DM 6018 HW 1

John Carpenter

January 2020

1 Exercise 1

For each of the parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method.

(a) The sample size of n is extremely large, and the number of predictors p is small.

Having a large sample size actually works to our advantage in this case. The larger the sample size the better we should be able to estimate the true nature of a phenomenon. If we want to fit a flexible learning method then we would require data that has a larger number of parameters. In this case we could actually do well having an inflexible method over an inflexible method.

(b) The number of predictors p is extremely large, and the number of observations is small.

In this case since the number of observations is small we are at a disadvantage as compared to problem (a). That being that we have a small number of samples and that does not help us understand the true statistics of our population. In this case having a flexible learning method will help us because it give many possible estimates of f as opposed to an inflexible method that may make assumptions about the functional form of f from the beginning. However, we do want to be careful because a flexible method can suffer from overfitting.

(c) The relationship between the predictors and response is highly non-linear

A flexible learning method would do better here than an inflexible method. For example, if we applied a parametric method such a linear regression $Y(X) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$ to data that is non-linear then you are applying a model that is already biased towards linear data. In this case your inflexible method(the parametric method) is the wrong choice and going with a flexible learning method is the better choice.

(d) The variance of the error terms, i.e., $\sigma^2 = Var(\epsilon)$, is extremely high

Having data that contains a lot of variance can cause issues in linear models. The linear model suffers significantly when data points are introduced that are well outside of the standard deviation. In case I believe that a flexible learning method would perform better than an inflexible learning method.

2 Exercise 2

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO Salary. We are interested in understanding which factors affect CEO salary.

In this case we want to know how the predictors affect the CEO's salary and therefore we are looking at a regression and inference problem. In this case n is the top 500 firms and p is three.

(b) We are considering launching a new product and want to know whether it will be a success or a failure. We collect data on 20 similar products that were preciously launched. For each product we have recorded whether the product was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

Since we want to know whether it is going to be a success or not we are interested in the classification problem(success or fail) and this would be done using prediction. The n is twenty and the p is thirteen.

(c) We are interested in predicting % change in the US dollar relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market.

Since we are interested in the % change in the US dollar in relation to the weekly change in the worlds stock markets we would be doing this by using regression. This is done using prediction because we want to know to know how the US dollar is affected be the worlds stock market. In this case n is fifty-two and p is three.

3 Exercise 6

Describe the difference between a parametric and non-parametric statistical learning approach. What are some advantages of a parametric approach to regression or classification? What are its disadvantages?

The differences between parametric approaches make very distinct assumptions about the functional form of f. For example, SLR(simple linear regression) assumes that the functional form of f is approximately linear and thus there is a bias towards treating data as if it's linear. On the flipside a non-parametric approach not only doesn't make assumptions about the functional form of f but it is also flexible in that it can create many different approximations about the functional form of f as opposed to a parametric method. The advantages of a parametric model are that if there is a bias in the data then a parametric approach will be the best option for the data. A disadvantage of the parametric approach is that if there are suddenly large variances in the data then approximating f can not only become difficult but can become very inaccurate very quickly. The advangtage of non-parametric methods are that they are flexible and can give many functional forms of f. However, non-parametric methods can suffer from overfitting and giving an inaccurate representation of f.

4 Exercise 8

See code(In the future if you would like I can put the plots here)

5 Exercise 10

See code(In the future if you would like I can put the plots here)