SHEET - An Introduction to Statistical Learning
Chapter 2 - Statistical Learning

PLAYE Nicolas

1 December 2023

1 Statistical Learning

1.1 Courses' Demonstrations

We suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, \ldots, X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$, which can be written in the very general form

$$Y = f(X) + \varepsilon.$$

Here f is some fixed but unknown function of $X_1, \ldots, X_p$, and ε is a random error term, which is independent of X and has mean zero. In this formulation, f represents the systematic information that X provides about Y.

Consider a given estimate $\hat{f}$ and a set of predictors X, which yields the prediction $\hat{Y} = \hat{f}(X)$. Assume for a moment that both $\hat{f}$ and X are fixed. Then, we show that

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \varepsilon - \hat{f}(X)]^2 \\ &= [f(X) - \hat{f}(X)]^2 + \mathrm{Var}(\varepsilon) \\ &= [f(X) - \hat{f}(X)]^2 + \mathrm{Var}(\varepsilon) \end{aligned} \tag{2.3}$$

Demonstration:

$$\begin{aligned} \mathrm{Var}[X] &= E[X^2] - E[X]^2 \\ E[X^2] &= \mathrm{Var}[X] + E[X]^2 \\ E[f] &= f \\ y &= f + \varepsilon \\ E[\varepsilon] &= 0 \\ E[y] &= E[f + \varepsilon] = E[f] = f \\ \mathrm{Var}[y] &= E[(y - E[y])^2] = E[(y - f)^2] \\ E[(y - \hat{f})^2] &= E[y^2 + \hat{f}^2 - 2y\hat{f}] \\ &= E[y^2] + E[\hat{f}^2] - E[2y\hat{f}] \\ &= \mathrm{Var}[y] + E[y]^2 + \mathrm{Var}[\hat{f}] + E[\hat{f}]^2 - 2fE[\hat{f}] \\ &= \mathrm{Var}[y] + \mathrm{Var}[\hat{f}] + (f - E[\hat{f}])^2 \\ &= \mathrm{Var}[y] + \mathrm{Var}[\hat{f}] + E[(f - \hat{f})]^2 \\ &= \mathrm{Var}[\varepsilon] + \mathrm{Var}[\hat{f}] + \mathrm{Biais}[\hat{f}]^2 \end{aligned}$$

1.2 Answers of Exercises

Question 1: For each of parts (a)-(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 the answer.

Definition:

Flexible models and **Inflexible models** generally refer to how adaptable or rigid a model is in learning from data a Flexible Model: has Adaptability (capable of adapting to a wide range of data inputs and variations in data patterns), Complexity (have a higher level of complexity), Overfitting Risk (more prone to overfitting). An Inflexible Model: has Limited Adaptability (more rigid structure and are less adaptive to a wide range of data inputs, Simplicity (generally simpler, with fewer parameters or less complexity), Generalizable (often better at generalizing from the training data to new, unseen data, as they are less likely to overfit.

(a) The sample size n is extremely large, and the number of predictors p is small - Flexible methods are great at capturing complex relationships in the data. With a large sample size, they have enough data to "learn" these complexities without overfitting, allowing them to potentially perform better by capturing more nuances in the data. - Inflexible methods are less likely to overfit since they don't model complex relationships as aggressively. If the true relationship between predictors and response is simple, these methods might perform better In general, opting for flexible models in the case of a sample size n extremely large, and a number of predictors p small is a preferred option.

(b) The number of predictors p is extremely large, and the number of observations n is small - When p is large and n is small, the choice generally leans towards inflexible methods. These methods are less likely to overfit and are more capable of providing a reasonable generalization from the limited data available.

(c) The relationship between the predictors and response is highly non-linear Flexible methods excel in modeling complex, non-linear relationships. They can adapt their parameters to fit the intricate patterns in the data that linear or less flexible methods might miss. Unlike Flexible models, Flexible models are typically designed to model linear relationships. When faced with non-linear data, these methods struggle because they cannot adapt their structure to fit the non-linear nature of the data effectively. Strictly most of the time, opting for flexible models in the case highly non-linear relationship between the predictors and response is a better choice.

(d) The variance of the error terms, i.e., $\sigma^2 = \mathbf{Var}(\varepsilon)$, is extremely high. The goal in such a scenario is to find a balance between bias and variance. Inflexible models, by not chasing complex patterns that might be noise, can offer better predictive performance and reliability under these conditions. In the presence of high variance in error terms, an inflexible model is generally a better choice. Its simplicity and robustness against overfitting make it more suitable for handling the uncertainty and noise associated with high error variance.

Question 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.

Definitions:

Classification: If the target is categorical, it's a classification problem.

Regression: If the outcome is a continuous value, it's a regression problem.

Inference: Inference is when we are more interested in understanding the relationships between variables.

Prediction: Prediction is about forecasting an outcome without necessarily understanding the exact relationships between variables.

n (Number of Observations) is the total number of data points you have.

p (Number of Predictors) refers to the number of variables that you're using to predict your outcome.

(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.