

Exercises week 8

To do one exercise, you need to get the `College` dataset, using the R commands

```
> data <- read.csv('https://www.statlearning.com/s/College.csv', header = T)
```

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

- Exercise 9.1: the goal is to understand a characteristic of the smoothing splines without digging into the eigenvalues of the smoothing matrix. See §9.1.1, §5.4.1 (and §6.1.1);

From James et al (2013):

- This question relates to the `College` data set.
 - (a) Split the data into a training set and a test set. Using out-of-state tuition as the response and the other variables as the predictors, perform forward stepwise selection on the training set in order to identify a satisfactory model that uses just a subset of the predictors.
 - (b) Fit a GAM on the training data, using out-of-state tuition as the response and the features selected in the previous step as the predictors. Plot the results, and explain your findings.
 - (c) Evaluate the model obtained on the test set, and explain the results obtained.
 - (d) For which variables, if any, is there evidence of a non-linear relationship with the response?
- In ~~Section 7.7~~ [class], it was mentioned that GAMs are generally fit using a backfitting approach [(see also §9.1.1)]. The idea behind backfitting is actually quite simple. We will now explore backfitting in the context of multiple linear regression.

Suppose that we would like to perform multiple linear regression, but we do not have software to do so. Instead, we only have software to perform simple linear regression. Therefore, we take the following iterative approach: we repeatedly hold all but one coefficient estimate fixed at its current value, and update only that coefficient estimate using a simple linear regression. The process is continued until convergence – that is, until the coefficient estimates stop changing. We now try this out on a toy example.

- (a) Generate a response Y and two predictors X_1 and X_2 , with $n = 100$.
- (b) Initialize $\hat{\beta}_1$ to take on a value of your choice. It does not matter what value you choose.
- (c) Keeping $\hat{\beta}_1$ fixed, fit the model

$$Y - \hat{\beta}_1 X_1 = \beta_0 + \beta_2 X_2 + \varepsilon.$$

You can do this as follows:

```
> a <- y - beta1 * x1
> beta2 <- lm(a ~ x2)$coef[2]
```

- (d) Keeping $\hat{\beta}_2$ fixed, fit the model

$$Y - \hat{\beta}_2 X_2 = \beta_0 + \beta_1 X_1 + \varepsilon.$$

You can do this as follows:

```
> a <- y - beta2 * x2
> beta2 <- lm(a ~~ x1)$coef[2]
```

- (e) Write a for loop to repeat (c) and (d) 1,000 times. Report the estimates of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ at each iteration of the for loop. Create a plot in which each of these values is displayed, with $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ each shown in a different color.
- (f) Compare your answer in (e) to the results of simply performing multiple linear regression to predict Y using X_1 and X_2 . Use the `abline()` function to overlay those multiple linear regression coefficient estimates on the plot obtained in (e).
- (g) On this data set, how many backfitting iterations were required in order to obtain a “good” approximation to the multiple regression coefficient estimates?
- Draw an example (of your own invention) of a partition of two-dimensional feature space that could result from recursive binary splitting. Your example should contain at least six regions. Draw a decision tree corresponding to this partition. Be sure to label all aspects of your figures, including the regions $R_1, R_2, \dots$, the cutpoints $t_1, t_2, \dots$, and so forth.