Compulsory exercise 2: Group 24

TMA4268 Statistical Learning V2019

Silje Anfindsen and Clara Panchaud

18 March, 2020

Help for equation set-up:

$$
\begin{aligned}
E[(y_0 - \hat{f}(x_0))^2] = & \quad \text{definition of y0} \\
E[(f(x_0) + \epsilon + \hat{f}(x_0))^2] = & \quad \text{by linearity of the expectation} \\
E[f(x_0)^2] + E[\epsilon^2] + E[\hat{f}(x_0)^2] - 2E[f(x_0)\hat{f}(x_0)] + 0 = & \quad \text{using the definition of variance} \\
f(x_0)^2 + Var(\epsilon) + Var(\hat{f}(x_0)) + E[\hat{f}(x_0)]^2 - 2f(x_0)E[\hat{f}(x_0)] = & \\
\underbrace{\text{Var}(\varepsilon)}_{\text{Irreducible error}} + \underbrace{\text{Var}(\hat{f}(x_0))}_{\text{Variance of prediction}} + \underbrace{\Big(f(x_0) - \text{E}[\hat{f}(x_0)]\Big)^2}_{\text{Squared bias}} &
\end{aligned}
$$

Problem 1

a)

b)

c)

d)

```
library(ISLR)
set.seed(1)
train.ind = sample(1:nrow(College), 0.5 * nrow(College))
college.train = College[train.ind, ]
college.test = College[-train.ind, ]
str(College)
```

```
## 'data.frame':    777 obs. of  18 variables:
##  $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Apps       : num  1660 2186 1428 417 193 ...
##  $ Accept     : num  1232 1924 1097 349 146 ...
##  $ Enroll     : num  721 512 336 137 55 158 103 489 227 172 ...
##  $ Top10perc  : num  23 16 22 60 16 38 17 37 30 21 ...
##  $ Top25perc  : num  52 29 50 89 44 62 45 68 63 44 ...
```

```
##  $ F.Undergrad: num  2885 2683 1036 510 249 ...
##  $ P.Undergrad: num  537 1227 99 63 869 ...
##  $ Outstate   : num  7440 12280 11250 12960 7560 ...
##  $ Room.Board : num  3300 6450 3750 5450 4120 ...
##  $ Books      : num  450 750 400 450 800 500 500 450 300 660 ...
##  $ Personal   : num  2200 1500 1165 875 1500 ...
##  $ PhD        : num  70 29 53 92 76 67 90 89 79 40 ...
##  $ Terminal   : num  78 30 66 97 72 73 93 100 84 41 ...
##  $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
##  $ perc.alumni: num  12 16 30 37 2 11 26 37 23 15 ...
##  $ Expend     : num  7041 10527 8735 19016 10922 ...
##  $ Grad.Rate  : num  60 56 54 59 15 55 63 73 80 52 ...
```

e)

Problem 2

a)

FALSE, FALSE, TRUE, TRUE

b)

The basis functions for a cubic spline with knots at each quartile, of variable X is,

\begin{allign} b_0(X) = 1 & b_4(X) = (X-q_1)*{+}*ˆ3 \ *b_1(X) = x & b_5(x) = (X-q_2)*{+}ˆ3 \ b_2(X) = xˆ2 & b_6(X) = (X-q_3)_{+}ˆ3 \ b_3(X)=xˆ3 \end{allign}

c)

d)

Problem 3

a)

b)

c)

Problem 4

a)

b)

c)

d)

Problem 5

a)

b)

c)

d)

e)

f)

References

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. An Introduction to Statistical Learning with Applications in R. New York: Springer.