Lab 2 - Stat 215A, Fall 2021

Due: Thursday October 07, 11:59 PM

Push a folder called "lab2" to your stat-215-a GitHub repository. This folder, "lab2", should contain the files below:

- **lab2.Rmd** or lab2.Rnw: the raw report + code with your name
- **lab2.pdf**: the output of lab2.Rnw/lab2.Rmd. This output should not contain any code.
- **lab2_blind.Rmd** or lab2_blind.Rnw: exactly the same as lab2.Rmd/lab2.Rnw but without name *anywhere* in the document (please double check).
- **lab2_blind.pdf**: the output of lab2_blind.Rnw/lab2_blind.Rmd. This output should not contain any code.
- **R**/: a folder containing .R scripts (e.g. load.R and clean.R) that will be sourced in lab2.Rmd
- **other**/: an optional folder containing other miscellaneous files required to reproduce your lab2.Rmd.

You should also have a local **data**/ (and an optional **documents**/) folder but please do not push these folders. Do not push any other files that are not needed for your report, and do not push multiple versions of the lab. Please make an effort to adhere to these filenames exactly, otherwise there is a chance that your lab will not be properly transferred for peer grading.

1 Linguistic Data

This section of the lab uses data from a Dialect Survey conducted by Bert Vaux, which you can take at https://www.dialectsofenglish.com/. The questions and answers can be found in the file `question_data.Rdata` (this information was found and processed from the http://dialect.redlog.net/index.html by an intrepid STAT215A student past). We will focus on the questions that look at lexical differences as opposed to phonetic differences, which are numbered 50-121. There two data sets on GitHub. `lingData` contains the answers to the questions for 47,471 respondents across the United States. The dataset contains the variables `ID`, `CITY`, `STATE`, `ZIP`, `Q50 - Q121` (a few questions in this range are left out), `lat` and `long`. `ID` is a number identifying the respondent. `CITY` and `STATE` were self reported by respondents. Former GSIs found the latitude and longitude for the center of each zipcode and added the `lat` and `long` variables based on the reported city and state. Note that there are missing values. The variables starting with `Q` are the responses to the corresponding question on the website. A value of 0 indicates no response. The other numbers should directly match the responses on the website, i.e. a value of 1 should match a response of (a).

For the second data set, `lingLocation`, the same categorical responses were turned into binary responses. Then the data was binned into one degree latitude by one degree longitude squares. Within each of these bins, the binary response vectors were summed over individuals. Please note that the rows are not normalized.

For example, say John and Paul take this questionnaire for two questions. The first question has three answer choices and the second question has four answer choices. If John answered A and D and Paul answered B and D, then `lingData` would encode two vectors: $(1, 4)$ and $(2, 4)$. If they lived in the same longitude and latitude box, then it would be encoded in `lingLocation` as one vector: $(1, 1, 0, 0, 0, 0, 2)$.

1.1 Your tasks

1. Have a look at the review papers Nerbonne and Kretzschmar [2003] and Nerbonne and Kretzschmar [2006] (both are posted in the `stat-215-a-gsi` repo). These will provide some information regarding the domain context.

2. As you begin exploring the data, pick two survey questions and investigate their relationship to each other and geography. You will need to use maps to examine the geographical relationships and may want to experiment with interactivity, e.g. using linked brushing (see the `crosstalk` R package https://rstudio.github.io/crosstalk/ or see an example using `shiny` https://jjallaire.shinyapps.io/shiny-ggplot2-brushing/). Do the answers to the two questions define any distinct geographical groups? Does a response to one question help predict the other? Try to analyze the categorical data for more than 2 questions.

3. Encode the data so that the response is binary instead of categorical. In the previous example of John and Paul, the encoded binary vectors would be $(1, 0, 0, 0, 0, 0, 1)$ for John and $(0, 1, 0, 0, 0, 0, 1)$ for Paul. (You might want to do this for the previous question as well.) This makes $p = 468$ and $n = 47,471$. Experiment with dimension reduction techniques. What do you see? If you do not see anything, change your projection. Does that make things look different? Did you center and/or scale your data before performing dimension reduction? Discuss your choice of centering/scaling. Why is it not a good idea to perform PCA or other dimension reduction techniques on the original `lingData` dataset?

4. Use the methods we learned in class for clustering to try to gain insights into the full dataset. Perform at least two different clustering methods (i.e K-means and NMF). Are there any groups/clusters of people? Do these groups relate to geography? Are the clusters completely separate or is there a continuum? From where to where? Which questions produce this continuum or separate the clusters? How did you choose the number of clusters? Does the mathematical model behind your dimension reduction strategy make sense for these clusters? What are the advantages and disadvantages of the clustering methods that you decided to use?

5. Choose one of your interesting clustering results. Analyze and discuss the robustness of the clusters. What happens when you perturb the data set (e.g., via bootstrap or subsampling)? What happens when you use different starting points in the algorithm? What do you conclude from your clustering and stability analysis?

6. Recall the three realms of data science (Figure 1.1): data, algorithms and analysis, and future data. Do you think this data is useful for future decision-making purposes? Why or why not? What about your clusters (the results of your algorithms and analysis)? Think of a reality check that would help you to verify your clustering. Given more time, is there anything you would have added or done differently?

1.2 Reproducibility

I will use the scf cluster to test the reproducibility of your code, by cloning into your repository and attempting to compile your reports. For your convenience, I provide you with the script I will be using to test your code. If you wish, you can use the script to test your code in the following way:

1. Go into your lab2 directory, for example using my directory:

   ```
   cd /Users/omerronen/Documents/Phd/215a2021/stat-215-a-gsi/lab2
   ```

2. Run

   ```
   bash test.sh .
   ```

To perform testing on scf, you may need to transfer the data file there which can be done using *SCP* (see detailed instructions on this link).

Lastly if you wish to use any R packages that require installation, please ask for approval here - Link. I will try to install it on scf, and approve if possible.

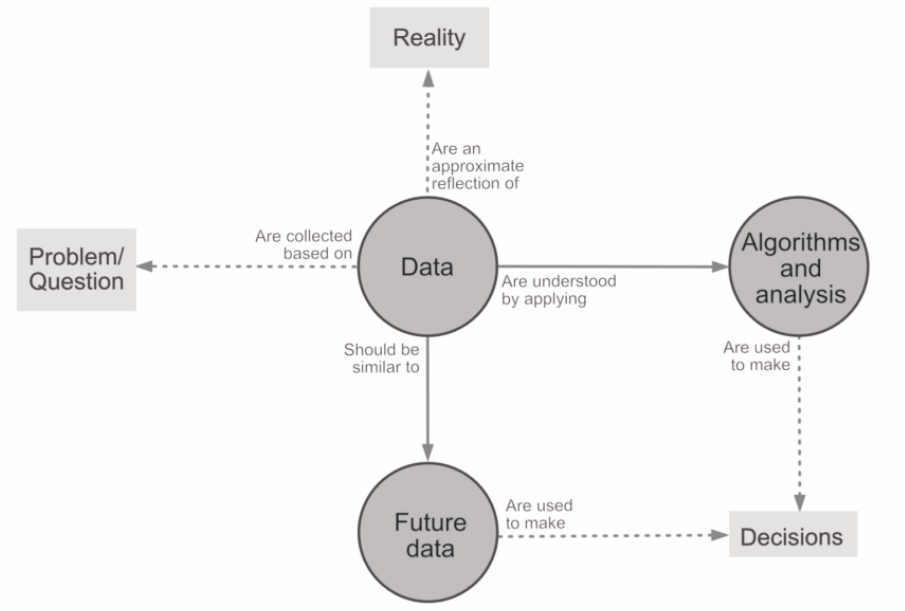

Figure 1.1: The three realms of data science.

References

John Nerbonne and William Kretzschmar. Introducing computational techniques in dialectometry. *Computers and the Humanities*, 37(3):245–255, 2003.

John Nerbonne and William Kretzschmar. Progress in dialectometry: toward explanation. *Literary and linguistic computing*, 21(4):387–397, 2006.