Compulsory exercise 1: Group 16

TMA4268 Statistical Learning V2022

Weicheng Hua, Emil Johannesen Haugstvedt, Torbjørn Baadsvik

28 March, 2022

Contents

Problem 1

a)

b)

c)

d)

Problem 2

a)

b)

Problem 3

a

Say for *each* of them if it is true or false.

(i) For the polynomial regression (where polynomial functions of features are used as predictors), variance increases when including predictor with a high order of the power.

True. The proof for this is pretty long and complicated. You can see it in the link

(ii) If the polynomial functions from (i) are replaced with step functions, then the regression model is too simple to be overfitted on a dataset even with multiple cutpoints.

False. By increasing the number of cutpoints each step function will get more and more affected by the points within their range, and thus overfit.

(iii) The smoothing spline ensures smoothness of its function, g, by having a penalty term $\int g^{\prime}(t)^2 dt$ in its loss.

False. The penalty term is $\int g''(t)^2 dt$.

(iv) The K-nearest neighbors regression (local regression) has a high bias when its parameter, k, is high.

True. With high k the number more neighbors are needed in order to classify a point, thus the variance will be low and the bias will increase.

b

Fit an additive model on ‘boston.train‘ using the function ‘gam()‘ from package ‘gam‘ with the following conditions, and plot the resulting curves.

- response: ‘medv‘; predictors: ‘rm‘, ‘ptratio‘, ‘lstat‘ (use these three predictors only).

- ‘rm‘ is a linear function

- ‘ptratio‘ is a smoothing spline with ‘df=3‘.

- ‘lstat‘ is a polynomial of degree 2.

```
# Fit model

model <- gam(medv ~ rm + s(ptratio, k = 3) + poly(lstat, df = 2), data = boston.train)

# Plot model with training data

plot(model, boston.train)
````
```

### a)

### b)

## Problem 4

### a)

### b)

### c)

### d)

## Problem 5

### a)

### b)

## Problem 6

### a)

### b)

### c)

### d)
```