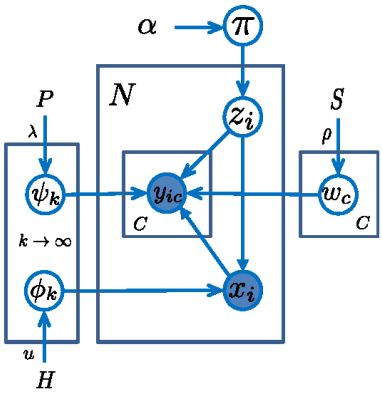

Figure 1: BNMC graphical model.

Figure 2: Generative process for BNMC

1: $\pi \sim \text{Stick}(\alpha)$
2: $\phi_k \overset{\text{iid}}{\sim} H(u),\ \forall k = 1, ..., \infty$
3: $\psi_{k,c} \overset{\text{iid}}{\sim} \text{Mult}(\lambda)\ \ \forall k, \forall c = 1...C$
4: $w_c \overset{\text{iid}}{\sim} S(\rho)\ \ \forall c = 1...C$
5: **for** each data point $i = 1, ..., N$ **do**
6: $\quad z_i \overset{\text{iid}}{\sim} \pi$ and $\boldsymbol{x}_i \sim F(\phi_{z_i})$
7: $\quad \boldsymbol{y}_{i,1....C} \sim \text{Mult}\left(\psi_{z_i,1....C} \times \sigma\left(\boldsymbol{x}_i^T w_{1...C}\right)\right)$
8: **end for**

Supplement for A Bayesian Nonparametric Approach for Multi-label Classification

Vu Nguyen V.NGUYEN@DEAKIN.EDU.AU
Sunil Gupta SUNIL.GUPTA@DEAKIN.EDU.AU
Santu Rana SANTU.RANA@DEAKIN.EDU.AU
Cheng Li CHENG.L@DEAKIN.EDU.AU
Svetha Venkatesh SVETHA.VENKATESH@DEAKIN.EDU.AU
Center for Pattern Recognition and Data Analytics, Deakin University

Editors: Robert J. Durrant and Kee-Eung Kim

1. Stochastic Variational Inference for BNMC

Recall that we need to estimate 5 variables (ϕ_k, ψ_k, π, w_c and z_i) each of which couples with the additional two parameters (one for variational parameter and one for natural gradient update). With a slight abuse of notation, we denote the parameters as follows: let consider ϕ_k be the variable in our model. $\tilde{\phi}_k$ is the parameter in variational distribution $q\left(\phi_k \mid \tilde{\phi}_k\right)$. $\hat{\phi}_k$ is the natural gradient to update $\tilde{\phi}_k$. Other variables are used in similar notations.

Estimating $\tilde{z}_i$ We have z_i is the local variable in the model. Each data point is a D-dimensional vector $\boldsymbol{x}_i = [x_{i1}, ...x_{iD}]$ and each label set is a C-dimensional vector $\boldsymbol{y}_i = [y_{i1}, ..., y_{iC}]$. Since the Multinomial conditional distribution for z_i^k in the original distribution p is defined as

$$p\left(z_i^k \mid .\right) \propto \exp\left\{\log \pi_k + \log p\left(\boldsymbol{x}_i \mid \phi_k\right) + \log p\left(\boldsymbol{y}_i \mid \psi_k\right)\right\}.$$

The variational distribution for z_i is $q(z_i \mid \tilde{z}_i) = \text{Mult}(\tilde{z}_i)$. The local variational parameter is set equally to the expected natural parameter of its complete conditional distribution that is

$$\begin{aligned}\tilde{z}_i^k &= \mathop{\mathbb{E}}_{q}[\eta_l(\boldsymbol{x}_i, \boldsymbol{y}_i, \phi, \psi, \pi)] \\ &= \exp\left\{\mathbb{E}[\log \pi_k] + \mathbb{E}[\log \phi_{k,\boldsymbol{x}_i}] + \mathbb{E}\left[\log \psi_{k,\boldsymbol{y}_i}\right]\right\}.\end{aligned}$$

We now go into details how to compute these expectation terms. We observe that $\log \pi_k$ is the vector of sufficient statistics of the Dirichlet distribution $q(\pi_k \mid \tilde{\pi}_k) = \text{Dir}(\tilde{\pi}_k)$, given the natural parameter $\tilde{\pi}_k$. Therefore, we can utilize the property of exponential family to compute $\mathbb{E}[\log \pi_k]$ by taking the gradient of the log-normalizer w.r.t. the natural parameter vector $\tilde{\pi}_k$ (Hoffman et al., 2013). Then, the expectation of its log can be computed using Ψ which is the first derivative of the log Gamma function. Thus, we write $\mathbb{E}[\log \pi_k] = \Psi(\tilde{\pi}_k) - \Psi(\sum \tilde{\pi})$.

Next, we are going to compute the expectations of $\mathbb{E}[\log \phi_{k,\boldsymbol{x}_i}]$ and $\mathbb{E}\left[\log \psi_{k,\boldsymbol{y}_i}\right]$. The feature and label vectors are generated as $\boldsymbol{x}_i \sim \text{Mult}(\phi_k)$ and $\boldsymbol{y}_i \sim \text{Mult}(\psi_k)$ that the variational distributions over the parameters are defined as follows $q\left(\phi_k \mid \tilde{\phi}_k\right) = \text{Dir}\left(\tilde{\phi}_k\right)$ and $q\left(\psi_k \mid \tilde{\psi}_k\right) = \text{Dir}\left(\tilde{\psi}_k\right)$. Therefore, we again use the property of the exponential family to compute $\mathbb{E}\left[\log \tilde{\phi}_{k,\boldsymbol{x}_i}\right]$ and $\mathbb{E}\left[\log \tilde{\psi}_{k,\boldsymbol{y}_i}\right]$ while we need to account for the contribution of each element in $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ toward the expectation,

$$\mathbb{E}[\log \phi_{k,\boldsymbol{x}_i}] = \sum_{d=1}^{D} x_{id} \times \mathbb{E}\left[\log \tilde{\phi}_{k,d}\right] = \sum_{d=1}^{D} x_{id}\left[\Psi\left(\tilde{\phi}_{k,d}\right) - \Psi\left(\tilde{\phi}_{k,*}\right)\right]$$

and similarly for $\tilde{\psi}_k$

$$\mathbb{E}\left[\log \psi_{k,\boldsymbol{y}_i}\right] = \sum_{c=1}^{C} y_{ic} \times \mathbb{E}\left[\log \tilde{\psi}_{k,c}\right] = \sum_{c=1}^{C} y_{ic}\left[\Psi\left(\tilde{\psi}_{k,c}\right) - \Psi\left(\tilde{\psi}_{k,*}\right)\right]$$

where $*$ denotes for the sum. Finally, we compute the variational parameter $\tilde{z}_i$ for a data point i as follows

$$\begin{aligned}\tilde{z}_i^k &\propto \exp\left\{\mathbb{E}[\log \tilde{\pi_k}] + \mathbb{E}\left[\log \tilde{\phi}_{k,\boldsymbol{x}_i}\right] + \mathbb{E}\left[\log \tilde{\psi}_{k,\boldsymbol{y}_i}\right]\right\} \\ &= \exp\left\{\Psi(\tilde{\pi}_k) - \Psi(\sum \tilde{\pi}) + \sum_{d=1}^{D} x_{id}\left[\Psi(\tilde{\phi}_{k,d}) - \Psi(\tilde{\phi}_{k,*})\right] + \sum_{c=1}^{C} y_{ic}\left[\Psi(\tilde{\psi}_{k,c}) - \Psi(\tilde{\psi}_{k,*})\right]\right\}.\end{aligned} \tag{1}$$

Estimating $\tilde{\phi}_k$ We now estimate variational parameter $\tilde{\phi}_k$ of the feature pattern ϕ_k. The conditional distribution for the feature topic is defined as

$$p(\phi_k \mid \boldsymbol{z}, \boldsymbol{x}, H) \propto \text{Dir}\left(\omega_\phi + \sum_{i=1}^{N} z_i^k x_i\right).$$

Each element ϕ_{kd} is the sum of the hyperparameter ω_ϕ and the number of times the term x_{id} are assigned to topic (or pattern) ϕ_k. This is a global variable that its complete conditional depends on the feature $\boldsymbol{x}$ and latent assignments $\boldsymbol{z}$. The variational distribution for each topic is a D-dimensional Dirichlet $q\left(\phi_k \mid \tilde{\phi}_k\right) = \text{Dir}\left(\tilde{\phi}_k\right)$. In the batch setting, the global variational parameter $\tilde{\phi}_k$ is computed as

$$\tilde{\phi}_k = \mathop{\mathbb{E}}_{q}\left[\eta_g\left(\boldsymbol{x}_i, \boldsymbol{y}_i, \phi, \psi, \pi\right)\right] = \omega_\phi + \sum_{i=1}^{N} \mathop{\mathbb{E}}_{q}\left[z_i^k\right] x_i = \omega_\phi + \sum_{i=1}^{N} \tilde{z}_i^k x_i.$$

In the stochastic setting, given a data point i, we incrementally update $\tilde{\phi}_k$ in an online setting using the natural gradient which is computed as $\hat{\phi}_k = \omega_\phi + N\tilde{z}_i^k \boldsymbol{x}_i$. Then, the variational parameter $\tilde{\phi}_k$ is updated

$$\tilde{\phi}_k^{(i+1)} = \left(1 - \rho_i\right)\tilde{\phi}_k^{(i)} + \rho_i \hat{\phi}_k. \tag{2}$$

Estimating $\tilde{\psi}_k$ In the similar spirit of estimating $\tilde{\phi}_k$, the variational distribution for label topic ψ_k is defined as $q\left(\psi_k \mid \tilde{\psi}_k\right) = \text{Dir}\left(\tilde{\psi}_k\right)$. We estimate $\tilde{\psi}_k$ similar to the case of $\tilde{\phi}_k$. First we compute the natural gradient $\hat{\psi}_k = \omega_\psi + N\tilde{z}_i^k \boldsymbol{y}_i$ and update $\tilde{\psi}_k$ as

$$\tilde{\psi}_k^{(i+1)} = \left(1 - \rho_i\right)\tilde{\psi}_k^{(i)} + \rho_i \hat{\psi}_k. \tag{3}$$

Estimating $\tilde{\pi}$ The full conditional for the proportions follows a standard stick-breaking construction

$$p\left(\pi_k \mid \alpha, z\right) = \text{Beta}\left(1 + \sum_{i=1}^{N} z_i^k, \alpha + \sum_{i=1}^{N}\sum_{j>k} z_i^j\right).$$

Then, the variational distribution is $q\left(\pi_K \mid \tilde{\pi}_K\right) = \text{Beta}\left(1 + \sum_{i=1}^{N} \tilde{z}_i^k, \alpha + \sum_{i=1}^{N}\sum_{j>k} \tilde{z}_i^j\right)$. In the batch setting, we set $\tilde{\pi}_K$ to the expected natural parameter of its complete conditional distribution that is

$$\begin{aligned}
\tilde{\pi}_K &= \mathop{\mathbb{E}}_{q}\left[\eta_g\left(\boldsymbol{x}_i, \boldsymbol{y}_i, \phi, \psi, \pi\right)\right] \\
&= \left(1 + \sum_{i=1}^{N} \mathop{\mathbb{E}}_{q}\left[z_i^k\right], \alpha + \sum_{i=1}^{N}\sum_{j>k} \mathop{\mathbb{E}}_{q}\left[z_i^j\right]\right) \\
&= \left(1 + \sum_{i=1}^{N} \tilde{z}_i^k, \alpha + \sum_{i=1}^{N}\sum_{j>k} \tilde{z}_i^j\right)
\end{aligned}$$

where the natural parameter in Beta distribution results in a 2-dimensional vector. For stochastic setting, we compute the natural gradient vector as $\hat{\pi} = \left(1 + N\tilde{z}_i^k, \alpha + N\sum_{j=k+1}^{K} \tilde{z}_i^j\right)$ and update

$$\tilde{\pi}_k^{(i+1)} = \left(1 - \rho_t\right)\tilde{\pi}^{(i)} + \rho_t \tilde{\pi}. \tag{4}$$

2. Posterior sampling for $\mathbf{w}_c$

We next present the posterior sampling for $\mathbf{w}_c$ using Bayesian settings of Support Vector Machine Polson et al. (2011); Nguyen et al. (2016a) and Logistic Regression Polson et al. (2013); Nguyen et al. (2016b). Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ be the training data that $\boldsymbol{x}_i \in \mathbb{R}^D$ represents the feature vector and $y_i \in \{-1, 1\}$ represents the corresponding observed label. Notation-wise we further write $\boldsymbol{x}$ to denote $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ and $\boldsymbol{y} = \{y_1, \ldots, y_N\}$.

2.1. Support Vector Machine

Minimizing the regularized Hinge-loss objective function for a standard SVM $\mathcal{L}(\mathbf{w}; C) = \sum_{i=1}^N 2\max\left\{1 - y_i \mathbf{w}^T \boldsymbol{x}_i, 0\right\} + C||\mathbf{w}||_2^2$ (where $\mathbf{w}$ is the vector of coefficient parameters and $C > 0$ is the regularization hyper-parameter) is equivalent to a MAP estimation for the following pseudo-posterior distribution defined for $\mathbf{w}$ due to the monotonic property of the exponential:

$$\begin{aligned}\hat{\mathbf{w}}_{\text{MAP}} &= \underset{\mathbf{w}}{\text{argmax}}\, p(\mathbf{w} \mid \boldsymbol{x}, \boldsymbol{y}, C) \\ &= \underset{\mathbf{w}}{\text{argmax}} \prod_{i=1}^N \exp\left\{-2\max\left(1 - y_i \mathbf{w}^T \boldsymbol{x}_i, 0\right)\right\} \exp\left\{-C||\mathbf{w}||_2^2\right\}. \end{aligned} \tag{5}$$

Thus, $p(y_i \mid \boldsymbol{x}_i, \mathbf{w}) \propto \exp\left\{-2\max\left(1 - y_i \boldsymbol{\eta}^T \boldsymbol{x}_i, 0\right)\right\}$ can be viewed as the likelihood and $p(\mathbf{w} \mid C) \propto \exp\left\{-C||\mathbf{w}||_2^2\right\}$ as a prior distribution over $\mathbf{w}$. Gaussian prior can be rewritten as $p(\mathbf{w} \mid C) = \mathcal{N}(\mu_0, \Sigma_0)$ with mean $\boldsymbol{\mu}_0 = 0$ and $\boldsymbol{\Sigma}_0 = w\mathbf{I}$ where $w = \frac{1}{2C}$.

In a Bayesian setting, we would like to sample from Eq. (5). However, the likelihood term renders it difficult to achieve this goal. The key idea from the work of (Polson et al., 2011) is to 'data augment' each data point $\boldsymbol{x}_i$ with an auxiliary variables $\lambda_i > 0$ so that the individual likelihood term can be written as

$$p(y_i \mid \boldsymbol{x}_i, \mathbf{w}) \propto \exp\left\{-2\max\left(1 - y_i \mathbf{w}^\top \boldsymbol{x}_i, 0\right)\right\} = \int \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left\{-\frac{\left[\lambda_i + \left(1 - y_i \mathbf{w}^\top \boldsymbol{x}_i\right)\right]^2}{2\lambda_i}\right\} d\lambda_i.$$

The term inside the integral can then be viewed as the joint distribution $p(y_i, \lambda_i \mid \boldsymbol{x}_i, \mathbf{w})$ (over which marginalized λ_i will recover the likelihood for y_i). Let $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_n\}$, then the posterior $p(\mathbf{w} \mid \boldsymbol{x}, \boldsymbol{y}, C)$ can be viewed as the marginal from a joint posterior with the auxiliary variables

$$\begin{aligned} p(\mathbf{w}, \boldsymbol{\lambda} \mid \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \propto \prod_{i=1}^N \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left\{-\frac{\left[\lambda_i + \left(1 - y_i \mathbf{w}^\top \boldsymbol{x}_i\right)\right]^2}{2\lambda_i}\right\} \\ \times \exp\left(-\frac{1}{2}\left[\mathbf{w} - \boldsymbol{\mu}_0\right]^T \boldsymbol{\Sigma}_0^{-1}\left[\mathbf{w} - \boldsymbol{\mu}_0\right]\right). \end{aligned}$$

Gibbs sampling can now be performed on this joint posterior by sequentially sampling $\mathbf{w}$ given λ_i and vice versa. Completing the square, we get $\mathbf{w} \mid \boldsymbol{\lambda} \sim \mathcal{N}(\mu_N, \Sigma_N)$ where $\boldsymbol{\Sigma}_N^{-1} = \sum_{i=1}^N \frac{1}{\lambda_i} \boldsymbol{x}_i \boldsymbol{x}_i^T + \boldsymbol{\Sigma}_0^{-1}$ and $\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N \left(\sum_{i=1}^N \frac{\lambda_i + 1}{\lambda_i} y_i \boldsymbol{x}_i\right)$. We sample λ_i given $\mathbf{w}$ as $\lambda_i^{-1} \mid \mathbf{w} \sim IG\left(|1 - y_i \mathbf{w}^T \boldsymbol{x}_i|^{-1}, 1\right)$, where we note that λ_i(s) are independent given $\mathbf{w}$.

2.2. Logistic Regression

Under a Bayesian setting, we define a prior distribution for $\mathbf{w}$: $p(\mathbf{w} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. Then, the posterior of $\mathbf{w}$ is computed as:

$$\log p(\mathbf{w}) \propto -\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\mathbf{w}-\boldsymbol{\mu}_0) + \sum_{i=1}^{N}[y_i \log \sigma_i + (1-y_i)(1-\log \sigma_i)] \quad (6)$$

We assume the final form of $f(\mathbf{w}) \sim \mathcal{N}(\mathbf{w}_{\text{MAP}}, \Sigma_{\text{MAP}})$. Since the posterior inference for Bayesian logistic regression (Bishop, 2006; Nguyen et al., 2016b) is not in closed-form update, Polson *et al* (Polson et al., 2013) utilize a new class of Polya-Gamma distribution to propose the augmentation strategy for closed-form Bayesian inference of logistic regression. We utilize the interesting property of sigmoid function (Polson et al., 2013):

$$\frac{\exp(\mathbf{w}^T\boldsymbol{x})^y}{1+\exp(\mathbf{w}^T\boldsymbol{x})} = \frac{1}{2}\exp(\mathbf{w}^T\boldsymbol{x}\kappa)\int_0^\infty \exp\left(-\frac{\lambda}{2}\left[\mathbf{w}^T\boldsymbol{x}\right]^2\right)p(\lambda)\,d\lambda \quad (7)$$

where $\kappa = y - \frac{1}{2}$ and $\lambda \sim PG(b,c)$ follows the Polya-Gamma distribution. We obtain the posterior of $\mathbf{w}$ by plugging Eq. (7) into Eq. (6),

$$\log p(\mathbf{w}) \propto -\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\mathbf{w}-\boldsymbol{\mu}_0) + \sum_{i=1}^{N}\left(\kappa_i \mathbf{w}^T\boldsymbol{x}_i - \int \frac{\lambda_i}{2}\left[\mathbf{w}^T\boldsymbol{x}_i\right]^2 \log p(\lambda_i)\,d\lambda_i\right).$$

Let $\boldsymbol{\lambda} = \{\lambda_1, ..., \lambda_N\}$, the posterior $\log p(\mathbf{w})$ in Eq. (6) can be viewed as the marginal from a joint posterior with the auxiliary variables as

$$\log p(\mathbf{w}, \boldsymbol{\lambda}) \propto \sum_{i=1}^{N}\left(\kappa_i \mathbf{w}^T\boldsymbol{x}_i - \frac{\lambda_i}{2}\mathbf{w}^T\boldsymbol{x}_i\right) - \frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\mathbf{w}-\boldsymbol{\mu}_0).$$

Gibbs sampling can be performed on this joint posterior by sequentially sampling $\mathbf{w}$ given λ_i and vice versa. Completing the square of the above equation, we get:

$$p(\lambda_i \mid \boldsymbol{x}_i, \mathbf{w}) \sim PG\left(1, \boldsymbol{x}_i^T\mathbf{w}\right) \quad , \forall i = 1...N \quad (8)$$

$$p(\mathbf{w} \mid \boldsymbol{y}, \boldsymbol{\lambda}) \sim \mathcal{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N) \quad (9)$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]^T$, $\boldsymbol{\Sigma}_N = \left(\boldsymbol{X}^{\boldsymbol{T}}\text{diag}(\boldsymbol{\lambda})\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$, $\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N\left(\boldsymbol{X}^T\kappa + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$ and $\kappa = [y_1 - 1/2, ..., y_N - 1/2]^T$.

References

C. M. Bishop. *Pattern recognition and machine learning*. springer New York, 2006.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

T. D. Nguyen, V. Nguyen, T. Le, and D. Phung. Distributed data augmented support vector machine on spark. In *Pattern Recognition (ICPR), 2016 23st International Conference on*. IEEE, 2016a.

V. Nguyen, T. D. Nguyen, T. Le, S. Venkatesh, and D. Phung. One-pass logistic regression for label-drift and large-scale classification on distributed systems. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, Spain, 2016b.

N. G. Polson, S. L. Scott, et al. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–23, 2011.

N. G. Polson, J. G. Scott, and J. Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504): 1339–1349, 2013.