Lab 2 - Stat 215A, Fall 2014

Due: Tuesday October 7, 4:00 PM

- Please bring hard copies of lab writeup and homework to lab on Oct 7. You can also leave hard copies in my mailbox in 367 Evans before 7 PM.
- By 4 PM on Oct 7, please submit in bSpace the .pdf file of your lab writeup, and all R code used to perform your analysis. You are welcome to submit your LaTeX or LyX files, but it is not required.

1 Kernel density plots and smoothing

These tasks use the redwood data from the previous lab. You may have already done similar things in your lab 1; these tasks are focused on experimenting with parameters in kernel smoothers.

1. Plot a density estimate for the distribution of temperature over the whole dataset. Experiment with different kernels and bandwidth. Explain your findings.
2. Choose a time of day and plot the temperature against the humidity for all nodes at that time for the entire project period (hint: there are 288 measurements per day so measurements where epoch mod 288 is constant will all be at the same time of day). Add a loess smoother to the plot. Experiment with bandwidth and the degree of the polynomials. Explain your findings.

2 Linguistic Data

This section of the lab uses data from a Dialect Survey conducted by Bert Vaux. Some limited information can be found at the original website `http://www4.uwm.edu/FLL/linguistics/dialect/index.html`. 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 STAT215 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 bSpace. `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: $(1, 1, 0, 0, 0, 0, 2)$.

2.1 Your tasks

1. Have a look at the review papers Nerbonne and Kretzschmar [2003] and Nerbonne and Kretzschmar [2006] (both are posted under Lab 2 in bSpace)

2. Pick two survey questions and investigate their relationship to each other and geography. You will need to use maps and should experiment with linked brushing (e.g. `iplots`). 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?

4. Use the methods we learned in class for dimension reduction and clustering to try to gain insight into the full dataset. Are there any groups? Do these groups relate to geography? What questions separate the groups? Is there a continuum? From where to where? Which questions produce this continuum? Does the mathematical model behind your dimension reduction strategy make sense for these clusters?

5. Choose one of your interesting finding. Analyze and discuss the robustness of the finding. What happens when you perturb the data set? Different starting points? What can be generalized from this finding?

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.