Notes on Collaborative Denoising Auto-Encoders (CDAE) for OCCF

Wei Dai, Weike Pan
College of Computer Science and Software Engineering, Shenzhen University

1 Reference

[1]

2 Notations

Table 1: Some notations.

n	user number
m	item number
$u \in \{1, 2, \ldots, n\}$	user ID
$i, i' \in \{1, 2, \ldots, m\}$	item ID
$r_{ui} \in \{0, 1\}$	observed rating of user u on item i
$\tilde{r}_{ui}$	corrupted rating of user u on item i
$\hat{r}_{ui}$	predicted rating of user u on item i
$\mathcal{P}$	the whole set of observed (user, item) pairs
$\mathcal{A}$, $\lvert\mathcal{A}_u\rvert = \rho\lvert\mathcal{P}_u\rvert$	a sampled set of unobserved (user, item) pairs
b_i	item bias
d	number of latent dimensions (i.e., number of neurons in the hidden layer)
$\boldsymbol{b}^h \in \mathbb{R}^{1\times d}$	bias vector of the hidden layer
$V_{i\cdot}, W_{i'\cdot} \in \mathbb{R}^{1\times d}$	item-specific latent feature vector of item i and i'
$U_{u\cdot} \in \mathbb{R}^{1\times d}$	user-specific latent feature vector of user u
$\bar{U}_{u\cdot}, \tilde{U}_{u\cdot}, Z_{u\cdot} \in \mathbb{R}^{1\times d}$	user-specific latent feature vector of virtual user u
I_u^{re}	recommended items for user u

3 Background

3.1 Prediction rule of AE

The predicted rating of user u on item i is as follows,

$$\hat{r}_{ui} = f(Z_{u\cdot}V_{i\cdot}^T + b_i) \tag{1}$$

where $f(\cdot)$ is the sigmoid function $\sigma(x) = 1/(1+\exp(-x))$, $b_i \in \mathbb{R}^{1\times 1}$ is the item bias of item i, and $Z_{u\cdot} \in \mathbb{R}^{1\times d}$ is the user-specific latent feature vector of **virtual user** u:

$$Z_{u\cdot} = h(\bar{U}_{u\cdot} + \boldsymbol{b}^h) \tag{2}$$

where $h(\cdot)$ is the sigmoid function $\sigma(x) = 1/(1+\exp(-x))$, $\boldsymbol{b}^h \in \mathbb{R}^{1\times d}$ is the bias vector of the hidden layer, and $\bar{U}_{u\cdot} \in \mathbb{R}^{1\times d}$ is the user-specific latent feature vector of **virtual user** u:

$$\bar{U}_{u\cdot} = \sum_{1\le i'\le m, r_{ui'}=1} W_{i'\cdot} = \sum_{i'\in\mathcal{P}_u} W_{i'\cdot} \tag{3}$$

where $W_{i'\cdot} \in \mathbb{R}^{1\times d}$ is the item-specific latent feature vector of item i'.

The model parameters to be learned in AE: $\{V_{i\cdot}, W_{i\cdot}, b_i, \boldsymbol{b}^h | i = 1, 2\ldots, m\}$

3.2 Prediction rule of DAE

The predicted rating of user u on item i is as follows,

$$\hat{r}_{ui} = f(Z_{u\cdot}V_{i\cdot}^T + b_i) \tag{4}$$

where $f(\cdot)$ is a mapping function such as an identity function $\delta(x) = x$ and a sigmoid function $\sigma(x) = 1/(1+\exp(-x))$, $b_i \in \mathbb{R}^{1\times 1}$ is the item bias of item i, and $Z_{u\cdot} \in \mathbb{R}^{1\times d}$ is the user-specific latent feature vector of **virtual user** u:

$$Z_{u\cdot} = h(\tilde{U}_{u\cdot} + \boldsymbol{b}^h) \tag{5}$$

where $h(\cdot)$ is a mapping function such as an identity function $\delta(x) = x$ and a sigmoid function $\sigma(x) = 1/(1+\exp(-x))$, $\boldsymbol{b}^h \in \mathbb{R}^{1\times d}$ is the bias vector of the hidden layer, and $\tilde{U}_{u\cdot} \in \mathbb{R}^{1\times d}$ is the user-specific latent feature vector of **virtual user** u:

$$\tilde{U}_{u\cdot} = \sum_{1\le i'\le m, \tilde{r}_{ui'}\neq 0} \tilde{r}_{ui'} W_{i'\cdot} \tag{6}$$

where $W_{i'\cdot} \in \mathbb{R}^{1\times d}$ is the item-specific latent feature vector of item i'.

The model parameters to be learned in DAE: $\{V_{i\cdot}, W_{i\cdot}, b_i, \boldsymbol{b}^h | i = 1, 2\ldots, m\}$

We can see that the major difference between DAE and AE is $\tilde{U}_{u\cdot}$ in Eq.(5).

3.2.1 Corruption

The rating $r_{ui} \in \{0, 1\}$ is corrupted to $\tilde{r}_{ui}$ is as follows:

$$P(\tilde{r}_{ui} = \frac{1}{1-q} r_{ui}) = 1 - q,\ P(\tilde{r}_{ui} = 0) = q \tag{7}$$

which means that for each r_{ui}, we will convert it to 0 with probability q, and convert it $\frac{1}{1-q} r_{ui}$ with probability $1 - q$.

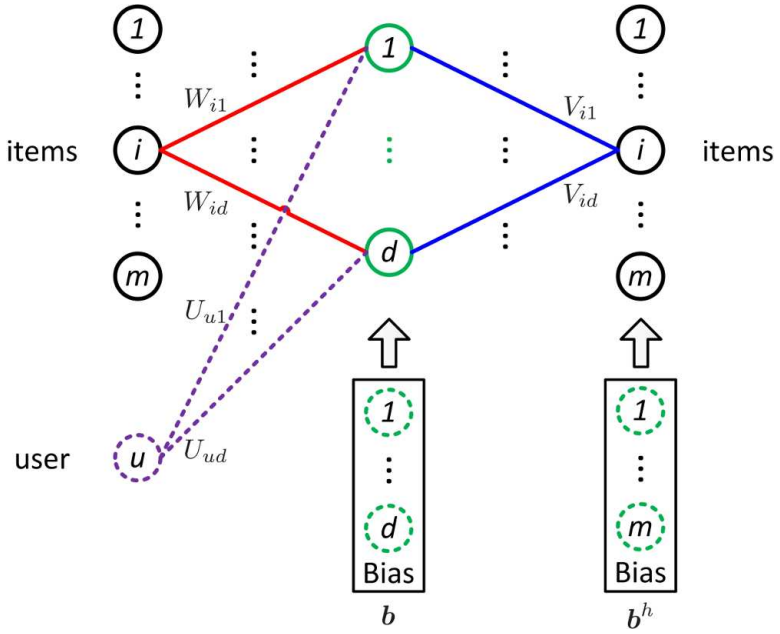

Figure 1: Illustration of CDAE.

4 CDAE

4.1 Structure

There are $m+1$ nodes in the input layer. Here m is the number of items. For the first m nodes, each of them represents the preference of user u to item i.

In the hidden layer, there are d nodes fully connected to the nodes in the input layer. Here d is a predefined constant which is usually much smaller than the size of the input vectors. The bias node is used to model the bias effect.

In the output layer, there are m nodes representing reconstructions of the input vector $\tilde{r}_{u\cdot}$. The nodes in the output layer are fully connected with nodes in the hidden layer.

The links between nodes are associated with different weights.

4.2 Prediction rule

The predicted rating of user u on item i is as follows,

$$\hat{r}_{ui} = f(Z_{u\cdot}V_{i\cdot}^T + b_i) \tag{8}$$

where $f(\cdot)$ is a mapping function such as an identity function $\delta(x) = x$ and a sigmoid function $\sigma(x) = 1/(1+\exp(-x))$, $b_i \in \mathbb{R}^{1\times 1}$ is the item bias of item i, and $Z_{u\cdot} \in \mathbb{R}^{1\times d}$ is the user-specific latent feature vector of **virtual user** u:

$$Z_{u\cdot} = h(\tilde{U}_{u\cdot} + U_{u\cdot} + \boldsymbol{b}^h) \tag{9}$$

where $h(\cdot)$ is a mapping function such as an identity function $\delta(x) = x$ and a sigmoid function $\sigma(x) = 1/(1+\exp(-x))$, $U_{u\cdot} \in \mathbb{R}^{1\times d}$ is the user-specific latent feature vector of user u, $\boldsymbol{b}^h \in \mathbb{R}^{1\times d}$ is the bias vector of the hidden layer, and $\tilde{U}_{u\cdot} \in \mathbb{R}^{1\times d}$ is the user-specific latent feature vector of **virtual user** u:

$$\tilde{U}_{u\cdot} = \sum_{1\le i'\le m, \tilde{r}_{ui'}\neq 0} \tilde{r}_{ui'} W_{i'\cdot} \tag{10}$$

where $W_{i'\cdot} \in \mathbb{R}^{1\times d}$ is the item-specific latent feature vector of item i'.

The model parameters to be learned in CDAE: $\Theta = \{U_{u\cdot}, V_{i\cdot}, W_{i\cdot}, b_i, \boldsymbol{b}^h | u = 1, 2\ldots, n; i = 1, 2\ldots, m\}$

We can see that the major difference between CDAE and DAE is $U_{u\cdot}$ in Eq.(9).

4.2.1 Discussion

- when $h(\cdot) = \delta(\cdot)$, $f(\cdot) = \delta(\cdot)$, $q = 0$ (i.e., no corruption), without $\boldsymbol{b}^h$ and b_i: $\hat{r}_{ui} = (\sum_{i' \in \mathcal{P}_u} W_{i'\cdot} + U_{u\cdot})V_{i\cdot}^T$
- when $h(\cdot) = \delta(\cdot)$, $h(\cdot) = \delta(\cdot)$, $q = 0$ (i.e., no corruption), without $\boldsymbol{b}^h$ and b_i, without $U_{u\cdot}$: $\hat{r}_{ui} = (\sum_{i' \in \mathcal{P}_u} W_{i'\cdot})V_{i\cdot}^T$
- when $h(\cdot) = \delta(\cdot)$, $f(\cdot) = \delta(\cdot)$, $q = 1$ (i.e., dropout), without $\boldsymbol{b}^h$ and b_i: $\hat{r}_{ui} = U_{u\cdot}V_{i\cdot}^T$

4.3 Objective function

$$\min_{\Theta} \frac{1}{n} \sum_{u=1}^{n} \sum_{i=1}^{m} \ell(\hat{r}_{ui}, \tilde{r}_{ui}) + \lambda \mathcal{R}(\Theta) \tag{11}$$

4.4 Gradients

The gradient of $V_{i\cdot}$:

$$\nabla V_{i\cdot} = \frac{\partial \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial V_{i\cdot}} = \frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} \frac{\partial \hat{r}_{ui}}{\partial V_{i\cdot}} + \lambda V_{i\cdot} \tag{12}$$

The gradient of b_i:

$$\nabla b_i = \frac{\partial \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial b_i} = \frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} \frac{\partial \hat{r}_{ui}}{\partial b_i} + \lambda b_i \tag{13}$$

The gradient of $Z_{u\cdot}$:

$$\nabla Z_{u\cdot} = \frac{\partial \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial Z_{u\cdot}} = \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} \frac{\partial \hat{r}_{ui}}{\partial Z_{u\cdot}} \tag{14}$$

The gradient of $W_{i\cdot}$:

$$\nabla W_{i'\cdot} = \nabla Z_{u\cdot} \frac{\partial Z_{u\cdot}}{\partial W_{i'\cdot}} + \lambda W_{i'\cdot} \tag{15}$$

The gradient of $U_{u\cdot}$:

$$\nabla U_{u\cdot} = \nabla Z_{u\cdot} \frac{\partial Z_{u\cdot}}{\partial U_{u\cdot}} + \lambda U_{u\cdot} \tag{16}$$

The gradient of $\boldsymbol{b}^h$:

$$\nabla \boldsymbol{b}^h = \nabla Z_{u\cdot} \frac{\partial Z_{u\cdot}}{\partial \boldsymbol{b}^h} + \lambda \boldsymbol{b}^h \tag{17}$$

4.4.1 Gradients: when $h(\cdot) = \sigma(\cdot)$, $f(\cdot) = \sigma(\cdot)$, $\ell(\hat{r}_{ui}, \tilde{r}_{ui})$ is the Logistic loss

The Logistic loss function:

$$\ell(\hat{r}_{ui}, \tilde{r}_{ui}) = -\log \sigma(\tilde{y}_{ui}\hat{r}_{ui}) \tag{18}$$

where $\tilde{y}_{ui} = 1$ if $\tilde{r}_{ui} > 0$, and $\tilde{y}_{ui} = -1$ if $\tilde{r}_{ui} = 0$.

We have

$$\frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} = -\sigma(-\tilde{y}_{ui}\hat{r}_{ui})\tilde{y}_{ui} \tag{19}$$

The gradient of $V_{i\cdot}$:

$$\begin{aligned} \nabla V_{i\cdot} &= \frac{\partial \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial V_{i\cdot}} \\ &= \frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} \frac{\partial \hat{r}_{ui}}{\partial V_{i\cdot}} + \lambda V_{i\cdot} \\ &= -\tilde{y}_{ui}\sigma(-\tilde{y}_{ui}\hat{r}_{ui}) Z_{u\cdot} \hat{r}_{ui}(1 - \hat{r}_{ui}) + \lambda V_{i\cdot} \end{aligned} \tag{20}$$

The gradient of b_i:

$$\begin{aligned} \nabla b_i &= \frac{\partial \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial b_i} \\ &= \frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} \frac{\partial \hat{r}_{ui}}{\partial b_i} + \lambda b_i \\ &= -\tilde{y}_{ui}\sigma(-\tilde{y}_{ui}\hat{r}_{ui})\hat{r}_{ui}(1 - \hat{r}_{ui}) + \lambda b_i \end{aligned} \tag{21}$$

The gradient of $Z_{u\cdot}$:

$$\begin{aligned} \nabla Z_{u\cdot} &= \frac{\partial \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial Z_{u\cdot}} \\ &= \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} \frac{\partial \hat{r}_{ui}}{\partial Z_{u\cdot}} \\ &= \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} -\tilde{y}_{ui}\sigma(-\tilde{y}_{ui}\hat{r}_{ui}) V_{i\cdot} \hat{r}_{ui}(1 - \hat{r}_{ui}) \end{aligned} \tag{22}$$

The gradient of $W_{i\cdot}$:

$$\begin{aligned} \nabla W_{i'\cdot} &= \nabla Z_{u\cdot} \frac{\partial Z_{u\cdot}}{\partial W_{i'\cdot}} + \lambda W_{i'\cdot} \\ &= \nabla Z_{u\cdot} \frac{\partial \tilde{U}_{u\cdot}}{\partial W_{i'\cdot}} \frac{\partial Z_{u\cdot}}{\partial \tilde{U}_{u\cdot}} + \lambda W_{i'\cdot} \\ &= \nabla Z_{u\cdot} \tilde{r}_{ui'} \frac{\partial Z_{u\cdot}}{\partial \tilde{U}_{u\cdot}} + \lambda W_{i'\cdot} \end{aligned} \tag{23}$$

The gradient of $U_{u\cdot}$:

$$\nabla U_{u\cdot} = \nabla Z_{u\cdot} \frac{\partial Z_{u\cdot}}{\partial U_{u\cdot}} + \lambda U_{u\cdot} \tag{24}$$

where $\frac{\partial Z_{u\cdot}}{\partial \tilde{U}_{u\cdot}} \in \mathbb{R}^{d \times d}$ is a diagonal matrix in which the entries outside the main diagonal are all zero, and the main diagonal entry $(\frac{\partial Z_{u\cdot}}{\partial \tilde{U}_{u\cdot}})_{ff} = Z_{uf}(1 - Z_{uf}), f = 1 \ldots d$.

The gradient of $\boldsymbol{b}^h$:

$$\nabla \boldsymbol{b}^h \quad = \quad \nabla Z_{u\cdot} \frac{\partial Z_{u\cdot}}{\partial \boldsymbol{b}^h} + \lambda \boldsymbol{b}^h \tag{25}$$

where $\frac{\partial Z_{u\cdot}}{\partial \boldsymbol{b}^h} \in \mathbb{R}^{d \times d}$ is a diagonal matrix in which the entries outside the main diagonal are all zero, and the main diagonal entry $(\frac{\partial Z_{u\cdot}}{\partial \boldsymbol{b}^h})_{ff} = Z_{uf}(1 - Z_{uf}), f = 1 \ldots d$.

4.4.2 Gradients: when $h(\cdot) = \sigma(\cdot)$, $f(\cdot) = I(\cdot)$, $\ell(\hat{r}_{ui}, \tilde{r}_{ui})$ is the Square loss

The Square loss function:

$$\ell(\hat{r}_{ui}, \tilde{r}_{ui}) = \frac{1}{2}(\tilde{r}_{ui} - \hat{r}_{ui})^2 \tag{26}$$

We have

$$\frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} = \hat{r}_{ui} - \tilde{r}_{ui} \tag{27}$$

The gradient of $V_{i\cdot}$:

$$\begin{aligned} \nabla V_{i\cdot} &= \frac{\partial \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial V_{i\cdot}} \\ &= \frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} \frac{\partial \hat{r}_{ui}}{\partial V_{i\cdot}} + \lambda V_{i\cdot} \\ &= (\hat{r}_{ui} - \tilde{r}_{ui}) Z_{u\cdot} + \lambda V_{i\cdot} \end{aligned} \tag{28}$$

The gradient of b_i:

$$\begin{aligned} \nabla b_i &= \frac{\partial \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial b_i} \\ &= \frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} \frac{\partial \hat{r}_{ui}}{\partial b_i} + \lambda b_i \\ &= \hat{r}_{ui} - \tilde{r}_{ui} + \lambda b_i \end{aligned} \tag{29}$$

The gradient of $Z_{u\cdot}$:

$$\begin{aligned} \nabla Z_{u\cdot} &= \frac{\partial \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial Z_{u\cdot}} \\ &= \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} \frac{\partial \hat{r}_{ui}}{\partial Z_{u\cdot}} \\ &= \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} (\hat{r}_{ui} - \tilde{r}_{ui}) V_{i\cdot} \end{aligned} \tag{30}$$

The gradient of $W_{i\cdot}$:

$$\begin{aligned} \nabla W_{i'\cdot} &= \nabla Z_{u\cdot} \frac{\partial Z_{u\cdot}}{\partial W_{i'\cdot}} + \lambda W_{i'\cdot} \\ &= \nabla Z_{u\cdot} \frac{\partial \tilde{U}_{u\cdot}}{\partial W_{i'\cdot}} \frac{\partial Z_{u\cdot}}{\partial \tilde{U}_{u\cdot}} + \lambda W_{i'\cdot} \\ &= \nabla Z_{u\cdot} \tilde{r}_{ui'} \frac{\partial Z_{u\cdot}}{\partial \tilde{U}_{u\cdot}} + \lambda W_{i'\cdot} \end{aligned} \tag{31}$$

where $\frac{\partial Z_{u\cdot}}{\partial \tilde{U}_{u\cdot}} \in \mathbb{R}^{d \times d}$ is a diagonal matrix in which the entries outside the main diagonal are all zero, and the main diagonal entry $(\frac{\partial Z_{u\cdot}}{\partial \tilde{U}_{u\cdot}})_{ff} = Z_{uf}(1 - Z_{uf}), f = 1 \dots d$.

The gradient of $U_{u\cdot}$:

$$\nabla U_{u\cdot} = \nabla Z_{u\cdot} \frac{\partial Z_{u\cdot}}{\partial U_{u\cdot}} + \lambda U_{u\cdot} \tag{32}$$

The gradient of $\boldsymbol{b}^h$:

$$\nabla \boldsymbol{b}^h = \nabla Z_{u\cdot} \frac{\partial Z_{u\cdot}}{\partial \boldsymbol{b}^h} + \lambda \boldsymbol{b}^h \tag{33}$$

where $\frac{\partial Z_{u\cdot}}{\partial \boldsymbol{b}^h} \in \mathbb{R}^{d \times d}$ is a diagonal matrix in which the entries outside the main diagonal are all zero, and the main diagonal entry $(\frac{\partial Z_{u\cdot}}{\partial \boldsymbol{b}^h})_{ff} = Z_{uf}(1 - Z_{uf}), f = 1 \dots d$.

4.4.3 Gradients: when $h(\cdot) = I(\cdot)$, $f(\cdot) = \sigma(\cdot)$, $\ell(\hat{r}_{ui}, \tilde{r}_{ui})$ is the Logistic loss

The Logistic loss function:

$$\ell(\hat{r}_{ui}, \tilde{r}_{ui}) = -\log \sigma(\tilde{y}_{ui}\hat{r}_{ui}) \tag{34}$$

where $\tilde{y}_{ui} = 1$ if $\tilde{r}_{ui} > 0$, and $\tilde{y}_{ui} = -1$ if $\tilde{r}_{ui} = 0$.

We have

$$\frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} = -\sigma(-\tilde{y}_{ui}\hat{r}_{ui})\tilde{y}_{ui} \tag{35}$$

The gradient of $V_{i\cdot}$:

$$\begin{aligned}
\nabla V_{i\cdot} &= \frac{\partial \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial V_{i\cdot}} \\
&= \frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} \frac{\partial \hat{r}_{ui}}{\partial V_{i\cdot}} + \lambda V_{i\cdot} \\
&= -\tilde{y}_{ui}\sigma(-\tilde{y}_{ui}\hat{r}_{ui}) Z_{u\cdot}\hat{r}_{ui}(1 - \hat{r}_{ui}) + \lambda V_{i\cdot}
\end{aligned} \tag{36}$$

The gradient of b_i:

$$\begin{aligned}
\nabla b_i &= \frac{\partial \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial b_i} \\
&= \frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} \frac{\partial \hat{r}_{ui}}{\partial b_i} + \lambda b_i \\
&= -\tilde{y}_{ui}\sigma(-\tilde{y}_{ui}\hat{r}_{ui})\hat{r}_{ui}(1 - \hat{r}_{ui}) + \lambda b_i
\end{aligned} \tag{37}$$

The gradient of $Z_{u\cdot}$:

$$\begin{aligned}
\nabla Z_{u\cdot} &= \frac{\partial \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial Z_{u\cdot}} \\
&= \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} \frac{\partial \hat{r}_{ui}}{\partial Z_{u\cdot}} \\
&= \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} -\tilde{y}_{ui}\sigma(-\tilde{y}_{ui}\hat{r}_{ui}) V_{i\cdot}\hat{r}_{ui}(1 - \hat{r}_{ui})
\end{aligned} \tag{38}$$

The gradient of $W_{i\cdot}$:

$$\begin{aligned}
\nabla W_{i'\cdot} &= \nabla Z_{u\cdot} \frac{\partial Z_{u\cdot}}{\partial W_{i'\cdot}} + \lambda W_{i'\cdot} \\
&= \nabla Z_{u\cdot} \frac{\partial \tilde{U}_{u\cdot}}{\partial W_{i'\cdot}} \frac{\partial Z_{u\cdot}}{\partial \tilde{U}_{u\cdot}} + \lambda W_{i'\cdot} \\
&= \nabla Z_{u\cdot}\tilde{r}_{ui'} + \lambda W_{i'\cdot}
\end{aligned} \tag{39}$$

The gradient of $U_{u\cdot}$:

$$\begin{aligned}
\nabla U_{u\cdot} &= \nabla Z_{u\cdot} \frac{\partial Z_{u\cdot}}{\partial U_{u\cdot}} + \lambda U_{u\cdot} \\
&= \nabla Z_{u\cdot} + \lambda U_{u\cdot}
\end{aligned} \tag{40}$$

The gradient of $\boldsymbol{b}^h$:

$$\begin{aligned}
\nabla \boldsymbol{b}^h &= \nabla Z_{u\cdot} \frac{\partial Z_{u\cdot}}{\partial \boldsymbol{b}^h} + \lambda \boldsymbol{b}^h \\
&= \nabla Z_{u\cdot} + \lambda \boldsymbol{b}^h
\end{aligned} \tag{41}$$

4.4.4 Gradients: when $h(\cdot) = I(\cdot)$, $f(\cdot) = I(\cdot)$, $\ell(\hat{r}_{ui}, \tilde{r}_{ui})$ is the Square loss

The Square loss function:

$$\ell(\hat{r}_{ui}, \tilde{r}_{ui}) = \frac{1}{2}(\tilde{r}_{ui} - \hat{r}_{ui})^2 \tag{42}$$

We have

$$\frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} = \hat{r}_{ui} - \tilde{r}_{ui} \tag{43}$$

The gradient of $V_{i\cdot}$:

$$\begin{aligned} \nabla V_{i\cdot} &= \frac{\partial \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial V_{i\cdot}} \\ &= \frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} \frac{\partial \hat{r}_{ui}}{\partial V_{i\cdot}} + \lambda V_{i\cdot} \\ &= (\hat{r}_{ui} - \tilde{r}_{ui}) Z_{u\cdot} + \lambda V_{i\cdot} \end{aligned} \tag{44}$$

The gradient of b_i:

$$\begin{aligned} \nabla b_i &= \frac{\partial \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial b_i} \\ &= \frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} \frac{\partial \hat{r}_{ui}}{\partial b_i} + \lambda b_i \\ &= \hat{r}_{ui} - \tilde{r}_{ui} + \lambda b_i \end{aligned} \tag{45}$$

The gradient of $Z_{u\cdot}$:

$$\begin{aligned} \nabla Z_{u\cdot} &= \frac{\partial \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial Z_{u\cdot}} \\ &= \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} \frac{\partial \ell(\hat{r}_{ui}, \tilde{r}_{ui})}{\partial \hat{r}_{ui}} \frac{\partial \hat{r}_{ui}}{\partial Z_{u\cdot}} \\ &= \sum_{i \in \mathcal{P}_u \cup \mathcal{A}_u} (\hat{r}_{ui} - \tilde{r}_{ui}) V_{i\cdot} \end{aligned} \tag{46}$$

The gradient of $W_{i\cdot}$:

$$\begin{aligned} \nabla W_{i'\cdot} &= \nabla Z_{u\cdot} \frac{\partial Z_{u\cdot}}{\partial W_{i'\cdot}} + \lambda W_{i'\cdot} \\ &= \nabla Z_{u\cdot} \frac{\partial \tilde{U}_{u\cdot}}{\partial W_{i'\cdot}} \frac{\partial Z_{u\cdot}}{\partial \tilde{U}_{u\cdot}} + \lambda W_{i'\cdot} \\ &= \nabla Z_{u\cdot} \tilde{r}_{ui'} + \lambda W_{i'\cdot} \end{aligned} \tag{47}$$

The gradient of $U_{u\cdot}$:

$$\begin{aligned} \nabla U_{u\cdot} &= \nabla Z_{u\cdot} \frac{\partial Z_{u\cdot}}{\partial U_{u\cdot}} + \lambda U_{u\cdot} \\ &= \nabla Z_{u\cdot} + \lambda U_{u\cdot} \end{aligned} \tag{48}$$

The gradient of $\boldsymbol{b}^h$:

$$\begin{aligned} \nabla \boldsymbol{b}^h &= \nabla Z_{u\cdot} \frac{\partial Z_{u\cdot}}{\partial \boldsymbol{b}^h} + \lambda \boldsymbol{b}^h \\ &= \nabla Z_{u\cdot} + \lambda \boldsymbol{b}^h \end{aligned} \tag{49}$$

5 Implementation Details

5.1 Initialization

We follow the implementation of the authors' code [1] and initialize each entry of the latent feature vector, i.e., W_{if}, V_{if} and U_{uf}, as follows,

$$r \times 4 \times \sqrt{\frac{6}{m+d}}, \tag{50}$$

where, $r \in [0, 1)$ is a random number, m is the number of nodes in the input layer and d is the number of nodes in the hidden layer.

5.2 Training

At training time, we apply AdaGrad to automatically adapt the learning rate during the learning procedure. The update formula is as follows:

$$\Theta^{(t+1)} = \Theta^{(t)} - \frac{\eta g_{\Theta}^{(t)}}{\sqrt{\beta + \sum_{s=1}^{t} (g_{\Theta}^{(s)})^2}} \tag{51}$$

where $\Theta^{(t)}$ is the value of the parameter Θ at the t-th SGD step and $g_{\Theta}^{(s)}$ is its gradient at step s.

We keep $\beta = 1$, and choose $\eta \in \{1, 0.1, 0.01, 0.001\}$ [1].

Algorithm 1 Training

```
iter ← 0
while iter<maxIter or error on validation set decrease do
  for all u ∈ {1, 2, ..., n} do
    Obtain r_{u·}.
    Sample r̃_{u·} using Eq.(7)
    Compute Ũ_{u·} using Eq.(10)
    Compute Z_{u·} using Eq.(9)
    Compute r̂_{u·} using Eq.(8)
    Sample negative samples A_u.
    for all i ∈ P_u ∪ A_u do
      Compute gradients ∇V_{i·} and ∇b_i
      Update V_{i·} and b_i
    end for
    Compute ∇Z_{u·}
    for all i' with r̃_{ui'} > 0 do
      Compute gradients ∇W_{i'·}
      Update W_{i'·}
    end for
    Compute ∇U_{u·}
    Update U_{u·}
    Compute ∇b^h
    Update b^h
  end for
end while
```

5.3 Prediction and Evaluation

At prediction time, CDAE takes user u's existing preference set (without corruption) as input

6 Question

- Can we include the user bias b_u in Eq.(8), i.e., $\hat{r}_{ui} = f(Z_{u\cdot}V_{i\cdot}^T + b_i + b_u)$?

References

[1] Yao Wu, Christopher DuBois, Alice X. Zheng, and Martin Ester. Collaborative denoising auto-encoders for top-n recommender systems. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, pages 153–162, 2016.