Homework 3

Jacob Thielemier

23 February 2024

Question 1

The right and Wrong way

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
## [1] 0
```

```
## [1] 0.5013586
```

```
## [1] 0
```

```
## [1] 0.3656465
```

- The Cross-Validation code focuses on demonstrating the best practices for predictive modeling, particularly in feature selection and cross-validation techniques. The code starts by generating a dataset with 50 observations and 5000 predictors. It employs an association-based method to rank predictors based on their relationship with the binary outcome. The "Right Way" involves performing feature selection separately within each training fold, while the "Wrong Way" performs feature selection on the entire dataset before cross-validation.

Monte Carlo simulation and the bootstrap

```
## [1] 0.6025703
```

```
## [1] 0.08227949
```

```
##      2.5%     97.5%
## 0.4426858 0.7675824
```

```
## [1] 0.6495215
```

```
## [1] 0.09183016
```

```
##      2.5%     97.5%
## 0.4772515 0.8318260
## [1] 0.6424349
```

```
## [1] 0.07919303
```

```
##      2.5%     97.5% 
## 0.4803213 0.7968199
```

- The Monte Carlo code focuses on estimating the properties of a specific statistical measure, alpha, which appears to be a function of variance and covariance from bivariate normal distributions. The first example simulates samples directly from a theoretical bivariate normal distribution with predefined parameters. The second simulates samples based on parameters estimated from observed data, mimicking real-world scenarios where true parameters are unknown. The code showcases the use of Monte Carlo simulations and bootstrap methods in statistical inference, specifically in estimating the distribution of complex statistics derived from sample data.

Question 2

(a)

- This is 1 - probability that it is the jth $= 1 - 1/n$.

(b)

- Each bootstrap observation is a random sample, so the probability is the same $(1 - 1/n)$.

(c)

- For the jth observation to not be in the sample, it would have to *not* be picked for each of n positions, so not picked for $1, 2, ..., n$, thus the probability is $(1 - 1/n)^n$

(d)

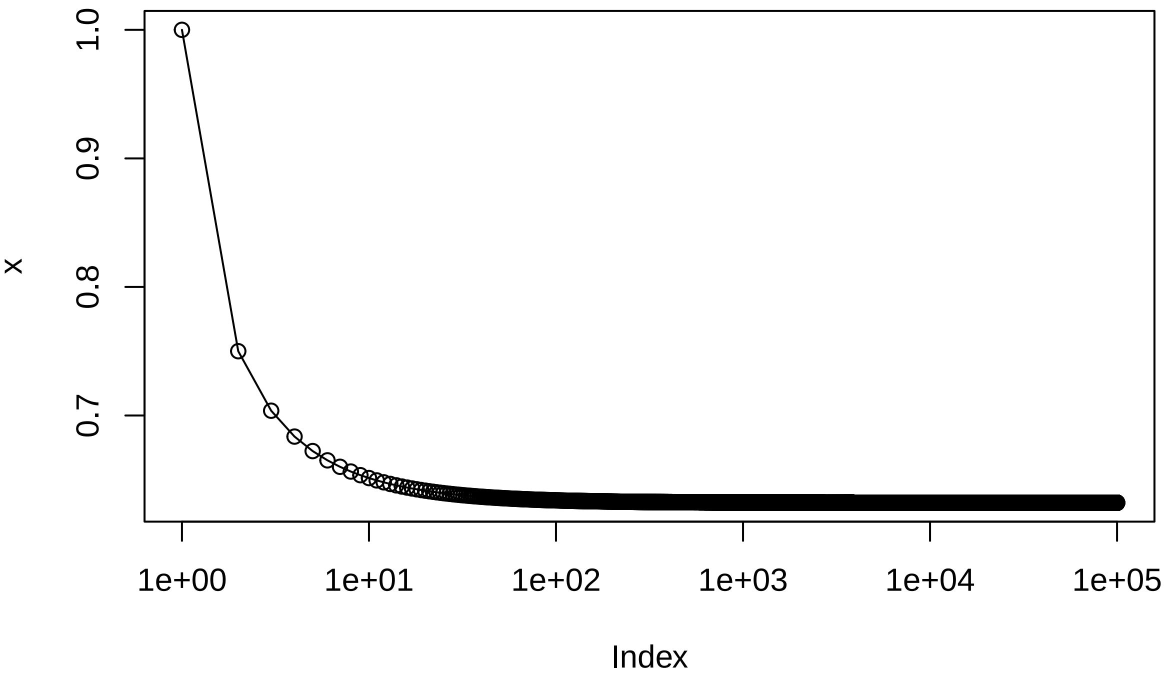
```
## [1] 0.67232
```

(e)

```
## [1] 0.6339677
```

(f)

```
## [1] 0.6321224
```



(g)

- The probability rapidly approaches 0.63 with increasing n.
- We know that: $e^x = \lim_{x \to \inf} \left(1 + \frac{x}{n}\right)^n$ and with $x = -1$, then the limit is $1 - e^{-1} = 1 - 1/e$.

(h)

```
## [1] 0.6291
```

- The probability of including 4 when resampling numbers 1...100 is close to the answer from 2(e).

Question 3

(a)

```
## Loading required package: ISLR2

## Warning: package 'ISLR2' was built under R version 4.3.2

##
## Attaching package: 'ISLR2'

## The following object is masked from 'package:MASS':
##
##     Boston
```

(b)

```
##
## pred    No  Yes
##   No  4817  110
##   Yes   20   53
```

```
## [1] 0.026
```

(c)

```
## [1] 0.0260 0.0294 0.0258
```

- The results are similar to each other (low variance). The minor differences can be explained by the fact that we used separate observations for each model.

(d)

```
## [1] 0.0278 0.0256 0.0250
```

- Including `student` does not seem to influence the test error.

Question 4

(a)

(b)

(c)

```
##    1
## TRUE
```

- Yes the observation was correct.

(d and e)

```
## [1] 0.4499541
```

- LOOCV error rate is 44.9%. This shows that the model is correct in 55% of its predictions, which is better than random guessing.

Question 5

(a)

```
## [1] 22.53281
```

(b)

```
## [1] 0.4088611
```

(c)

```
## Warning: package 'boot' was built under R version 4.3.2

##
## Attaching package: 'boot'

## The following object is masked from 'package:lattice':
##
##     melanoma

##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = Boston$medv, statistic = function(v, i) mean(v[i]),
##     R = 10000)
##
##
## Bootstrap Statistics :
##     original      bias    std. error
## t1* 22.53281 0.002175751   0.4029139
```

- The standard error using the bootstrap (0.403) is very close to that obtained from the formula above (0.409).

(d)

```
## [1] 21.72698 23.33863
```

(e)

```
## [1] 21.2
```

(f)

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = Boston$medv, statistic = function(v, i) median(v[i]),
##     R = 10000)
##
##
## Bootstrap Statistics :
##     original   bias    std. error
## t1*     21.2 -0.01331   0.3744634
```

- The estimated standard error of the median is 0.374. This is lower than the standard error of the mean, so we can be reasonably confident of the estimate.

(g)

```
##   10%
## 12.75
```

(h)

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = Boston$medv, statistic = function(v, i) quantile(v[i],
##     0.1), R = 10000)
##
##
## Bootstrap Statistics :
##     original   bias    std. error
## t1*    12.75 0.013405    0.497298
```

- We get a standard error of 0.497. This is higher than the standard error of the median, but still quite small. We can still be confident about the value of the 10th percentile.