Appendix for *"Learning to Aggregate Ordinal Labels by Maximizing Separating Width"*

Guangyong Chen [1] **Shengyu Zhang** [1] **Di Lin** [2] **Hui Huang** [2] **Pheng Ann Heng** [1]

Abstract

In this Appendix, we present the derivations of procedures outlined in Algorithm 1, and discuss the performance of formulating $f_t(R^i)$ as $\text{Tr}(W_t R^i)$ with $W_t \in \mathbb{R}^{K\times M}$, instead of $a_t^T R^i b_t$.

1. Derivations of Algorithm 1

In the manuscript, we have presented the derivations of sampling the true label z_i, as well as its corresponding augmented variable $\{r_{it}\}_{t=1}^{K-1}$. In this supplementary file, we should discuss the derivations of sampling other model parameters, and the gradient method employed to optimize the $K-1$ decision boundaries and update prior distributions.

For the confusion matrices, $\boldsymbol{A}$. Given the prior distribution over the k-th row of A^j, which is a conjugate Dirichlet distribution, we can get its conditional distribution also follows a Dirichlet distribution, and can be reformulated a,

$$A_{kd}^j \sim D(A_{kd}^j|\alpha_j + \sum_{i=1}^{N} R_{jd}^i \mathbb{I}(z_i = k)). \tag{1}$$

For the items' difficulties, $\boldsymbol{\omega}$. The posterior distribution of ω_i can be derived in a similar way as A^j, which is also Dirichlet distribution as,

$$\omega_k^i \sim D(\omega_k^i|\beta_i + \mathbb{I}(z_i = k)). \tag{2}$$

For $a_t, b_t, \forall t \in [K-1]$. Given the updating procedures of a_t, the updating procedures associated with b_t can be achieved by altering the notations. Thus, we only discuss the updating of a_t here. Given the estimations of all other parameters, we can rewrite the objective function with respect to a_t as follows,

$$\begin{aligned}\mathcal{L}(a_t) &\propto -\mathbb{E}_q \ln \phi(\boldsymbol{z}, \boldsymbol{\gamma}|\boldsymbol{R}) + \lambda_1 a_t^T a_t b_t^T b_t \\ &\propto -\sum_{i=1}^{N} \mathbb{E}_q \frac{1}{2\gamma_{it}}(\gamma_i + \lambda_2 \zeta_{it})^2 + \lambda_1 a_t^T a_t b_t^T b_t.\end{aligned} \tag{3}$$

Thus, by setting $\frac{\partial \mathcal{L}(a_t)}{\partial a_t} = 0$, we can obtain the optimal solution of a_t as follows,

$$\begin{aligned}\Sigma_{a_t} &= 2\lambda_1 \|b_t\|_2^2 \boldsymbol{I} + \sum_{i=1}^{N} \frac{\lambda_2^2}{\langle \gamma_{it} \rangle} R^i b_t b_t^T R^{i^T}, \\ a_t &= \Sigma_{a_t}^{-1} (\sum_{i=1}^{N} (\lambda_2 + \frac{\lambda_2^2}{\langle \gamma_{it} \rangle}) \langle \text{sgn}_t(z_i) \rangle R^i b_t),\end{aligned} \tag{4}$$

where $\langle f(x) \rangle = \int q(x) f(x) dx$ can be estimated during the sampling steps.

For the updating of prior distributions. Given the updating procedures of α, the updating procedures associated with β can be achieved by altering the notations. Thus, we only discuss the updating of α here. By fixing all other parameters, we can obtain an objective function with respect to α as follows,

$$\begin{aligned}\mathcal{L}(\alpha_j) =& K(\ln \Gamma(K\alpha_j) - K \ln \Gamma(\alpha_j)) \\ &+ \sum_{k,d} (\alpha - 1) \mathbb{E}_{q(\boldsymbol{A})} \ln A_{kd}^j.\end{aligned} \tag{5}$$

Thus, we have the gradient $\frac{\partial \mathcal{L}(\alpha_j)}{\partial \alpha_j}$ as

$$\frac{\partial \mathcal{L}(\alpha_j)}{\partial \alpha_j} = K^2(\psi(K\alpha_j) - \psi(\alpha_j)) + \sum_{k,d} \langle \ln A_{kd}^j \rangle,$$

where $\psi(\cdot)$ is a digamma function. Thus, we can minimize $\mathcal{L}$ by setting

$$\alpha_j = \alpha_j - \eta \frac{\partial \mathcal{L}(\alpha_j)}{\partial \alpha_j}, \tag{6}$$

with η as a learning rate.

Thus, we can get the procedures outlined in Algorithm 1.

*Equal contribution [1]The Chinese University of Hong Kong, Hong Kong, China. [2]Shenzhen University, China. Correspondence to: Shengyu Zhang <syzhang@cse.cuhk.edu.hk>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

Table 1. Errors in predicting the latent labels on the Web dataset.

Ordinal Dataset		Ours $f_t(R^i) = a_t^T R^i b_t$	Ours $f_t(R^i) = \text{Tr}(W_t R^i)$	G-CrowdSVM	Entropy(O)	MV-DS	MV
Web	l_0	**0.0322**±0.0013	0.1153±0.0027	0.0799±0.0026	0.1040	0.1574	0.2693
	l_1	**0.0369**±0.0032	0.1432±0.0015	0.0940±0.0057	0.1173	0.2149	0.4251
	l_2	**0.2153**±0.0019	0.4594±0.0062	0.3629±0.0044	0.3816	0.5358	0.9247

2. Discussions on Setting $f_t(R^i) = \textbf{Tr}(W_t R^i)$

In the manuscript, we assume $f_t(R^i) = a_t^T R^i b_t$ for $t \in [K-1]$, which is inspired by the rank-1 formulas used in MV and WMV. However, one may wonder why confining to such rank-1 measurements. In this part, we give the optimization of the separating width with $f_t(R^i) = \text{Tr}(W_t R^i)$, and present the experiments on the Web dataset in comparison with $f_t(R^i) = a_t^T R^i b_t$.

For $W_t, \forall t \in [K-1]$. Let vec$(\cdot)$ denote a linear transformation which converts the matrix into a column vector. For example, for the 2×2 matrix $B = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, its vectorization is $\text{vec}(B) = \begin{bmatrix} a \\ c \\ b \\ d \end{bmatrix}$. Then, we can rewrite $f_t(R^i) = \text{vec}(W_t)^T \text{vec}(R^i)$, leading to an objective function with respect to W_t as follows,

$$\begin{aligned} \mathcal{L}(W_t) &\propto -\mathbb{E}_q \ln \phi(\boldsymbol{z}, \boldsymbol{\gamma} | \boldsymbol{R}) + \lambda_1 \|W_t\|_F \\ &\propto \sum_{i=1}^{N} \mathbb{E}_q \frac{1}{2\gamma_{it}} (\gamma_i + \lambda_2 \zeta_{it})^2 + \lambda_1 \text{vec}(W_t)^T \text{vec}(W_t), \end{aligned}$$

where $\zeta_{it} = 1 - \text{sgn}_t(z_i) \text{vec}(W_t)^T \text{vec}(R^i)$. Thus, by setting $\frac{\partial \mathcal{L}(W_t)}{\partial \text{vec}(W_t)} = 0$, we can obtain the optimal solution of W_t as follows,

$$\begin{aligned} \Pi_{W_t} &= 2\lambda_1 \boldsymbol{I} + \sum_{i=1}^{N} \frac{\lambda_2^2}{\langle \gamma_{it} \rangle} \text{vec}(R^i) \text{vec}(R^i)^T, \\ \text{vec}(W_t) &= \Sigma_{W_t}^{-1} (\sum_{i=1}^{N} (\lambda_2 + \frac{\lambda_2^2}{\langle \gamma_{it} \rangle}) \langle \text{sgn}_t(z_i) \rangle \text{vec}(R^i)). \end{aligned} \tag{7}$$

Other procedures are similar with the ones outlined in Algorithm 1.

To evaluate the performance of formulating $f_t(R^i) = \text{Tr}(W_t R^i)$, we implement its algorithm on the Web dataset. Its comparisons with other competitive ones can be found in Table 1, which shows that using higher rank measurements actually makes the performance worse. It can be explained that each W_t introduces MK free parameters, which is more than the free parameters introduced by a_t and b_t, only $M + K$. More free parameters may face an overfitting problem.