Machine Learning for Economics and Finance

Problem Set 2

Ole Wilms

July 29, 2024

Important Instructions

- In this problem set you are asked to apply the machine learning techniques we covered in the past weeks
- In case you struggle with some problems, please post your questions on the OpenOlat discussion board.
- We will discuss the solutions for the problem set on `MONTH DAY`

Setup

Assume the same setup as in *Problem Set 1* but now you try to improve the return predictions using the machine learning approaches we have discussed in class. For this you are asked to use the same training and test datasets we constructed in *Problem Set 1*.

Question 1: Shrinkage Methods

1. Fit a ridge regression using the training data. Determine the optimal penalty parameter λ using 5-fold cross validation (set the seed to 2 before you run the CV). Provide a plot of the cross-validation MSE as a function of $\log(\lambda)$ and interpret the outome.

[ ]:

2. Prepare a slide with a table that reports training MSE and test MSE for different models. Fill in the MSE from the linear model using all features from Problem Set 1. Now compute the training and test MSE for the ridge regression with the optimal penalty parameter λ from *Q1.1*.

[ ]:

3. Redo the two tasks above using Lasso instead of Ridge. Again fix the seed to 2. Provide a plot of the cross-validation MSE as a function of $\log(\lambda)$ and interpret. Provide a table that shows the coefficient of the Lasso with the optimal penalty parameter λ. Compute the training and test MSE of this Lasso model and add it to the table from *Q1.2*.

[ ]:

4. Now suppose your boss tells you that he only trusts sparse models with few variables. Use the Lasso and choose the tuning parameter λ such that the model only considers 3 out of the six variables. Report the coefficients and compare them to the coefficients from the optimal model from *Q1.3* and interpret. Compute the training and test MSE of this Lasso model and add it to the table from *Q1.2*. Interpret.

[ ]:

Question 2: Tree-Based Methods

1. Fit a large regression tree using the training data. Report the number of terminal nodes as well as the most important variables for splitting the tree.

[ ]:

2. Compute the training and test MSE of the tree and add it to the table from *Q1.2*.

[ ]:

3. Again set the seed to 2 and use 5-fold cross validation to determine the optimal pruning parameter for the large tree. Provide a plot of the prediction error against the size of the tree. Report the optimal tree size and provide a plot of the pruned tree. Which variables are important for splitting the pruned tree?

[ ]:

4. Compute the training and test MSE of the pruned tree and add it to the table from *Q1.2*.

[ ]:

5. Finally, use random forest to improve the predictions. Motivate your choice for the tuning parameters. Report the training and test MSE and add it to the table from *Q1.2*. Which variables are most important in the random forest?

[ ]:

6. Supposed it is the beginning of 2020 and you have access to both the in-sample and out-of-sample errors for the different methods. Which model do you choose to predict stock markets in the future and why?

[ ]:

Appendix

The dataset contains the following variables:

- **ret**: the quarterly return of the US stock market (a number of 0.01 is a 1% return per quarter)
- **date**: the date in format *yyyyq* (19941 means the first quarter of 1994)
- **DP**: the dividend to price ratio of the stock market (a valuation measure whether prices are high or low relative to the dividends payed)
- **CS**: the credit spread defined as the difference in yields between high rated corporate bonds (save investments) and low rated corporate bonds (corporations that might go bankrupt). CS measures the additional return investors require to invest in risky firms compared to well established firms with lower risks
- **ntis**: A measure for corporate issuing activity (IPO's, stock repurchases,...)
- **cay**: a measure of the wealth-to-consumption ratio (how much is consumed relative to total wealth)
- **TS**: the term spread is the difference between the long term yield on government bonds and short term yields.
- **svar**: a measure for the stock market variance

For a full description of the data, see *Welch und Goyal* (2007). Google is also very helpful if you are interested in obtaining more intuition about the variables.

References

Welch, I. and A. Goyal (2007, 03). A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies 21* (4), 1455 – 1508.