STAT 4710: Homework 2

Name

Due: October 4, 2022 at 12:00pm

Contents

Instructions

Materials

The allowed materials are as stated on the Syllabus:

> "Students may consult all course materials, including course textbooks, for all assignments and assessments. For programming-based assignments (homeworks and exams), students may also consult the internet (e.g. Stack Overflow) for help with general programming tasks (e.g. how to add a dashed line to a plot). Students may not search the internet for help with specific questions or specific datasets on any homework or exam. In particular, students may not use solutions to problems that may be available online and/or from past iterations of the course."

Collaboration

The collaboration policy is as stated on the Syllabus:

> "Students are permitted to work together on homework assignments, but must write up and submit solutions individually. In particular, students may not copy each others' solutions. Furthermore, students must disclose all classmates with whom they collaborated on a given homework assignment."

In accordance with this policy,

Please list anyone you discussed this homework with:

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. In particular, if the instructions ask you to "print a table", you should use `kable`. If the instructions ask you to "print a tibble", you should not use `kable` and instead print the tibble directly.

Programming

The `tidyverse` paradigm for data visualization, manipulation, and wrangling is required. No points will be awarded for code written in base R.

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

```
library(tidyverse)  # tidyverse
library(readxl)     # for reading Excel files
library(knitr)      # for include_graphics()
library(kableExtra) # for printing tables
library(cowplot)    # for side by side plots
library(FNN)        # for K-nearest-neighbors regression
library(stat471)    # for cross_validate_spline()
```

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 4 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 (given in `bmd-data.xlsx`) on spinal bone mineral density, a physiological indicator that increases during puberty when a child grows. In this dataset, `idnum` is an identifier for each child and `spnbmd` represents the relative change in spinal bone mineral density between consecutive doctor's visits.

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

1.1 Import (2 points)

Since the data are in Excel format, the functions in `readr` are insufficient to import it. Instead, you must use `readxl`, another tidyverse package. Familiarize yourself with `readxl` by referring to the data import cheat sheet or the package website.

1. Using the `readxl` package, import the data into a tibble called `bmd_raw`.

2. Print the imported tibble (no need to use `kable`).

1.2 Tidy

1. Comment on the layout of the data in the tibble. What should be the variables in the data? What operation is necessary to get it into tidy format?
2. Apply this operation to the data, storing the result in a tibble called `bmd`.

1.3 Explore (10 points)

1. What is the total number of children in this dataset? What are the number of boys and girls? What are the median ages of these boys and girls?
2. Produce plots 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?
3. Create a scatter plot of `spnbmd` (y axis) versus `age` (x axis), faceting by `gender`. What trends do you see in this data?

1.4 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.

1.4.1 Split

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

1. Split `bmd` into training (80%) and test (20%) sets, using the rows in `train_samples` below for training. Store these in tibbles called `bmd_train` and `bmd_test`, respectively.

```
set.seed(5) # seed set for reproducibility (DO NOT CHANGE)
n <- nrow(bmd)
train_samples <- sample(1:n, round(0.8*n))
```

1.4.2 Tune

1. 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`.
2. Using `cross_validate_spline` from the `stat471` R package, 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.
3. 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?
4. 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`.
5. Given your intuition for what growth curves look like, which of these two values of the degrees of freedom makes more sense?

1.4.3 Final fit

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

1.5 Evaluate (6 points)

1. Using the final models above, answer the following questions for boys and girls separately: What is the training RMSE? What is the test RMSE? Print these metrics in a nice table.
2. How do the training and test errors compare? What does this suggest about the extent of overfitting that has occurred?

1.6 Interpret (7 points)

1. 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.
2. 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 $Y = f(X) + \epsilon$, where

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

The data you collected are as in Figure 1 (compile to PDF or see to view the figure).

The underlying trend is depicted in Figure 2, with the top right-hand corner of the orchard being more fruitful (compile to PDF or see to view the figure).

NOTE: Some of your answers for this question will include mathematical expressions. Please see this page for a quick guide on how to write mathematical expressions in R Markdown. Alternatively, you may write any mathematical derivations by hand and then include them in your writeup via `include_graphics()`.

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:

1. What is the training error of such a rule?
2. Averaged across all trees, what is the squared bias, variance, and expected test error of this prediction rule?

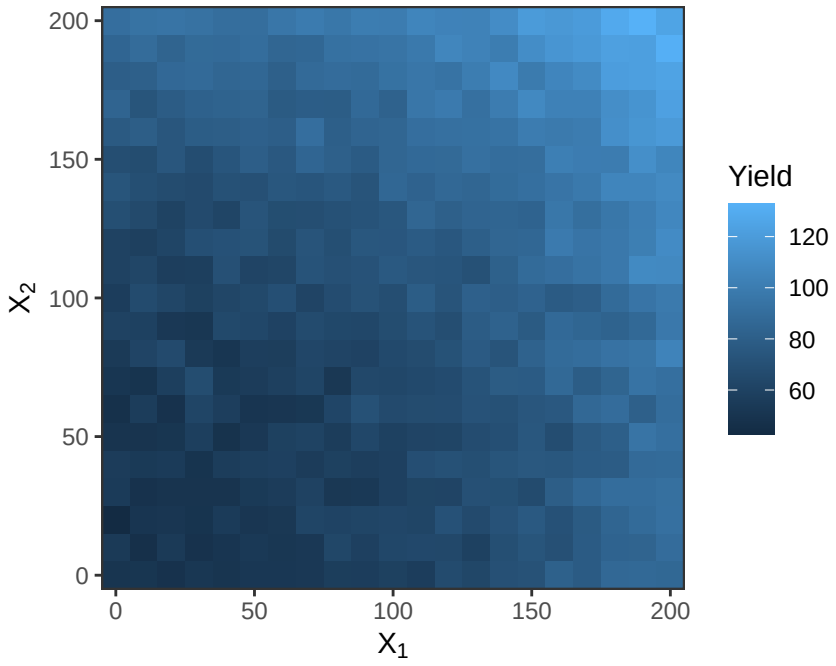

Figure 1: Apple tree yield for each 10m by 10m block of the orchard in a given year.

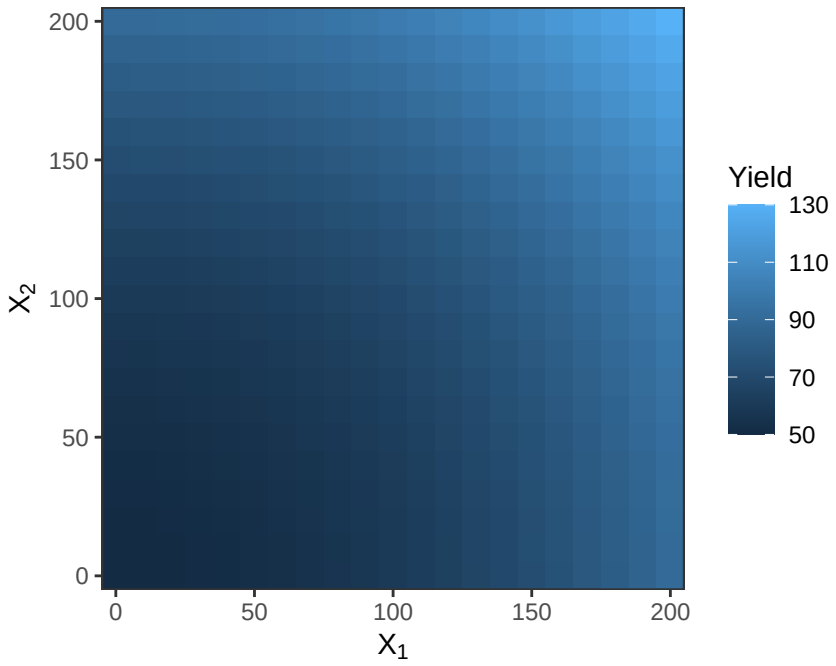

Figure 2: Underlying trend in apple yield for each 10m by 10m block of the orchard.

3. 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.

1. What happens to the model complexity as K increases? Why?
2. 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.]
3. Conceptually, why might increasing K tend to improve the prediction rule? What does this have to do with the bias-variance tradeoff?
4. 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 value of K, we use a numerical simulation to compute the bias and variance for every tree in the orchard. These results are contained in `training_results_summary` below.

```
training_results_summary <- readRDS("training_results_summary.rds")
training_results_summary
```

```
## # A tibble: 6,174 x 5
##        K    X1    X2    bias variance
##    <int> <dbl> <dbl>   <dbl>    <dbl>
##  1     1     0     0 -0.250      16.2
##  2     1     0    10  0.140      12.2
##  3     1     0    20 -0.523      20.4
##  4     1     0    30  0.109      15.6
##  5     1     0    40 -0.566      21.4
##  6     1     0    50 -0.336      15.9
##  7     1     0    60 -1.04       12.4
##  8     1     0    70 -0.0213     12.4
##  9     1     0    80 -0.884      13.5
## 10     1     0    90 -0.342      14.6
## # ... with 6,164 more rows
## # i Use `print(n = ...)` to see more rows
```

1. Create a new tibble called `overall_results` the contains the mean squared bias, mean variance, and expected test error for each value of `K`. This tibble should have four columns: `K`, `mean_sq_bias`, `mean_variance`, and `expected_test_error`.
2. Plot the mean squared bias, mean variance, and expected test error on the same axes as a function of K. Based on this plot, what is the optimal value of `K`?
3. 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?

4. 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.]

5. Based on the information in `training_results_summary`, 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?

6. Redo the bias-variance plot from part 2, this time putting `df = n/K` on the x-axis. What do we notice about the variance as a function of `df`?

7. Derive a formula for the KNN mean variance. [Hint: First, write down an expression for the KNN prediction for a given tree. Then, compute the variance of this quantity using the fact that the variance of the average of N independent random variables each with variance s^2 is s^2/N. Finally, compute the mean variance by averaging over trees.]

8. Create a plot like that in part 6, but with the KNN formula from part 7 superimposed as a dashed curve. Do these two variance curves match?