An Introduction to Statistical Learning with Applications in R

—

Gareth James, Daniela Witten, Trevor Hastie, & Robert Tibshirani

—

Chapter 2: Exercises

Graham Strickland

November 4, 2024

Question 1

(a) Given a very large sample size n and a small number of predictors p, an inflexible method would be better than a flexible one, since the risk of overfitting is less.

(b) For the same reasons as (a), a flexible method would yield better results for small n and large p.

(c) A flexible method would yield better results, since non-linear functions cannot be accurately modelled by linear functions.

(d) If there is high variance in the error terms, an inflexible method would be better, since a flexible method would introduce even more variance in the values of $\hat{f}$.

Question 2

(a) This is a classification problem, since we are trying to identify a qualitative trend in the data. It is an inference problem, since we are not trying to estimate future values of f. In this case, we have $n = 500$ and $p = 4$.

(b) This is a classification problem, since we are trying to classify the product as either a success or a failure. It is also a prediction problem, since we are looking to estimate a future output. We have $n = 20$ and $p = 14$.

(c) This is a regression problem since we have quantitative data and assume that it fits some function f, which we are attempting to estimate. Since this is a future estimate, it is a prediction problem. We have $n = 52$ and $p = 4$.

Question 3

(a) We have the following R plot of various error curves: Arbitrary polynomi-

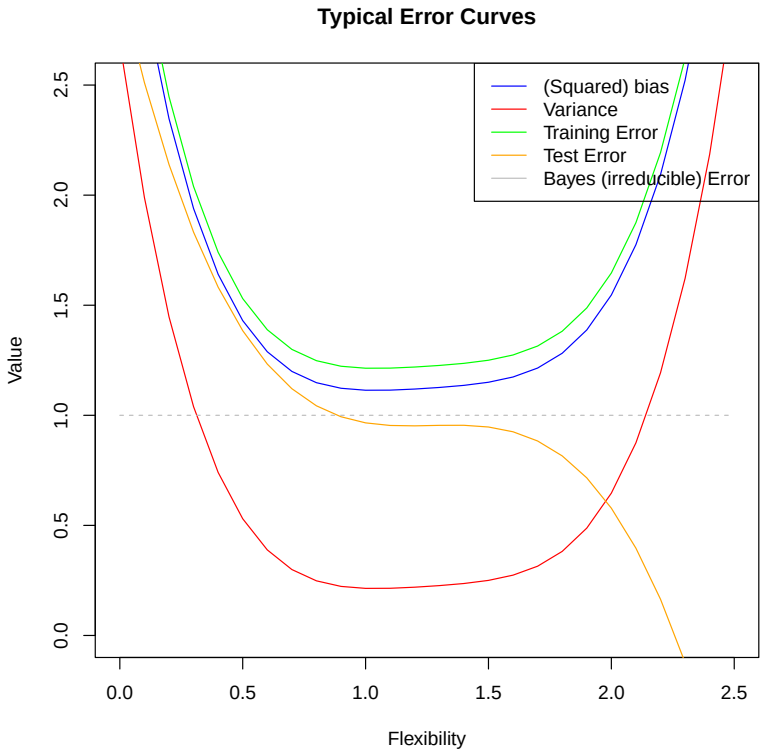

Figure 1: Typical (squared) bias, variance, training error, test error, and Bayes (irreducible) error

als of the appropriate degree were used to generate these plots, since they are approximations, using the following script:

```
# Exercise 2.3(a) - Plots of error curves

squared_bias <- function(x) {
  return((x - 1.25)^4 + 0.01 * x^3 + 0.05 * x^2 - 0.1 * x + 1.15)
```

```
}

variance <- function(x) {
  return((x - 1.25)^4 + 0.01 * x^3 + 0.05 * x^2 - 0.1 * x + 0.25)
}

training_err <- function(x) {
  return((x - 1.25)^4 + 0.01 * x^3 + 0.05 * x^2 - 0.1 * x + 1.25)
}

test_err <- function(x) {
  return(-(x - 1.25)^3 + 0.05 * x^2 - 0.1 * x + 1.0)
}

bayes_err <- function(x) {
  return(rep(1.0, length(x)))
}

x <- seq(0, 2.5, by = 0.1)

y1 <- squared_bias(x)
y2 <- variance(x)
y3 <- training_err(x)
y4 <- test_err(x)
y5 <- bayes_err(x)

pdf("ex2_3_a.pdf")

plot(x, y1,
  type = "l", col = "blue", lwd = 1, ylim = c(0, 2.5), xlab = "Flexibility",
  ylab = "Value", main = "Typical Error Curves"
)

lines(x, y2, col = "red", lwd = 1)
lines(x, y3, col = "green", lwd = 1)
lines(x, y4, col = "orange", lwd = 1)
lines(x, y5, col = "gray", lwd = 1, lty = 2)

legend("topright",
  legend = c(
    "(Squared) bias", "Variance", "Training Error", "Test Error",
    "Bayes (irreducible) Error"
  ),
  col = c("blue", "red", "green", "orange", "gray"), lwd = 1
)
```

```
dev.off()
```

(b) We are given that the training error will always be greater than the (squared) bias, and that both will be u-shaped. The variance will always be less than the (squared) bias, while the test error will decrease as the model becomes more flexible. The Bayes error is irreducible, so it remains constant.

Question 4

(a) Three real-life applications in which classification might be useful include:

(i) Determining a voter's party preference based upon the results of a census. The response would be a classification of a voter as being likely to vote for one of the possible political parties and the predictors would be features such as age, demographic, or location, which could be quantitative or qualitative. The goal in this case would be prediction, since we would most likely want to determine their vote in the next election.

(ii) Classification could also be used to determine the medical condition causing certain symptoms via medical imaging analysis. The response would be any of a number of medical conditions and the predictors would be features of the image, like abnormal textures or shapes within images provided by medical scanning. The goal would be inference, given that the condition already exists and we would like to determine its nature.

(iii) Another application would be determining a credit score for an individual. The response would be a natural number within a predetermined range and the predictors would be factors such as time taken to repay debts, number of credit lines awarded to the individual, and total accumulated debt. The goal would be prediction, since the credit score is used to determine the likelihood that the individual will repay their debts on time.

(b) Three real-life applications in which regression might be useful include:

(i) Attempting to predict the performance of a stock, given its past history would be a suitable application for regression. The response would be a numerical value indicating the expected price at a certain time and the predictors would be past values on a set of times. The goal would be prediction, given that we are estimating future performance.

(ii) Regression could also be used to determine the probability that a person will purchase a certain item from a retailer, given their past purchase history. The response would be a probability ($\Pr(X) \in [0, 1]$) and the predictors would be factors such as number of past purchases from that same retailer, spending history, credit rating, etc. The goal would also be prediction.

(iii) Another application would be determining the levels of a contaminant in a water supply, based upon readings from sources which are not directly drawn from the water supply itself, e.g., taps and waste water. The response would be a numeric value (say in mg/L) and the predictors would be the equivalent values in the other sources. The goal would be inference, since we are attempting to determine a current value.

(c) Three real-life applications in which cluster analysis might be useful include:

(i) A useful application of cluster analysis would be attempting to determine the species of certain related organisms given the degree to which they exhibit certain features.

(ii) Cluster analysis could also be used to determine whether certain geological samples fit within distinct groups based upon their chemical composition.

(iii) Another application of cluster analysis could be using segmenting consumers into certain target markets based upon their response to marketing surveys.

Question 5

The advantages of a very flexible approach for regression or classification include accuracy if over-fitting has not been exhibited, while the disadvantages include interpretability, large variance in errors, and tendency for over-fitting to occur. A more flexible approach might be preferred when a set of data contains a large variation in the response and a less flexible approach would be preferable if the data tends to contain few outliers.

Question 6

When utilising a parametric statistical learning approach, we attempt to estimate a countable number of parameters β_i for $i \in \{1, \ldots, p\}$ s.t. the observation (X, Y) has Y s.t.

$$Y \approx f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

This means we need to select the number of parameters used in order to minimize the error. In doing so, we risk overfitting the model to the data, so that the approximation does not closely match the form of f.

In contrast, a non-parametric method makes no assumption about the parameters used, with the advantage that we do not need to concern ourselves with the number of parameters used, but with the disadvantage that we must select a level of smoothness which makes the approximation to f easy to calculate without introducing unnecessary variation in the shape of the approximation relative to f.

Question 7

(a) If we denote the number of observations by n and the number of variables by p, then we let x_{ij} denote the ith observation of the jth variable. We calculate the Euclidian distances using the formula

$$\begin{aligned} &d\big((X_1 = 0, X_2 = 0, X_3 = 0), (x_{i1}, x_{i2}, x_{i3})\big) \\ &= \sqrt{(0 - x_{i1})^2 + (0 - x_{i2})^2 + (0 - x_{i3})^2} \\ &= \sqrt{(-x_{i1})^2 + (-x_{i2})^2 + (-x_{i3})^2}, \end{aligned}$$

for $i \in \{1, \ldots, 6\}$.

Thus we have

$$\begin{aligned} &d\big((X_1 = 0, X_2 = 0, X_3 = 0), (x_{11}, x_{12}, x_{13})\big) \\ &= \sqrt{9^2} \\ &= 9, \end{aligned}$$

$$\begin{aligned} &d\big((X_1 = 0, X_2 = 0, X_3 = 0), (x_{21}, x_{22}, x_{23})\big) \\ &= \sqrt{2^2} \\ &= 2, \end{aligned}$$

$$\begin{aligned} &d\big((X_1 = 0, X_2 = 0, X_3 = 0), (x_{31}, x_{32}, x_{33})\big) \\ &= \sqrt{1^2 + 3^2} \\ &= \sqrt{10} \\ &\approx 3.162278, \end{aligned}$$

$$\begin{aligned} &d\big((X_1 = 0, X_2 = 0, X_3 = 0), (x_{41}, x_{42}, x_{43})\big) \\ &= \sqrt{1^2 + 2^2} \\ &= \sqrt{5} \\ &\approx 2.236068, \end{aligned}$$

$$
\begin{aligned}
& d\big((X_1 = 0, X_2 = 0, X_3 = 0), (x_{51}, x_{52}, x_{53})\big) \\
&= \sqrt{(-1)^2 + 1^2} \\
&= \sqrt{2} \\
&\approx 1.414214,
\end{aligned}
$$

and

$$
\begin{aligned}
& d\big((X_1 = 0, X_2 = 0, X_3 = 0), (x_{61}, x_{62}, x_{63})\big) \\
&= \sqrt{1^2 + 1^2 + 1^2} \\
&= \sqrt{3} \\
&\approx 1.732051.
\end{aligned}
$$

(b) Our prediction with $K = 1$ is that $Y =$ Green since the nearest neighbour to the point $(X_1 = 0, X_2 = 0, X_3 = 0)$ is the point given by (x_{51}, x_{52}, x_{53}) corresponding to observation 5, which yields the value $y =$ Green.

(c) With $K = 3$, we have $Y =$ Red, since of the 3 nearest neighbours calculated in (a), we have two with value $y =$ Red and only one with value $y =$ Green.

(d) We would expect the best value to be small, since this would yield a decision boundary which is highly flexible and more easily approximates the non-linear nature of the Bayes decision boundary.