A Conditional Multinomial Mixture Model for Superset Label Learning (Supplementary Materials)

Li-Ping Liu
EECS, Oregon State University
Corvallis, OR 97331
liuli@eecs.oregonstate.edu

Thomas G. Dietterich
EECS, Oregon State University
Corvallis, OR 97331
tgd@cs.orst.edu

1 The Model

In this supplement paper, we show the detailed derivation of LSB-CMM.

The generative process of the whole model is as below and the plate representation is shown in (1).

$$\begin{aligned}
\mathbf{w}_k &\sim \text{Normal}(0, \Sigma), 1 \le k \le K-1, \;\; \mathbf{w}_K = (+\infty, 0, \cdots, 0) && (1)\\
z_n &\sim \text{Mult}(\phi_n), \quad \phi_{nk} = \text{expit}(\mathbf{w}_k^T \mathbf{x}_n) \prod_{i=1}^{k-1} (1 - \text{expit}(\mathbf{w}_i^T \mathbf{x}_n)) && (2)\\
\theta_k &\sim \text{Dirichlet}(\alpha) && (3)\\
y_n &\sim \text{Mult}(\theta_{z_n}) && (4)\\
Y_n &\sim \text{Dist1}(y_n) \quad \text{(Dist1 is some distribution satisfying the assumption in the paper} && (5)
\end{aligned}$$

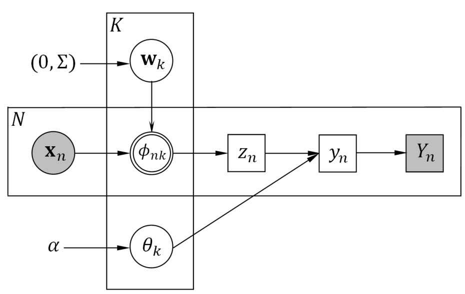

Figure 1: The LSB-CMM. Square nodes are discrete, circle nodes are continuous, and double-circle nodes are deterministic.

The model needs to maximize the likelihood that each y_n is in Y_n. After incorporating the priors, we can write the penalized maximum likelihood objective as

$$\max LL = \sum_{n=1}^{N} \log \left(\sum_{y_n \in Y_n} p(y_n | \mathbf{x}_n, \mathbf{w}, \alpha) \right) + \log(p(\mathbf{w}|0, \Sigma)). \tag{6}$$

This cannot be solved directly, so we apply variational EM.

1.1 Variational EM

The hidden variables in the model are y, z, and θ. For these hidden variables, we introduce the variational distribution $q(y, z, \theta | \hat{\phi}, \hat{\alpha})$, where $\hat{\phi} = \{\hat{\phi}_n\}_{n=1}^N$ and $\hat{\alpha} = \{\hat{\alpha}_k\}_{k=1}^K$ are the parameters.

Then we factorize q as

$$q(z,y,\theta|\hat{\phi},\hat{\alpha}) = \prod_{n=1}^{N} q(z_n, y_n|\hat{\phi}_n) \prod_{k=1}^{K} q(\theta_k|\hat{\alpha}_k), \tag{7}$$

where $\hat{\phi}_n$ is a $K \times L$ matrix and $q(z_n, y_n|\hat{\phi}_n)$ is a multinomial distribution in which $p(z_n = k, y_n = l) = \hat{\phi}_{nkl}$. This distribution is constrained by the candidate label set: if a label $l \notin Y_n$, then $\hat{\phi}_{nkl} = 0$ for any value of k. The distribution $q(\theta_k|\hat{\alpha}_k)$ is a Dirichlet distribution with parameter $\hat{\alpha}_k$.

With Jensen's inequality, the lower bound of the log likelihood is

$$\begin{aligned} LL &\geq E[\log p(z,y,\theta|\mathbf{x},\mathbf{w},\alpha)] - E[\log q(z,y,\theta|\hat{\phi},\hat{\alpha})] + \log(p(\mathbf{w}|0,\Sigma)) \\ &= \sum_{n=1}^{N} E[\log p(z_n|\mathbf{x}_n,\mathbf{w})] + \sum_{k=1}^{K} E[\log p(\theta_k|\alpha)] + \sum_{n=1}^{N} E[\log p(y_n|z_n,\theta)] \\ &\quad - \sum_{n=1}^{N} E[\log q(y_n, z_n|\hat{\phi}_n)] - \sum_{k=1}^{K} E[\log q(\theta_k|\hat{\alpha}_k)] + \log(p(\mathbf{w}|0,\Sigma)), \end{aligned} \tag{8}$$

where $E[\cdot]$ is the expectation under the variational distribution $q(z,y,\theta|\hat{\phi},\hat{\alpha})$.

Expand the expectation in the first, second and third term.

$$E[\log p(z_n|\mathbf{x}_n,\mathbf{w})] = \sum_{k=1}^{K}\sum_{l=1}^{L} \hat{\phi}_{nkl} \log(\phi_{nk}), \tag{9}$$

$$E[\log p(y_n|z_n,\theta)] = \sum_{k=1}^{K}\sum_{l=1}^{L} \hat{\phi}_{nkl} \int_{\theta_k} Dir(\theta_k;\hat{\alpha}_k) \log \theta_{kl} d\theta_k, \tag{10}$$

$$E[\log p(\theta_k|\alpha)] \propto \int_{\theta_k} Dir(\theta_k;\hat{\alpha}_k) \sum_{l=1}^{L} (\alpha - 1) \log \theta_{kl} d\theta_k, \tag{11}$$

where $Dir(\theta_k;\hat{\alpha}_k)$ is the density at θ_k of the Dirichlet distribution with $\hat{\alpha}_k$.

In the E step, this lower bound is maximized with respect to $\hat{\phi}$ and $\hat{\alpha}$. Each $\hat{\phi}_n$ can be optimized separately. Adding all terms involving $\hat{\phi}_n$ (i.e. the first, third and the fourth terms), we obtain

$$\sum_{k=1}^{K}\sum_{l=1}^{L} \hat{\phi}_{nkl} \log\left(\phi_{nk} \exp(E_{q(\theta_k|\hat{\alpha}_k)}[\log(\theta_{kl})])\right) - \hat{\phi}_{nkl}\log(\hat{\phi}_{nkl}), \tag{12}$$

Maximizing the term (12) is equivalent to minimizing the KL divergence between $\hat{\phi}_n$ and the term in the first logarithm function. With the constraint imposed by the candidate label set, the updating formula for $\hat{\phi}_n$ is (13). The update of $\hat{\alpha}_k$ for each k follows the standard procedure for variational inference in the exponential family and is shown in (14).

$$\hat{\phi}_{nkl} \propto \begin{cases} \phi_{nk} \ \exp\left(E_{q(\theta_k|\hat{\alpha}_k)}\left[\log(\theta_{kl})\right]\right), & \text{if } l \in Y_n \\ 0, & \text{if } l \notin Y_n \end{cases} \tag{13}$$

$$\hat{\alpha}_k = \alpha + \sum_{n=1}^{N} \hat{\phi}_{nkl}, \tag{14}$$

We calculate the expectation of $\log(\theta_{kl})$ via Monte Carlo sampling.

In the M step, the lower bound is maximized with respect to $\mathbf{w}$. Only the first and the last terms in the lower bound are related to $\mathbf{w}$, and each $\mathbf{w}_k, 1 \leq k \leq K-1$, can be maximized separately. After some derivation, we obtain the optimization problem in Eq. (15), which is similar to the problem of logistic regression. It is a concave maximization problem, so any gradient based method, such as BFGS, can find the global optimum.

$$\max_{\mathbf{w}_k} \ -\frac{1}{2}\mathbf{w}_k^T \Sigma^{-1} \mathbf{w}_k + \sum_{n=1}^{N} \Big[\hat{\phi}_{nk} \log(\text{expit}(\mathbf{w}_k^T \mathbf{x}_n)) + \hat{\psi}_{nk} \log(1 - \text{expit}(\mathbf{w}_k^T \mathbf{x}_n))\Big], \tag{15}$$

where $\hat{\phi}_{nk} = \sum_{l=1}^{L} \hat{\phi}_{nkl}$ and $\hat{\psi}_{nk} = \sum_{j=k+1}^{K} \hat{\phi}_{nj}$.

1.2 Prediction

For a test instance $\mathbf{x}_t$, we predict the label with maximum posterior probability. The test instance can be mapped to a topic, but there is no coding matrix θ from the EM solution. We use the variational distribution $p(\theta_k|\hat{\alpha}_k)$ as the prior of each θ_k and integrate out all θ_ks. Given a test sample $\mathbf{x}_t$, the prediction l that maximizes the probability $p(y_t = l|\mathbf{x}_t, \mathbf{w}, \hat{\alpha})$ can be calculated as

$$\begin{aligned} p(y_t = l|\mathbf{x}_t, \mathbf{w}, \hat{\alpha}) &= \sum_{k=1}^{K} \int_{\theta_k} p(y_t = l, z_t = k, \theta_k|\mathbf{x}_t, \mathbf{w}, \hat{\alpha}) d\theta_k \\ &= \sum_{k=1}^{K} p(z_t = k|\mathbf{x}_t, \mathbf{w}) \int_{\theta_k} p(\theta_k|\hat{\alpha}_k) p(y_t = l|\theta_k) d\theta_k \\ &= \sum_{k=1}^{K} \phi_{tk} \frac{\hat{\alpha}_{kl}}{\sum_l \hat{\alpha}_{kl}} , \end{aligned} \tag{16}$$

where $\phi_{tk} = \Big(\text{expit}(\mathbf{w}_k^T \mathbf{x}_t) \prod_{i=1}^{k-1} (1 - \text{expit}(\mathbf{w}_i^T \mathbf{x}_t))\Big)$.