Stat 432 Homework 5

Assigned: Feb 23, 2019; Due: 11:59pm Mar 1, 2019

Question 1 (linear regression)

[2 points] On page 23 of lecture note "LinearReg", what are the irreducible error, bias and variance? Provide a brief explanation of your answer.

[2 points] For Ridge regression, how does the tuning parameter trades bias and varaince of the prediction error? Provide a brief and non-technical explanation (within 100 words).

Question 2 (model selection criteria)

The Boston Housing data is a classical dataset that models the median house values `medv` of different areas of Boston. Because a lot of variables exhibit an asymmetry, we will use some transformations.

```
data(Boston, package="MASS")
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat medv
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33 36.2
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7
```

```
useLog = c(1,3,5,6,8,9,10,14)
Boston[,useLog] = log(Boston[,useLog])
Boston[,2] = Boston[,2] / 10
Boston[,7] = Boston[,7]^2.5 / 10^4
Boston[,11] = exp(0.4 * Boston[,11])/1000
Boston[,12] = Boston[,12] / 100
Boston[,13] = sqrt(Boston[,13])
```

part a)

[1 point] Fit a linear regression that models `medv` using all other covariates, including an intercept term.

part b)

[3 points] You cannot use existing statistical functions, e.g. `AIC()`, for the first two questions.

- Calculate the Mallow's Cp statistic of this model fitting.
- Based on the parameter estiamtes, if we assume that the errors follow i.i.d. Normal distribution, calculate the -2 log-likelihood of this model fitting based on the maximum likelihood estimators of σ^2. Count σ^2 in the Normal density function as one additional parameter, calculate the AIC and BIC statistics of this model fitting.
- Select the best models based on Mallow's Cp, AIC and BIC respectively. Are they the same?

Question 3 (ridge regression)

[2 points] Use the ridge regression to fit this dataset. You should consider a range of penalty levels and use the generalized cross-validation criteria to select the best tuning. Report sufficient infromation of your final model fitting results, such as parameter estiamtes and the best penalty level.