Final Project Report

Jiaxing Qiu, Mark Tenzer, Shenghua Wu

1. Problem description

Our project was to use Bayesian Machine Learning to find predictive relationships between recipes and customer reviews and mine interesting insights from recipes. We aimed to answer three specific questions: 1. Classify customer reviews to two categories (high or low), based on minutes to cook, number of steps, number of ingredients, calories and six nutrition facts (total fat, sugar, sodium, protein, saturated_fat, carbohydrates) of a recipe. 2. Predict calories of a recipe, using these six nutrition facts. 3. Predict how many minutes a recipe takes to prepare, using the number of steps, number of ingredients, calories and six nutrition facts. Such insights could predict fundamental recipe properties, such as calories and cooking/preparation time, and understand users' behavior and preferences. This knowledge could be used to encourage healthy dietary choices.

2. Feature Engineering

With the two raw datasets consisting of 180K+ recipes and 700K+ recipe reviews from Kaggle, we selected features related to nutritional information, customers' ratings and cooking complexity. For computational efficiency, we subset the original datasets by randomly sampling 100,000 observations to our training dataset, on top of dropping all of the missing values and meaningless rows. For question 1, to predict the reviews' rating of the recipes, we merged these two datasets by 'recipe_id'. On the new merged dataset, we re-binned the response variable 'rating' into two levels: high (5) and low (lower than 5). For predictors, we used the nutrition facts of recipes, number of steps, number of ingredients and minutes to cook, and split the nutrition facts (originally formed as a list) into separate columns.

For question 2 and 3, feature engineering only was concerned with the raw dataset of recipes, which was processed the same as in question 1. Before each prediction, we standardized the response variable and predictors by subtracting the mean and dividing by the standard deviation.

3. Bayesian Analysis

GENERAL APPROACH

When selecting methods for this project, we required approaches that were flexible enough to handle both classification (question 1) and regression (questions 2 and 3) problems, and that would be able to accommodate a broad range of uninformative and informative priors. The obvious candidates for such an approach would be Markov Chain Monte Carlo (MCMC), to directly sample the relevant distributions, and variational inference (to approximate them). However, the dataset involved in this project was extremely large in size, and MCMC was found to struggle with the size of these data. Although approximation with variational inference might not fit the data as precisely as Hamiltonian Monte Carlo-based approaches, such as the No U-Turn Sampler (NUTS), the questions presented in this report represent a case where variational inference's weaknesses are of less concern. In particular, variational approximations are typically unimodal, regardless of the underlying distributions, but all continuous prior distributions in this report were Gaussian or Cauchy, which are in fact unimodal.

QUESTION 1

For this problem, we compared two models inspired by logistic regression. The first was a relatively simple model (Figure 1), with 11 coefficients **β** (one intercept, ten predictors) sampled from a univariate unit Normal (mean 0, variance 1) distribution. Their dot product with the observed features formed the logit of the parameter *p* for the outcome, a Bernoulli random variable (whether the review was for 5 stars, or not). For the second model, hierarchical priors were placed on both the mean μ_θ and standard deviation σ_θ of the coefficients' Normal. The mean was itself a unit Normal, while the standard deviation was estimated form a Half-Cauchy distribution, i.e., the positive half of a Cauchy distribution (Figure 1).

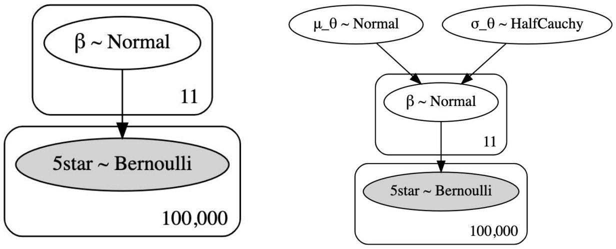


Figure 1. Model architectures for the question 1. Left: a simple logistic regression. Right: the same model, with hierarchical priors added.

QUESTION 2

For this problem, we compared five models with varying priors. For this first approach, we attempted a simple linear regression (Figure 2.1), with multivariate Normal coefficients **β**, a Half-Cauchy error term σ_η, and a Normal output variable for the calorie count. For the second approach, we replicated the same architecture but altered the parameters for **β** to use the known relationship between calories and nutritional information: 4 calories per gram of protein, plus 4 per gram of carbohydrates, plus 9 per gram of fat (and 0 for all other predictors). (Note that sugars are already included in carbohydrates, and saturated fat in total fat.) For our latter three approaches, we used hierarchical modeling. Initially (our third approach), we placed priors on μ_θ and σ_θ exactly as in question 1, Normal and Half-Cauchy (Figure 2.1). However, this disregarded our subjective prior, by using one prior for all predictors. We therefore attempted two more hierarchical models (Figure 2.2), one with a hierarchical μ_θ only for the three "true" predictors (protein, carbohydrates, total fat), and one where a Multivariate Normal prior created an independent prior for each β-coefficient.

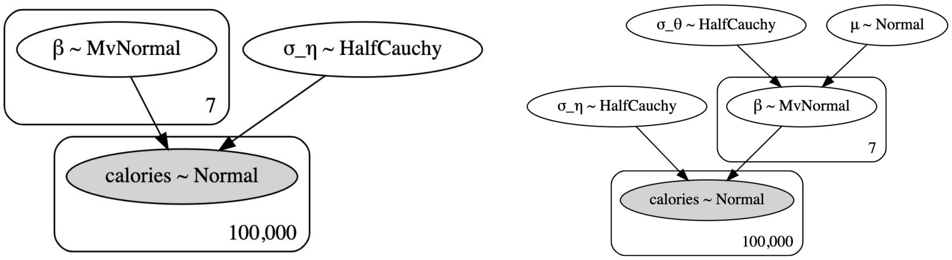


Figure 2.1. Model architectures for question 2, adapted from question 1. Left: architecture used for the first two models, simple linear regression (centered or not centered on 0 for all predictors). Right: basic hierarchical model, as in question 1.

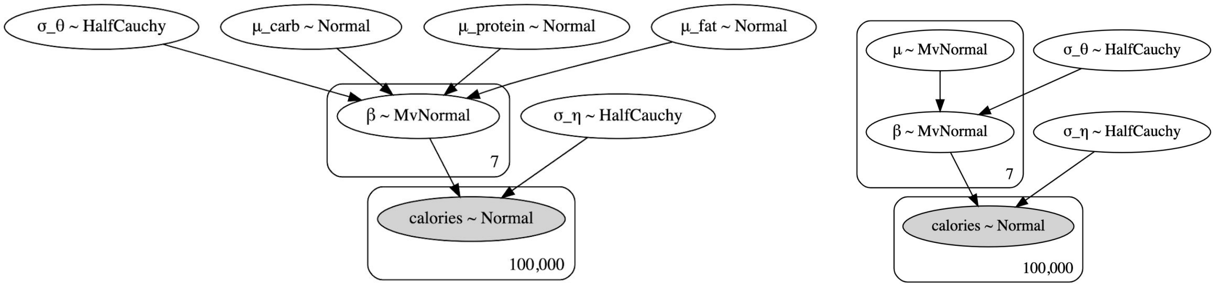


Figure 2.2. Unique model architectures for question 2. Left: Separate hierarchical priors for μ_θ for each of the three "true" predictors only. Right: separate hierarchical priors for μ_θ for each predictor.

QUESTION 3

For this problem, we compared four models to predict the length of time a recipe requires to complete. The first two approaches followed the simple linear regression approach from questions 1 and 2 (Figure 3.1), with multivariate Normal coefficients **β**, a Half-Cauchy error term σ_η, and a Normal output variable. The final two models adopted a hierarchical approach (Figure 3.2), placing priors on both the mean μ_θ and standard deviation σ_θ of the coefficients' Normal (as in both questions 1 and 2). For each approach, one model was trained to predict minutes, and one was trained to predict the *log* of minutes, based on our observation that minutes' true distribution appeared more similar to a log-Normal than a Normal.

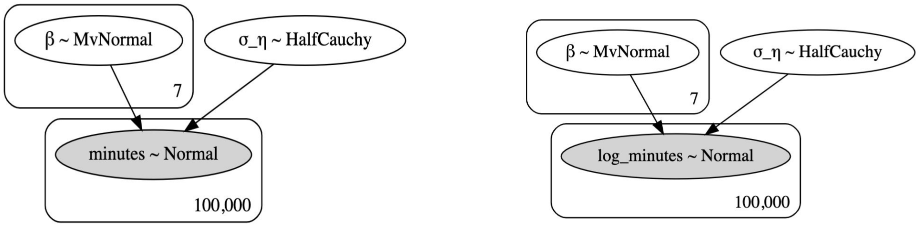


Figure 3.1. Simple Linear Regression model architectures for question 3. Left: original response. Right: log-transformed.

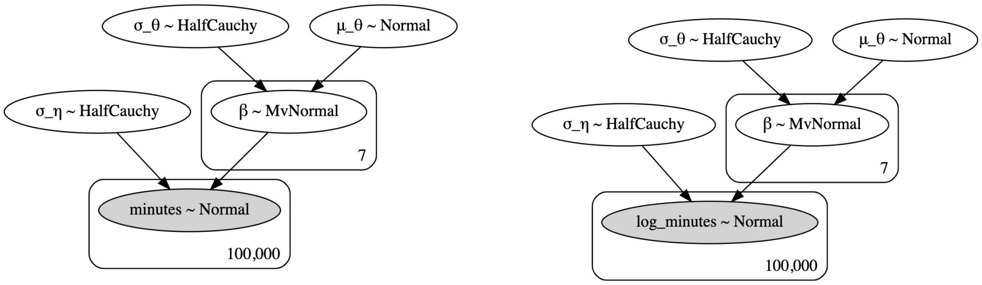


Figure 3.2. Linear Regression with Hierarchical Priors for question 3. Left: original response. Right: log-transformed.

MODEL SELECTION WITH WAIC

For each question, models were compared using the Widely-Applicable Information Criterion (WAIC). The model that minimizes the WAIC is estimated to be the superior model. In addition, the same Python package also calculates a *weight* for each model, which the documentation indicates is the estimated probability that the given model is the best (of the models tested).

HIERARCHICAL MODELING AND RIDGE REGRESSION

Originally, this project considered using ridge regression to predict the quantitative response variables (calories and cooking duration, questions 2 and 3). However, a similar procedure is already performed by the hierarchical models. It can be shown that the regularization parameter in ridge regression, λ, is the quotient of two prior variances: σ_η^2 divided by σ_θ^2. Both of these terms are included in our graphical models, using a similar model architecture. Unlike traditional ridge regression, which typically selects a single λ by cross-validation, the variational inference approach estimates distributions for both variances, providing a *distribution* of λ.

4. Results and conclusions

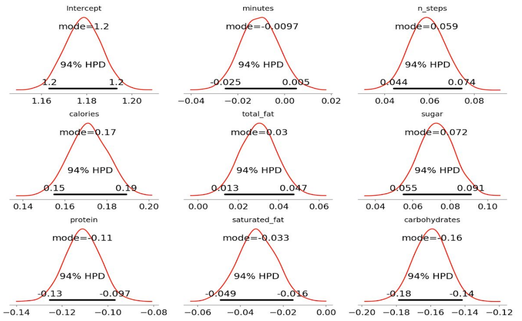
MODEL SELECTION WITH WAIC

Of the two models considered for question 1 (classifying high reviews), both models have near-identical WAIC values, to at least six significant digits. No difference can be determined by consulting the table, because the differences are too close for rounding. Using the weight column, we select the simpler model (Figure 1, left) for generating predictions, but only very slightly: the probability that this is the better model, is only roughly 52 percent (to the hierarchical model's 48 percent). We do not have a strong degree of belief that this model is superior.

For question 2, the hierarchical model with seven subjective priors (Figure 3.2, right), one for each predictor, strongly outperforms all other models. Its WAIC is much lower than the others', and using its weight, we are about 88 percent confident that it is the best of the five models.

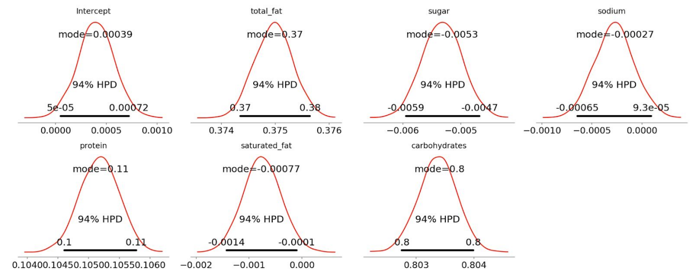
For question 3, both models using log-transformed response variables are strongly preferred over the untransformed models: their WAIC values are much smaller. The simple model with a log-transformed response (Figure 3.1, right) is WAIC's overall preference, with a WAIC dramatically lower than all other models. Using its weight, we are about 98 percent confident that it is the best of the four models.

ANALYSIS AND INTERPRETATION OF COEFFICIENTS



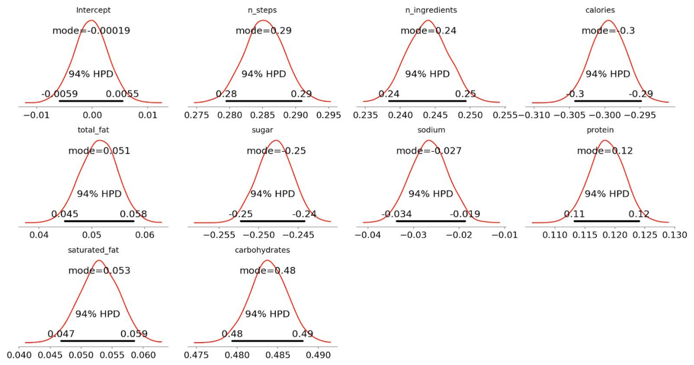
QUESTION 1

Based on the results and their credibility intervals, we are more than 94% confident that a recipe's number of recipe steps, calories, total fat, sugar, and sodium values are linked to a higher probability of a 5-star review, while number of ingredients, protein, saturated fat, and carbohydrates values could have negative effects on a customer's review. The model is uncertain whether cooking/preparation time has any effect; it appears that this coefficient may be zero.



QUESTION 2

The model has a very strong degree of belief (greater than 94 percent) that total fat, protein, and carbohydrates increase calories, matching the known relationship. The model is also very confident that sugar (for a fixed amount of carbs) and saturated fat (for a fixed amount of total fat) vary negatively with calories. The model is uncertain whether sodium has any effect; it appears (correctly) that this coefficient may be zero.



QUESTION 3

Based on the 94% credible intervals, the model has a very strong degree of belief that more steps and more ingredients increase the cooking/preparation time, which matches our expectations. All other factors held equal, the model is confident that cooking/preparation time tends to be higher for higher total fat, protein, saturated fat, and carbohydrates, and lower calories, sugar, and sodium. We have a strong belief that none of the coefficients are zero.

HIERARCHICAL MODELING AND RIDGE REGRESSION

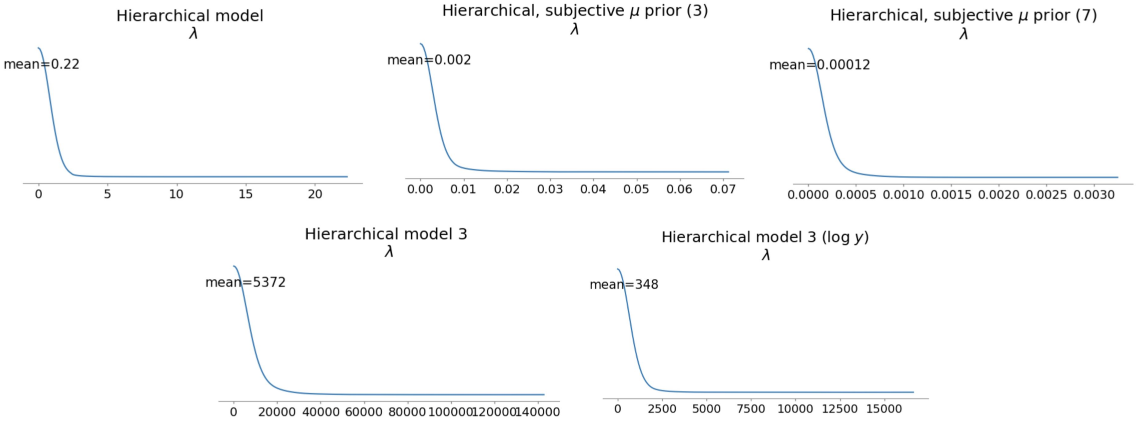


Figure 5. Choosing a better-specified model, either by using a good prior (question 2, calories, top row) or by choosing the correct form for the response variable (question 3, cooking/preparation time, bottom row), seems to decrease the amount of regularization needed. Since λ decreases, the error term σ_η^2 decreases relative to the prior variance σ_θ^2 when the model is correctly specified.

CONCLUSION

Using variational inference, we have developed 11 potential models to estimate probabilities of five-star reviews, calorie counts, and cooking/preparation time from 100,000 recipes. We have also compared non-hierarchical and hierarchical models, as well as a range of different subjective and non-subjective priors, to select the best approach for each problem. The results explain the probability and uncertainty at each stage of the process, quantifying the uncertainty of selecting each model, of the coefficient values for each model, and of the amount of regularization (λ). The results could yield powerful insights on recipes' intrinsic properties and users' food preferences and behavior, which could produce impressive benefits for public health.

CITATIONS:

Dataset: https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions