

STAT 341: Assignment 3 - Fall 2020

Ryan Browne

58 Marks, Due: Friday, October 30 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 receive zero points.
 - **Exception** any functions used in the notes or function glossary can be loaded using `echo=FALSE` but any other code chunks should have `echo=TRUE`. e.g. the code chunk 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 (40 Marks): The brightness relationship between digit 1 and digit 2

- Here we explore the relationship between the amount of brightness for digit 1 and 2.
- a) [2 Marks] Data wrangling; From the two files `one180.csv` and `two120.csv` make a single matrix or data frame with 300 rows (units or digits) and two columns (variables);
- **Brightness** is the average pixel brightness for each observation (or each digit) and
 - **Digit1** is an indicator which is 1 for the one's and 0 for the two's.
- b) [2 Marks] Construct two histograms (side-by-side) of average pixel brightness by
- using **equal bin widths** with bin width equal to 6 over the range 8 to 68 and
 - using **varying bin widths** with 10 bins.

c) [4 Marks] We can model the relationship between brightness and the digit type non-parametrically by calculating the proportion of ones for a given brightness range. Construct two plots (side-by-side) of **Brightness** variable (x-axis) versus the **Digit1** variable (y-axis).

- Then using the two brightness partitions (**equal bin widths** and **varying bin widths**) from part b), add points using the mid-point of the brightness interval and the proportion of ones within each brightness interval.
- In addition, for each brightness partition construct a table consisting of the following columns: brightness range, total number of digits within that range, number of one's, and the proportion of one's. It might be helpful to write a function that takes in the brightness partitions and then creates a plot and outputs the table.
- Briefly compare and contrast the two tables and plots.

d) Another approach is to model the proportion using a parametric model. i.e. Use a function to model the proportions that is bounded to the range $[0, 1]$. One example is the logistic function but here we will consider using the the cumulative distribution function of the standard normal distribution denoted by Φ and is also known as the **inverse probit function**. It is given by

$$\Phi(z) = \int_{-\infty}^z \phi(u) du = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

- [2 Marks] Using the R-function **pnorm** and plot the Φ function over the range $[-6, 6]$
- [2 Marks] Recreate the plot from c) using **varying bin widths** and overlay the following function

$$\Phi(\hat{y}) \quad \text{where} \quad \hat{y} = 1/2 - 0.03 \times (\text{Average Pixel Brightness}).$$

e) [2 Marks] To find a model from part d) that fits the population well, we will use the log-likelihood of N bernoulli trials with varying probability of success which is given by

$$l(\theta) = l(\alpha, \beta) = \frac{1}{N} \sum_{i=1}^N \left[y_i \log \frac{p_i}{1-p_i} + \log(1-p_i) \right]$$

where

$$p_i = \Phi(\hat{y}) = \Phi(\alpha + \beta [x_i - \bar{x}])$$

Differentiate the above with respect to α and β . In particular, show that

$$\frac{\partial l}{\partial(\alpha, \beta)} = \frac{1}{N} \sum_{i=1}^N \frac{y_i - p_i}{p_i(1-p_i)} \times \phi(\hat{y}) \times \begin{bmatrix} 1 \\ x_i - \bar{x} \end{bmatrix}$$

where $\phi(\cdot)$ is the probability density function for the standard normal density. i.e.

$$\phi(\hat{y}) = \frac{1}{\sqrt{2\pi}} e^{-\hat{y}^2/2}$$

Hint: Use the chain rule.

f) Here we will fit the model given by the equation from part e) via gradient descent. **Note**, under maximum likelihood we would maximize the function given e), so we want to minimize the negative of the log-likelihood function.

- [2 Marks] Modify the **createLeastSquaresRho** function to calculate the objective or negative of the log-likelihood function, $-l(\theta)$. Call this new function **createObjProbit**.
- [2 Marks] Modify the **createLeastSquaresGradient** function to calculate the gradient of the objective function or the gradient of the negative of the log-likelihood. Call this new function **createGradientProbit**
- [2 Marks] Using the functions **gradientDescent**, **gridLineSearch**, and **testConvergence** from notes and the functions you created in part i) and ii), perform gradient descent until convergence with **theta=c(0,0)** and **lambdaStepsize = 0.001**, **lambdaMax = 1**

- iv) [2 Marks] What are the values of α and β that correspond to brightness having no effect on proportion of ones's (digit type)? Use these parameter values as a starting value for gradient descent. Is there any improvement?
- g) Now, we will assess the fitted model.
- i) [2 Marks] Recreate the plot from c) with **varying bin widths** and overlay the fitted model.
 - ii) [2 Marks] Compare the fitted values from the model to the points added to the plot.
 - iii) [2 Marks] What is implicit assumption behind the parametric model and non-parametric model?
 - iv) [2 Marks] At what average pixel brightness would the model report as a 50-50 chance of the digit being an one.
- h) Comparing different starting values
- i) [2 Marks] Generate a contour plot superimposed with the function $-l(\alpha, \beta)$ for the region for $\alpha \in (0, 2)$ and $\beta \in (-0.3, 0)$. You may find the functions `outer`, `image`, and `contour` useful for this task, but you do not have to use them.
 - ii) [1 Mark] Modify the `gradientDescent` function to use the **unnormalized gradient** for the search direction and in addition outputs the sample path, i.e. the sequence of updates.
 - iii) [3 Marks] Using the initial values $(\alpha^0, \beta^0) = \{(0, 0), (0.25, 0), (2, -0.3)\}$ perform gradient descent using the function from part ii). As part of the solution,
 - Use the argument `lambdaStepsize = 0.001`, `lambdaMax = 1` and `tolerance = 1e-3`.
 - Plot the contour plot and add the three solution paths based on the initial values.
 - Summarized the output from the three algorithms in a table.
 - **Note:** There are marks allocated to towards organized and presentation.
 - iv) [1 Marks] Write a function similar to `createGradientProbit` called `createStochasticGrad` which estimates the gradient of the objective function using a sample. The function will have
 - inputs the variables `x` and `y` & the size of the sample used for approximation `nsize`, and
 - output is a function that has α & β as inputs and returns an estimate of the gradient using a sample.
 - v) [3 Marks] Perform random sample stochastic gradient descent from part iv) using samples of size 25, a fixed step size equal to 0.001, 500 iterations and the three initial values $(\alpha^0, \beta^0) = \{(0, 0), (0.25, 0), (2, -0.3)\}$. As part of the solution give three plots in a row;
 - The contour plot with the three solution paths based on the initial values.
 - A plot of solution paths for α versus the iteration number.
 - A plot of solution paths for β versus the iteration number.
 - To the plots by iteration number add a horizontal line to represent the value of α or β at corresponds to the minimum of the objective function.
 - **Note:** There are marks allocated to organization and presentation.

Question 2 - 10 Marks

In your own words summarize the subsections **3.0-Samples**, **3.1-All_Possible_Samples** and **3.1.1-Consistency_and_Sample_Size**.

- You are limited to 1 to 2 pages.
- You summarize should include a combination of formulas, full sentences an example.

Rubric

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

Question 3 - 8 Marks - Test Question

Using the material from Section 2.1, 2.2 and 2.3, construct one exam question worth approximately 2 marks. Notes;

- There is a one page limit.
- Provide a full solution with mark allocations.
- The question can be about a single topic or two disjoint topics.
- **Handwritten** questions and answers are acceptable, but it must be legible.
- TRUE/FALSE or multiple choice questions are not allowed.
- Note: Do not expect full marks for reproducing examples or questions from the exercises, assignments, lecture notes or webpages.

However, you may copy and paste from the notes the definitions, or generic R codes.

Rubric

| Criteria | Descriptor | Marks |
|----------|--|-------|
| Question | Concept, Difficulty, Clarity and Creativity | /4 |
| Solution | Explanation, Correct Justification and Clarity | /4 |