Lab 2 - Linguistics Data, Stat 215A, Fall 2024

September 20, 2024

Please use this structure for your report, but you do not have to slavishly follow this template. All bullet points are merely suggestions and potential points to discuss in your writeup. Your report should be no more than 12 pages, including figures. Do not include *any* code or code output in your report. Indicate your informal collaborators on the assignment, if you had any.

1 Introduction

Things to potentially include in your introduction:

- Describe the problem of interest and put your analysis in the domain context. Read the introduction of the two Nerbonne and Kertzschmar papers for some help here.
- What do you aim to learn from this data?
- Outline what you will be doing in the rest of the report/analysis

2 The Data

- What is the data that you will be looking at?
- Provide a brief overview of the data
- How is this data relevant to the problem of interest? In other words, make the link between the data and the domain problem

2.1 Data Cleaning

- This dataset isn't as bad as the TBI data, but there are still some issues. You should discuss them here and describe your strategies for dealing with them.
- Remember to record your preprocessing steps and to be transparent!

2.2 Exploratory Data Analysis

- This is where you compare pairs of questions with discussion and plots.

3 Dimension Reduction

- This is where you discuss and show plots about the results of whatever dimension reduction techniques you tried—PCA, variants of PCA, t-SNE, NMF, random projections, etc.
- What do you learn from your dimension reduction outputs
- Discuss centering and scaling decisions

4 Clustering

- This is where you discuss and show plots about the results of whatever clustering methods you tried—k-means, hierarchical clustering, NMF, etc.

5 Stability of findings to perturbation

- What happens to your clusters when you perturb the data set?
- What happens when you re-run the algorithm with different starting points?

6 Conclusion

- Discuss the three realms of data science by answering the questions in the instructions pdf.

- Come up with a reality check that would help you to verify your clustering. You do not necessarily have to perform this reality check, but you can if doable.
- What are the main takeaways from your exploration/clustering/stability analysis?

7 Academic Honesty

7.1 Statement

Please include your academic honesty statement here. Do NOT include your name.

7.2 LLM Usage

If, in accordance with the policy in `lab2-instructions.pdf`, you used the one exception to the LLM ban to help complete the lab, please see the instructions for what to write here.

7.3 Collaborators

List your collaborators here.

8 Bibliography

Include any references you used in your report here.