STATS215: Assignment 2

Your Name

Due: February 14, 2020

Problem 1: *Bernoulli GLMs as a latent variable models.*

Consider a Bernoulli regression model,

$$\begin{aligned} w &\sim \mathcal{N}(\mu, \Sigma) \\ y_n \mid x_n, w &\sim \text{Bern}(f(w^\top x_n)) \quad \text{for } n = 1, \ldots, N, \end{aligned}$$

where w and x_n are vectors in $\mathbb{R}^D$, $y_n \in \{0, 1\}$, and $f : \mathbb{R} \to [0, 1]$ is the mean function. In class we studied Newton's method for finding the maximum a posteriori (MAP) estimate $w^\star = \arg\max p(w \mid \{x_n, y_n\}_{n=1}^N)$. Now we will consider methods for approximating the full posterior distribution.

(a) Rather than using the logistic function, let the mean function be the normal cumulative distribution function (CDF), or "probit" function,

$$\begin{aligned} f(u) &= \Pr(z \le u) \text{ where } z \sim \mathcal{N}(0, 1) \\ &= \int_{-\infty}^{u} \mathcal{N}(z; 0, 1)\, \mathrm{d}z. \end{aligned}$$

This is called the probit regression model. Show that the likelihood $p(y_n \mid x_n, w)$ is a marginal of a joint distribution,

$$p(y_n, z_n \mid x_n, w) = \mathbb{I}[z_n \ge 0]^{\mathbb{I}[y_n = 1]}\, \mathbb{I}[z_n < 0]^{\mathbb{I}[y_n = 0]}\, \mathcal{N}(z_n \mid x_n^\top w, 1).$$

«Your answer here.»

(b) Derive the conditional distributions $p(w \mid \{x_n, y_n, z_n\}_{n=1}^N)$ and $p(z_n \mid x_n, y_n, w)$.[1]

«Your answer here.»

(c) *Gibbs sampling* is a Markov chain Monte Carlo (MCMC) method for approximate posterior inference. It works by repeatedly sampling from the conditional distribution of one variable, holding all others fixed. For the probit regression model, this means iteratively performing these two steps:

1. Sample $z_n \sim p(z_n \mid x_n, y_n, w)$ for $n = 1, \ldots, N$ holding w fixed;
2. Sample $w \sim p(w \mid \{x_n, y_n, z_n\}_{n=1}^N)$ holding $\{z_n\}_{n=1}^N$ fixed.

[1] Observe that z_n is conditionally independent of $\{x_{n'}, y_{n'}, z_{n'}\}_{n' \neq n}$ given w.

Note the similarity to EM: rather than computing a posterior distribution over z_n, we draw a sample from it; rather than setting w to maximize the ELBO, we draw a sample from its conditional distribution. It can be shown that this algorithm defines a Markov chain on the space of $(w, \{z_n\}_{n=1}^N)$ whose stationary distribution is the posterior $p(w, \{z_n\}_{n=1}^N \mid \{x_n, y_n\}_{n=1}^N)$. In other words, repeating these steps infinitely many times would yield samples of w and $\{z_n\}_{n=1}^N$ drawn from their posterior distribution.

Implement this Gibbs sampling algorithm and test it on a synthetic dataset with $D = 2$ dimensional covariates and $N = 100$ data points. Scatter plot your samples of w and, for comparison, plot the true value of w that generated the data. Do your samples look approximately Gaussian distributed? How does the posterior distribution change when you vary N?

«Your figures and captions here.»

(d) **Bonus.** There are also auxiliary variable methods for logistic regression, where $f(u) = e^u/(1+e^u)$. Specifically, we have that,

$$\frac{e^{y_n \cdot w^\mathsf{T} x_n}}{1 + e^{w^\mathsf{T} x_n}} = \int_0^\infty \tfrac{1}{2} \exp\left\{\left(y_n - \tfrac{1}{2}\right) x_n^\mathsf{T} w - \tfrac{1}{2} z_n (w^\mathsf{T} x_n)^2\right\} \mathrm{PG}(z_n; 1, 0)\, \mathrm{d}z_n,$$

where $\mathrm{PG}(z; b, c)$ is the density function of the *Pólya-gamma* (PG) distribution over $z \in \mathbb{R}_+$ with parameters b and c. The PG distribution has a number of nice properties: it is closed under exponential tilting so that,

$$e^{-\frac{1}{2} z c^2} \mathrm{PG}(z; b, 0) \propto \mathrm{PG}(z; b, c),$$

and its expectation is available in closed form,

$$\mathbb{E}_{z \sim \mathrm{PG}(b,c)}[z] = \frac{b}{2c} \tanh\left(\frac{c}{2}\right).$$

Use these properties to derive an EM algorithm for finding $w^\star = \arg\max p(\{y_n\} \mid \{x_n\}, w)$. How do the EM updates compare to Newton's method?

Problem 2: *Spike sorting with mixture models*

As discussed in class, "spike sorting" is ultimately a mixture modeling problem. Here we will study the problem in more detail. Let $\{y_n\}_{n=1}^N$ represent a collection of spikes. Each $y_n \in \mathbb{R}^D$ is a vector containing features of the n-th spike waveform. For example, the features may be projections of the spike waveform onto the top D principal components. We have the following, general model,

$$\begin{aligned} z_n \mid \pi &\sim \pi \\ y_n \mid z_n, \theta &\sim p(y_n \mid \theta_{z_n}). \end{aligned}$$

The label $z_n \in \{1, \ldots, K\}$ indicates which of the K neurons generated the n-th spike waveform. The probability vector $\pi \in \Delta_K$ specifies a prior distribution on spike labels, and the parameters $\theta = \{\theta_k\}_{k=1}^K$ determine the likelihood of the spike waveforms y_n for each of the K neurons. The goal is to infer a posterior distribution $p(z_n \mid y_n, \pi, \theta)$ over labels for each observed spike, and to learn the parameters $\pi^\star$ and $\theta^\star$ that maximize the likelihood of the data.

(a) Start with a Gaussian observation model,

$$y_n \mid z_n, \theta \sim \mathcal{N}(y_n \mid \mu_{z_n}, \Sigma_{z_n}),$$

where $\theta_k = (\mu_k, \Sigma_k)$ includes the mean and covariance for the k-th neuron.

Derive an EM algorithm to compute $\pi^\star, \theta^\star = \arg\max p(\{y_n\}_{n=1}^N \mid \pi, \theta)$. Start by deriving the "responsibilities" $w_{nk} = p(z_n = k \mid y_n, \pi', \theta')$ for fixed parameters π' and θ'. Then use the responsibilities to compute the expected log joint probability,

$$\mathcal{L}(\pi, \theta) = \sum_{n=1}^N \mathbb{E}_{p(z_n \mid y_n, \pi', \theta')}[\log p(y_n, z_n \mid \pi, \theta)].$$

Finally, find closed-form expressions for $\pi^\star$ and $\theta^\star$ that optimize $\mathcal{L}(\pi, \theta)$.

«Your answer here.»

(b) The Gaussian model can be sensitive to outliers and lead spikes from one neuron to be split into two clusters. One way to side-step this issue is to replace the Gaussian with a heavier-tailed distribution like the multivariate Student's t, which has probability density,

$$p(y_n \mid \theta_{z_n}) = \frac{\Gamma[(\alpha_0 + D)/2]}{\Gamma(\alpha_0/2)\alpha_0^{D/2}\pi^{D/2}\left|\Sigma_{z_n}\right|^{1/2}}\left[1 + \frac{1}{\alpha_0}(y_n - \mu_{z_n})^\top \Sigma_{z_n}^{-1}(y_n - \mu_{z_n})\right]^{-(\alpha_0+D)/2}.$$

We will treat α_0 as a fixed hyperparameter.

Like the negative binomial distribution studied in HW1, the multivariate Student's t can also be represented as an infinite mixture,

$$p(y_n \mid \theta_{z_n}) = \int p(y_n, \tau_n \mid \theta_{z_n})\, \mathrm{d}\tau_n = \int \mathcal{N}(y_n; \mu_{z_n}, \tau_n^{-1}\Sigma_{z_n})\, \mathrm{Gamma}(\tau_n; \tfrac{\alpha_0}{2}, \tfrac{1}{2})\, \mathrm{d}\tau_n.$$

We will derive an EM algorithm to find $\pi^\star, \theta^\star$ in this model.

First, show that the posterior takes the form

$$\begin{aligned} p(\tau_n, z_n \mid y_n, \pi, \theta) &= p(z_n \mid y_n, \pi, \theta)\, p(\tau_n \mid z_n, y_n, \theta) \\ &= \prod_{k=1}^{K} \Big[w_{nk}\, \text{Gamma}(\tau_n \mid a_{nk}, b_{nk}) \Big]^{\mathbb{I}[z_n = k]}, \end{aligned}$$

and solve for the parameters w_{nk}, a_{nk}, b_{nk} in terms of y_n, π, and θ.

«Your answer here.»

(c) Now compute the expected log joint probability,

$$\mathscr{L}(\pi, \theta) = \sum_{n=1}^{N} \mathbb{E}_{p(\tau_n, z_n \mid y_n, \pi', \theta')} [\log p(y_n, z_n, \tau_n \mid \pi, \theta)],$$

using the fact that $\mathbb{E}[X] = a/b$ for $X \sim \text{Gamma}(a, b)$. You may omit terms that are constant with respect to π and θ.

«Your answer here.»

(d) Finally, solve for $\pi^\star$ and $\theta^\star$ that maximize the expected log joint probability. How does your answer compare to the solution you found in part (a)?

«Your answer here.»

Problem 3: *Poisson matrix factorization*

Many biological datasets come in the form of matrices of non-negative counts. RNA sequencing data, neural spike trains, and network data (where each entry indicate the number of connections between a pair of nodes) are all good examples. It is common to model these counts as a function of some latent features of the corresponding row and column. Here we consider one such model, which decomposes a count matrix into a superposition of non-negative row and column factors.

Let $Y \in \mathbb{N}^{M \times N}$ denote an observed $M \times N$ matrix of non-negative count data. We model this matrix as a function of non-negative row factors $U \in \mathbb{R}_+^{M \times K}$ and column factors $V \in \mathbb{R}_+^{N \times K}$. Let $u_m \in \mathbb{R}_+^K$ and $v_n \in \mathbb{R}_+^K$ denote the m-th and n-th rows of U and V, respectively. We assume that each observed count y_{mn} is conditionally independent of the others given its corresponding row and column factors. Moreover, we assume a linear Poisson model,

$$y_{mn} \mid u_m, v_n \sim \text{Poisson}(u_m^\top v_n).$$

(Since u_m and v_n are non-negative, the mean parameter is valid.) Finally, assume gamma priors,

$$u_{mk} \sim \text{Gamma}(\alpha_0, \beta_0),$$
$$v_{nk} \sim \text{Gamma}(\alpha_0, \beta_0).$$

Note that even though the gamma distribution is conjugate to the Poisson, here we have an inner product of two gamma vectors producing one Poisson random variable. The posterior distribution is more complicated. The entries of u_m are not independent under the posterior due to the "explaining away" effect. Nevertheless, we will derive a mean-field variational inference algorithm to approximate the posterior distribution.

(a) First we will use an augmentation trick based on the additivity of Poisson random variables; i.e. the fact that

$$y \sim \text{Poisson}\left(\sum_k \lambda_k\right) \iff y = \sum_k y_k \text{ where } y_k \sim \text{Poisson}(\lambda_k) \text{ independently},$$

for any collection of non-negative rates $\lambda_1, \ldots, \lambda_K \in \mathbb{R}_+$. Use this fact to write the likelihood $p(y_{mn} \mid u_m, v_n)$ as a marginal of a joint distribution $p(y_{mn}, \bar{y}_{mn} \mid u_m, v_n)$ where $\bar{y}_{mn} = (y_{mn1}, \ldots, y_{mnK})$ is a length-K vector of non-negative counts. (Hint: this is similar to Problem 1 in that y_{mn} is deterministic given $\bar{y}_{mn}$.)

«Your answer here.»

(b) Let $\bar{Y} \in \mathbb{N}^{M \times N \times K}$ denote the augmented data matrix with entries y_{mnk} as above. We will use mean field variational inference to approximate the posterior as,

$$p(\bar{Y}, U, V \mid Y) \approx q(\bar{Y})\, q(U)\, q(V) = \left[\prod_{m=1}^M \prod_{n=1}^N q(\bar{y}_{mn})\right]\left[\prod_{m=1}^M \prod_{k=1}^K q(u_{mk})\right]\left[\prod_{n=1}^N \prod_{k=1}^K q(v_{nk})\right].$$

We will solve for the optimal posterior approximation via coordinate descent on the KL divergence to the true posterior. Recall that holding all factors except for $q(\bar{y}_{mn})$ fixed, the KL is minimized when

$$q(\bar{y}_{mn}) \propto \exp\left\{\mathbb{E}_{q(\bar{Y}_{\neg mn})q(U)q(V)}\left[\log p(Y, \bar{Y}, U, V)\right]\right\},$$

where $q(\bar{Y}_{\neg mn}) = \prod_{(m',n') \neq (m,n)} q(\bar{y}_{m'n'})$ denotes all variational factors except for the (m,n)-th.

Show that the optimal $q(\bar{y}_{mn})$ is a multinomial of the form,

$$q(\bar{y}_{mn}) = \text{Mult}(\bar{y}_{mn}; y_{mn}, \pi_{mn}),$$

and solve for $\pi_{mn} \in \Delta_K$. You should write your answer in terms of expectations with respect to the other variational factors.

«Your answer here.»

(c) Holding all factors but $q(u_{mk})$ fixed, show that optimal distribution is

$$q(u_{mk}) = \text{Gamma}(u_{mk}; \alpha_{mk}, \beta_{mk}).$$

Solve for α_{mk}, β_{mk}; write your answer in terms of expectations with respect to $q(\bar{y}_{mn})$ and $q(v_{nk})$.

«Your answer here.»

(d) Use the symmetry of the model to determine the parameters of the optimal gamma distribution for $q(v_{nk})$, holding $q(\bar{y}_{mn})$ and $q(u_{mk})$ fixed,

$$q(v_{nk}) = \text{Gamma}(v_{nk}; \alpha_{nk}, \beta_{nk}).$$

Solve for α_{nk}, β_{nk}; write your answer in terms of expectations with respect to $q(\bar{y}_{mn})$ and $q(u_{mk})$.

«Your answer here.»

(e) Now that the form of all variational factors has been determined, compute the required expectations (in closed form) to write the coordinate descent updates in terms of the other variational parameters. Use the fact that $\mathbb{E}[\log X] = \psi(\alpha) - \log \beta$ for $X \sim \text{Gamma}(\alpha, \beta)$, where ψ is the digamma function.

«Your answer here.»

(f) Suppose that Y is a sparse matrix with only $S \ll MN$ non-zero entries. What is the complexity of this mean-field coordinate descent algorithm?

«Your answer here.»

Problem 4: *Apply Poisson matrix factorization to C. elegans connectomics data*

Make a copy of this Colab notebook:

`https://colab.research.google.com/drive/1ZMwcB6vzVaXz4WJiNT514b7zB5s3_SBk`

Use your solutions from Problem 3 to finish the incomplete code cells. Once you're done, run all the code cells, save the notebook in `.ipynb` format, print a copy in `.pdf` format, and submit these files along with the rest of your written assignment.