NYU | COURANT

MATH-GA.2047-001 Data Science in Quantitative Finance
MATH-GA.2070-001 Data Science and Data-Driven Modeling

Homework 6

Profs. Ivailo Dimov and Petter Kolm
Due: October 20, 2020

Instruction

For students in *MATH-GA.2070-001 Data Science and Data-Driven Modeling*: This homework is optional for you. However, if you missed a previous homework or both quizzes[1], then this homework is an opportunity to "make up" for those.

This homework is to be done individually. No collaboration and/or code sharing permitted.

Objective

In this assignment, you will:

- Find the optimal predictor under L_1 loss.
- Perform feature engineering on the Housing Dataset.

Methodology and Deliverables

1. Bias-Variance Trade-off of Ridge Regression.
   (a) Derive the variance-bias trade-off (of MSE) for *predictions* produced by ridge regression.
   (b) What can you say about its minimum and where it is obtained?

2. Beyond Quadratic Loss.
   In class we showed that $\mathbb{E}[y \mid \mathbf{x}] = \operatorname{argmin}_{f(\mathbf{x})} \mathbb{E}[(y - f(\mathbf{x}))^2]$, or equivalently, the minimizer of the square loss $L_2(y, \hat{y}) := \mathbb{E}[(y - \hat{y})^2 \mid \mathbf{x}]$ is the conditional expectation $\mathbb{E}[y \mid \mathbf{x}]$.

[1] As previously announced, we will drop the quiz with the lowest score for the purposes of grading the course.

(a) Consider $\mathbf{y}$ and $\mathbf{x}$ both one dimensional real random variables. The L_1 loss is defined as $L_1(y, \hat{y}) := \mathbb{E}[|y - \hat{y}| \mid \mathbf{x}]$. What is the minimizer of the L_1 loss? *Hint:* Rewrite $\mathbb{E}[|y - a| \mid \mathbf{x}] = \int_{-\infty}^{\infty} |y - a| p_{y|\mathbf{x}}(y) dy$ and optimize with respect to a.

(b) Suppose $\mathbf{x} = S_t$ and $y = S_T$ be the stock price at time t and T respectively. Provide a financial interpretation of your result in the previous part.

(c) *Extra Credit:* How does the result in (a) change if $\mathbf{x}$ is a multi-dimensional random vector?

3. Feature Engineering of the Housing Dataset.
In class we used elastic net to regularize the OLS regression with an L_1-L_2 penalty term. This type of regularization becomes quite powerful when we have many regressors.

(a) Add more regressors to the problem by applying non-linear transformations of your choice to the features in the dataset. Add at least 10 more regressors by modifying appropriately the pipeline of your code.

(b) Fit the elastic net model and see if you will improve on the MSE obtained in class.

- Run a 10-fold cross validation on the training set to find the MSE distribution of your model, and compare it to the MSE distribution of the OLS model with the features used in class.
- Can you improve both on the bias and variance?
- For each set of features you consider, find the optimal elastic net model using a grid search of the parameter space. Does your best model remain the best in terms of MSE of the *test set*? Explain.