Problem Set 4

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 5 at 1:00 pm. The deadline for predictions is November 4 at 8:00 pm.

The repo link to create your submission is
https://classroom.github.com/g/Ky83zuit

1 Theory Exercises

1. Consider the regression model $y_i = \alpha + \beta x_i + \epsilon, i = 1, ..., N$, a model with a constant and a single regressor. Assume that the classical assumptions hold. Let $\hat{\alpha}_N$ and $\hat{\beta}_N$ be the OLS estimators for α and β respectively. Suppose that these N observations are your train set. You test set consist in one observation and you make a prediction $\hat{y}_{test} = \hat{\alpha}_N + \hat{\beta}_N x_{test}$. Show that $E(\hat{y}_{test} - y_{test}) = 0$ and $Var(\hat{y}_{test} - y_{test}) = \sigma\left(1 + \frac{1}{N} + \frac{(x_{test}-\bar{x})^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}\right)$

2. Consider the regression model $y_i = \beta x_i + \epsilon, i = 1, ..., N$, that is a model with only one regressor and **no** intercept. Moreover, x_i has been standardized to have mean zero and variance one. All the classical assumptions hold and x_i is fixed (x_i is a non random variable. Consider estimating $\hat{\beta}^r = \frac{\hat{\beta}^{OLS}}{1+\gamma}$ where $\hat{\beta}^{OLS}$ is the OLS estimator and γ is a positive scalar. This is known as the ridge estimator.

   (a) Calculate the bias and the variance of $\hat{\beta}^r$,for a fixed γ. Compare it to the bias and variance of the OLS estimator

   (b) Calculate the MSE and compare it to the MSE of the OLS estimator. Show that there is a γ for which $MSE(\hat{\beta}^r) < MSE(\hat{\beta})$

   (c) What conclusion can you take with respect of the classical econometric practice of strictly preferring unbiased estimators?

3. Consider the regression model $y = X\beta + \epsilon$ with $\epsilon \sim N(0, \sigma^2 I)$ furthermore assume that β has a normal prior, i.e. $\beta \sim N(0, \tau^2 I)$.

   (a) Find the posterior distribution.

(b) Compare it to Ridge.

(c) What is the relationship between λ in the ridge model and σ^2 and τ^2?

(d) Intuitively which prior would you choose to get back Lasso?

4. *LDA vs QDA*. I showed in class where the term "linear" comes from LDA. However, I skipped steps in the lecture.

   (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 Empirical Problem

The main objective of this section 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 the following section of the problem set in a way that resembles a paper. As such, I expect graphs, tables, and writing to be as neat as possible. You can write it in Spanish or English, and either language is acceptable. For students in the Ph.D., it would be a good practice to do it in English.

Don't forget to upload everything to your repository and follow the template repository.

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, are 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.

The data.zip file 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 make things a bit more challenging. The file is available here.

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

Note that a household is classified as $Poor = I(Inc < Pl)$, if the family income is below a certain poverty line. This suggests two ways to go about predicting poverty. First, approach it as a classification problem: predict zeros (no poor), and ones (poor). Second, as an income prediction problem. With the predicted income you can use the poverty line and get the classification. Whatever route you choose, describe in detail what you did, and the model you used for your final prediction.

An important dimension for policy makers is that it can be *rapidly and cheaply measured*. In building your model aim to have a model that uses the minimum amount of variables.

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

1. Describe the data and any "cleaning" or transformation that you've done. You are free to add external data, but you have to explain why, and how you are using it.

2. Explain in detail the final model chosen, how it was trained, the selection of hyperparameters, and any other relevant information about the model.

3. Describe the variables that you used in the model and a measure of their relative importance in the prediction.

4. Include comparisons to at least 5 other models. You can compare them in terms of ROC, AUC, False Positives, or False Negatives.

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. More weight (75%) will be given to false-negatives, i.e. poor families classified as non poor.

- On the file name please include the number of variables that you used for prediction, e.g., `predictions-problem_set_4_sarmiento-cano-4.csv`, where `4` indicates the number of variables in your final model. The more variables you use, the lower your score.

- On presentation day, I will announce the "winning" team, which will present first and get extra credit.