Bayesian models for Large-scale Hierarchical Classification (Supplementary Material)

Anonymous Author(s)
Affiliation
Address
email

1 VARIATIONAL INFERENCE FOR MODELS

1.1 Details of Variational Inference for HBLR-M2

Generally in Bayesian Inference, computing the posterior distribution of parameters in closed form solution might not be possible. In such situations, one often resorts to computing an approximate posterior which is 'close' to the true posterior distribution. For HBLR-M2, the posterior of the model parameters is given by,

$$
\begin{aligned}
p(\mathbf{W}, \boldsymbol{\alpha}|\mathbf{D}) \propto & p(\mathbf{D}|\mathbf{W}, \boldsymbol{\alpha}) p(\mathbf{W}, \boldsymbol{\alpha}) \\
p(\mathbf{D}|\mathbf{W}, \boldsymbol{\alpha}) = & \prod_{(x,t)\in D} \frac{\exp(w_t^\top x)}{\sum_{t'\in T} \exp(w_{t'}^\top x)} \qquad (1) \\
p(\mathbf{W}, \boldsymbol{\alpha}) = & \prod_{y\in Y\setminus T} \prod_{i=1}^{d} p(\alpha_y^{(i)}|a_y^{(i)}, b_y^{(i)}) \prod_{y\in Y} p(w_y|w_{\pi(y)}, \Sigma_{\pi(y)}) \\
= & \prod_{y\in Y\setminus T} \prod_{i=1}^{d} \Gamma(\alpha_y^{(i)}|a_y^{(i)}, b_y^{(i)}) \prod_{y\in Y} \mathcal{N}(w_y|w_{\pi(y)}, \Sigma_{\pi(y)}) \qquad (2)
\end{aligned}
$$

The posterior has a logistic likelihood term with the $\boldsymbol{W}$ in (1) and with a Gamma and Normal prior over $\boldsymbol{\alpha}$, $\boldsymbol{W}$ in (2). The convolution between a normal-gamma prior and logistic likelihood cannot be computed in closed form; therefore one has to resort to approximate methods to calculate the posterior.

Variational methods try to compute an approximate posterior having a simplified factored form which is closest in KL divergence to the true posterior. They rely on the following bound for the log-marginal probability of D. For any distribution $q(\boldsymbol{W}, \boldsymbol{\alpha})$,

$$
\begin{aligned}
P(\boldsymbol{D}) = & \int q(\boldsymbol{W}, \boldsymbol{\alpha}) \log \frac{p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D})}{q(\boldsymbol{W}, \boldsymbol{\alpha})} d\boldsymbol{W} d\boldsymbol{\alpha} - \int q(\boldsymbol{W}, \boldsymbol{\alpha}) \log \frac{p(\boldsymbol{W}, \boldsymbol{\alpha}|\boldsymbol{D})}{q(\boldsymbol{W}, \boldsymbol{\alpha})} d\boldsymbol{W} d\boldsymbol{\alpha} \qquad (3) \\
\geq & \int q(\boldsymbol{W}, \boldsymbol{\alpha}) \log \frac{p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D})}{q(\boldsymbol{W}, \boldsymbol{\alpha})} d\boldsymbol{W} d\boldsymbol{\alpha} \\
= & \int q(\boldsymbol{W}, \boldsymbol{\alpha}) \log p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D}) d\boldsymbol{W} d\boldsymbol{\alpha} - \int q(\boldsymbol{W}, \boldsymbol{\alpha}) \log q(\boldsymbol{W}, \boldsymbol{\alpha}) d\boldsymbol{W} d\boldsymbol{\alpha} \qquad (4) \\
= & E_q \left[\log p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D})\right] + H(q) \qquad (5)
\end{aligned}
$$

The (3) can be easily verified by combining the RHS terms.

In order to compute such a q, we start by assuming a simplified factored form using independent distributions for each parameter. Note that this does not neglect the dependence between the various parameters in the model; it just finds the suitable factored form which best approximates the dependencies.

$$
\begin{aligned}
q(\mathbf{W}, \boldsymbol{\alpha}) &= \prod_{y \in Y \setminus T} q(\alpha_y) \prod_{y \in Y} q(w_y) \\
q(w_y) &\propto \mathcal{N}(.|\mu_y, \Psi_y) \\
q(\alpha_y) &= \prod_{i=1}^{d} q(\alpha_y^{(i)}) \propto \prod_{i=1}^{d} \Gamma(.|\tau_y^{(i)}, \upsilon_y^{(i)})
\end{aligned}
$$

Here $q(\mathbf{W}, \boldsymbol{\alpha}) \equiv q(\mathbf{W}|\boldsymbol{\mu}, \boldsymbol{\Psi}) q(\boldsymbol{\alpha}|\boldsymbol{\tau}, \boldsymbol{\upsilon})$ where $\boldsymbol{\tau}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ are variational parameters which we can optimize one at a time to maximize (4).

For example, to optimize w_y, we differentiate $P(D)$ w.r.t $q(w_y)$,

$$
\begin{aligned}
P(D) &= \int q(w_y|\mu_y, \Psi_y) \left[q(\boldsymbol{W}^{-w_y}, \boldsymbol{\alpha}) \log p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D}) d\boldsymbol{W} d\boldsymbol{\alpha} \right] - \int q(w_y|\mu_y, \Psi_y) \log q(w_y|\mu_y, \Psi_y) dw_y \\
\log q^*(w_y|\mu_y, \Psi_y) &= E_{q^{-w_y}} \left[\log p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D}) d\boldsymbol{W} d\boldsymbol{\alpha} \right] \quad + constant
\end{aligned} \tag{6}
$$

where $-w_y$ denotes all parameters other than w_y. Similary,

$$
\log q^*(\alpha_y|\tau_y, \upsilon_y) = E_{q^{-\alpha_y}} \left[\log p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D}) d\boldsymbol{W} d\boldsymbol{\alpha} \right] \quad + constant \tag{7}
$$

In (7) and (6), to update the parameters we need to calculate the expected log-likelihood under the variational posterior i.e. $E_q[\log p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D})]$.

$$
E_q[\log p(\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{D})] = E_q \left[\log P(\boldsymbol{D}|\boldsymbol{W}, \boldsymbol{\alpha}) \right] + E_q \left[\log P(\boldsymbol{W}, \boldsymbol{\alpha}) \right]
$$

Where,

$$
\begin{aligned}
E_q \left[\log P(\boldsymbol{W}, \boldsymbol{\alpha}) \right] = & \sum_{Y \in Y} -\frac{1}{2} E_q \left[(w_y - w_{\pi(y)})^\top \Sigma_{\pi(y)}^{-1} (w_y - w_{\pi(y)}) \right] \\
& - \sum_{y \in Y} \frac{1}{2} E_q \left[\log |\Sigma_y| \right] - \sum_{y \in Y/T} \sum_{i=1}^{d} E_q \left[\alpha_y^{(i)} \right] b_y + E_q \left[(\alpha_y^{(i)} - 1) \log(\alpha_y^{(i)}) \right]
\end{aligned} \tag{8}
$$

$$
E_q \left[\log P(\boldsymbol{D}|\boldsymbol{W}, \boldsymbol{\alpha}) \right] = \sum_{(x,t) \in D} \mu_y^\top x - \sum_{(x,t) \in D} E_q \left[\log \left(\sum_{y \in T} \exp(w_y^\top x) \right) \right] \tag{9}
$$

The RHS term in the above equation is not directly computable. Therefore we use a suitable lower-bound proposed in [1], to bound the log-likelihood term from below.

$$
\log \left(\sum_{y \in T} \exp(w_y^\top x) \right) \leq \beta_x + \sum_{y \in T} \frac{(w_y^\top x - \beta_x - \xi_{xy})}{2} + \sum_{y \in T} \lambda(\xi_{xy}) \left(\left(w_y^\top x - \beta_x \right)^2 - \xi_{xy}^2 \right) \tag{10}
$$

$$
+ \sum_{y \in T} \log \left(1 + \xi_{xy}^2 \right) \tag{11}
$$

$$
= -(\frac{|T|}{2} - 1)\beta_x + \sum_{y \in T} \frac{w_y^\top x}{2} - \sum_{y \in T} \frac{\xi_{xy}}{2} \tag{12}
$$

$$
+ \sum_{y \in T} \lambda(\xi_{xy}) \left(w_y^\top (x x^\top) w_y - 2\beta_x (w_y^\top x) \lambda(\xi_{xy}) + \beta_x^2 - \xi_{xy}^2 \right) + \sum_{y \in T} \log(1 + e^{\xi_{xy}}) \tag{13}
$$

Here we have introduced variational parameter β_x and ξ_{xy} for every $x \in D, y \in Y$. In order to get the tightest possible bound, we can optimize over these variational parameters. We discuss this optimization later in the text.

$$E_q\left[\log\left(\sum_{y\in T}\exp(w_y^\top x)\right)\right] = \sum_{y\in Y}\left(\frac{1}{2} - 2\lambda(\xi_{xy})\beta_x\right)\mu_y^\top x + \sum_{y\in Y}\mu_y^\top(2\lambda(\xi_{xy})xx^\top)\mu_y \\ -(\frac{|T|}{2}-1)\beta_x - \sum_{y\in Y}\left(\lambda(\xi_{xy})\beta_x^2 - \frac{\xi_{xy}}{2} - \frac{\xi_{xy}^2}{2} + \log(1+e^{\xi_{xy}})\right) \tag{14}$$

1.1.1 Optimizing $q^*(w_y)$

Combining the above equation with (8) and (6), the update for parameter $w_y, y \in T$ can be written as

$$\log q^*(w_y|\mu_y,\Psi_y) = (\sum_{(x,t)\in D}(I(t=y) - \frac{1}{2} + 2\lambda(\xi_{xy})\beta_x)x + E_{q^{-w_y}}\left[\Sigma_{\pi(y)}^{-1}\right]\mu_y) - w_y)^\top \\ \left(\sum_{(x,t)\in D} 2\lambda(\xi_{xy})xx^\top + E_{q^{-w_y}}\left[\Sigma_{\pi(y)}^{-1}\right]\right) \\ (\sum_{(x,t)\in D}(I(t=y) - \frac{1}{2} + 2\lambda(\xi_{xy})\beta_x)x + E_{q^{-w_y}}\left[\Sigma_{\pi(y)}^{-1}\right]\mu_y) - w_y) + constant$$

Since $q^*(w_y)$ is assumed to be a normal distribution, we can directly match the sufficient statistics i.e. mean and the covariance matrix and set μ_y and Ψ_y as in the paper. Note that we have used the fact that

$$E_{q^{-w_y}}\left[\Sigma_y^{-1}\right] = \text{diag}\left(E_{q^{-w_y}}\left[\alpha_y^{(1)}\right], E_{q^{-w_y}}\left[\alpha_y^{(2)}\right], \ldots, E_{q^{-w_y}}\left[\alpha_y^{(d)}\right])\right) \\ = \text{diag}\left(\frac{\tau_y^{(1)}}{\upsilon_y^{(1)}}, \frac{\tau_y^{(2)}}{\upsilon_y^{(2)}}, \ldots, \frac{\tau_y^{(d)}}{\upsilon_y^{(d)}})\right)$$

For $q^*(w_y), y \notin T$, fortunately the w_y is not present in the logistic function, therefore, we can use just (8) to match the likelihood with (6) and get the equations as in the paper.

1.1.2 Optimizing $q^*(\alpha_y)$

For optimizing $q^*(\alpha_y)$, we essentially follow the same strategy as above. Each optimize each $\alpha_i^{(i)}$ by matching (8) and (7)

$$\log q^*(\alpha_y^{(i)}|\tau_y^{(i)},\upsilon_y^{(i)}) = -\sum_{y\in C_y}\frac{\alpha_y^{(i)}}{2}E_q\left[(w^{(i)} - w_{\pi(y)}^{(i)})^2\right] + \sum_{y\in C_y}\frac{\alpha_y^{(i)}}{2} - \alpha_y^{(i)}b_y^{(i)} + (\alpha_y^{(i)}-1)\log(\alpha_y^{(i)}) \\ = -\alpha_y^{(i)}\left(b_y^{(i)} + \sum_{y\in C_y}\frac{1}{2}\left((\mu^{(i)} - \mu_{\pi(y)}^{(i)})^2 + \Psi^{(i,i)} + \Psi_{\pi(y)}^{(i,i)}\right)\right) \\ + (\alpha_y^{(i)} + \frac{|C_y|}{2} - 1)\log(\alpha_y^{(i)}) + constant$$

Since $q^*(\alpha_y)$ is a gamma distribution, we simply match the sufficient statistcs and set update the as in the paper. Note that we have used the fact that

$$-\log|\Sigma_y| = \log|\Sigma_y^{-1}| = \log\prod_{i=1}^d \alpha_y^{(i)} = \sum_{i=1}^d \log\alpha_y^{(i)}$$

1.2 Extension to HBLR-M1

Most of the derivation simply goes through changed except that $\alpha_y^{(1)} = \alpha_y^{(2)} = \ldots = \alpha_y$. Although, this makes a difference only in how the α's are updated; we present the full update equations.

If $y \in T$

$$\Psi_y^{-1} = \sum_{(x,t)\in D} 2\lambda(\xi_{xy}) xx^\top + E_{q^{-w_y}}\left[\Sigma_{\pi(y)}^{-1}\right]$$

$$\mu_y = \Psi_y \left(\sum_{(x,t)\in D} (I(t=y) - \frac{1}{2} + 2\lambda(\xi_{xy})\beta_x)x + E_{q^{-w_y}}\left[\Sigma_{\pi(y)}^{-1}\right]\mu_{\pi(y)} \right)$$

If $y \notin T$

$$\Psi_y^{-1} = E_{q^{-w_y}}\left[\Sigma_{\pi(y)}^{-1}\right] + |C_y| E_{q^{-w_y}}\left[\Sigma_y^{-1}\right]$$

$$\mu_y = \Psi_y \left(E_{q^{-w_y}}\left[\Sigma_{\pi(y)}^{-1}\right]\mu_{\pi(y)} + E_{q^{-w_y}}\left[\Sigma_y^{-1}\right]) \sum_{c\in C_y} \mu_c \right)$$

$$\upsilon_y = (b_y + \sum_{c\in C_y} \mathrm{diag}(\Psi_y) + diag(\Psi_c) + (\mu_y - \mu_c)^\top (\mu_y - \mu_c))^\top e$$

$$\tau_y = a_y + \frac{|C_y|d}{2}$$

where

$$E_{q^{-w_y}}\left[\Sigma_y^{-1}\right] = \mathrm{diag}\left(\frac{\tau_y}{\upsilon_y}, \frac{\tau_y}{\upsilon_y}, \ldots, \frac{\tau_y}{\upsilon_y})\right)$$

and e is a unit vector.

1.3 Extension to HBLR-M3

The extension to HBLR-M3 follows on similar lines.

If $y \in T$

$$\Psi_y^{-1} = \sum_{(x,t)\in D} 2\lambda(\xi_{xy}) xx^\top + E_{q^{-w_y}}\left[\Sigma_y^{-1}\right]$$

$$\mu_y = \Psi_y \left(\sum_{(x,t)\in D} (I(t=y) - \frac{1}{2} + 2\lambda(\xi_{xy})\beta_x)x + E_{q^{-w_y}}\left[\Sigma_y^{-1}\right]\mu_{\pi(y)} \right)$$

If $y \notin T$

$$\Psi_y^{-1} = E_{q^{-w_y}}\left[\Sigma_{\pi(y)}^{-1}\right] + |C_y| E_{q^{-w_y}}\left[\Sigma_y^{-1}\right]$$

$$\mu_y = \Psi_y \left(E_{q^{-w_y}}\left[\Sigma_{\pi(y)}^{-1}\right]\mu_{\pi(y)} + E_{q^{-w_y}}\left[\Sigma_y^{-1}\right] \sum_{c\in C_y} \mu_c \right)$$

$$\upsilon_y = (b_y + \mathrm{diag}(\Psi_{\pi(y)}) + diag(\Psi_y) + (\mu_y - \mu_{\pi(y)})^\top (\mu_y - \mu_{\pi(y)}))^\top e$$

$$\tau_y = a_y + \frac{d}{2}$$

where

$$E_{q^{-w_y}}\left[\Sigma_y^{-1}\right] = \mathrm{diag}\left(\frac{\tau_y}{\upsilon_y}, \frac{\tau_y}{\upsilon_y}, \ldots, \frac{\tau_y}{\upsilon_y})\right)$$

2 Empirical Bayes

In this section, we show why Empirical Bayes route for learning the hyperparameters in our model does not work. Let us model M1 for instance. The general procedure for Empirical Bayes is to maximize the marginal likelihood w.r.t to the hyperparameters and get point estimates for them (another approach would be use MCMC, which we do not pursue in the interested of scalability).

In M1, the marginal likelihood of the data can be computed as

$$P(D|a,b) \propto \int_{\boldsymbol{\alpha}} P(\boldsymbol{\alpha}|a,b) \int_{\boldsymbol{W}} P(D|\boldsymbol{W})P(\boldsymbol{W}|\boldsymbol{\alpha})$$

As before the integral cannot be computed in closed form and hence cannot be maximized analytically. Therefore we use the variational lower-bound as a proxy which is more amenable to maximization.

$$\begin{aligned}
\log P(D|a,b) &\geq E_q[logP(D|a,b)] \\
&\geq E_q[logP(\alpha|a,b)] + \text{ terms that do not depend on } a,b \\
&\geq \sum_{Y \in Y \setminus T} E_q[log(P(\alpha_y|a_y,b_y)] \\
&\geq \sum_{Y \in Y \setminus T} E_q\left[b_y^{a_y} \frac{1}{\Gamma(a_y)} \alpha_y^{a_y-1} e^{-b_y \alpha}\right]
\end{aligned}$$

Note that the maximization can be carried out independently for all the a_y's. This leads to a Gamma distribution type MLE of the following form,

$$(a_y^*, b_y^*) = \arg\max a_y \log(b_y) - \log\Gamma(a_y) + (a_y - 1)E_q[log(\alpha_y)] - b_y E_q[\alpha_y]$$

Note that $E_q[log(\alpha_y)] = \Psi(\tau_y) - \log(\upsilon_y)$ and $E_q[log(\alpha_y)] = \frac{\tau_y}{\upsilon_y}$; both of twhich can be considered as constants in the maximization. But since there is exactly only one sample, one cannot learn the a_y, b_y effectively [2].

One way to overcome this, is to assume that all the α_y's are commonly drawn from a single a, b. This enables a larger number of samples to succesfully estimate the the single a, b. The downside is that, this commonly shrinks all the α_y's to the expected value of the distribution $\frac{a}{b}$, which might be not a good thing.

We conducted several preliminary experiments were we tried sharing all α_y's under single a, b as well sharing the α_y's of sinling nodes etc. None of the models seemed to achieve competitive performance. For example, using a common a, b on the best performing model $M3$ on the CLEF dataset achieved a Micro-F_1 of of 80.32 and Macro-F_1 of 54.82 both of which are lower than M3-var. Further investigation is required to establish how empirical can be succesfully applied.

References

[1] G. Bouchard. Efficient bounds for the softmax function. 2007.

[2] George Casella. Empirical bayes method - a tutorial. Technical report.