Math 280: Project 1

(linear algebra)

Overview

We've seen in class that eigenvectors of Markov matrices describe their equilibrium behavior. In particular: every Markov M matrix has 1 as its largest eigenvalue, under mild assumptions this is the unique largest eigenvalue, and that for almost all vectors x the iterates $M^n x/|M^n x|$ converge to an associated eigenvector. So this eigenvector describes the expected distribution of states when the process runs "forever".

Markov's first application was to linguistics: he took the first chapter of *Eugene Onegin* (a well-known Russian novel written in verse) and determined probability of a noun/vowel following a noun/vowel. The equilibrium vector is the proportion of nouns/vowels in the chapter.

In this project, we will explore a similar linguistic question in poetry: are there patterns to the sequences of tones in classical Chinese poetry? There should be, according to the tonal requirements of the forms in which they composed their work, but the language's evolution has changed many pronunciations. Careful analyses of this sort: tone patterns, rhyme, and so on, relative to descriptions of the forms back then are an important linguistic tool in tracking language change, and play an important part in efforts to reconstruct the pronunciations of ancient languages – for example, modern pronunciations of Latin are approximations based in part on combining contemporaneous comparisons to ancient Greek with our understanding of how ancient Greek has morphed into modern.

No language or poetry knowledge required: you can think of tones as analogues of the light/dark colors for English consonants and vowels, but tones furnish a data set that's both more interesting and easier to process.

If you're interested in learning more about our text corpus, it will be drawn from the work of 朱淑真 (Zhu Shuzhen) and 杜甫 (Tu Fu, often Du Fu).

The primary steps of the project are:

1. Process files of *pinyin* text into tone data,
2. Understand the mathematics to develop a method for validating your results and testing them against other texts,
3. Build a Markov matrix for the data,
4. Validate your matrix against other text samples,
5. Compare the results and use them to guess the writer/genre of a mystery text.

A TeX template for the project will be provided.

Purpose Statement

1. Gain hands-on experience implementing with data-cleanup and analysis
   (a) Knowledge: understand file and text manipulation in Sage, conceptualize data structures.

(b) Skills: programming (primarily I/O, loops, strings, lists).

2. Learn a little about computational linguistics.

   (a) Knowledge: why might Markov matrices be appropriate for understanding structured verse?

   (b) Knowledge: what are the limits of these techniques?

   (c) Skills: assess how linear algebra can help explore a given problem arising outside of mathematics.

   (d) Skills: communication, interpreting problems from other fields and producing an explanatory report of your results at a level your peers can appreciate and understand.

3. Practice reflecting on quantitative and qualitative aspects of problems in the sciences, in order to to anticipate their behavior w/r/t numerical methods. Then, assess and enrich your understanding by comparing your expectations to your results and the results of standard numerical tools.

   (a) Knowledge: more detailed quantitative and qualitative aspects of the problems you've selected.

   (b) Knowledge: typical challenges for the techniques we've implemented.

   (c) Knowledge: familiarity with using standard tools and libraries.

   (d) Skills: hypothesize about mathematical phenomena.

   (e) Skills: distinguish suitable/unsuitable/less suitable techniques for a given application.

   (f) Skills: analyze data (program output) for patterns,

   (g) Skills: organizing a report, mechanics of writing and grammar, using LaTeX to typeset mathematical documents.

Task

1. Summary [< 1 page]. For Project 1, the problem is assigned.

   Summarize your own understanding of the project. In particular, read about 律 (lü shī) and perhaps draw an analogy to the analysis of stressed/unstressed syllables in metered English poetry. Suggest one or more other linguistic features that might be interesting to analyze with Markov methods, in poetry, prose or other kinds of writing.

   Submit your summary by October 4th 10pm.

2. Pre-port [< 1 page]. Write a ¡1 page preliminary analysis/expectations.

   Mainly, address the following: how would you use a Markov matrix to distinguish authors or genres? After constructing a matrix from the initial corpus, how can you validate it by testing against other texts? How could you try to identify an author or genre from this information? Do you think we will be able to distinguish the authors we are studying?

   We may discuss the second question in class. Can you validate it without, for example, doing the work of constructing a new matrix from the test text? Consider the following: what do the entries of the equilibrium vector *mean* in this case?

   Submit the pre-port by October 7th 10pm.

3. Implementation. Using SageMath, write a small program which

   (a) Processes text files `df.txt` and `zsz.txt` consisting of lines of poetry in the form of numbered pinyin syllables (`hao3`,`mou2`, etc) to extract the tones into a list, array, or other data structure you find convenient.

   (b) Uses that data to construct, for each writer, a Markov matrix for the 5-state system of tones 1, 2, 3, 4, 5.

   (c) Validates each matrix against the data from the reference texts `df-test.txt` or `df-zsz.txt`.

   (d) Has function which, given a text, guesses the author.

   Implementations should be finished by October 16th, and we will have a lab day on October 10th.

4. Post-port [1+ pages].

   Summarize your results. Compare to your expectations in the pre-port. Does anything surprise you?

5. Report. Reflect on the process and summarize your results in a brief 3-4 page report using the provided LaTeX style file and template. Upload your source code separately. The report is a compilation-of and expansion-to the preceding parts of the task.

   In particular, it should consist of:

   (a) ($<$ 1 pages, Proposal) Background on the examples you chose.

   (b) ($<$ 1 pages, Pre-port) Summary of your expectations for those examples. Be sure to make clear *why* you predict certain behavior.

   (c) (1 pages, Testing/Implementation) Summary of results, including tables, matrices, numerical data, and other figures/information you deem relevant.

   (d) (1+ pages, Post-port) Analysis of the results.

   Submit a draft by October 21 and the final version by 25th.

Criteria

Several of the task items above are due before the main report. Turn them in on time for full credit and so that I am able to provide you feedback! Excellent drafts will require very little work to get them to their final form, and the final report is, in part, a compilation of these components, so it is in your best interest to polish them as much as possible.

- **(15%)** Summary. Explains the background in general terms. Your target audience is a peer not already familiar with the problem (poetry) or method (Markov matrices).
- **(35%)** Pre-port. States expectations for each examples (10%), shows serious mathematical thought about the methods we are using, especially validation (20%) and how it might be used to distinguish data sets (10%).
- **(20%)** Implementation. Accurate implementation (15%), readable source augmented by clarifying comments (5%).
- **(10%)** Post-port. Analysis of the results. Compare to your thinking at the pre-port stage – did things work or not? Why or why not?

- **(10%)** Report. Combines all the described components (2%), and shows evidence of reflection on the process (6%). Incorporates draft feedback (2%).
- **(10%)** Writing quality (grammar, style).

There are not necessarily "right" or "wrong" analyses: objective (did/did not work) and numerical results are largely unambiguous, but I want to read ***your own, original, thoughts*** about what you think happened. You just need to reflect, thoughtfully and critically, on your observations and make an effort to explain how they arose, then *clearly communicate those explanations, with justification,* to your reader. This is challenging.