Exercises for ISLP*

Kevin Ma

May 2025

2 Statistical Learning

Conceptual

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) When the sample size n is extremely large and the number of predictors p is small, a flexible method is **better**, since it is better able to account for the patterns present in the more abundant training data, with less risk of overfitting, since outliers are balanced out.

   (b) When p is extremely large and n is small, a flexible method is **worse**, since it overfits to the hyper-specificity of the many parameters in the data, without a large sample size to support any inferred patterns. In these scenarios, flexible models are highly inconsistent and have high variance; that is, they depend strongly on the amount—or lack thereof—of training data.

   (c) When the relationship between the predictors and response is highly non-linear, a flexible method is **better**, since it is able to account for nuance, while inflexible methods remain rigid.

   (d) When the variance of the error terms, i.e. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high, a flexible method is **worse**, since it will be overly influenced by the noise from the training data and prone to overfitting, while an inflexible method is more stable and better at generalizing to unseen data.

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)

An Introduction to Statistical Learning with Applications in Python: `https://www.statlearning.com/`

(b)

(c)

3. (a)

(b)

(c)

4. (a)

(b)

(c)

5. (a)

(b)

(c)

6. (a)

(b)

(c)

7. (a)

(b)

(c)

8. (a)

(b)

(c)