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. $\sigma^2 = \mathrm{Var}(\epsilon)$, 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) 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 a **regression** problem dealing with a quantitative `salary`. n is 500 and p is 3.

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

(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.

    This is a **classification** problem with two categories as output, `success` and `failure`. n is 20 and p is 13.

(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.

    This is a **regression** problem dealing with qualitative percent change. n is 52 (weeks in 2012), and p is 3.

3. (a)
   (b)
   (c)
4. (a)
   (b)
   (c)
5. (a)
   (b)
   (c)
6. (a)
   (b)
   (c)
7. (a)
   (b)
   (c)

Applied

8. (a)
   (b)
   (c)