Homework 1
STCS6701
Lincole Jiang (lj2575)
October 10, 2023

Problem 1

Introduction

For this problem, we implemented a stochastic MAP estimation for the IMDB dataset that contains 50k movie reviews and sentiment labels with the goal of predicting the binary sentiments (positive/negative) from the text of the reviews. Since the response variable is binary, we considered Bayesian logistic regression. With hyperparameters p (the number of input features), λ^2 (prior variance), and feature vectors $x_i \in \mathbb{R}^p$, we assume the generative process of parameter

$$\beta_k \sim \text{Normal}(0, \lambda^2)$$

for coefficients $k \in \{1, ..., p\}$, and for each data point

$$y_i \mid x_i \sim \text{Bernoulli}(\sigma(\beta \cdot x_i))$$

, where $\sigma(\beta \cdot x) = \dfrac{1}{1 + \exp(-\beta \cdot x)}$ is the sigmoid function. With the intention of maximizing the log posterior probability,

$$\log p(\beta \mid \mathbf{x}, \mathbf{y}) \propto \sum_{i=1}^{n} y_i \log(\sigma(\beta \cdot x_i)) + (1 - y_i) \log(\sigma(-\beta \cdot x_i)) - \frac{1}{2\lambda^2} \sum_{k=1}^{p} \beta_k^2$$

we make use of the stochastic gradient ascent algorithm following a Robbins-Monro step-size schedule, described in detail in the following section.

Methods

From the IMDB movie review dataset from kaggle, 10k data points are randomly selected from the original dataset as training data and 10k from the rest of the 40k data points were selected as test data. A random seed was set globally to ensure reproducibility. The `R` package `tm` was used to perform a bag-of-word representation for each review, upon normalizing the features using a ti-idf. The common stop words and the rarest words (those that appeared in less than 0.001 of all the documents) were removed. The final number of words considered were 9714 (i.e., p = 9714), which is only slightly smaller than the sample size. A "positive" sentiment is assigned 1 and a "negative" sentiment is assigned 0. On the other hand, to implement stochastic optimization to estimate the MAP estimator for the coefficients that maximizes the posterior likelihood, we selected a the step-size schedule of $\rho_t = 1/t$ (where t is the number of iterations), following a step-size schedule of Robbins and Monro. For the prior β_k's, a standard variation of 1000 was assumed following results from lecture notes. Then, we initialized the p-dimensional coefficients to 0, and

while the algorithm hasn't converged, we draw subsamples with batch-size of 1024 at random from the training dataset, for which the stochastic gradient is calculated as

$$g = \frac{n}{B}\sum_{b=1}^{B}(y_b - \hat{y}_b)x_b - \frac{1}{\lambda}\beta[t]$$

and the β_k's updated as $\beta[t+1] = \beta[t] + \rho_t g$. For convergence diagnostics, we monitored the norm of the gradient and each iteration, and set the tolerance for declaring convergence as 10^{-5}.

Results

In terms of training error, the prediction accuracy is 0.8951. The top 10 and bottom 10 coefficients found by MAP estimation were shown in Table 1, which not surprisingly, the model places large positive coefficients in terms like "wonderful", "excellent", "superb" and large negative coefficients in terms like "worst", "waste", and "awful". On the other land, we plotted the log joint likelihood function w.r.t. iteration in Figure 1, and we note that it plateaus after a the initial 100 or so iterations.

Discussion

Given time constraints, we didn't manage to produce graphics to experiment with batch sizes and step-size schedules; althoguh a small experimentation with different batch sizes has shown that the gradient norm is usually much noisier with smaller batch sizes, albeit the algorithm runs much faster.

Problem 2

a) For this problem, we consider the `recovery.Rdata` dataset, which is from a study that combined three existing cohort studies that track participants for several years aiming to study the factors that predict recovery time from COVID-19. This dataset contains 15 variables of participant ID, gender, race/ethnicity (white, Asian, black, or Hispanic), smoking status, height, weight, BMI, hypertension, diabetes, systolic blood pressure, LDL cholesterol, vaccination status at time of infection, severity of COVID infection, the study cohort the participants were assigned to (A, B, or C); as well as time to recovery (Figures 1). From rudimentary exploratory analysis, we expect some moderate correlation between gender, hypertension, vaccination status, and severity of infection with recovery time, the response variable for which we wish to predict. The explantory variables are not expected to be independent in this dataset, however: e.g., BMI is calculated directly from weight and height; hypertension is evaluated from SBP level; diabetes is known to be associated with higher blood pressure; and vaccination status at time of infection is known to be associated with the severity of infection.

b) Some latent variables to consider for this dataset would be socioeconoimc status and immune response efficacy. Both of these could be considered as "hidden" in the dataset: immune response efficacy–which is determined by various factors including age, presence of chronic illnesses, genetics, as well as nutritional status–is multifaceted in nature and can be only inferred from the dataset we have at hand. Similarly, socioeconomic status, which may reflect financial and housing stability as well as accessibility to healthcare that may be important factors for COVID-19 recovery, is also multifaceted in nature and can only be inferred from several metrics from this dataset. On the other hand, One trivial latent variable that summarizes an aspect of the data is BMI, which can be calculated directly from weight and height.

c) (1) Which hyperparameters should be considered in this model? Which other variables (e.g., age, anemic status) may be relevant to the latent variables proposed that could improve model efficacy upon consideration and how can socioeconomic conditions be extrapolated effectively?

(2) Following the previous question, which probabilistic model would be the optimal for this particular dataset? i.e., should we consider a Bayesian regression model or a mixture model?

(3) Upon conducting a prediction analysis, what are the types of statistical inference questions can we ask for this particular dataset?

Appendix: Figures & Tables

Table 1

most positive coeffcients:

worst	waste	awful	terrible	horrible	bad	worse	boring poor	wasted
-47.52	-40.403	-39.68	-31.09 -30.199	-29.568	-29.42	-25.29	-25.02	-24.95

most negative coefficients:

wonderful	excellent	superb	perfect	great	powerful	outstanding	favorite	underrated
24.16	23.54	23.03	21.10	20.38	19.77	19.47	18.77	18.596

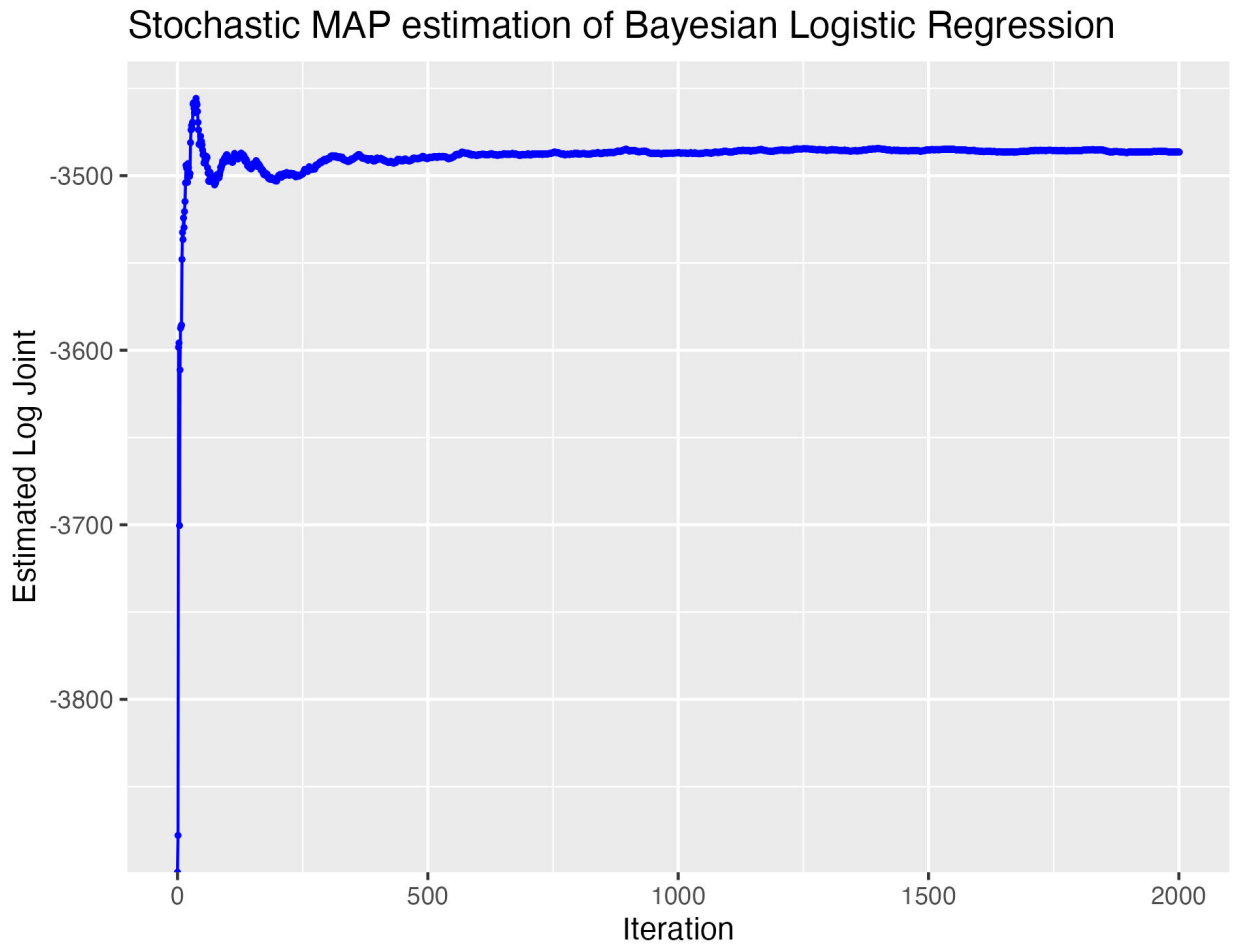

Figure 1: Scatterplot of continuous predictors against recovery time: none of which seem to display strong correlation

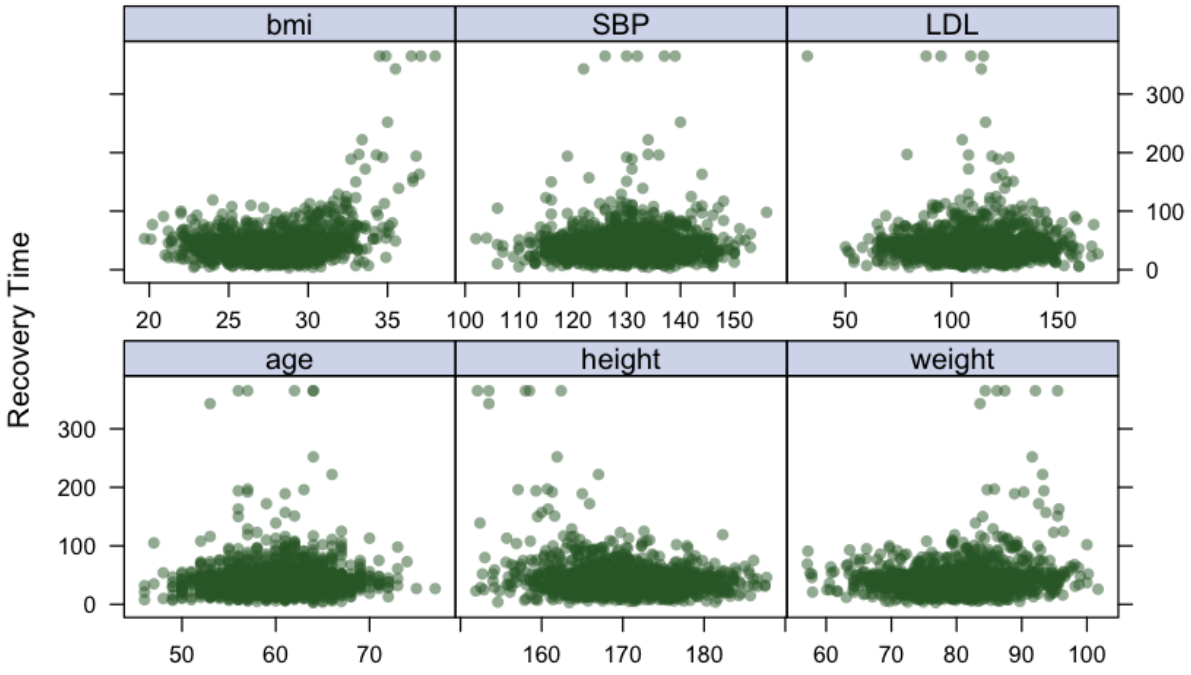
bmi
SBP
LDL
age
height
weight
Recovery Time
0
100
200
300
20
25
30
35
100
110
120
130
140
150
50
100
150
50
60
70
160
170
180
60
70
80
90
100

Figure 2