Biostat 682 Homework 4

Due: Wednesday, November 26, 2025 (23:59 pm)

Please use JAGS and/or PyMC to complete this homework.

1. Revisit the `UScrime.csv` data from Canvas. The dataset contains crime rates (y) and data on 15 explanatory variables for 47 U.S. states.

   (a) Fit a Bayesian neural networks with one hidden layer with $q \in \{2, 3, 4, 5, 6\}$ hidden units using spike and slab priors. Compare the DIC and choose

   (b) To test how well neural networks models can predict crime rates based on the explanatory variables, randomly divide the data roughly in half, into training set and a test set. Use the training dataset to fit the model and generate the posterior predictive median of the crime rates given the explanatory variables in the test dataset. Compare the posterior predictive median and the actual crime rate in the test dataset for different models in part (a).

   (c) Compare the prediction performance of Bayesian neural networks and Bayesian linear regression with Spike and Slab priors in homework 3.

2. On Canvas, you are provided with two CSV files, `spam_train.csv` and `spam_test_0.csv`, derived from a dataset of 4601 emails. A description of the variables is provided in Table 1.

Table 1: Description of variables in the spam dataset

Variable	**Description**
`crl.tot`	Total length of words written in capital letters
`dollar`	Number of occurrences of the "$" symbol
`money`	Number of occurrences of the word "money"
`n000`	Number of occurrences of the string "000"
`make`	Number of occurrences of the word "make"
`yesno`	Outcome variable: "y" = spam, "n" = not spam (training data only)

The two provided files are structured as follows:

- `spam_train.csv`: complete data (predictors + outcome) for 2500 emails, of which 997 are spam.
- `spam_test_0.csv`: predictor variables only, for the remaining emails. The outcome variable `yesno` has been removed.

Your task is to use Bayesian methods to estimate the probability that each email in the test set is spam.

(a) Using `spam_train.csv`, fit a Bayesian classification model. You may choose one of the following:

- Bayesian logistic regression,
- Bayesian probit regression,
- Bayesian neural network

Choose priors appropriate for your model (e.g., weakly informative, shrinkage, or spike-and-slab) and briefly describe your prior choices.

(b) Using your fitted model, compute the posterior predictive probability

$$\Pr(\text{spam} \mid \text{predictors})$$

for each email in `spam_test_0.csv`.

Prepare a CSV file containing *one column* with the posterior predictive spam probability. The order of the probabilities **must match the row order in** `spam_test_0.csv`.

(c) Submit the following:

- A short description of your Bayesian model and priors.
- Your source code to fit the model
- A CSV file containing the posterior predictive probabilities for all emails in `spam_test_0.csv`.

We will evaluate predictive accuracy using the withheld true labels.