Columbia GSB: Machine Learning
Homework 1

Dr. George A. Lentzas

This Homework is due by 11:59pm on Sunday Feb 12th. Sumbit your results in .pdf format in Canvas. Good Luck!

1. Go to Wharton WRDS and register for an account using your Columbia email. You should be approved almost immediately as everyone in my class has now been pre-approved (if you have any issues please contact me asap). Once logged in, go to "CRSP → Annual Update Stock / Security Files → CRSP Daily Stock.

   Figure out the constituents of the Dow Jones Industrial Average on January 1st 2000 and for these companies download and import into a data-frame historical daily total returns from January 1st 2000 to December 31st 2016. Your data-frame should have columns that represent each stock's historical daily returns and rows that represent days. Print out some descriptive statistics for your data.

   [5 Points]

2. Split your sample into a Training Set 1 from January 1st 2000 to December 31st 2005, Training Set 2 from January 1st 2006 to December 31st 2010 and a Test Set from January 1st 2011 to December 31st 2016.

   (a) (Here you use Training Set 1): Perform PCA on the stock returns in the Training Set 1. (If you need help on how to do this in R, section 10.4 in ISLR in an excellent reference). Print the Principal Component loadings you calculated.

   (b) (Here you use Training Set 2): Then use the estimated Principal Components loadings and apply them to Training Set 2 to create daily data for all the Principal Components for the dates in Training Set 2. Then create a data-frame where the Y variable is the first stock's return at time $t+1$ and the X variables are all the lagged Principal Components from time t to time $t-30$.

   (c) Repeat this for all the stocks and stack these data-frames vertically (across stocks) to produce one such big data frame. Add a dummy variable describing the different stocks. What is the dimensionality of your data-frame? Provide a screenshot of the top 3 and bottom 3 rows and a printout of its 'summary()'.

   [30 Points]

3. Still using only the Training Set 2:

   (a) Fit a Lasso model to predict the $t+1$ return using the Principal Components from t to $t-30$ as explanatory variables. In your data-frame above, for each each row the "Y" should be the return of a stock at $t+1$ and the "X"s should be all the principal components from t to $t-30$ plus the stock dummy variable.

   (b) Use 5-fold cross validation to do feature selection. Create a plot of the Lasso λ parameter vs. the MSE and report your optimal Lasso parameter. Fit the model using the optimal Lasso λ parameter calculated above to the whole training data and report your results.

   (c) Are there any issues with using cross validation in such a time series setting?

   [50 Points]

4. Use the fitted model to predict returns in the Test Set. At each time period rank the forecasted returns and create an long/short portfolio comprising of the top 5 and bottom 5 forecasted returns in absolute value. You may assume that there are no transaction costs and that at each period you allocate 1/10 of your capital to each of the 5 top and 5 bottom forecasts. Your starting capital is 100 Dollars, how does your strategy perform?

   [15 Points]