

Exercises week 9

To do the exercises, you need to get the `Boston` and the `Carseats` datasets, using the R commands

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

and, after having installed the package `ISLR`,

```
> library(ISLR)
> attach(Carseats)
```

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

- Exercise 11.1 (excluding the last sentence): to understand the similarities between PPR and NN. See §11.2 and §11.3;

From James et al (2013):

- ~~In the lab, we applied~~ [Apply] random forests to the Boston data using `mtry = 6` and using `ntree = 25` and `ntree = 500`. Create a plot displaying the test error resulting from random forests on this data set for a more comprehensive range of values for `mtry` and `ntree`. You can model your plot after ~~Figure 8.10~~ [15.3 of the textbook]. Describe the results obtained.
- ~~In the lab, a classification tree was applied to the Carseats data set after converting Sales into a qualitative response variable.~~ Now we will seek to predict `Sales` [in the `Carseats` dataset] using regression trees and related approaches, treating the response as a quantitative variable.
 - (a) Split the data set into a training set and a test set.
 - (b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test error rate do you obtain?
 - (c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test error rate?
 - (d) Use the bagging approach in order to analyze this data. What test error rate do you obtain? Use the `importance()` function to determine which variables are most important.
 - (e) Use random forests to analyze this data. What test error rate do you obtain? Use the `importance()` function to determine which variables are most important. Describe the effect of `m`, the number of variables considered at each split, on the error rate obtained.
- Apply boosting, bagging, and random forests to a data set of your choice. Be sure to fit the models on a training set and to evaluate their performance on a test set. How accurate are the results compared to simple methods like linear or logistic regression? Which of these approaches yields the best performance?