Compulsory exercise 1: Group 21

TMA4268 Statistical Learning V2018

Jørgen Opheim, Ole-Andreas Sandnes and Sander Coates

12 February, 2018

Take a look at the cheat sheets for R Markdown here: File > Help > Cheatsheets > R Markdown Cheat Sheet in RStudio, or here http://www.rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf

or the lessons: http://rmarkdown.rstudio.com/lesson-1.html

Problem 1 - Core concepts in statistical learning [2 points]

a) Training and test MSE [1 point]

- *Figure 2* shows that variance is reduced for increased values of K, but at the cost of increased bias.
- A low value of K gives the most flexible fit.
- As expected from lower flexibility for higher K, the training MSE increases with K. This is due to the reduced fitting of the model to the specific data. However, when introducing the test data, overfitting by a too flexible model leads to increased test MSE for the lowest values of K. This suggest a slightly less flexible and thus less biased model (more on the bias-variance trade-off later) is better.
- By observation, it seems that $K = 3$ gives the lowest test MSE, and hence is the best choice of K for modelling $f(x)$ based on observed values $y = f(x) + \epsilon$.

b) Bias-variance trade-off [1 point]

- Explain how that is done. Hint: this is what the M repeated training data sets are used for. The variance is calculated by use of R's own function `var` over all experiments M for a given x and K. The squared bias is then found by squaring the difference between the mean over all M experiments for a given x and K and the true value of y for that particular x (equal for all values of K). Add some formulae?
- Focus on Figure 4. As the flexibility of the model increases (K decreases), what happens with
  - the squared bias,
  - the variance, and
  - the irreducible error?
- What would you recommend is the optimal value of K? Is this in agreement with what you found in a)?

Problem 2 - Linear regression [4 points]

Here you see an R chunk that is evaluated (when knitting) and code is displayed.

```r
library(ggplot2)
data = read.table("https://www.math.ntnu.no/emner/TMA4268/2018v/data/SYSBPreg3uid.txt")
dim(data)
colnames(data)
modelA=lm(-1/sqrt(SYSBP) ~ .,data = data)
summary(modelA)
```

a) Understanding model output [1 point]

b) Model fit [1 point]

c) Confidence interval and hypothesis test [1 points]

d) Prediction [1 point]