ISLap - Conceptual exercises from chapter 2

Gustavo S. Garone

2025-04-23

Exercise 1

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

Part a

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

In this case, we'd expect a flexible model to perform better in this case since overfitting is less of a concern with few predictors.

Part b

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

In contrast to (a), with a lot of parameters, we should use an inflexible model as to avoid overfitting.

Part c

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

With a highly non-linear relationship, we should use more flexible models as to lower the bias.

Part d

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

In contrast to (c), with a high variance of the error terms, we should aim at using less flexible models as to account for the high error term variance and lower our total MSE (Figure 2.9).

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.

Part 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.

This is an inference and regression problem, as we are interested in understanding factors (inference) that affect a numerical value (CEO's salaries). We are using data from $n = 500$ companies to learn how the $p = 3$ parameters (profit, number of employees, and industry) affect CEO salary.

Part b

> We are considering launching a new product and wish to know whether it will be a success or afailure. 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.

This is a classification problem concerned with prediction whether the company will fail or not. We are using data from $n = 20$ similar products and for each of them $p = 13$ parameters.