DSA 595 Bayesian computations for machine learning
Problem set 7

March 5, 2025

1. Similar to the previous problem set, derive the likelihood function for data drawn independently from the logistic regression model:

$$Y_i \mid \beta \sim \text{Bernoulli}\left(\frac{1}{1+e^{-x_i^\top \beta}}\right),$$

for $i \in \{1, \ldots, n\}$, but this time assume $\beta, x_i \in \mathbb{R}^p$ for some $p > 1$.

2. Taking the prior distribution on β to be $\text{normal}_p(0, \tau^2 I_p)$, derive the kernel of the multivariate posterior distribution of β.

3. Write a Metropolis-Hastings algorithm to draw samples from the posterior $\pi(\beta \mid y_1, \ldots, y_n)$. Generate synthetic data from the logistic regression model, for some fixed choice of β, and some fixed covariate vectors $x_1, \ldots, x_n$ (you can generate these as well, but then they are regarded as fixed). Take $p = 4$.

4. Modify your Metropolis-Hastings algorithm in problem 3 with the pre-burn-in adaptive proposal scaling strategy discussed in lecture this week. Demonstrate that you are able to target an acceptance rate close to the range of $(.4, .5)$, post-burn-in.

5. Using the real data set that you have found to be used for your course project, propose a Bayesian statistical model with population features that may be appropriate to estimate from your data. For example, if your data set includes housing prices in US along with a variety of other housing related covariates, then you might be interested in learning regression parameters for explaining the variation in housing prices. Specifically, if Y denotes the price of a given house, X_1 is the square footage of the house, X_2 is the city where the house is located, etc., then perhaps you could formulate the linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + U,$$

where U is a random error variable, and $\beta_0, \beta_1, \beta_2, \ldots$ are the coefficient parameters (i.e., the population features) that you will estimate with your real data. Note that your

statistical model need not be a linear model; this is simply an example for illustration. Please also discuss your choice of prior distribution specifications.

6. Using pseudocode, describe how you will be able to generate synthetic data from your posited statistical model for your course project.