

Exercises week 5

To do the exercises, you need to get the `Boston` dataset, using the R commands

```
> library(MASS)
> data('Boston')
```

From the course book (Hastie et al, 2009):

- Exercise 3.13 and 3.16 at page 96

Exercises

From James et al (2013):

- Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^p |\beta_j| \leq s$$

for a particular value of s . For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

(a) As we increase s from 0, the training RSS will:

- Increase initially, and then eventually start decreasing in an inverted U shape.
- Decrease initially, and then eventually start increasing in a U shape.
- Steadily increase.
- Steadily decrease.
- Remain constant.

(b) Repeat (a) for test RSS.

(c) Repeat (a) for variance.

(d) Repeat (a) for (squared) bias.

(e) Repeat (a) for the irreducible error.

- Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

for a particular value of λ . For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

(a) As we increase λ from 0, the training RSS will:

- Increase initially, and then eventually start decreasing in an inverted U shape.
- Decrease initially, and then eventually start increasing in a U shape.
- Steadily increase.
- Steadily decrease.

- v. Remain constant.
 - (b) Repeat (a) for test RSS.
 - (c) Repeat (a) for variance.
 - (d) Repeat (a) for (squared) bias.
 - (e) Repeat (a) for the irreducible error.
- We will now try to predict per capita crime rate in the **Boston** data set.
 - (a) Try out some of the regression methods explored in ~~this chapter~~ [the last lectures], such as best subset selection, the lasso, ridge regression, and PCR. Present and discuss results for the approaches that you consider.
 - (b) Propose a model (or set of models) that seem to perform well on this data set, and justify your answer. Make sure that you are evaluating model performance using validation set error, cross-validation, or some other reasonable alternative, as opposed to using training error.
 - (c) Does your chosen model involve all of the features in the data set? Why or why not?

References

- Hastie, T., Tibshirani, R., & Friedman, J.** (2009). *The Elements of Statistical Learning*. Springer, New York.
- James, G., Witten, D., Hastie, T., & Tibshirani, R.** (2013). *An Introduction to Statistical Learning*. Springer, New York.