1^i - \eta \frac{\partial f}{\partial \beta_1}$
- $\beta_2^{i+1} = \beta_2^i - \eta \frac{\partial f}{\partial \beta_2}$
- compute $|f^{i+1} - f^i|$
- if $|f^{i+1} - f^i| < tol$ stop, otherwise continue
- $i \leftarrow i + 1$

iii. here you need to define the step size η and what *"does not change much do"*.
   A. Pick a "small" step and a "big" step.
   B. Set a high tolerance rate (tol) and a small tolerance rate to define *"does not change much do"*.

iv. Graphically illustrate these results

2 Empirical Problems

The main objective of these sections is to apply the concepts we learned using "real" world data. With these, I also expect that you sharpen your data collection and wrangling skills. Finally, you should pay attention to your writing.

I encourage you to turn each of the following two parts of the problem set in a way that resembles paper. As such, I expect graphs, tables, and writing to be as neat as possible. You can write it in Spanish or English, either language is fine. For students in the Ph.D., it would be a good practice to do it in English.

These parts also involve a lot of coding. Please attach your code with your responses or point to your repo. In coding, like in writing, a good coding style is critical for readable code. I encourage you to follow the tidyverse style guide.

In this problem set, we add a new wrinkle. I'll base part of your grade on your following Jacob's rubric posted in the course communication channel. The rubric details how to structure the repository and the document. The repo link to create your submission is https://classroom.github.com/g/_k73dlEL.

2.1 *"Wars of nations are fought to change maps. But wars of poverty are fought to map change"* M. Ali

This section was inspired by a recent competition hosted by the world bank: Pover-T Tests: Predicting Poverty. The idea is to predict poverty in Colombia. As the competition states *"measuring poverty is hard, time consuming, and expensive. By building better models, we can run surveys with fewer, more targeted questions that rapidly and cheaply measure the effectiveness of new policies and interventions. The more accurate our models, the more accurately we can target interventions and iterate on policies, maximizing the impact and cost-effectiveness of these strategies."* Today this is our objective.

Predictions have to be made at the household level only. Data, however, is provided at the household and individual levels. You can use individual-level information to build extra variables to improve your prediction. You can use the variable `id` to merge households with individuals.

A document describing the variables is available in the data folder. You can also check them on the DANE website.

The data folder contains four data sets. Training and testing data sets at the household and the individual level. You will note that some variables are missing in the testing data sets. This is by design, to prevent "cheating", and make things a bit more challenging.

The expected output of this section are a document and a `.csv` file of predictions.

1. The document should:
   (a) Explain the data used, and describe the variables in it you are planning on using on your model. Mention if you made any adjustments/transformations.
   (b) Describe the model you used for your final prediction.
      i. You should explain how you reached that model and the variables that you used.
      ii. You should include comparisons to at least 4 other models. You can compare them in terms of ROC, AUC, or in terms of the intended scoring function, i.e., false-positive rates, false-negative rates, and model sparsity. Remember that a portion of the judging depends on the sparsity of your model.
2. Besides writing up your results, you should submit a `.csv`. The submission file format is the variable id, and a 0-1 Poor prediction, where 1 denotes Poor and 0 otherwise. An example of how the submission file should look like is in the data folder. I will judge predictions based on false-positive rates, false-negative rates, and the model's sparsity. On presentation day, I will announce the "winning," which will present first and get extra credit.