Data Tidying and Reporting – Task 1

MSc in Statistics for Data Science at Carlos III University of Madrid

Eduardo García-Portugués ©

2023-01-31, v1.1

Research

Download and import the dataset `qmnist_nist.RData`. It contains the data frames `train_nist` and `test_nist`, both with variables `digit` (a vector with digit labels), `writer` (a vector with the writer's ID), and `px` (a matrix with $28 \times 28 = 784$ columns of pixel gray levels). The digit images can be visualized with the following function:

```
show_digit <- function(x, col = gray(255:1 / 255), ...) {
  l <- sqrt(length(x))
  image(matrix(as.numeric(x), nrow = l)[, l:1], col = col, ...)
}
```

Do the following:

1. Transform `train_nist` to do a ridge logistic model (refresher here) for classifying the digits 4 and 9.
2. Fit ridge a model, with a cross-validated-chosen λ penalty. Do not `standardize` the predictors (the default; why not?).
3. Plot the estimated β in a way that delivers insights about the classification.
4. Using the `test_nist` dataset, evaluate the prediction accuracy of the model.
5. [Optional] Tackle the 45 classification problems of one versus another digit. Report in a 10×10 matrix the classification accuracy on each problem. Visualize the estimated β's in a way similar to Point 3.

Report

Based on your results for the above points, write a report on the problem of separating hand-written digits. Do not write it for your professor of *Data Tidying and Reporting*, nor as if you where following a class exercise in which you are doing the previous tasks. Rather, write the report for a **general audience** to which you present the results of your own investigations (i.e., tasks 1–5).

The report **must**:

a. Be a **single** `.Rmd` (or `.qmd`) file.
b. Compile in **any environment** with `qmnist_nist.RData` and the required libraries present.
c. Compile in **less than 60 seconds**. *Beware*: `cache` is not going to cut it; all the information must be in a single file!
d. Have, at most, **2 pages**. You will need to compromise and prioritize what to report.

Grading **rubric** for the report:

1. Checks the requirements a–d?
2. Is it rich in information?
3. Gives context on the data and the techniques employed, as well as justifies them?
4. Contains high-quality plots with transparent background?
5. Has informative code comments and formats the code according to tidyverse style?
6. Is it written in a direct, clear way?

7. Has a tidied BibTeX-based bibliography citing, at least, this paper and the used packages?
8. Gives clear and informative interpretations of the results?
9. Deviates substantially from the simplest R Markdown document?
10. Addresses the [Optional] step?