STAT 471: Homework 2

Name

Due: October 4, 2021 at 11:59pm

Contents

Instructions

Setup

Pull the latest version of this assignment from Github and set your working directory to `stat-471-fall-2021/homework/homework-2`. Consult the getting started guide if you need to brush up on `R` or `Git`.

Collaboration

The collaboration policy is as stated on the Syllabus:

> "Students are permitted to work together on homework assignments, but solutions must be written up and submitted individually. Students must disclose any sources of assistance they received; furthermore, they are prohibited from verbatim copying from any source and from consulting solutions to problems that may be available online and/or from past iterations of the course."

In accordance with this policy,

Please list anyone you discussed this homework with:

Please list what external references you consulted (e.g. articles, books, or websites):

Writeup

Use this document as a starting point for your writeup, adding your solutions after "**Solution**". Add your R code using code chunks and add your text answers using **bold text**. Consult the preparing reports guide for guidance on compilation, creation of figures and tables, and presentation quality.

Programming

The `tidyverse` paradigm for data wrangling, manipulation, and visualization is strongly encouraged, but points will not be deducted for using base `R`.

We'll need to use the following `R` packages:

```
library(tidyverse)  # tidyverse
library(kableExtra) # for printing tables
library(cowplot)    # for side by side plots
library(FNN)        # for K-nearest-neighbors regression
```

We'll also need the `cross_validate_spline` function from Unit 2 Lecture 3:

```
source("../../functions/cross_validate_spline.R")
```

Grading

The point value for each problem sub-part is indicated. Additionally, the presentation quality of the solution for each problem (as exemplified by the guidelines in Section 3 of the preparing reports guide will be evaluated on a per-problem basis (e.g. in this homework, there are three problems). There are 100 points possible on this homework, 85 of which are for correctness and 15 of which are for presentation.

Submission

Compile your writeup to PDF and submit to Gradescope.

1 Case study: Bone mineral density (40 points for correctness; 10 points for presentation)

In this exercise, we will be looking at a data set (available online) on spinal bone mineral density, a physiological indicator that increases during puberty when a child grows.

Below is the data description:

> "Relative spinal bone mineral density measurements on 261 North American adolescents. Each value is the difference in spnbmd taken on two consecutive visits, divided by the average. The age is the average age over the two visits."
>
> Variables:
>
> `idnum`: identifies the child, and hence the repeat measurements
>
> `age`: average age of child when measurements were taken
>
> `gender`: male or female
>
> `spnbmd`: Relative Spinal bone mineral density measurement

The goal is to learn about the typical trends of bone mineral density during puberty for boys and girls.

1.1 Import (2 points)

- Using `readr`, import the data from the above URL into a tibble called `bmd`. Specify the column types using the `col_types` argument.
- Print the imported tibble (no need to use `kable`).

1.2 Explore (10 points)

- To keep things simple, let's ignore the fact that we have repeated measurements on children. To this end, remote the `idnum` column from `bmd`.
- What is the number of boys and girls in this dataset (ignoring the fact that there are repeated measurements)? What are the median ages of these boys and girls?
- Produce boxplots to compare the distributions of `spnbmd` and `age` between boys and girls (display these as two plots side by side, one for `spnbmd` and one for `age`). Are there apparent differences in either `spnbmd` or `age` between these two groups?
- Create a scatter plot of `spnbmd` (y axis) versus `age` (x axis), faceting by `gender`. What trends do you see in this data?

1.3 Model (15 points)

There are clearly some trends in this data, but they are somewhat hard to see given the substantial amount of variability. This is where splines come in handy.

1.3.1 Split

To ensure unbiased assessment of predictive models, let's split the data before we start modeling it.

- Split `bmd` into training (80%) and test (20%) sets. Store these in tibbles called `bmd_train` and `bmd_test`, respectively.

```r
set.seed(5) # seed set for reproducibility (DO NOT CHANGE)
```

1.3.2 Tune

- Since the trends in `spnbmd` look somewhat different for boys than for girls, we might want to fit separate splines to these two groups. Separate `bmd_train` into `bmd_train_male` and `bmd_train_female`, and likewise for `bmd_test`.
- Using `cross_validate_spline` from Lecture 3, perform 10-fold cross-validation on `bmd_train_male` and `bmd_train_female`, trying degrees of freedom 1,2,. . . ,15. Display the two resulting CV plots side by side.
- What are the degrees of freedom values minimizing the CV curve for boys and girls, and what are the values obtained from the one standard error rule?
- For the sake of simplicity, let's use the same degrees of freedom for males as well as females. Define `df.min` to be the maximum of the two `df.min` values for males and females, and define `df.1se` likewise. Add these two spline fits to the scatter plot of `spnbmd` (y axis) versus `age` (x axis), faceting by `gender`.
- Given our intuition for what growth curves look like, which of these two values of the degrees of freedom makes more sense?

1.3.3 Final fit

- Using the degrees of freedom chosen above, fit final spline models to `bmd_train_male` and `bmd_train_female`.

1.4 Evaluate (6 points)

- Using the final models above, answer the following questions for boys and girls separately: What percent of the variation in `spnbmd` is explained by the spline fit in the training data? What is the training RMSE? What is the test RMSE? Print these three metrics in a nice table.
- How do the training and test errors compare? What does this suggest about the extent of overfitting that has occurred?

1.5 Interpret (7 points)

- Using the degrees of freedom chosen above, redo the scatter plot with the overlaid spline fits, this time without faceting in order to directly compare the spline fits for boys and girls. Instead of faceting, distinguish the genders by color.
- The splines help us see the trend in the data much more clearly. Eyeballing these fitted curves, answer the following questions. At what ages (approximately) do boys and girls reach the peaks of their growth spurts? At what ages does growth largely level off for boys and girls? Do these seem in the right ballpark?

2 KNN and bias-variance tradeoff (45 points for correctness; 5 points for presentation)

Setup: Apple farming

You own a square apple orchard, measuring 200 meters on each side. You have planted trees in a grid ten meters apart from each other. Last apple season, you measured the yield of each tree in your orchard (in average apples per week). You noticed that the yield of the different trees seems to be higher in some places of the orchard and lower in others, perhaps due to differences in sunlight and soil fertility across the orchard.

Unbeknownst to you, the yield Y of the tree planted X_1 meters to the right and X_2 meters up from the bottom left-hand corner of the orchard has distribution

$$Y = 50 + 0.001X_1^2 + 0.001X_2^2 + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \quad \sigma = 4.$$

The data you collected are as in Figure 1.

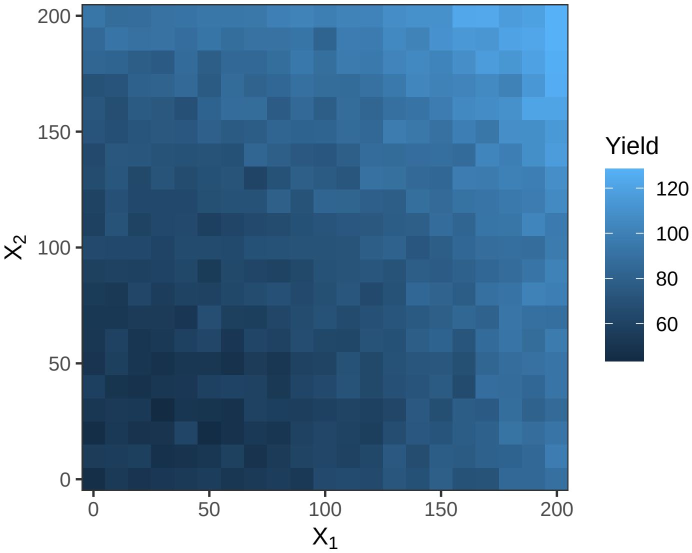

Figure 1: Apple tree yield for each 10m by 10m block of the orchard in a given year.

The underlying trend is depicted in Figure 2, with the top right-hand corner of the orchard being more fruitful.

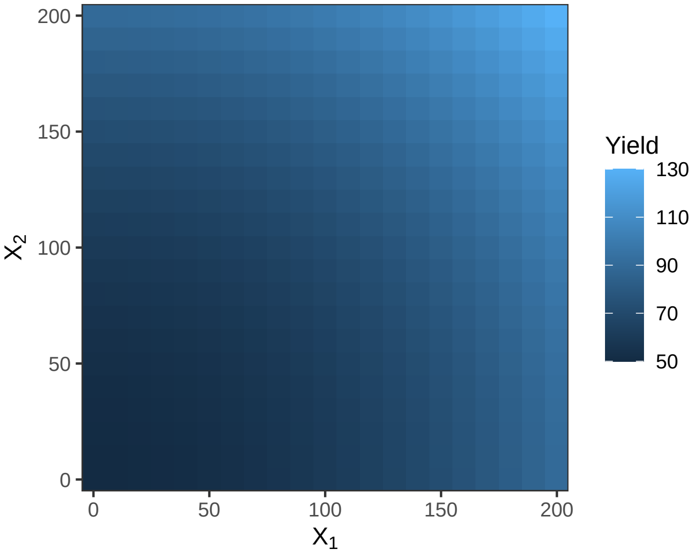

Figure 2: Underlying trend in apple yield for each 10m by 10m block of the orchard.

2.1 A simple rule to predict this season's yield (15 points)

This apple season is right around the corner, and you'd like to predict the yield of each tree. You come up with perhaps the simplest possible prediction rule: predict this year's yield for any given tree based on last year's yield from that same tree. Without doing any programming, answer the following questions:

- What is the expected training error of such a rule?
- Averaged across all trees, what is the squared bias, variance, and ETE of this prediction rule?
- Why is this not the best possible prediction rule?

2.2 K-nearest neighbors regression (conceptual) (15 points)

As a second attempt to predict a yield for each tree, you average together last year's yields of the K trees closest to it (including itself, and breaking ties randomly if necessary). So if you choose $K = 1$, you get back the simple rule from the previous section. This more general rule is called *K-nearest neighbors (KNN) regression* (see ISLR p. 105).

KNN is not a parametric model like linear or logistic regression, so it is a little harder to pin down its degrees of freedom.

- What happens to the model complexity as K increases? Why?
- The degrees of freedom for KNN is sometimes considered n/K, where n is the training set size. Why might this be the case? [Hint: consider a situation where the data are clumped in groups of K.]
- Conceptually, why might increasing K tend to improve the prediction rule? What does this have to do with the bias-variance tradeoff?
- Conceptually, why might increasing K tend to worsen the prediction rule? What does this have to do with the bias-variance tradeoff?

2.3 K-nearest neighbors regression (simulation) (15 points)

Now, we try KNN for several values of K. For each, we compute the bias, variance, and ETE for each value based on 50 resamples. The code for this simulation, provided for you below (see Rmd file; code omitted from PDF for brevity), results in Figure 3.

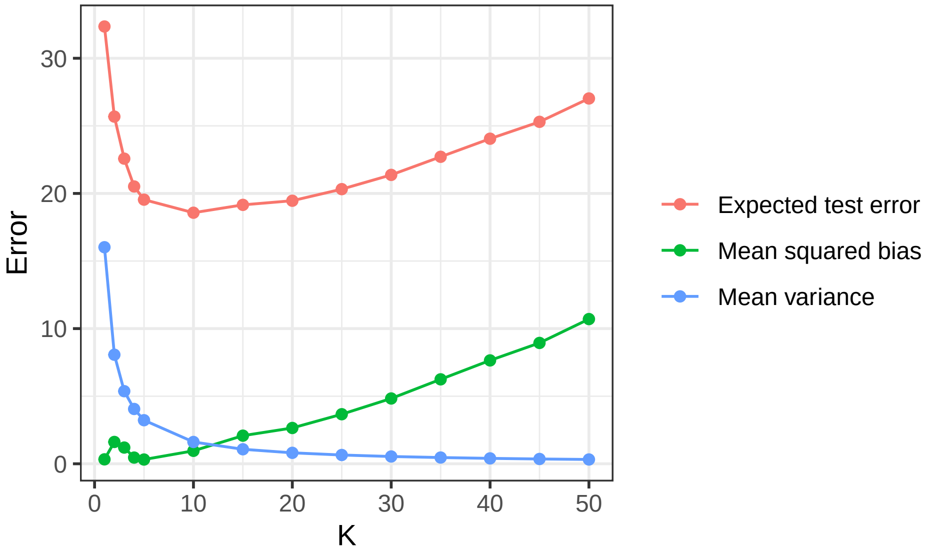

Figure 3: Bias-variance trade-off for KNN regression.

- Based on Figure 3, what is the optimal value of `K`?
- We are used to the bias decreasing and the variance increasing when going from left to right in the plot. Here, the trend seems to be reversed. Why is this the case?

- The squared bias has a strange bump between `K` $= 1$ and `K` $= 5$, increasing from `K` $= 1$ to `K` $= 2$ but then decreasing from `K` $= 2$ to `K` $= 5$. Why does this bump occur? [Hint: Think about the rectangular grid configuration of the trees. So for a given tree, the closest tree is itself, and then the next closest four trees are the ones that are one tree up, down, left, and right from it.]
- The data frame `training_results_summary` contains the bias and variance for every tree in the orchard, for every value of `K`. Which tree and which value of `K` gives the overall highest absolute bias? Does the sign of the bias make sense? Why do this particular tree and this particular value of `K` give us the largest absolute bias?
- Redo the bias-variance plot above, this time putting `df = n/K` on the x-axis. What do we notice about the variance as a function of `df`? Derive a formula for the KNN variance and superimpose this formula onto the plot as a dashed curve. Do these two variance curves match? [Hint: To derive the KNN variance, focus first on the prediction of a single tree. Recall the facts that the variance of the sum of independent random variables is the sum of their variances, and that the variance of a constant multiple of a random variable is the square of that constant times its variance.]