Cognitive Modeling: Homework Assignment 3
Discrete Models and Model Criticism

Caleb Carr & Maryellen Marino

March 28, 2024

Team GitHub: https://github.com/maryellenmarino/Cognitive_ModelingS24

All answers and solutions to non-programming questions should be submitted to LMS as a legible write-up (either fully digital or a scan). All code should be committed to and merged into the main branch of your team's GitHub repository.

Problem 1: True-False Questions (6 points)

Mark all statements which are **FALSE**

1. Direct K-fold cross-validation requires K model re-fits, which may be computationally demanding, especially when inverse inference is costly.

   True. In K-fold cross-validation, each subset of the data is used exactly once as the test set, while the remaining data is used for training, leading to K separate model fittings. This process can be computationally intensive, particularly for complex models or large datasets.

2. Bayes factors (BFs) are *relative measures*, that is, they cannot differentiate between "equally good" and "equally bad" models.

   True. Bayes factors provide a comparison of the evidence for two models but do not indicate absolute goodness of fit. They can indicate whether one model is more supported by the data than another but cannot distinguish if both models are equally valid or equally poor.

3. Marginal likelihoods and, by extension, Bayes factors (BFs) cannot be used to compare models with different likelihoods.

   False. Marginal likelihoods and Bayes factors are specifically used for comparing models with different likelihoods. They integrate out the parameters of each model to allow for a comparison based on the data.

4. Both the Binomial and the Dirichlet distribution can be formulated as special cases of the Multinomial distribution.

False. The Binomial distribution is indeed a special case of the Multinomial distribution for two outcomes. However, the Dirichlet distribution is a separate entity, serving as a prior for Multinomial distributions in Bayesian analysis, but it is not a special case of the Multinomial distribution itself.

5. Bayesian leave-one-out cross-validation (LOO-CV) relies on the posterior predictive distribution of left-out data points.

   True. Bayesian LOO-CV involves leaving out one observation at a time and predicting it using a model fitted to the remaining data. This process heavily relies on the posterior predictive distribution for the left-out data point.

6. The Akaike Information Criterion (AIC) penalizes model complexity indirectly through the variance of a model's marginal likelihood.

   False. The AIC penalizes model complexity by considering the number of parameters. The formula for AIC is $2k - 2\ln(\hat{L})$, where k is the number of parameters and $\hat{L}$ is the maximum likelihood of the model. It does not involve the variance of the model's marginal likelihood.

7. The log-predictive density (LPD) is a relative metric of model complexity.

   False. The LPD is a measure of a model's predictive accuracy, not its complexity. It assesses how well a model predicts new data, usually via the average log-likelihood of observed data under the model.

8. The LPD can be approximated by evaluating the likelihood of each posterior draw (e.g., as provided by an MCMC sampler) and taking the average of all resulting likelihood values.

   True. The LPD for a set of observed data can be approximated in a Bayesian framework by averaging the log likelihoods of these data across samples from the posterior distribution.

9. Bayes factors do not depend on the prior odds, that is, the ratio of prior model probabilities $p(\mathcal{M}_1)/p(\mathcal{M}_2)$.

   True. Bayes factors compare the likelihood of the data under two different models, independent of the prior odds of the models. They measure the evidence provided by the data in favor of one model over another.

10. You should always prefer information criteria to cross-validation in terms of estimation predictive performance.

    False. The choice between information criteria and cross-validation depends on the specific context and the nature of the data and models. Each method has its advantages and disadvantages, and one is not universally superior to the other in all situations.

Problem 2: Simple Multinomial Processing Trees (MPTs) (10 points)

Collect some data (e.g., from a friend or your teammate) on the recognition memory task from the slides (or construct your own task) and fit the following two models using `Stan`:

- The One-High-Threshold Model (1HT)
- The Two-High-Threshold Model (2HT)

The models are depcited on Slide 12 (MPT models). As usual, inspect the convergence of the MCMC samplers and report the estimation results. Do the two models suggest different estimates for the two key parameters? Describe and interpret the results.

Solution is present here: https://github.com/maryellenmarino/Cognitive_ModelingS24/blob/main/hw3/MPT_Model_Comparison.ipynb

Problem 3: A More Complicated MPT Model (10 points)

Write down the model equations for the MPT model depicted in **Figure 1** in the paper (avaliable on LMS):

- Walk1er, G. M., Hickok, G., & Fridriksson, J. (2018). A cognitive psychometric model for assessment of picture naming abilities in aphasia. Psychological assessment, 30(6), 809.

Then, write a Stan program for the MPT model featuring the following five blocks: `data` – for passing the hypothetical categorical data; `parameters` – for defining the latent model parameters; `transformed parameters` – for transforming latent model parameters into probabilities; `model` – for formulating the Bayesian joint model; `generated quantities` – for sampling new frequency data given the posterior draws (generative performance).
Bonus (6 points): Simulate a data set according to the forward model and inspect parameter recovery.

Solution is present here: https://github.com/maryellenmarino/Cognitive_ModelingS24/blob/main/hw3/aphasia_mpt.ipynb

Problem 4: A Discrete Conjugate Model (6 points)

Derive the analytic posterior for the conjugate Dirichlet-Multinomial model (no ChatGPT):

$$\theta \sim \text{Dirichlet}(\alpha) \tag{1}$$

$$y \sim \text{Multinomial}(y|\theta; N) \tag{2}$$

$$\text{posterior} \propto \text{prior} \times \text{likelihood} \tag{3}$$

$$\text{posterior} \propto \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{a_k - 1} \times \frac{N!}{y_1!...y_K!} \prod_{k=1}^{K} \theta_k^{y_k} \tag{4}$$

$$\text{posterior} \propto \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \theta_k^{y_k} \times \frac{N!}{y_1!...y_K!} \tag{5}$$

$$\text{posterior} \propto \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1 + y_k} \times \frac{N!}{y_1!...y_K!} \tag{6}$$

$$\tag{7}$$

Drop the constant $\frac{1}{B(\alpha)}$

$$\prod_{k=1}^{K} \theta_k^{\alpha_k - 1 + y_k} \times \frac{N!}{y_1!...y_K!} \tag{8}$$

$$\tag{9}$$

Drop the constant $\frac{N!}{y_1!...y_K!}$

$$\prod_{k=1}^{K} \theta_k^{\alpha_k - 1 + y_k} \tag{10}$$

$$\tag{11}$$

This itself is another Dirichlet distribution Posterior $\propto$ Dirichlet$(\alpha + y)$

Problem 5: Multiple Regression (8 points)

Extend your simple Bayesian regression model from the previous exercise into a multiple regression model:

$$\sigma \sim \text{Inv-Gamma}(\tau_0, \tau_1) \tag{12}$$

$$\alpha \sim \text{Normal}(0, \sigma_\alpha) \tag{13}$$

$$\beta \sim \text{Multivariate-Normal}(0, \sigma_\beta) \tag{14}$$

$$y_n \sim \text{Normal}(\alpha + \beta x_n^T, \sigma) \quad \text{for} \quad n = 1, ..., N \tag{15}$$

where you need to set the hyperparameters of the prior (i.e., τ_0, τ_1, σ_β, σ_α) to some reasonable values. Next, use your Stan program to fit a Bayesian multiple regression model for the Insurance Costs data set: https://github.com/stedy/Machine-Learning-with-R-datasets/blob/2master/insurance.csv.

Your goal is to predict the insurance charges (`charges`) from a patient's BMI (`bmi`), age (`age`), and number of children (`children`). Thus, you need to estimate three regression weights (β_1, β_2, and β_3), along with the intercept (α), and the noise parameter (σ). It is also recommended that you standardize your predictors (i.e., subtract the means from the input variables and divide by their standard deviations) in order to bring them to a common scale. Split the data into a training set and a test set and fit the model only to the training set. Perform the usual convergence checks and describe your results. Which of the three variables is the best predictor of Insurance Charges?

Alternative: Use the Bayesian Ridge regression implementation from scikit-learn:https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.BayesianRidge.html, also for the next task. If you go this path, include a small description on how the Bayesian ridge differs from the model implementation suggested above

Solution is present here: https://github.com/maryellenmarino/Cognitive_ModelingS24/blob/main/hw3/problem5/problem5.ipynb

Problem 6: Predictive Distribution (5 points)

Use the `generated quantities` block in the `Stan` program to also pass the test data and sample from the predictive distribution. Extract the samples from the predictive distribution, compute the means predictive means from the samples, calculate the root-mean-squared error (RMSE) between the predictive means and the actual charges in the test set:

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{y}_m - y_m)^2}, \tag{16}$$

where M denotes the number of test instances and $\hat{y}_m$ denotes the predictive means. How good are your predictions? What information did you lose by computing the predictive means? How could you possibly propagate the uncertainty information encoded in the predictive distribution to obtain a distribution over the test RMSE values?

Solution is present here: https://github.com/maryellenmarino/Cognitive_ModelingS24/blob/main/hw3/problem6.ipynb