Compulsory exercise 1: Group 21

TMA4268 Statistical Learning V2018

Jørgen Opheim, Ole-Andreas Sandnes and Sander Coates

12 February, 2018

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]

- 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?
- As flexibility increases (K decreases)
  - the squared bias decreases (as expected by less fitting to the specific training data),
  - variance increases (as expected from closer fitting to specific training data),
  - and the irreducible error is left unchanged. The irreducible error is caused by variance of the underlying data and is not affected by modelling.
- By observation of the total test MSE the optimal value of K seems to be $K = 3$ (for which the total error is smallest), as suggested in a).
- [Extra] In $Figure 5$ the optimal value of K seems to be greater than that previously identified. This is however for four specific values of x, none of which are at the boundaries of x's domain ($x \in [-3, 3]$) Possible to do analysis for more values of x to test validity of reasoning?

Problem 2 - Linear regression [4 points]

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

```
library(ggplot2)
data = read.table("https://www.math.ntnu.no/emner/TMA4268/2018v/data/SYSBPreg3uid.txt")
dim(data)
colnames(data)
```

```r
modelA=lm(-1/sqrt(SYSBP) ~ .,data = data)
summary(modelA)
```

a) Understanding model output [1 point]

Hva skjer bri?

b) Model fit [1 point]

Fitting models is my forte

c) Confidence interval and hypothesis test [1 points]

d) Prediction [1 point]