Homework 3: Bayesian Analysis

Harvard CS 109B, Spring 2018

Feb 20, 2018

Homework 3 is due March 5, 2018 11:59 PM

LDA & Bayes

In the first part of this assignment, you will be working with text from @realDonaldTrump Twitter. The text was taken from all tweets Donald Trump sent between 01/19/2016 and 01/19/2018. The goal is to use Latent Dirichlet Allocation in order to model the topics that the president tweeted about during this time.

In the second part of this assignment, you are provided with data sets *dataset-2-train.txt* and *dataset-2-test.txt* containing details of contraceptive usage by 1934 Bangladeshi women. There are four attributes for each woman, along with a label indicating if she uses contraceptives. The attributes include:

- district: identifying code for the district the woman lives in
- urban: type of region of residence
- living.children: number of living children
- age-mean: age of women (in years, centred around mean)

The women are grouped into 60 districts. The task is to build a classification model that can predict if a given woman uses contraceptives.

1. Data Preparation

The tweet data is provided for you as `trump-tibble.csv`. After you read the data into R, you'll see that there are only two columns: document and text. The *document* column contains the date and time of the tweet, and the *text* column contains the actual text of the tweet. Before you begin, you'll want to cast the columns as characters rather than factors. You can do this with the following code:

```
# cast factors to characters
trump_tibble$document <- as.character(trump_tibble$document)
trump_tibble$text <- as.character(trump_tibble$text)
```

The following libraries will be of use for this problem:

```
# load libraries
library(topicmodels) #topic modeling functions
library(stringr) #common string functions
library(tidytext) #tidy text analysis
suppressMessages(library(tidyverse)) #data manipulation and visualization
#messages give R markdown compile error so we need to suppress it

## Source topicmodels2LDAvis & optimal_k functions
invisible(lapply(file.path("https://raw.githubusercontent.com/trinker/topicmodels_learning/
master/functions",
c("topicmodels2LDAvis.R", "optimal_k.R")),
devtools::source_url))
```

(a) Use the `unnest-tokens` function to extract words from the tweets text

(b) Create a dataframe consisting of the document-word counts

(c) Create a document-term matrix using the `cast-dtm` function

2. LDA

(a) Using the following control parameters, run the optimal-k function to search for the optimal number of topics. Be sure to set the "max.k" parameter equal to 30. `control <- list(burnin = 500, iter = 1000, keep = 100, seed = 46)`

(b) Plot the results of the optimal-k function. What does this plot suggest about the number of topics in the text?

(c) Run LDA on the document-term matrix using the optimal value of k. Print out the top 10 words for each of the k topics. Comment on the results and their plausibility.

3. Bayesian Logistic Regression

The first model we will fit to the contraceptives data is a varying-intercept logistic regression model, where the intercept varies by district.

Prior distribution:

$\beta_{0j} \sim N(\mu_0, \sigma_0)$, with $\mu_0 \sim N(0, 100)$ and $\sigma_0 \sim \text{Exponential}(.1)$

$\beta_1 \sim N(0, 100)$, $\beta_2 \sim N(0, 100)$, $\beta_3 \sim N(0, 100)$

Model for data:

$Y_{ij} \sim Bernoulli(p_{ij})$

$logit\ p_{ij} = \beta_{0j} + \beta_1 * urban + \beta_2 * \text{living-children} + \beta_3 * \text{age-mean}$

where Y_{ij} is 1 if woman i in district j uses contraception, and 0 otherwise, and where $i = 1, \ldots, N$ and $j = 1, \ldots, J$ (N is the number of observations in the data, and J is the number of districts). The above notation assumes $N(\mu, \sigma)$ is a normal distribution with mean μ and standard deviation σ. Also, the above notation assumes Exponential(λ) has mean $1/\lambda$. These are consistent with the parameterizations in Stan.

After you read the train and test data into R, the following code will help with formatting:

```
# convert everything to numeric
for (i in 1:ncol(train)) {
  train[,i] <- as.numeric(as.character(train[,i]))
  test[,i] <- as.numeric(as.character(test[,i]))
}

# map district 61 to 54 (so that districts are in order)
train_bad_indices <- which(train$district == 61)
train[train_bad_indices, 1] <- 54
test_bad_indices <- which(test$district == 61)
test[test_bad_indices, 1] <- 54
```

(a) To verify the procedure, simulate binary response data (using the `rbinom` function) assuming the following parameter values (and using the existing features and district information from the training data):

```
mu_beta_0 = 2
sigma_beta_0 = 1
set.seed(123)  # to ensure the next line is common to everyone
beta_0 = rnorm(n=60, mean=mu_beta_0, sd=sigma_beta_0)
beta_1 = 4
```

```
beta_2 = -3
beta_3 = -2
```

(b) Fit the varying-intercept model specified above to your simulated data

(c) Plot the trace plots of the MCMC sampler for the parameters $\mu_{\beta_0}, \sigma_{\beta_0}, \beta_1, \beta_2, \beta_3$. Does it look like the samplers converged?

(d) Plot histograms of the posterior distributions for the parameters $\beta_{0,10}, \beta_{0,20}....\beta_{0,60}$. Are the actual parameters that you generated contained within these posterior distributions?

We now fit our model to the actual data.

(e) Fit the varying-intercept model to the real train data. Make sure to set a seed at 46 within the Stan function, to ensure that you will get the same results if you fit your model correctly.

(f) Check the convergence by examining the trace plots, as you did with the simulated data.

(g) Based on the posterior means, women belonging to which district are most likely to use contraceptives? Women belonging to which district are least likely to use contraceptives?

(h) What are the posterior means of μ_{β_0} and σ_{β_0}? Do these values offer any evidence in support of or against the varying-intercept model?

4. Varying-Coefficients Model

In the next model we will fit to the contraceptives data is a varying-coefficients logistic regression model, where the coefficient on living-children varies by district:

Prior distribution:

$\beta_{0j} \sim N(\mu_0, \sigma_0)$, with $\mu_0 \sim N(0, 100)$ and $\sigma_0 \sim \text{Exponential}(0.1)$

$\beta_{1j} \sim N(0, \sigma_1)$, with $\sigma_1 \sim \text{Exponential}(0.1)$

$\beta_{2j} \sim N(0, \sigma_2)$, with $\sigma_2 \sim \text{Exponential}(0.1)$

$\beta_{3j} \sim N(0, \sigma_3)$, with $\sigma_3 \sim \text{Exponential}(0.1)$

Model for data:

$Y_{ij} \sim \text{Bernoulli}(p_{ij})$

$\text{logit } p_{ij} = \beta_{0j} + \beta_{1j}\text{urban} + \beta_{2j}\text{age-mean} + \beta_{3j}\text{living-children}$

where $i = 1, \ldots, N$ and $j = 1, \ldots, J$ (N is the number of observations in the data, and J is the number of districts).

(a) Fit the model to the real data. For each of the three coefficients to the predictors, plot vertical segments corresponding to the 95% central posterior intervals for the coefficient within each district. Thus you should have 54 parallel segments on each graph. If the segments are overlapping on the vertical scale, then the model fit suggests that the coefficient does not differ by district. What do you conclude from these graphs?

(b) Use all of the information you've gleaned thus far to build a final Bayesian logistic regression classifier on the train set. Then, use your model to make predictions on the test set. Report your model's classification percentage.

5. "Bayesball" (AC 209b students only)

In Major League Baseball (MLB), each team plays 162 games in the regular season. The data in `Bayesball.txt` contains information about every regular season game in 2017. The data can be read in by the command

```
games2017 = read.table("Bayesball.txt", sep=',')
```

The relevant columns of the data are

```
data2017 = games2017[,c(7,4,11,10)]
names(data_2017) = c("Home","Away","Home_Score","Away_Score")
```

Because we are going to focus on wins versus losses, you will need to convert these data into binary outcomes.

Under the Bradley-Terry model, each team i is assumed to have some underlying talent parameter π_i. The model states that the probability that team i defeats opponent j in any game is:

$$\Pr(\text{team } i \text{ defeats team } j) = \frac{\pi_i}{\pi_i + \pi_j}.$$

where $i, j \in (1, 2, ...30)$, since there are 30 teams in MLB. The parameter π_i is team i's "strength" parameter, and is required to be positive.

If we let V_{ij} be the number of times in a season that team i defeats team j, and n_{ij} to be the number of games between them, an entire season of MLB can be described with the following density:

$$p(V \mid \pi) = \prod_{i=1}^{N-1} \prod_{j=i+1}^{N} \binom{n_{ij}}{V_{ij}} (\frac{\pi_i}{\pi_i + \pi_j})^{V_{ij}} (\frac{\pi_j}{\pi_i + \pi_j})^{V_{ji}}$$

Team i's victories against team j follows a binomial distribution governed by the Bradley-Terry probability with the given strength parameters.

Rather than work with the π_i, we will transform the model by letting $\lambda_i = \log \pi_i$ for all i. Thus the probability i defeats j is

$$\Pr(\text{team } i \text{ defeats team } j) = \frac{\exp(\lambda_i)}{\exp(\lambda_i) + \exp(\lambda_j)}.$$

The advantage of this parameterization is that the λ_i are unconstrained real-valued parameters.

We now carry out a Bayesian analysis of the Bradley-Terry model. We will assume a hierarchical normal prior distribution on the λ_i, that is

$$\lambda_i \sim N(0, \sigma).$$

for all $i = 1, \ldots, N$. We will also assume that the standard deviation σ has a uniform prior distribution,

$$\sigma \sim \text{Uniform}(0, 50)$$

with a maximum value of 50.

Thus the full model is:

Prior distribution:

$\lambda_i \mid \sigma \sim N(0, \sigma)$, with $\sigma \sim \text{Uniform}(0, 50)$

Model for data: $V_{ij} \mid \lambda_i, \lambda_j, n_{ij} \sim \text{Binomial}\left(n_{ij}, \frac{\exp(\lambda_i)}{\exp(\lambda_i)+\exp(\lambda_j)}\right)$

(a) Why does this prior distribution on the λ_i and σ make sense? Briefly explain.

(b) Implement the model in Stan.

(c) Report the posterior means for each team's exponentiated strength parameters (that is, $\exp(\lambda_i)$).

(d) Using the posterior predictive distribution for the strengths of the Dodgers and Astros, simulate 1000 recreations of the 2017 World Series. That is, simulate 1000 separate series between the teams, where the series ends when either team gets to 4 wins. Based on your simulation, what was the probability that the Astros would have won the World Series last year?