STAT 578 – Advanced Bayesian Modeling – Fall 2018

Assignment 1

1. The film review aggregator website `rottentomatoes.com` publishes ranked lists of movies based on the number of positive critical reviews out of a total number counted for each movie. See, for example, `https://www.rottentomatoes.com/top/bestofrt/?year=2017`. Because the site uses an "Adjusted Score," a movie with a higher approval percentage sometimes ranks lower on the list.

   Consider the following hypothetical scenario:

   Movie 1: 150 positive reviews out of 200 (75%)
   Movie 2: 4 positive reviews out of 5 (80%)

   Assume that reviews of a given movie are independent with a common probability p of being positive (depending on the movie). Assume a uniform prior on p.

   (a) [4 pts] Determine the posterior distribution of p for Movie 1 and for Movie 2 (separately).

   (b) [3 pts] Which movie ranks higher according to posterior mean? According to posterior median? According to posterior mode? Show your computations. (For median, use R function `qbeta`. For mean and mode, use formulas in BDA3, Table A.1.)

2. File `randomwikipedia.txt` contains the ID numbers and number of `bytes` in length for 20 randomly selected English Wikipedia articles.

   (a) (i) [2 pts] Draw a histogram of article length, and describe it.
   (ii) [2 pts] Transform article length to the log scale. Then re-draw the histogram and describe it.
   (iii) [2 pts] Based on parts (i) and (ii), which scale would be better to use for the remainder of the analysis? (Read below.) Explain.

   (b) [2 pts] Based on your decision in the previous part, let y_i be length of article i on the scale you chose (original or log). Compute the sample mean and sample standard deviation.

   Now assume each y_i is normally distributed with unknown mean μ but <u>known</u> variance equal to the sample variance.

   (c) Determine a natural conjugate prior for μ such that the prior's mean equals the sample <u>median</u> of y_i and the prior's variance equals the sample variance of y_i. Then use it to:
   (i) [3 pts] Compute the posterior mean, posterior variance, and posterior precision of μ.
   (ii) [2 pts] Plot the prior and posterior densities in a single plot.
   (iii) [2 pts] Compute a <u>90%</u> central posterior interval for μ.

   (d) Consider a flat prior for μ. Use it to:
   (i) [3 pts] Compute the posterior mean, posterior variance, and posterior precision of μ.

(ii) [2 pts] Plot the prior and posterior densities in a single plot.

(iii) [2 pts] Compute a 90% central posterior interval for μ.

(e) Assume a flat prior for μ, as in the previous part. Using R, simulate 1000 samples from the posterior predictive distribution of article length (in bytes). (Note that, if you chose the log scale for the analysis, you will need to transform back to the original scale.)

(i) [2 pts] Approximate the mean and variance of the posterior predictive distribution of (original-scale) article length.

(ii) [1 pt] Given that there are about 5.7 million articles on the English Wikipedia, estimate the total number of bytes that represents, based on the posterior predictive distribution.

Reminder: Show the R code you used and also a summary of the approximate inference results that you used to answer the preceding parts.

Total: 32 pts