STAT 341: Assignment 1 - Fall 2020

Ryan Browne

54 Marks, Due: Friday, September 25 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. 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 received zero points.
- 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 One - 17 Marks - MNIST Database

Here we will use a subset of the Modified National Institute of Standards and Technology database (MNIST) as an introduction to R and data exploration. A description of the data can be found at http://yann.lecun.com/exdb/mnist/

> The MNIST database of handwritten digits, available from this page, has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image.
>
> It is a good database for people who want to try learning techniques and pattern recognition methods on real-world data while spending minimal efforts on preprocessing and formatting.

a) **[2 Marks]** Write a function `initializeDigitPlot` that accepts no parameters and creates a blank plot with x and y coordinate ranges of 0.5 to 28.5, no borders, and no axes. Test this function.

b) **[2 Marks]** Write a function drawBox that accepts three parameters (x, y, brightness) that draws a solid-filled 1x1 box centred at (x, y) onto a preexisting plot with colour equal to the greyscale value. The grayscale value is between zero and 1, where 0 is white, 1 is black and the values inbetween are differents levels of grey. Test this function using the part a) and the parameters `(1,1,0.5)`, `(1,2,0)` and `(1,3,1)`. Use the commands `polygon` and `gray`.

c) **[2 Marks]** Write a function `drawDigit` that accepts a numeric matrix of dimension 28×28 that calls initializeDigitPlot and drawBox to draw the matrix. The values of the matrix are brightness values.

d) **[2 Marks]** Import the files `one100.csv` and `two100.csv` digits file. These files contain one hundred 1's and one hundred 2's with $28 \times 28 = 784$ columans. Each row has brightness values of a digit made from the consecutive columns of a matrix. Use the `drawDigit` function to draw the first digit from each of the files `one100.csv` and `two100.csv` side by side in a 1×2 grid.

e) **[1 Mark]** Plot all the digits from `one100.csv` in a 10×10 grid.

f) **[1 Mark]** Plot all the digits from `two100.csv` in a 10×10 grid.

g) **[4 Marks]** Summarize each of these populations by taking the average and median of the brightness within each pixel. Then plot the four resulting 28×28 matrices in a 2×2 grid using your `drawDigit` function. Compare and constrast using the average and median to summarize the images.

h) **[3 Marks]** Summarize the variability of these populations by taking the standard deviation of the brightness within each pixel. Then plot the two resulting 28×28 matrices in a 1×2 grid using your `drawDigit` function. Comment on the variability.

Question Two - 17 Marks - Properties of Pearson's moment coefficient of skewness

Consider the population $\mathcal{P} = \{y_1, \ldots, y_N\}$. Pearson's moment coefficient of skewness is

$$a(\mathcal{P}) = \frac{\frac{1}{N}\sum_{u \in \mathcal{P}} (y_u - \overline{y})^3}{[SD_{\mathcal{P}}(y)]^3}$$

and hence a measure of skewness. In this question you will investigate several of its properties.

- Note: Some marks in each part allocated to formating and organization.

(a) **[3 points]** Determine whether this skewness coefficient is location invariant, location equivariant, or neither.

(b) **[3 points]** Determine whether this skewness coefficient is scale invariant, scale equivariant, or neither.

(c) **[3 points]** Determine whether this skewness coefficient is location-scale invariant, location-scale equivariant, or neither.

(d) **[3 points]** Determine whether this skewness coefficient is replication invariant, replication equivariant, or neither.

(e) **[3 points]** For the population below, plot the sensitivity curve of $a(\mathcal{P})$ for $y \in [-8, 8]$. You may find the `sc()` function from lecture useful.

```
set.seed(341)
pop <- rnorm(1001)
```

(f) **[2 points]** Given all that you have learned in parts (a) - (f), state one thing that is *good* about the pearson's moment coefficient of skewness attribute and one thing that is *bad* about the skewness attribute.

Question Three - 10 Marks - Finding and describing a population.

By searching the web, find a public dataset that constitutes a population. Provide the following:

- Give a description of the data **in your own words**. Then justify why the dataset is indeed a population as opposed to a sample.
- A URL to access the data.
- Define what is an unit and describe two variate(s) that have been recorded.
- Give a single graphical display of the populaton.
- Describe some interesting attribute or feature of this population.

Some places you might consider looking:

- Kaggle
- UCI Machine Learning Repository
- r/datasets
- data.gov
- KDnuggets

Rubric

Criteria	Descriptor	Marks
Data/URL	Creativity: Was an interesting or unique dataset chosen, provided?	/2
Description	Clarity & Justification	/2
Unit/Variate	Description, Correctness and Graphic	/4
Feature	Description and Justification	/2

Question Four - 10 Marks - Review

A student (Ryan Browne) was given the following question.

In your own words summarize the concept of sensitivity based on subsection 2.2.3 - Influence & Sensitivity.

- You are limited to 1 to 2 pages.

- Your solution should use a combination of formulas, full sentences and an example.

Rubric

Criteria	Descriptor	Marks
Format	Organization & LaTex	/3
Writing	Clarity & Grammar	/2
Content	Coverage, Depth, Relevant Terminology used and Example	/5

- The student's solution is provided on LEARN in the file "Sensitivity_Question.pdf". Line numbers have been added to each page to help refer to certain parts of the report.

a) **[3 Marks]** Give 3 examples of improper formatting, LaTex or grammar

b) **[2 Marks]** Give 2 examples of errors in content or gaps in knowledge.

c) **[2 Marks]** Did the student summarize the concept of sensitivity in their own words? Briefy explain how and provide some evidence or examples.

d) **[1 Mark]** Give one suggestion on how to improve the report.

e) **[2 Marks]** Give this report a grade and briefy describe why this grade is appropriate.