Problem set 4
Stat 154, SP 21; N. El Karoui
Due date: 4/5/21, 11:59pm

Problem 1 [Sparse methods in practice: Tracking an index] In the Data/SP Data Old/, there are two datasets, one with the daily value of the SP500 index a few years back. One with the daily stock price of around 500 major stocks at the same time period. (They are the components of the SP500 index as of 9/17/13).

One concern for individual investors is the high transaction costs associated with buying stocks. Hence, one might be interested in portfolios which contain few different stocks.

Please do the following:

1. Using methods seen in class, construct a sparse/parsimonious portfolio that tracks/replicates approximately the SP500 index. Explain your methods and how you judge the quality of your replicating portfolio.

2. Suppose you are allowed to change your portfolio every 60 days. Find sparse portfolios that track the SP500 index well over 60 days. How stable are your portfolios? Explain what penalty you could use to try to get portfolios that change little over time. (You do not need to implement that second idea.)

3. Suppose now that you are not allowed to owe a negative amount of stock. (In practice, having a negative amount of stock is allowed, this is called "shorting" a stock. Though for various reasons, you might not want to be short a stock.) Explain how this change the optimization problems you are trying to solve.

4. Suppose now that you are only interested in replicating the returns of the SP500. (The daily return of a stock at time t is $R_t = S_t/S_{t-1} - 1$, where S_t is the price of the stock.) How does this change your answers to questions 1 and 2?

5. Suppose now that you are not concerned about transaction costs and hence do not limit yourselves to sparse portfolios. Use methods seen in class to track the SP500 as best you can. Explain your methods and how you judge the quality of your replicating portfolio.

NOTE: the data was obtained from CRSP through Wharton's WRDS. This is intended only for scholarly research and use in this class. Using it for other purposes would be a violation of their end-user agreement.

Problem 2 CART, bagging and random forests Use the prostate cancer dataset you used in HW3 (from the `spls` package).

- Use CART to analyze this dataset. Compare its results to your sparse logistic regression classifier in HW4.
- Use bagging to analyze this dataset. Is there a notable improvement upon the CART results?
- Use random forests to analyze this dataset. Is there a notable improvement upon the CART results?

Problem 3 Create a low-dimensional linear regression dataset. Fit CART to it. How good is the fit? More generally, how well do you think CART deals with data that has an inherently linear relationship b/t responses and predictors?

Problem 4 Create a low-dimensional (e.g $n = 100$ observations, $p = 5$ predictors) dataset where $y_i = \beta_0 + \beta' x_i + \epsilon_i$. Fit a CART tree to it. Now add 5, 10, 30 independent Gaussian (say variance 1) predictors to your "meaningful" predictors.

What is the impact on your CART tree? Is CART able to pick the "right" variables? How does prediction error change when you add more predictors?

Do the same analysis for random forests.

Problem 5 Using the same ideas as CART, how would you go about fitting a piecewise linear function to the data using tree-based methods?

Write pseudo-code for that and explain your ideas in plain English.