Chapter2-Exercises

Anthony Chau

12/8/2019

Contents

Conceptual Exercises

Question 1

For each of 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. Justify your answer.

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

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

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

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

Solution:

(a) We would expect the performance of a flexible statistical learning method to be better than an inflexible method because the flexible methods can model the data more accurately given the large amount of observations. Also, since the number of predictors is small, the flexible methods would not estimate a large number of parameters - increasing overall interprebility.

(b) The performance of flexible methods is worse than inflexible methods. Given the large number of predictors and low number of observations, flexible methods would most likely result in overfitting as well as have low interprebility.

(c) The performance of flexible methods is better than inflexible methods because the flexible methods can better capture non-linear relationships over inflexible methods (such as linear regression).

(d) Flexible methods would perform better than inflexible methods because the flexible methods could include more potential variables that could be useful in predicting the response. The inclusion of those variables can offset the presence of a high variance for the error terms.

Question 2

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.

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

(c) We are interested in predicting the % change in the USD/Euro exchange rate in 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 USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

Solution:

(a) This scenario is a regression problem because we are predicting a quantitative variable, CEO salary. We are mainly interested in inference - understanding the factos that affect CEO salary. For this case, n = 500 and p = 3.

(b) This scenario is a classification problem because we are predicting a qualitative category - whether the product is a success or a failure. We are mainly interested in predicting - correctly determining if a new product will be a success or a failure. For this case, n = 20 and p = 13.

(c) This scenario is a regression problem because we are predicting a quantitative variable, USD/Euro exchange rate. We are mainly interested in prediction - what the USD/Euro exchange rate is given some predictor variables. For this case, n = 52 (52 weeks in a year) and p = 4.

Question 3