Data Analysis Prelim Exam

Department of Statistics, University of Chicago

Assigned Tuesday Sept 17 2024 at 12pm (noon); Due Thursday Sept 19 2024 at 12pm (noon).

The exam is "open book" but **you must not use AI tools like ChatGPT nor discuss any aspect of this exam with anyone else.** Questions of clarification should be addressed to `ignat@uchicago.edu`, `jingshuw@uchicago.edu`, `schein@uchicago.edu`, and `mcpeek@uchicago.edu`.

Remember to pace yourself throughout the exam. Take regular breaks, and ensure you eat and sleep. If time becomes limited, focus on delivering complete answers, even if it means leaving some sections unfinished, rather than offering incomplete responses for many questions.

1 Data

The data set (provided `yelp.csv` file) has been preprocessed from the Yelp dataset available at `https://www.yelp.com/dataset/documentation/main`. Descriptions of the dataset's columns are shown in Table 1 at the end of this file. The dataset is structured as a longitudinal panel, with multiple rows for each restaurant. Specifically, there is one row for each month in which a restaurant receives at least one review, and an additional row is included for the month following the last review.

2 Questions

Your report should include code for all plots and results that you present. The report should be *concise, well-organized, and easy to read.* For example, if you examine plots across a long list of different subsets of the data, you may choose to show only a small selection of representative plots.

1. Reorganize the dataset so that each row represents a unique restaurant. For each restaurant, record the following information: the earliest month the restaurant appeared on Yelp, the latest month, the duration (calculated as the difference between the earliest and latest months), the total number of reviews, the number of rating at each star and the average star rating. Additionally, retain all other restaurant-specific columns from the original data.

   (a) After reorganizing the data, visually explore it by creating histogram for the following four variables: duration, total number of reviews, average star rating and the price category, together with their pairwise scatter plots. Ensure that the visualizations are clear, well-labeled, and provide meaningful insights. Summarize the key observations from these plots.

   (b) Visually examine any differences in the four variables mentioned above across different states.

2. We seek to conduct formal inference for differences in the average star rating across states (without adjusting for any other covariates).

   (a) To carry out the analysis, we first want to compute the average star rating for each state. When averaging, we are faced with two options: weight all restaurants in each state equally or weight them proportionally to the number of reviews they received. Before conducting any computations, what is the conceptual difference and how would you explain it to a lay person? Your answer should focus on the quantities being estimated (and not on issues of statistical efficiency).

   (b) Now consider weighting all restaurants equally. Compute the per-state average and report 95% confidence intervals. Please state your assumptions and briefly explain why you deemed these reasonable. Accounting for uncertainty, can you confidently say which state has the highest average rating?

   (c) Repeat the task from (b), this time weighting restaurants proportionally to number of reviews.

   (d) For each state, compute the difference of the two averages (weighted by the total number of reviews vs equal weighting). Which weighting method results in a larger average rating for each state? Is the difference statistically significant? Can you think of a reason for the observed differences?

3. Here our task is to develop a more quantitative understanding of how the average star rating of a restaurant varies by features going beyond the states.

(a) Develop a linear model to assess how the average star rating is influenced by various restaurant features. Briefly explain your choices and summarize your findings. Which features are important? Accompany your analysis by diagnostic plots for your model (e.g., plotting residuals vs fitted values) and interpret the diagnostic plots.

(b) At dinner, two friends debate whether restaurant ratings on Yelp tend to increase in a restaurant's second month of operation as compared to the restaurant's opening month. Use the original longitudinal dataset (retaining multiple rows for each restaurant) to provide data-driven evidence for this debate.

4. Average star ratings might not tell the full story. A restaurant might be highly polarizing, with diners either hating it or loving it, and often rating it 1- or 5-stars but seldom anything in between. Generally speaking, average star ratings might hide interesting heterogeneity in the distributions of restaurants' star ratings.

This question asks you to analyze that heterogeneity using the following mixture model:

$$
\begin{aligned}
&\text{for restaurant } i = 1, \dots, n\text{:}\\
&\qquad z_i \sim \text{Categorical}(\pi_1, \dots, \pi_K)\\
&\qquad \text{for review } r = 1, \dots, R_i\text{:}\\
&\qquad\qquad w_{i,r} \sim \text{Categorical}(\phi_{z_i}^{(1)}, \dots, \phi_{z_i}^{(5)})
\end{aligned}
$$

The following terms are defined:

- There are n unique restaurants in the data set, each indexed by i.
- The data set contains different numbers of reviews for different restaurants; restaurant i has R_i reviews.
- $w_{i,r} \in \{1, \dots, 5\}$ is the star-rating assigned by the r^{th} review of restaurant i.
- $z_i \in \{1, \dots, K\}$ is the assigned mixture component of restaurant i.
- $\pi_k \in [0, 1]$ is the frequency of mixture component k where $\sum_{k=1}^{K} \pi_k = 1$.
- $\phi_k^{(x)} \in [0, 1]$ is the frequency of x-star ratings in mixture component k, where $\sum_{x=1}^{5} \phi_k^{(x)} = 1$.

For convenience, we will also assume the following non-informative prior distributions:

$$
\begin{aligned}
&(\pi_1, \dots, \pi_K) \sim \text{Dirichlet}(1.15, \dots, 1.15)\\
&\text{for mixture component } k = 1, \dots, K\text{:}\\
&\qquad (\phi_k^{(1)}, \dots, \phi_k^{(5)}) \sim \text{Dirichlet}(1.15, \dots, 1.15)
\end{aligned}
$$

(a) Derive an expectation-maximization (EM) algorithm to perform maximum-a-posteriori (MAP) estimation of the π_k and $\phi_k^{(x)}$ parameters. You should give a clear description of the full algorithm and provide a complete derivation which clearly explains and justifies each step.

(b) The data set provided to you only contains the total number of 1-, 2-, 3-, 4-, and 5-star reviews for each restaurant. Write a few sentences about why this information is sufficient to fit the given mixture model.

(c) Implement the EM algorithm in a programming language of your choice. Your implementation should be numerically stable and reasonably efficient. It should also include comments that explain the code.

(d) Design and perform a procedure to select K based on a train-test split of the data.

   i. Create a train-test split of the data by randomly assigning each review $w_{i,r}$ to a training set with probability 0.7 (and to the test set with probability 0.3).

   ii. Run your EM algorithm on the training set for values of $K \in \{5, 8, 12, 15, 20, 25, 30, 40\}$.

   iii. Design a principled way to measure how well each of the models fits the test set. Describe and justify your approach, and use it to make a plot with K on the x-axis and test-set performance on the y-axis.

   iv. Based on this plot, choose the smallest value of K that seems to fit the data well enough.

(e) With your selected value of K, run your EM algorithm on the full data set, and perform an exploratory analysis of the inferred latent structure. Your analysis should highlight certain components, and visualize patterns in the data that the model seems to uncover. You may also choose to incorporate other covariates, for instance to explore what kinds of restaurants are assigned to which clusters. (This question is intentionally left open-ended; use your creativity.)

5. Restaurants may be interested in understanding how their popularity is influenced by various features. Using Yelp data, we can approximate popularity by measuring the frequency of reviews, defined as the average number of reviews per month. Please conduct an appropriate regression analysis to determine which restaurant features are significantly associated with their popularity. Based on your findings, provide insights on how restaurants might leverage these features to enhance their popularity. Additionally, discuss any potential limitations of your analysis.

Column Name	Description
business_id	Unique identifier for each restaurant
name	Name of the restaurant
city	City where the restaurant is located (the dataset was filtered to only include cities with at least 900 restaurants in the dataset)
state	State where the restaurant is located
postal_code	Zip code of the restaurant location
population	Population of the zip code area
population_density	Population density of the zip code area (persons per square mile)
median_household_income	Median household income within the zip code area
latitude	Latitude coordinate of the business
longitude	Longitude coordinate of the business
price_category	Price range category of the business (i.e., $, $$, $$$, $$$$)
outdoor_seating	Whether the restaurant has outdoor seating
delivery	Whether the restaurant offers delivery
good_for_dessert	Whether the restaurant is good for dessert
good_for_latenight	Whether the restaurant is good for late night dining
good_for_lunch	Whether the restaurant is good for lunch
good_for_dinner	Whether the restaurant is good for dinner
good_for_brunch	Whether the restaurant is good for brunch
good_for_breakfast	Whether the restaurant is good for breakfast
year_month	Year and month of the review data
monthly_reviews	Number of reviews throughout the month
monthly_stars	Average star rating throughout the month (each user can give a rating of 1,2, 3, 4, 5)
monthly_stars_1	Number of 1-star ratings received throughout the month
monthly_stars_2	Number of 2-star ratings received throughout the month
monthly_stars_3	Number of 3-star ratings received throughout the month
monthly_stars_4	Number of 4-star ratings received throughout the month
monthly_stars_5	Number of 5-star ratings received throughout the month
total_reviews	Total number of all past reviews up to the end of the current month
total_stars	Average star rating of all past reviews up to the end of the current month
total_stars_1	Cumulative number of 1-star ratings up to the end of the current month
total_stars_2	Cumulative number of 2-star ratings up to the end of the current month
total_stars_3	Cumulative number of 3-star ratings up to the end of the current month
total_stars_4	Cumulative number of 4-star ratings up to the end of the current month
total_stars_5	Cumulative number of 5-star ratings up to the end of the current month
is_last_month	Whether this is the month after the restaurant received the last review in the dataset
is_open	Whether the restaurant is open in the given month (this indicator is crowd-sourced by Yelp)

Table 1: Description of variables in each column