Prediction, estimation and attribution

Exercises

1 CASI

Many of these exercises use data used in the book. These datasets can be found on the book webpage `https://web.stanford.edu/~hastie/CASI`.

Chapter 1 Exercises

1. (a) Fit a cubic regression, as a function of age, to the `kidney` data of Figures 1.1 and 1.2, calculating estimates and standard errors at ages 20, 30, 40, 50, 60, 70, 80.

   (b) How do the results compare with those in Table 1.1?

2. The `lowess` curve in Figure 1.2 has a flat spot between ages 25 and 35. Discuss how one might use bootstrap replications like those in Figure 1.3 to suggest whether the flat spot is genuine or just a statistical artifact.

3. Suppose that there were no differences between AML and ALL patients for any gene, so that t in (1.6) exactly followed a student-t distribution with 70 degrees of freedom in all 7128 cases. *About* how big might you expect the largest observed t value to be? Hint: $1/7128 = 0.00014$.

2 Materiale didattico

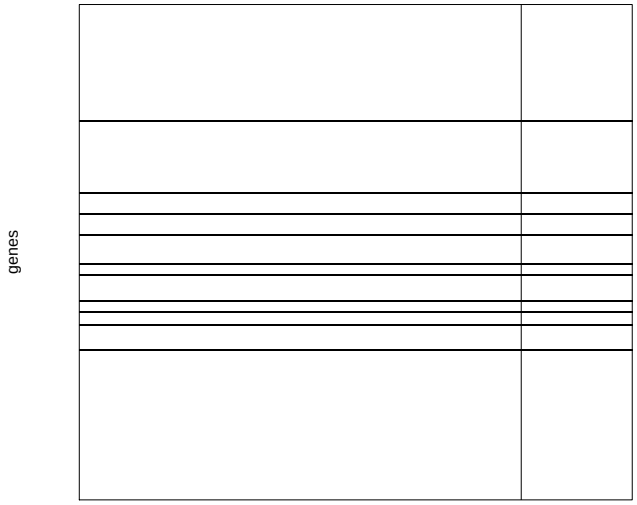
Consider the following hypothetical microarray study: $n = 400$ subjects participate in the study, arriving one per day in alternation between Treatment and Control (day 1 Treatment, day 2 Control, day 3 Treatment, etc.). Each subject is measured on a microarray of $p = 200$ genes. The 400×200 data matrix X has independent normal entries

$$X_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_{ij}, 1)$$

1. Suppose that most of μ_{ij} are 0, only for $j = 30, 48, 57, 65, 84, 92, 113, 128, 143, 195$

$$\mu_{ij} = 0.5 \quad i \text{ odd (Treatment)} \quad \mu_{ij} = -0.5 \quad i \text{ even (Control)}$$

See the following Figure, where the lines correspond to genes with average gene expression of 0.5 for Treatments and -0.5 for Controls:



In the first random Forest analysis (RF-I), the 400 subjects were randomly divided into a training set of 320 and a test set of 80.

The second random Forest analysis (RF-II), uses the subjects from days 1 to 320 for the training set and from days 321 to 400 for the test set.

(a) Do you expect the test error of RF-II to be lower than 50%?

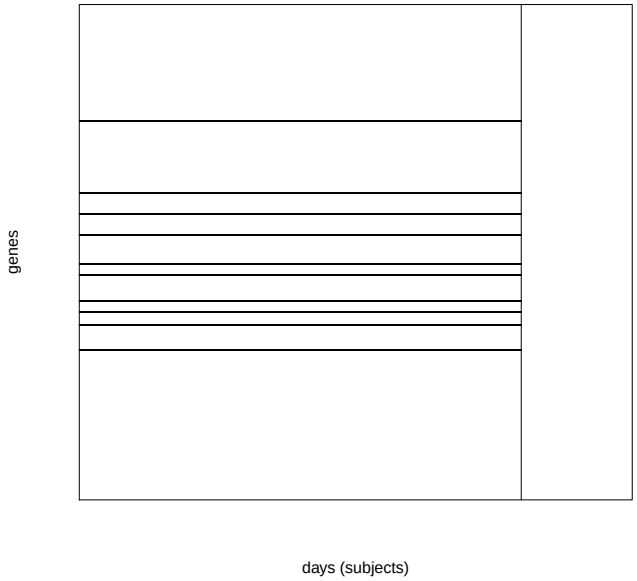
(b) Do you expect any difference in training prediction error between RF-I and RF-II ?

(c) How many genes large Variable Importance score do you expect to find with RF-I?

2. Now suppose that for $j = 30, 48, 57, 65, 84, 92, 113, 128, 143, 195$

$\mu_{ij} = 2 \quad i = 1, 3, \ldots, 317, 319$ (Treatment) $\quad \mu_{ij} = -2 \quad i = 2, 4, \ldots, 318, 320,$ (Control)

and for everything else $\mu_{ij} = 0$. See the following Figure:



(d) Do you expect the test prediction error of RF-II to be lower than 50%?

(e) Do you expect the training prediction error of RF-I to be higher than the training prediction error of RF-II ?