Supplementary Material: Relational Stacked Denoising Autoencoder for Tag Recommendation

Multi-Relational Stacked Denoising Autoencoder

Here we present a generalized version of RSDAE called multi-relational stacked denoising autoencoder (MRSDAE). This generalization allows the new model to handle multi-relational data. We assume that there are Q types of relational data (Q networks) and use q to denote any one type. The graphical model of MRSDAE is shown in Figure 1 and the generative process is listed as follows:

1. For each type of relational data (each of the Q networks), draw the *relational latent matrix* $\mathbf{S}^{(q)} = [\mathbf{s}_1^{(q)}, \mathbf{s}_2^{(q)}, \cdots, \mathbf{s}_J^{(q)}]$ from a *matrix variate normal distribution* (Gupta and Nagar 2000):
$$\mathbf{S}^{(q)} \sim \mathcal{N}_{K,J}(0, \mathbf{I}_K \otimes (\lambda_l \mathscr{L}_{aq})^{-1}). \tag{1}$$
2. For layer l of the SDAE network where $l = 1, 2, \dots, \frac{L}{2} - 1$,
   (a) For each column n of the weight matrix $\mathbf{W}_l$, draw $\mathbf{W}_{l,*n} \sim \mathcal{N}(0, \lambda_w^{-1}\mathbf{I}_{K_l})$.
   (b) Draw the bias vector $\mathbf{b}_l \sim \mathcal{N}(0, \lambda_w^{-1}\mathbf{I}_{K_l})$.
   (c) For each row j of $\mathbf{X}_l$, draw
$$\mathbf{X}_{l,j*} \sim \mathcal{N}(\sigma(\mathbf{X}_{l-1,j*}\mathbf{W}_l + \mathbf{b}_l), \lambda_s^{-1}\mathbf{I}_{K_l}).$$
3. For layer $\frac{L}{2}$ of the SDAE network, draw the representation vector for item j from the product of $Q+1$ Gaussians (PoG) (Gales and Airey 2006):
$$\mathbf{X}_{\frac{L}{2},j*} \sim \text{PoG}(\sigma(\mathbf{X}_{\frac{L}{2}-1,j*}\mathbf{W}_l + \mathbf{b}_l), (\mathbf{s}_j^1)^T, \dots, (\mathbf{s}_j^Q)^T, \lambda_s^{-1}\mathbf{I}_K, \lambda_r^{-1}\mathbf{I}_K, \dots, \lambda_r^{-1}\mathbf{I}_K). \tag{2}$$
4. For layer l of the SDAE network where $l = \frac{L}{2}+1, \frac{L}{2}+2, \dots, L$,
   (a) For each column n of the weight matrix $\mathbf{W}_l$, draw $\mathbf{W}_{l,*n} \sim \mathcal{N}(0, \lambda_w^{-1}\mathbf{I}_{K_l})$.
   (b) Draw the bias vector $\mathbf{b}_l \sim \mathcal{N}(0, \lambda_w^{-1}\mathbf{I}_{K_l})$.
   (c) For each row j of $\mathbf{X}_l$, draw
$$\mathbf{X}_{l,j*} \sim \mathcal{N}(\sigma(\mathbf{X}_{l-1,j*}\mathbf{W}_l + \mathbf{b}_l), \lambda_s^{-1}\mathbf{I}_{K_l}).$$
5. For each item j, draw a clean input
$$\mathbf{X}_{c,j*} \sim \mathcal{N}(\mathbf{X}_{L,j*}, \lambda_n^{-1}\mathbf{I}_B).$$

Here $K = K_{\frac{L}{2}}$ is the dimensionality of the learned representation vector for each item. $\mathbf{S}^{(q)}$ denotes the $K \times J$ relational latent matrix in which column j is the *relational latent vector* $\mathbf{s}_j^{(q)}$ for item j. Note that $\mathcal{N}_{K,J}(0, \mathbf{I}_K \otimes (\lambda_l \mathscr{L}_{aq})^{-1})$ in (1) is a *matrix variate normal distribution* defined as (Gupta and Nagar 2000):

$$\begin{aligned} p(\mathbf{S}^{(q)}) &= \mathcal{N}_{K,J}(0, \mathbf{I}_K \otimes (\lambda_l \mathscr{L}_{aq})^{-1}) \\ &= \frac{\exp\{\text{tr}[-\frac{\lambda_l}{2}\mathbf{S}^{(q)}\mathscr{L}_{aq}(\mathbf{S}^{(q)})^T]\}}{(2\pi)^{JK/2}|\mathbf{I}_K|^{J/2}|\lambda_l\mathscr{L}_{aq}|^{-K/2}}, \end{aligned} \tag{3}$$

where the operator $\otimes$ denotes the Kronecker product of two matrices (Gupta and Nagar 2000), $\text{tr}(\cdot)$ denotes the trace of a matrix, and $\mathscr{L}_{aq}$ is the Laplacian matrix incorporating the qth type of relational data. $\mathscr{L}_{aq} = \mathbf{D}^{(q)} - \mathbf{A}^{(q)}$, where $\mathbf{D}^{(q)}$ is a diagonal matrix whose diagonal elements $\mathbf{D}_{ii}^{(q)} = \sum_j \mathbf{A}_{ij}^{(q)}$ and $\mathbf{A}^{(q)}$ is the adjacency matrix of the qth type of relational data with binary entries indicating the links (or relations) between items. $\mathbf{A}_{jj'}^{(q)} = 1$ indicates that there is a link between item j and item j' and $\mathbf{A}_{jj'}^{(q)} = 0$ otherwise. Equation (2) denotes the product of the Gaussian $\mathcal{N}(\sigma(\mathbf{X}_{\frac{L}{2}-1,j*}\mathbf{W}_l + \mathbf{b}_l), \lambda_s^{-1}\mathbf{I}_K)$ and Q Gaussians of the form $\mathcal{N}((\mathbf{s}_j^{(q)})^T, \lambda_r^{-1}\mathbf{I}_K)$, which is also a Gaussian (Gales and Airey 2006).

According to the generative process above, maximizing the posterior probability is equivalent to maximizing the joint log-likelihood of $\{\mathbf{X}_l\}$, $\mathbf{X}_c$, $\{\mathbf{S}^{(q)}\}$, $\{\mathbf{W}_l\}$, and $\{\mathbf{b}_l\}$

given λ_s, λ_w, λ_l, λ_r, and λ_n:

$$\begin{aligned}\mathscr{L} = &-\frac{\lambda_l}{2}\sum_q \mathrm{tr}(\mathbf{S}^{(q)}\mathscr{L}_{aq}(\mathbf{S}^{(q)})^T) \\ &-\frac{\lambda_r}{2}\sum_q\sum_j \|((\mathbf{s}_j^{(q)})^T - \mathbf{X}_{\frac{L}{2},j*})\|_2^2 \\ &-\frac{\lambda_w}{2}\sum_l(\|\mathbf{W}_l\|_F^2 + \|\mathbf{b}_l\|_2^2) \\ &-\frac{\lambda_n}{2}\sum_j \|\mathbf{X}_{L,j*} - \mathbf{X}_{c,j*}\|_2^2 \\ &-\frac{\lambda_s}{2}\sum_l\sum_j \|\sigma(\mathbf{X}_{l-1,j*}\mathbf{W}_l + \mathbf{b}_l) - \mathbf{X}_{l,j*}\|_2^2.\end{aligned}$$

Similar to the generalized SDAE, taking λ_s to infinity, the joint log-likelihood becomes:

$$\begin{aligned}\mathscr{L} = &-\frac{\lambda_l}{2}\sum_q \mathrm{tr}(\mathbf{S}^{(q)}\mathscr{L}_{aq}(\mathbf{S}^{(q)})^T) \\ &-\frac{\lambda_r}{2}\sum_q\sum_j \|((\mathbf{s}_j^{(q)})^T - \mathbf{X}_{\frac{L}{2},j*})\|_2^2 \\ &-\frac{\lambda_w}{2}\sum_l(\|\mathbf{W}_l\|_F^2 + \|\mathbf{b}_l\|_2^2) \\ &-\frac{\lambda_n}{2}\sum_j \|\mathbf{X}_{L,j*} - \mathbf{X}_{c,j*}\|_2^2, \end{aligned} \tag{4}$$

where $\mathbf{X}_{l,j*} = \sigma(\mathbf{X}_{l-1,j*}\mathbf{W}_l + \mathbf{b}_l)$. Note that by simple manipulation, we have

$$\mathrm{tr}(\mathbf{S}^{(q)}\mathscr{L}_{aq}(\mathbf{S}^{(q)})^T) = \frac{1}{2}\sum_{j=1}^{J}\sum_{j'=1}^{J}\mathbf{A}_{jj'}\|\mathbf{S}_{*j}^{(q)} - \mathbf{S}_{*j'}^{(q)}\|^2 \tag{5}$$

$$\begin{aligned}&= \frac{1}{2}\sum_{j=1}^{J}\sum_{j'=1}^{J}[\mathbf{A}_{jj'}\sum_{k=1}^{K}(\mathbf{S}_{kj}^{(q)} - \mathbf{S}_{kj'}^{(q)})^2] \\ &= \frac{1}{2}\sum_{k=1}^{K}[\sum_{j=1}^{J}\sum_{j'=1}^{J}\mathbf{A}_{jj'}(\mathbf{S}_{kj}^{(q)} - \mathbf{S}_{kj'}^{(q)})^2] \\ &= \sum_{k=1}^{K}(\mathbf{S}_{k*}^{(q)})^T\mathscr{L}_{aq}\mathbf{S}_{k*}^{(q)},\end{aligned}$$

where $\mathbf{S}_{r*}^{(q)}$ denotes the rth row of $\mathbf{S}^{(q)}$ and $\mathbf{S}_{*c}^{(q)}$ denotes the cth column of $\mathbf{S}^{(q)}$. As we can see, maximizing $-\frac{\lambda_l}{2}\mathrm{tr}((\mathbf{S}^{(q)})^T\mathscr{L}_{aq}\mathbf{S}^{(q)})$ is equivalent to making $\mathbf{s}_j^{(q)}$ closer to $\mathbf{s}_{j'}^{(q)}$ if item j and item j' are linked (namely $\mathbf{A}_{jj'} = 1$).

The learning procedure of MRDSAE can also be derived similarly.

Sensitivity to Hyperparameters

Figure 2 shows the sensitivity of RSDAE's performance to the hyperparameter λ_l for *movielens-plot* in the sparse setting ($P = 1$). $\lambda_r = 1$ and $\lambda_n = 1$. As we can see, the recall is not very sensitive over a wide range of values either.

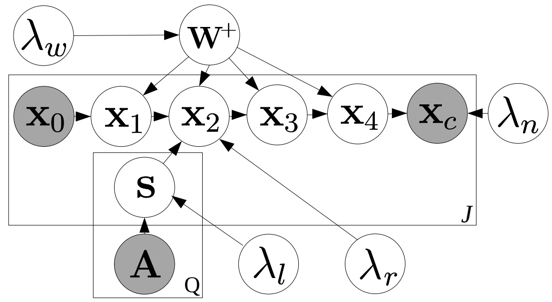

Figure 1: Graphical model of MRSDAE when $L = 4$ and there are two types of relational data. λ_s is not shown here to prevent clutter.

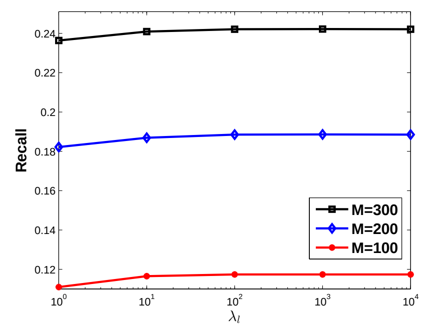

Figure 2: Effect of λ_l in RSDAE.

References

Gales, M. J. F., and Airey, S. S. 2006. Product of gaussians for speech recognition. *CSL* 20(1):22–40.

Gupta, A., and Nagar, D. 2000. *Matrix Variate Distributions*. Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics. Chapman & Hall.