STAT 341: Assignment 5 - Spring 2020

Name

111 Marks, Due: Friday, August 7 at 10:00am

NOTES

Your assignment must be submitted by the due date listed at the top of this document, and it must be submitted electronically in .pdf format via Crowdmark/LEARN. This means that your responses for different questions should be in separate .pdf files. Your .pdf solution files must have been generated by R Markdown unless otherwise specified. Additionally:

- For mathematical questions: your solutions must be produced by LaTeX (from within R Markdown). Handwritten and scanned/photographed solutions will not be accepted and you will receive zero points.
- For computational questions: R code should always be included in your solution (via code chunks in R Markdown). If code is required and you provide none, you will receive zero points.
  - **Exception** any functions used in the notes or function glossary can loaded using `echo=FALSE` but any other code chunks should have `echo=TRUE`. e.g. the code chuck loading `gradientDescent` can use `echo=FALSE` but chunks that call `gradientDescent` should have `echo=TRUE`.
- For interpretation question: plain text (within R Markdown) is fine.

Organization and comprehensibility is part of a full solution. Consequently, points will be deducted for solutions that are not organized and incomprehensible.

- You will submit your solutions in the form of one pdf file per question through LEARN For example, for Q1 you should submit one pdf file containing the solution to the first question only. Failing to follow the formatting instructions may result in your whole paper or individual questions receiving a grade of 0%.

Question 1 - 21 Marks - Daily temperature data for Ottawa

- To explore the nature of the predictive accuracy of various polynomials, we will use some daily average air temperatures in Ottawa.
  - The file `ottawaTemp1995.csv` contains the temperatures for ottawa during the Year 1995.
  - The file `ottawaOtherYears.csv` contains the temperatures for ottawa during the Years 1996 to 2014.
  - We interested in using the day in the year from 1 to 365 to predict the daily temperature. I have removed the leap days to avoid confusion.
  - We will built and assess models using temperatures from 1995 in parts a) to g) then evaluate it using the other tears in part h).

a) **[3 Marks]** Generate the scatter plot of the data and overlay fitted polynomials with degrees 2 and 9 to the data.

b) **[2 Marks]** Generate $m = 25$ samples of size $n = 50$. Fit polynomials of degree 2 and 9 to every sample.

c) **[4 Marks]** Using `par(mfrow=c(1,2))` plot all the fitted polynomials with degree 2 and 9 on two different figures. Overlay the two fitted polynomials of degree 2 and 9 based on the whole population (make the colour of the population curves different from the others to make them stand out).

d) **[1 Mark]** Using `var_mutilde` function, calculate the sampling variability of the function of the polynomials with degree equal to 2 and 9.

e) **[1 Mark]** Using `bias2_mutilde` function, calculate the squared bias of the polynomials with degree equal to 2 and 9

f) **[2 Marks]** Generate $m = 25$ samples of size $n = 50$, and using `apse_all` function, calculate the APSE for complexities equal to `0:10`.

- Summarize the results with a table and a graphical display.
- Give a conclusion.

g) Instead of randomly constructing sample and test sets we can use k-fold cross-validation.

i) **[2 Marks]** Create a function creates the k-fold samples from a given population. i.e.

- The function has arguements `k` the number of k-fold, `pop` a population, `xvarname` the name of the x variable, and `yvarname` the of the y variable.
- The function outputs a list containing the k-fold samples and test samples labelled as `Ssamples` and `Tsamples`.
- The function `rep_len` might be helpful.

ii) **[1 Mark]** Use the function from part i) and the `apse` function to find an estimate of the APSE using $k = 5$ fold cross-validation when the `complexity=2`.

iii) **[2 Marks]** Perform $k = 10$-fold cross-validation to estimate the complexity parameter from the set $0 : 10$. Plot APSE by the complexity and give a conclusion.

h) **[3 Marks]** Load the data for temperatures in ottawa during the years 1996 to 2014 and calculate the APSE for complexities equal to `0:10` when using the temperatures from 1995 to fit the model.

- Summarize the results with a table and a graphical display.
- Give a conclusion in comparison to parts f) and g).

Question 2 - 14 Marks

Pick 4 topics within this course. For each topic give a brief description and explain why is it important.

- There is 1 to 2 page limit.

Rubric

Criteria	Descriptor	Marks
Topics	Description and Explanation	/12
Format	Organization and LaTeX	/2

Exam - 76 Marks

Complete the exam posted on LEARN. There is no time limit but a well-prepared student should be able to complete it within 2.5 hours.

- **Notes:**
  - Your solution should be **handwritten** or constructed using ipad or drawing tablet.
  - In Crowdmark, these questions are labelled ExamQ1 to ExamQ11.
  - Please submit 1 page per question and take extra care in posting each solution to the correct question.
  - You not required to print out the exam and write on it but you are required to make your solution clear and legible.

Bonus Question - 1 Bonus Mark

- The bonus mark will be applied to your final grade. Pick one of the following questions
  - Assignment 1, Question 6,
  - Assignment 2, Question 4,
  - Assignment 3, Question 2, and
  - Assignment 4, Question 2.

Modify your answer to incorporate the feedback provided. Use an initial page to explain and summarize how you modified your solution in relation to the feedback. Then provide the updated solution.