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.

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

좋다. 표본이 많기 때문에 flexibility가 높아도 좋다.

1. The number of predictors p is extremely large, and the number of observations n is small.

안 좋다. 위의 경우와 반대

1. The relationship between the predictors and response is highly non-linear.

좋다.

1. The variance of the error terms, i.e. σ2 = Var(), is extremely high.

엡실론은 irreducible error이므로 우리가 컨트롤 할 수 없다. 따라서 모델이 overfitting하다면 error에도 fit하게 되므로 좋지 않다.

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.

1. 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.

Regression

Inference

N = 500, p = 3

1. 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.

Classification, 성공 or 실패를 분류

Prediction

N = 20, p = 13

1. 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.

Regression

Prediction

N = 52, p = 3

6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

1. 함수의 형태를 가정하느냐 안하느냐의 차이

7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

테이블이(가) 표시된 사진

자동 생성된 설명

1. Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.
2. 3
3. 2
4. 3.16
5. 2.24
6. 1.41
7. 1.73
8. What is our prediction with K = 1? Why?

가장 가까운 green

1. What is our prediction with K = 3? Why?

3번째로 가까운 red

1. If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

K를 크게 잡으면 경계가 선형적으로 변하므로 nonlinear한 문제에 적절하지 않다. 따라서 k가 작은것이 좋다.