A Hybrid Conditional Variational Autoencoder Model for Personalised Top-n Recommendation

Yaxiong Wu
University of Glasgow
y.wu.4@research.gla.ac.uk

Craig Macdonald, Iadh Ounis
University of Glasgow
{firstname.lastname}@research.gla.ac.uk

ABSTRACT

The interactions of users with a recommendation system are in general sparse, leading to the well-known *cold-start* problem. Side information, such as age, occupation, genre and category, have been widely used to learn latent representations for users and items in order to address the sparsity of users' interactions. Conditional Variational Autoencoders (CVAEs) have recently been adapted for integrating side information as *conditions* to constrain the learned latent factors and to thereby generate personalised recommendations. However, the learning of effective latent representations that encapsulate both user (e.g. demographic information) and item side information (e.g. item categories) is still challenging. In this paper, we propose a new recommendation model, called Hybrid Conditional Variational Autoencoder (HCVAE) model, for personalised top-n recommendation, which effectively integrates both user and item side information to tackle the *cold-start* problem. Two CVAE-based methods – using conditions on the learned latent factors, or conditions on the encoders and decoders – are compared for integrating side information as conditions. Our HCVAE model leverages user and item side information as part of the optimisation objective to help the model construct more expressive latent representations and to better capture attributes of the users and items (such as demographic, category preferences) within the personalised item probability distributions. Thorough and extensive experiments conducted on both the MovieLens and Ta-feng datasets demonstrate that the HCVAE model conditioned on user category preferences with conditions on the learned latent factors can significantly outperform common existing top-n recommendation approaches such as MF-based and VAE/CVAE-based models.

ACM Reference Format:
Yaxiong Wu and Craig Macdonald, Iadh Ounis. 2020. A Hybrid Conditional Variational Autoencoder Model for Personalised Top-n Recommendation. In *Proceedings of the 2020 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '20), September 14–17, 2020, Virtual Event, Norway*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3409256.3409835

1 INTRODUCTION

Recommender systems are increasingly being used to mitigate information overload on e-commerce and social media platforms. Such recommender systems attempt to recognise the users' preferences from their behavior patterns and use such preferences to suggest a ranked list of personalised top-n recommendations.

Model-based Collaborative Filtering (CF) approaches, including *linear* and *non-linear* models, have been widely studied in top-n recommendation scenarios [1, 6, 13, 17, 22]. Matrix Factorisation (MF) [13], a typical linear CF model, predicts users' preferences over items by exploiting low-dimensional subspace representations of the users and items, also known as *embeddings*. Recently, non-linear models based on neural networks, especially Variational-Autoencoder-based (VAE-based) models [16] with non-linear activation functions, have been shown to be superior to MF-based models for top-n recommendation tasks [16, 21] due to their expressive representation learning abilities. A typical VAE-based model encodes a user's interactions with all items as input, then decodes and reconstructs the user's preference probability over all items from the latent representations as output.

Despite their generally superior performances, a notable limitation of the VAE-based recommenders is the manner in which they integrate *side information*. This so-called side information can include metadata about the users – such as age, gender or other demographic information – or item metadata – such as popularity, genre, or item description – that can be integrated into the CF models to augment the user-item interactions and mitigate the sparsity problem [1, 2, 5, 9, 18]. The latter leads to the well-known *cold-start* problem. However, when integrating side information into VAE-based recommenders, some past studies have resorted to complicated structures that result in a lack of flexibility when integrating additional types of side information. For instance, cVAE [2] collectively encodes and decodes both interactions and user/item side information through the same inference and generation network. As a consequence, the structure of the cVAE model, such as the dimension of the latent factors, i.e. the mean and variance of latent user representations, has to be revised when considering various types of side information.

Conditional-Variational-Autoencoder (CVAE) approaches are VAE-based methods that apply *conditions* to constrain the learned latent factors [4, 9, 14, 18]. These conditions are often applied for integrating external evidence. In particular, various CVAE approaches have been applied in recommendations to leverage the user and item side information. For instance, Pang et al. [18] proposed a CVAE-based model (which we denote as CVAE$_{clf}$) that integrates multiple user features as conditions on the latent factors, while Iqbal et al. [9] proposed a SCR model (denoted as CVAE$_{ced}$) where the styles of the items clicked by each user are introduced as conditions on both encoders and decoders. These models are more flexible than the aforementioned cVAE model at integrating additional types of side information into the models as conditions. However, the CVAE$_{clf}$ model, as well as the cVAE model, can only integrate either the user side information or item side information separately, while the CVAE$_{ced}$ model only considers features of the five most recently clicked items as side information due to the complexity of the model's user profile extraction process. These

ICTIR '20, September 14–17, 2020, Virtual Event, Norway

restrictions constitute obstacles to the further enhancement of recommendation models that aim to effectively integrate both user and item side information into expressive latent representations.

In our paper, we make use of both user and item side information, including: user age as demographical user side information; and categories (i.e. movie genre or grocery category) as item side information. To learn more expressive representations and tackle the *cold-start* problem, while ensuring effective recommendations for all users, we address the integration of both user demographic information and item categorical information into a single model by leveraging them as part of the optimisation objective. To this end, we propose a CVAE-based model, called the Hybrid Conditional Variational Autoencoder (HCVAE) model, which effectively integrates both user and item side information for personalised top-n recommendation. Our comprehensive experiments use 3 different recommendation datasets (MovieLens 100K & 1M, and Ta-Feng), consistently applying the same user and item side information across those datasets. Our contributions are summarised below:

- We propose a CVAE-based model, the Hybrid Conditional Variational Autoencoder (HCVAE) model, which integrates both user and item side information for personalised top-n recommendation tasks. Our proposed model differs from the existing CVAE$_{clf}$ [18] and CVAE$_{ced}$ [9] models as follows: HCVAE maps the user's item preferences into distributions over the item categorical information as user category preferences and integrates both user demographics and category preferences, while the CVAE$_{clf}$ model only considers user demographic information and the CVAE$_{ced}$ model only considers the features of recently clicked items.
- Within the HCVAE model, we propose an AE network to extract various side information into lower dimensional embeddings and integrate them together via concatenation, to enhance the effectiveness of the model, as well as to alleviate the *cold-start* problem.
- Within the HCVAE model, we compare two types of conditioning methods, namely, conditions on the learned latent factors, or conditions on the encoders and decoders.
- We conduct a set of comprehensive experiments on 3 datasets from MovieLens and Ta-Feng to demonstrate the recommendation accuracy of our proposed HCVAE model. The experimental results demonstrate that HCVAE with the user category preferences and conditions on the learned latent factors consistently and significantly outperforms various state-of-the-art top-n recommendation approaches across the three used datasets.

The remainder of the paper is organised as follows: In Section 2, we present related work, and position our contributions in comparison to the existing literature. Section 3 defined the problem statements, while Section 4 presents our proposed HCVAE model. Our experimental setup and results are presented in Sections 5 and 6 respectively. Section 7 summarises our findings.

2 RELATED WORK

Autoencoder (AE) [7] is a type of feedforward neural networks and an unsupervised model. Typically, an Autoencoder consists of three layers: the input layer, the hidden layer, and the output layer. During the learning process, the network can be divided into two mappings: encoder and decoder. While the encoder maps the input data from the input layer into a hidden layer, the decoder maps the encoded data from the hidden layer to the output layer. In general, the hidden layer is usually used as a salient feature representation of the input data for the purposes of dimensionality reduction and latent features extraction [1]. Autoencoders have been applied to recommender systems for a variety of purposes, such as feature extraction, dimensionality reduction, or generating predictions, as well as integrating user or item side information, and thereby addressing the sparsity problem [1, 15, 16, 20, 21].

Variational Autoencoder (VAE) [4, 11, 12], a variant of an Autoencoder, is also an approach that consists of two parts: a recognition model (known as an encoder) and a generative model (known as a decoder). The recognition model encodes input data into latent representations. The generative model then decodes the latent representations to generate meaningful outputs. VAE uses a Variational Bayesian method for training with an optimisation objective containing the sum of the reconstruction loss of input data and the KL-divergence between the variational posterior and the prior. VAEs have been used in top-n recommendation to mitigate the sparsity of user-item interactions by blending together various side information [2, 3, 5, 9, 10, 18]. For instance, Chen and de Rijke [2] proposed a collective Variational Autoencoder (cVAE) model, to learn feature representations from side information, which simultaneously reconstruct the user rating matrix and user or item side information. Both the user ratings and side information were encoded and decoded collectively through the same inference and generation network. The effectiveness of top-n recommendations generated by cVAE has been shown to improve by complementing the sparse interactions with side information, of which user side information was found to be more effective than item side information. However, the cVAE model is not sufficiently flexible to be extended with more user and item features, since they are theoretically constrained by the dimension of the hidden middle layer, thereby restricting the representation learning ability of the model.

A more flexible model with various item side information, such as genres, plots, or reviews, was introduced by Gupta et al. [5], which used VAE as a salient feature representation of the input data for the purposes of dimensionality reduction and latent features extraction. They proposed a hybrid, multi-modal approach, which they named Hybrid Variational Autoencoder (H-VAE), by integrating the extracted item latent features as embeddings. However, according to their reported results, the effectiveness of the H-VAE model was not significantly improved. Furthermore, the cVAE and the H-VAE models have not addressed the integration of *both* user and item side information in a single CF model.

Conditional Variational Autoencoder (CVAE) [4] is an extension of Variational Autoencoder (VAE). As a generative model, the data generation process of the VAE model can be controlled to generate some specific data by conditioning the encoder and decoder. There are two variants of CVAEs: CVAE with conditions on the learned latent factors and CVAE with conditions on the encoders and decoders. CVAEs have also been adapted for integrating user side information into recommendation frameworks as conditions [9, 18]. For instance, Pang et al. [18] proposed an extended Variational Autoencoder recommendation framework based on multiple conditional user features and considered them as a learning condition, which they called Conditional Variational Autoencoder (CVAE). We denote this model as CVAE$_{clf}$, due to the conditions on the learned latent factors. However, the existing CVAE$_{clf}$ model for user-based recommendations only considers

demographic information, which is generally sparse, and ignores item information.

Meanwhile, Iqbal et al. [9] proposed Style Conditioned Recommendations (SCR), called CVAE$_{ced}$, where the styles of the items clicked by each user are introduced into the user's latent representations as conditions. Conditioning is achieved in the VAE by simply concatenating the user profiles to both the input of the encoder and the input of the decoder. However, due to the complexity of the user profile extraction process, only the 5 most recent clicked items were considered for each user, thereby limiting the model's capability to effectively represent users' interests. These two aforementioned CVAE-based models have also not integrated *both* user and item side information into a single CF model.

As a consequence, in this paper, we argue that the existing VAE-based and CVAE-based models are not able to make a full use of both user and item side information, which limits these models' representation learning ability. These models typically consider sparse user demographic information as user side information, which limits the applicability of those models to datasets without such information. Inspired by these previous VAE-based and CVAE-based works, we propose an approach that integrates both user and item side information – along with the users' historical interactions – into a single model, by leveraging user and item side information as part of the optimisation objective.

3 PROBLEM STATEMENT

The task of personalised top-n recommendation is to generate a ranked list of items that a user might be interested in, given the users' historical interactions. We use $u \in \{1, ..., U\}$ to index users and $i \in \{1, ..., I\}$ to index items. Let the matrix $\mathbf{X} \in \mathbb{R}^{U \times I}$ denote the user-item interaction matrix. The matrix $\mathbf{X}$ is filled with binarised values as implicit feedback, where $x_{ui} = 1$ denotes that user u has clicked on or has reviewed item i while $x_{ui} = 0$ denotes that the user has not interacted with the item. The lower case $\mathbf{x_u} = [x_{u1}, ..., x_{uI}] \in \mathbf{X}$ is a vector with the binarised value of interactions for each item from user u. The full user-item interaction matrix $\mathbf{R} \in \mathbb{R}^{U \times I}$ is divided into a training set $\mathbf{R_{train}} \in \mathbb{R}^{U \times I}$, a validation set $\mathbf{R_{val}} \in \mathbb{R}^{U \times I}$ and a test set $\mathbf{R_{test}} \in \mathbb{R}^{U \times I}$. The recommender model is trained based on $\mathbf{R_{train}}$ while the tuning of its hyperparameters is performed using the validation set $\mathbf{R_{val}}$. The final recommender model with its tuned hyperparameters is evaluated by assessing the accuracy with which the model can correctly predict the interactions in $\mathbf{R_{test}}$. In the following, we introduce our proposed CVAE-based recommender model to predict those interactions in $\mathbf{R_{test}}$.

4 HYBRID CONDITIONAL VARIATIONAL AUTOENCODER ARCHITECTURE

We propose a novel Hybrid Conditional Variational Autoencoder (HCVAE) model that effectively incorporates both user and item side information along with the implicit interactions to model the users' preferences. Our proposed HCVAE model is illustrated in Figure 1. The architecture consists of three parts: user category preferences, an Autoencoder (AE) network , and a CVAE-based network. We introduce two variants of the HCVAE model, namely HCVAE$_{clf}$ and HCVAE$_{ced}$. The HCVAE$_{clf}$ variant adopts a CVAE$_{clf}$ [18] model with conditions on the learned latent factors in the CVAE-based network, while the HCVAE$_{ced}$ variant adopts a CVAE$_{ced}$ [9] model with conditions on the encoders and decoders.

4.1 User and Item Side Information

To encode side information, we initially consider user information. Let the user side information such as user age, or other demographic information, be denoted as $\mathbf{W_{dem}} \in \mathbb{R}^{U \times S}$, where S is the size of the user demographic vectors – for instance S might be the number of distinct age intervals when using a one-hot encoding.

We now turn to the item side information. Each item that a user interacts with will be associated to different categories. If we count the distinct categories, rather than the items, this expresses the preferences of the user in terms of categories, but does not encompass any actual *user* side information (such as user demographics). Indeed, the historical interactions of users and item side information are provided in most of the existing recommendation datasets. Therefore, the categorical information associated to the items that a user has interacted with, such as the genres of movies or the subclasses of groceries, can be seen as the user's preferences over all categories. Let the item one-hot categorical side information (e.g. genre, subclass) be denoted as $\mathbf{I_{cat}} \in \mathbb{R}^{I \times K}$, where K is the number of item categories. Thus, we can merge the user-item interaction matrix $\mathbf{R_{train}}$ with the item category matrix $\mathbf{I_{cat}}$ to form a matrix of the user's category preferences $\mathbf{W_{pref}} \in \mathbb{R}^{U \times K}$, as shown in Figure 1:

$$\mathbf{W_{pref}} = \mathbf{R_{train}} \times \mathbf{I_{cat}} \tag{1}$$

Then, we encode the user's category preferences in the same manner as for the other types of user demographic side information, such as age, gender or occupation, by using an AE network (Section 4.2).

4.2 An Autoencoder for User Preferences

Incorporating a high-dimensional feature vector for each user – such as $\mathbf{W_{dem}}$ and $\mathbf{W_{pref}}$ – can be computationally expensive. Similar to the previous work of Gupta et al. [5], an AE network is used to encode the user feature vectors into a dense low-dimensional latent space as shown in Figure 1. The AE can be trained to convert the user feature vectors $\mathbf{W_{dem}}$ and $\mathbf{W_{pref}}$ into dense feature embeddings $\mathbf{E_{dem}}$ and $\mathbf{E_{pref}}$:

$$\mathbf{E_{dem}} = g_\varphi(\mathbf{W_{dem}}); \qquad \mathbf{E_{pref}} = g_\varphi(\mathbf{W_{pref}}) \tag{2}$$

where the $g_\varphi()$ function represents the encoder of the AE network.

In the following, we use the embedded user features as conditions during learning. Let $\mathbf{C}$ denote a matrix of conditions - we can use the embeddings as conditions directly, i.e. $\mathbf{C} = \mathbf{E_{dem}}$ or $\mathbf{C} = \mathbf{E_{pref}}$. However, as the categorical item side information is represented at the user level, we can use both, through the row-wise concatenations of the matrices:

$$\mathbf{C} = [\mathbf{E_{dem}}; \mathbf{E_{pref}}] \tag{3}$$

4.3 CVAE-based Network

The CVAE-based network is a fundamental component of our proposed HCVAE model, in order to learn user representations and reconstruct the input vector $\mathbf{x} \in \mathbf{X}$ (where $\mathbf{X} = \mathbf{R_{train}}$). There are two types of CVAE models, namely those who have conditions on the learned latent factors, denoted CVAE$_{clf}$ [18], or those who add conditions on the encoders and decoders, denoted CVAE$_{ced}$ [9].

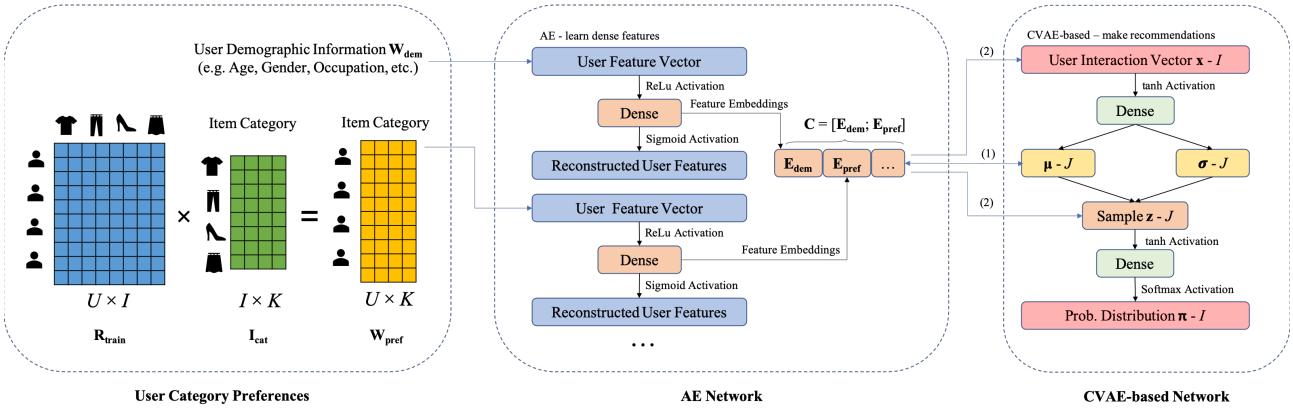

Figure 1: Architecture of the HCVAE model, which consists of three parts: user category preferences, an AE network, and a CVAE-based network. In the latter, there are two types of CVAE models: (1) CVAE with conditions on the learned latent factors, denoted CVAE$_{clf}$; or (2) CVAE with conditions on the encoders and decoders, denoted CVAE$_{ced}$.

The CVAE model with conditions on the learned latent factors considers conditions by learning means of latent user representations that are closer to the conditions. In contrast, the CVAE model with conditions on the encoders and decoders directly concatenates conditions to the input layer and the latent sampling layer. In the following, we describe these two CVAE variants and derive their loss functions from a classical VAE loss function.

CVAE with conditions on the learned latent factors (CLF). In a VAE network, the encoder transforms the input user-item interactions $\mathbf{x}$ to low-dimensional latent representations $\mathbf{z}$ for each user. The decoder part then decodes the latent representations $\mathbf{z}$ to generate a probability distribution across all items for each user. The optimisation objective is the evidence lower bound (ELBO) [12], which is the sum of the reconstruction loss of the input data and the negated KL-divergence between the variational posterior and the prior. By optimising the model's variational lower bound, the objective function is transformed into maximising the likelihood estimation of the mappings from latent variable $\mathbf{z_u}$ to data $\mathbf{x_u}$ for each user u, $\log p_\theta(\mathbf{x_u}|\mathbf{z_u})$, and minimising the differences between the predefined simple distribution $q_\phi(\mathbf{z_u}|\mathbf{x_u})$ and the true latent distribution $p(\mathbf{z_u})$ [12]:

$$\mathcal{L}(\mathbf{x_u}; \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z_u}|\mathbf{x_u})}[\log p_\theta(\mathbf{x_u}|\mathbf{z_u})] - D_{KL}[q_\phi(\mathbf{z_u}|\mathbf{x_u})||p(\mathbf{z_u})] \tag{4}$$

When the VAE model is applied with a condition $\mathbf{c_u} \in \mathbf{C}$ on the learned latent factors, the VAE model becomes a CVAE$_{clf}$ model. In this case, the input $\mathbf{x_u}$ is encoded into a distribution $q_\phi(\mathbf{z_u}|\mathbf{x_u}, \mathbf{c_u})$. The loss function can be rewritten as follows [18]:

$$\begin{aligned}\mathcal{L}_{clf} = \mathcal{L}(\mathbf{x_u}, \mathbf{c_u}; \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z_u}|\mathbf{x_u}, \mathbf{c_u})}[\log p_\theta(\mathbf{x_u}|\mathbf{z_u}, \mathbf{c_u})] \\ - D_{KL}[q_\phi(\mathbf{z_u}|\mathbf{x_u}, \mathbf{c_u})||p(\mathbf{z_u})]\end{aligned} \tag{5}$$

In the output layer, a softmax function is applied as the activation function to map the constructed outputs into the range [0,1]. This is suitable for modeling the top-n recommendation task with binarised interactions, since these values can be recognised as the degree of preference on all items for each specific user [18]. The reconstruction loss of the input data $\mathbf{x_u}$ for user u can be computed with a categorical cross-entropy:

$$\mathcal{L}_{CCE} = \sum_{i=1}^{I} x_{ui} \log \pi_{ui} \tag{6}$$

where I is the number of items, and $\pi_i = e^{x_i} / \sum_{i=1}^{I} e^{x_i}$ is the softmax function. The KL divergence loss is integrated with the dense embeddings encoded with side information, i.e. E_{age} and/or E_{pref}, to force the model to learn from the conditions matrix $\mathbf{C}$:

$$\mathcal{L}_{KL_{clf}} = -\frac{1}{2} \sum_{j=1}^{J} [\sigma_{uj}^2 + (\mu_{uj} - c_{uj})^2 - \log \sigma_{uj}^2 - 1] \tag{7}$$

where J is the dimension of the latent sampled factor, $\mu_{uj} \in \boldsymbol{\mu_u}$ and $\sigma_{uj} \in \boldsymbol{\sigma_u}$ are the latent mean and standard deviation of the approximate posterior, and $c_{uj} \in \mathbf{C}$ is the j-th condition for user u.

Therefore, the final optimisation objective of the CVAE$_{clf}$ model with conditions on the learned latent factors is:

$$\mathcal{L}_{clf} = \mathcal{L}_{CCE} + \mathcal{L}_{KL_{clf}} \tag{8}$$

CVAE with conditions on the encoders and decoders (CED). When the VAE model concatenates a condition $\mathbf{c_u}$ to both the input layer and the latent representations $\mathbf{z}$, the VAE model becomes a CVAE$_{ced}$ model. In terms of the condition $\mathbf{c}$, the variational lower bound objective can be rewritten as follows [9]:

$$\begin{aligned}\mathcal{L}_{ced} = \mathcal{L}(\mathbf{x_u}, \mathbf{c_u}; \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z_u}|\mathbf{x_u}, \mathbf{c_u})}[\log p_\theta(\mathbf{x_u}|\mathbf{z_u}, \mathbf{c_u})] \\ - D_{KL}[q_\phi(\mathbf{z_u}|\mathbf{x_u}, \mathbf{c_u})||p(\mathbf{z_u}|\mathbf{c_u})]\end{aligned} \tag{9}$$

Then the real latent variable is distributed under $p(\mathbf{z_u}|\mathbf{c_u})$, which is a conditional probability distribution. This is also the same for the decoder. The reconstruction loss of the input data $\mathbf{x_u}$ can also be computed with a categorical cross-entropy, as $\mathcal{L}_{CCE}$ in Equation (6). The KL divergence loss is as follows:

$$\mathcal{L}_{KL_{ced}} = -\frac{1}{2} \sum_{j=1}^{J} [\sigma_{uj}^2 + \mu_{uj}^2 - \log \sigma_{uj}^2 - 1] \tag{10}$$

where J is the dimension of the latent sampled factor, and μ_{uj} & σ_{uj} are the latent mean and standard deviation of the approximate posterior, respectively. Therefore, the final optimisation objective

Table 1: Statistics of Datasets

Dataset	ML-100K	ML-1M	Ta-Feng
User	943	6040	32,266
Item	1682	3883	23,812
Interaction	100,000	1,000,209	817,741
Density	6.3%	4.3%	0.1%
User Information	Age	Age	Age
Item Information	Genre	Genre	Subclass

of the CVAE model with conditions on the encoders and decoders is as follows:

$$\mathcal{L}_{ced} = \mathcal{L}_{CCE} + \mathcal{L}_{KL_{ced}} \tag{11}$$

Overall, the combination of the AE network and the CVAE-based network allows our HCVAE model, including HCVAE$_{clf}$ and HCVAE$_{ced}$, the ability to integrate both user information and user category preferences over all categories of items, to learn expressive latent representations. To the best of our knowledge, the novel structure of these HCVAE models constitutes the first work based on CVAE with the capability to seamlessly combine user and item side information. Our proposed HCVAE model can also be more generally applied with other types of side information, just by changing the item categorical information (e.g. movie certificates) when constructing the user category preferences in Equation (1), and/or by changing the user demographic information (e.g. user occupation) when generating the dense feature embeddings in Equation (2). We leave this to future work.

5 EXPERIMENTAL SETUP

In this section, we evaluate the effectiveness of our proposed Hybrid Conditional Variational Autoencoder (HCVAE) model in comparison to the existing approaches from the literature, namely the MF-based and VAE-based/CVAE-based models. In particular, we address four research questions:

- RQ1 *Can our proposed HCVAE model outperform the MF-based, VAE-based and CVAE-based baseline models? (Section 6.1)*
- RQ2 *Can the HCVAE model mitigate the cold-start user problem? (Section 6.2)*
- RQ3 *What are the observed effects when considering different types of side information as conditions in the HCVAE model? (Section 6.3)*
- RQ4 *What are the impacts of the two conditioning methods, namely when considering conditions on the learned latent factors or when considering conditions on the encoders and decoders? (Section 6.4)*

5.1 Datasets & Measures

We perform experiments on three public datasets: MovieLens-100K (ML-100K)[1], MovieLens-1M (ML-1M)[2] and Ta-Feng[3], which include both user demographic information and item categorical information, and which vary markedly in data sparsity. The MovieLens ML-100K and ML-1M datasets are popular movie rating datasets while Ta-Feng is an implicit grocery transaction dataset. The statistics of the MovieLens and Ta-Feng datasets are shown in Table 1.

In the MovieLens datasets, the explicit data is transformed to an implicit form following [16] by binarising the ratings to 0 and 1. Items rated ≥4 are marked as 1, or 0 otherwise. The age information of users in ML-100K is divided into 7 intervals according to the method provided by the ML-1M dataset. There are 18 movie genres in the MovieLens datasets and each movie may belong to one or multiple genres, otherwise its genre can be tagged as "unknown". A binary encoding of all genres for each movie is applied to obtain a feature vector for each movie. The Ta-Feng dataset provides implicit grocery transaction data with the users' age groups and the items' grocery subclasses.

We measure the effectiveness of the personalised top-n recommendations in terms of Normalised Discounted Cumulative Gain (NDCG@n) and Recall@n. Indeed, while Recall considers all items ranked within the first n items to be equally important, NDCG uses a monotonically increasing discount function to emphasise the importance of higher ranks in relation to the lower ones.

5.2 Baselines

We compare the HCVAE models with two traditional MF methods (i.e. WMF, BPR-MF), two state-of-the-art VAE-based models (i.e. VAECF, H-VAE) and two state-of-the-art CVAE-based models (i.e. CVAE$_{ced}$, CVAE$_{clf}$):

WMF: Weighted Matrix Factorisation [8] is a linear low-rank factorisation model, which models implicit feedback with alternative least square (ALS).

BPR-MF: Bayesian Personalised Ranking [19] is a classical method for learning personalised rankings from implicit feedback. It is a matrix factorisation method that optimises a pairwise ranking function using negative sampling, through stochastic gradient descent.

VAECF: Variational Autoencoder Collaborative Filtering [16] extends the VAE model to collaborative filtering for implicit feedback and shows a state-of-the-art performance over other neural network approaches on several real-world datasets. It introduces a generative model with multinomial likelihood and uses Bayesian inference for the parameter estimation.

H-VAE: Hybrid Variational Autoencoder [5] extends the VAE model for collaborative filtering with implicit feedback by incorporating item embeddings, which are learned from a VAE network with item side information such as item categories.

CVAE$_{ced}$: Conditional Variational Autoencoder with conditions on the encoders and decoders [9], called Style Conditioned Recommendations (SCR), uses a CVAE architecture, where both the encoder and decoder are conditioned on a user profile learned from the items' content data. Here the user demographic information, such as age or occupation, are used as a user profile.

CVAE$_{clf}$: Conditional Variational Autoencoder with conditions on the learned latent factors [18] extends the VAE-based recommendation framework based on multiple condition labels. This type of CVAE concentrates on learning with condition verification signals to ensure an exclusive latent mean factor for users with the same conditions. In our experiments, the exclusive latent mean factor is derived from the user demographic information, i.e. age.

[1] https://grouplens.org/datasets/movielens/100k/
[2] https://grouplens.org/datasets/movielens/1m/
[3] https://www.kaggle.com/chiranjivdas09/ta-feng-grocery-dataset

5.3 Experimental Settings

All datasets are randomly split into a training set $\mathbf{R_{train}}$ (80% of interactions), a validation set $\mathbf{R_{val}}$ (10%) and a test set $\mathbf{R_{test}}$ (10%). The models are trained on the training set of each dataset and tested on the test set, while the tuning of their hyperparameters is performed using the validation set. Predictions are made by using the training sets as input to the VAE-based and CVAE-based models. In the test sets, users are divided into cold-start users and normal users according to the number of historical interactions in the training set. In the MovieLens datasets, cold-start users are those with less than 20 historical interactions in the training set, while normal users are those with no less than 20 historical interactions in the training set. In the Tafeng dataset, cold-start users are those with less than 10 historical interactions in the training set, while normal users are those with no less than 10 historical interactions in the training set. We notice that the normal users' average lengths of interactions are naturally longer than those of the cold start users. Therefore, to ensure fair comparisons, we only measure the effectiveness of top-n recommendations for cold-start users in terms of NDCG@1. This means that we only consider the first clicked item and the items at the top-1 position. In each dataset, only the user age information and the item categorical information are considered as side information in both the baseline models and our proposed HCVAE variants.

An Adam optimiser is applied for the model optimisation and parameter update. The hyperparameters are tuned on the validation set by applying a grid search. The hyperparameters of the HCVAE models, namely the size of the latent dimensions, J, in {5, 10, 15, 20}, and the number of latent factors in the embeddings matrices, $\mathbf{E}_{pref}$ and $\mathbf{E}_{dem}$, are both varied in the range [2..10]. Similarly, for the training hyperparameters, the learning rate is varied in {0.1, 0.01, 0.001, 0.0001}, and the batch size in {1, 8, 16, 32}.[4]

6 EXPERIMENTAL RESULTS

In this section, we analyse the experimental results with respect to the four research questions stated in Section 5, concerning recommendation effectiveness (Section 6.1), cold-start vs. normal users (Section 6.2), impacts of side information (Section 6.3) and impacts of the conditioning methods with conditions on the learned latent factors or conditions on the encoders and decoders (Section 6.4).

6.1 HCVAEs vs. Baselines

Table 2 shows the obtained performances of the models on our used datasets in term of NDCG@1, NDCG@5 and Recall@5. For each dataset, we compare the performances of our proposed HCVAE model with conditions on the learned latent factors, or with conditions on the encoders and decoders, i.e. HCVAE$_{ced}$ and HCVAE$_{clf}$, with the performances of the MF-based, VAE-based and CVAE-based baseline models.

More specifically, Table 2 contains three parts: The first part reports the effectiveness of the baselines (i.e., the MF-based, VAE-based and CVAE-based models). The second part reports the performances of our proposed HCVAE model variants (i.e. HCVAE$_{ced}$ and HCVAE$_{clf}$) with different user and item side information. For each dataset and metric, the third part of the table reports the improvements of the best performing model in the second part of the table in comparison to the best baseline model in the first part of the table. The best performing results in the first and second parts of the table are underlined, while the best overall performing results are highlighted in bold in Table 2. * and † respectively denote significant differences in comparison to the HCVAE$_{clf}$(C) and (A+C) models for the given metric, according to the paired t-test, $p < 0.05$. Furthermore, we vary the applied side information: for both of the Movielens datasets and the Ta-Feng dataset, the available user side information pertains to the demographic characteristics of users, i.e. age, denoted (A); The user categorical preferences – derived from the users' historical interactions and the item categorical information, as per Section 4.1 – are denoted by (C); Finally, recall that H-VAE uses categorical *item* side information, but does not convert it into *user* categorical preferences – we denote it as H-VAE (C*).

Comparing the results in the first and second parts of the table, we observe that when using the conditioning method with conditions on the learned latent factors together with the integration of the user category preferences, our HCVAE$_{clf}$ (C) and HCVAE$_{clf}$ (A + C) models achieve the best overall performances compared to the baseline models and the other HCVAE models. Indeed, the HCVAE$_{clf}$ (C) model is significantly (paired t-test, $p < 0.05$) more effective than the highest performing baselines for each metric and dataset. The third part of the table indicates that the effectiveness of our proposed HCVAE$_{clf}$ (C) model can be improved by 2% - 8% across all metrics and datasets. The HCVAE$_{clf}$ (A + C) model outperforms all baseline models - and by a significant margin (paired t-test, $p < 0.05$) for all datasets and metrics except ML-100K. HCVAE$_{clf}$ (A + C) also significantly outperforms all HCVAE$_{ced}$ models, although it exhibits significantly lower performance than the HCVAE$_{clf}$ (C) model for the ML-1M and Ta-Feng datasets. The observed high effectiveness of the HCVAE$_{clf}$ (C) and HCVAE$_{clf}$ (A + C) models across all three datasets demonstrates that the dense user embeddings of the user category preferences with conditions on the learned latent factors can effectively augment the historical interactions with a better representation of the interactions between users and items. Overall, the results demonstrate that our proposed HCVAE model can not only significantly outperform the MF-based models (WMF and BPR-MF), but it can also significantly outperform recent strong VAE-based recommenders such as the VAE-based models (VAECF, H-VAE) and the CVAE-based models (CVAE$_{ced}$, CVAE$_{clf}$). These results address research question RQ1.

6.2 Cold-Start vs. Normal Users

Table 3 shows the performances of the VAE-based and CVAE-based baseline models in comparison to our proposed models on cold-start users on each dataset in terms of NDCG@1, while Table 4 shows the performances of the same models for normal users on each dataset. For assessing the performances of the models with both cold-start and normal users, we select the VAECF, CVAE$_{ced}$ and CVAE$_{clf}$ models as baselines, due to their best performances among all of the baseline models shown in Table 2.

To address RQ2, we compare our proposed HCVAE models with the selected three baseline models in the first part of both Table 3 and Table 4. From the tables, we observe that – as expected from Table 2 – HCVAE$_{clf}$ (C) consistently and significantly outperforms VAECF, CVAE$_{ced}$ (A), CVAE$_{clf}$ (A), HCVAE$_{ced}$ (C) and HCVAE$_{ced}$

[4]For replicability, the supplementary material reports the used hyperparameter values, see http://dx.doi.org/10.5525/gla.researchdata.1043.

Table 2: HCVAEs and baselines ranking performances. The best performing results in each section are underlined. % Improv. indicates the improvements by the best performing model on each metric and dataset in the second part over the best one in the first part of the table. The best overall performing results are highlighted in bold. * and †denotes a significant difference in terms of paired t-test with $p < 0.05$, compared to HCVAE$_{clf}$ (C) and HCVAE$_{clf}$ (A + C), respectively.

Model	ML-100K			ML-1M			Ta-Feng		
	NDCG@1	NDCG@5	Recall@5	NDCG@1	NDCG@5	Recall@5	NDCG@1	NDCG@5	Recall@5
WMF	0.0566*†	0.0709*†	0.0670*†	0.0904*†	0.0886*†	0.0582*†	0.0208*†	0.0246*†	0.0284*†
BPR-MF	0.1505*	0.1512*	0.1179*†	0.1660*†	0.1520*†	0.0869*†	0.0282*†	0.0262*†	0.0265*†
VAECF	0.1508*	0.1587*	0.1329*	0.1702*†	0.1521*†	0.0917*†	0.0287*†	0.0295*†	0.0315*†
H-VAE (C*)	0.1529*	0.1569*	0.1303*	0.1663*†	0.1488*†	0.0889*†	0.0287*†	0.0293*†	0.0311*†
CVAE$_{ced}$ (A)	0.1479*†	0.1478*†	0.1198*†	0.1743*†	0.1546*†	0.0875*†	0.0300*†	0.0301*†	0.0315*†
CVAE$_{clf}$ (A)	0.1456*†	0.1582*	0.1341	0.1704*†	0.1523*†	0.0913*†	0.0302*†	0.0313*†	0.0335*
HCVAE$_{ced}$ (C)	0.1462*†	0.1443*†	0.1171*†	0.1673*†	0.1507*†	0.0872*†	0.0284*†	0.0301*†	0.0322*†
HCVAE$_{ced}$ (A + C)	0.1479*†	0.1486*†	0.1212*†	0.1667*†	0.1495*†	0.0849*†	0.0280*†	0.0301*†	0.0322*†
HCVAE$_{clf}$ (C)	**0.1590**	**0.1643**	**0.1377**	**0.1811**	**0.1626**	**0.0990**	**0.0309**	**0.0322**	**0.0346**
HCVAE$_{clf}$ (A + C)	0.1549	0.1607	0.1363	0.1773	0.1588*	0.0967*	0.0307	0.0318*	0.0338*
% Improv.	3.99	3.53	2.68	3.90	5.17	7.96	2.32	2.88	3.28

(A + C), in terms of NDCG@1, across all datasets, and for both cold-start and normal users. The bottom rows of Table 3 and Table 4 show the percentage improvements of the best HCVAE approach over the strongest baseline approach. In general, for two of the three datasets, the observed improvements are larger for normal users (the exception is the ML-1M dataset). This suggests that, as expected, resorting to content features, such as representing users' preferences using categories can benefit cold-start users. Therefore, in response to research question RQ2, we find that our proposed HCVAE$_{clf}$ (C) and HCVAE$_{clf}$ (A + C) models, which take the user category preferences into account, can alleviate the cold-start problem while ensuring better recommendations for all users.

Table 3: As per Table 2, but only for cold-start users.

Model	ML-100K	ML-1M	Ta-Feng
	NDCG@1	NDCG@1	NDCG@1
VAECF	0.1164	0.0791	0.0271*†
CVAE$_{ced}$ (A)	0.0982*†	0.0580*†	0.0261*†
CVAE$_{clf}$ (A)	0.1073*†	0.0780*	0.0277*†
HCVAE$_{ced}$ (C)	0.0719*†	0.0538*†	0.0273*†
HCVAE$_{ced}$ (A + C)	0.0881*†	0.0503*†	0.0271*†
HCVAE$_{clf}$ (C)	**0.1286**	**0.0792**	**0.0285**
HCVAE$_{clf}$ (A + C)	0.1205	0.0782	0.0283
% Improv.	10.48	0.13	2.89

6.3 Impact of Side Information

To address RQ3, the second parts of Table 2 - 4 examine the comparative performances of the HCVAE$_{ced}$ and HCVAE$_{clf}$ models with different user and item side information. In the first part of these tables, the CVAE$_{ced}$ (A) model is equivalent to HCVAE$_{ced}$ (A) because it uses the same structure when integrating the age information only. Similarly, the CVAE$_{clf}$ (A) model is also equivalent to the HCVAE$_{clf}$ (A) model for the same reason as for the CVAE$_{ced}$ (A) model. Firstly, focussing on HCVAE$_{clf}$, we observe that the HCVAE$_{clf}$ (C) model performs significantly better than the CVAE$_{clf}$ (A) model in terms of all metrics on the three datasets, except for Recall@5 on the ML-100K dataset. We can also observe the same significant improvements in terms of NDCG@1 in both Table 3 and Table 4. Meanwhile, the HCVAE$_{clf}$ (A + C) model also performs better than the CVAE$_{clf}$ (A) model in terms of all metrics on the three datasets across Tables 2, 3 and 4.

Based on these observations, we conclude that using the item categorical information as a means of representing user preferences (denoted C) is more informative than using the user age demographics (denoted A). This is further supported by the fact that HCVAE$_{clf}$ (A + C) is statistically indistinguishable from HCVAE$_{clf}$ (C). Moreover, we note that HCVAE$_{clf}$ (C) significantly outperforms H-VAE (C*), suggesting that integrating the item categorical information as a means of representing user preferences is more effective than using the item categories as item side information alone.

On the other hand, the HCVAE$_{ced}$ models present a notably different situation compared to HCVAE$_{clf}$. Indeed, in Table 2, the HCVAE$_{ced}$ (C) and HCVAE$_{ced}$ (A + C) models perform worse than the VAECF model for all metrics on the ML-100K and ML-1M datasets. We observe a similar conclusions on the Ta-Feng dataset for all metrics except NDCG@1. This indicates that the proposed user category preferences are not effective at enhancing the HCVAE$_{ced}$ (C) and HCVAE$_{ced}$ (A + C) models to learn more expressive and personalised representations.

Overall, for RQ3, based on the observed results for HCVAE$_{clf}$, we conclude that integrating the item categorical information as a means of representing user preferences is more effective than using the item categories as item side information alone, and is more effective than the user demographic side information.

6.4 Impact of Conditioning Methods

Finally, to address RQ4, we compare the effectiveness of our proposed HCVAE models with different conditioning methods with the effectiveness of all the CVAE-based baseline models in Tables 2-4. As mentioned above, our HCVAE$_{clf}$ (C) models achieve the best overall performance compared to the baseline models and other HCVAE models, across all users (Section 6.1) as well as for both the cold-start and normal users (Section 6.2). The HCVAE$_{clf}$ (A + C) model also significantly outperforms the CVAE$_{ced}$ and HCVAE$_{ced}$ models with various side information. We conclude that the CLF conditioning method – which uses conditions on the learned latent

Table 4: As per Table 2, but only for normal users.

Model	ML-100K	ML-1M	Ta-Feng
	NDCG@1	NDCG@1	NDCG@1
VAECF	0.1647*	0.1910*†	0.0292*†
CVAE$_{ced}$ (A)	0.1705	0.2012*	0.0314*
CVAE$_{clf}$ (A)	0.1611*	0.1915*†	0.0312*
HCVAE$_{ced}$ (C)	0.1611*	0.1928*†	0.0286*†
HCVAE$_{ced}$ (A + C)	0.1672*	0.1903*†	0.0287*†
HCVAE$_{clf}$ (C)	**0.1713**	**0.2045**	**0.0318**
HCVAE$_{clf}$ (A + C)	0.1689	0.2000	0.0316
% Improv.	0.47	1.64	1.27

factors – is significantly better than the CEF method – which conditions instead on the encoders and decoders – when integrated into our proposed HCVAE models with user category preferences.

If there is only demographic information available, we need however to be more cautious about the selection between these two methods. In Table 2, CVAE$_{clf}$ (A) shows a better overall performance on both the ML-100K and Ta-Feng datasets, while it performs worse than CVAE$_{ced}$ (A) in terms of NDCGs on the ML-1M dataset. Table 3 and Table 4 provide a clear comparison between the cold-start and normal users with different sparsities. When the interactions of users are sparse, the CVAE$_{clf}$ (A) model performs better than CVAE$_{ced}$ (A) overall on all metrics and datasets, as shown in Table 3. However, we also note that the CVAE$_{clf}$ (A) model performs worse than the VAECF model without side information on ML-100K and ML-1M in Table 3. Moreover, when only normal users are considered, CVAE$_{ced}$ (A) performs better than CVAE$_{clf}$ (A) in terms of NDCG@1 on the ML-100K, ML-1M and Ta-Feng datasets.

Therefore, in our proposed HCVAE models, we argue that it is better to integrate conditions, i.e. user demographic information and user category preferences, by using conditions on the learned latent factors. Indeed, the combination of the HCVAE model with conditions on the learned latent factors and using the user category preferences, i.e. HCVAE$_{clf}$ (C), shows more advantages over other proposed models with various side information, in that significant improvements over all baseline models can be achieved without the need for sparse demographic information.

7 CONCLUSIONS

In this paper, we proposed a new recommendation model, called Hybrid Conditional Variational Autoencoder (HCVAE), which is a CVAE-based model that can integrate both user and item side information. In particular, the user category preferences are computed by mapping the user's item preferences into distributions over the item category side information. The user feature embeddings are extracted from the user demographics by an AE network. Moreover, the dense user feature vectors are concatenated together as conditions of the CVAE-based network to control the encoding and decoding processes.

Our experiments on the MovieLens-100K, MovieLens-1M and Ta-Feng datasets showed that our HCVAE$_{clf}$ (C) model with additional side information achieves significantly better performances by 2% - 8% than other baseline models for personalised top-n recommendation. Our reported results also showed that the HCVAE$_{clf}$ (C) model with conditions on the learned latent factors and user category preferences can alleviate the cold-start problem and avoid the sparse demographic information.

For future work, we plan to integrate more types of user demographic information (e.g. occupation) and item categorical information (e.g. movie certificates, product functions) to further enrich the latent representations of users. We also plan to extend the HCVAE model to incorporate additional neural networks such as Convolutional Neural Networks which would allow to capture the semantic properties of the users' reviews information and thereby further enhance the quality of recommendations.

ACKNOWLEDGMENTS

EPSRC grant EP/R018634/1: Closed-Loop Data Science for Complex, Computationally- and Data-Intensive Analytics.

REFERENCES

[1] Zeynep Batmaz, Ali Yurekli, Alper Bilge, and Cihan Kaleli. 2019. A review on deep learning for recommender systems: challenges and remedies. *Artificial Intelligence Review* 52, 1 (2019), 1–37.
[2] Yifan Chen and Maarten de Rijke. 2018. A collective variational autoencoder for top-n recommendation with side information. In *Proceedings of DLRS*. 3–9.
[3] Kenan Cui, Xu Chen, Jiangchao Yao, and Ya Zhang. 2018. Variational collaborative learning for user probabilistic representation. *arXiv preprint arXiv:1809.08400* (2018).
[4] Carl Doersch. 2016. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908* (2016).
[5] Kilol Gupta, Mukund Yelahanka Raghuprasad, and Pankhuri Kumar. 2018. A Hybrid Variational Autoencoder for Collaborative Filtering. *arXiv preprint arXiv:1808.01006* (2018).
[6] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of WWW*. 173–182.
[7] Geoffrey E Hinton and Ruslan R Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science* 313, 5786 (2006), 504–507.
[8] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Proceedings of ICDM*. 263–272.
[9] Murium Iqbal, Kamelia Aryafar, and Timothy Anderton. 2019. Style conditioned recommendations. In *Proceedings of RecSys*. 128–136.
[10] Giannis Karamanolakis, Kevin Raji Cherian, Ananth Ravi Narayan, Jie Yuan, Da Tang, and Tony Jebara. 2018. Item recommendation with variational autoencoders and heterogeneous priors. In *Proceedings of DLRS*. 10–14.
[11] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
[12] Diederik P Kingma and Max Welling. 2019. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691* (2019).
[13] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 8 (2009), 30–37.
[14] Wonsung Lee, Kyungwoo Song, and Il-Chul Moon. 2017. Augmented variational autoencoders for collaborative filtering with auxiliary information. In *Proceedings of CIKM*. 1139–1148.
[15] Xiaopeng Li and James She. 2017. Collaborative variational autoencoder for recommender systems. In *Proceedings of KDD*. 305–314.
[16] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *Proceedings of WWW*. 689–698.
[17] Jarana Manotumruksa, Craig Macdonald, and Iadh Ounis. 2017. A deep recurrent collaborative filtering framework for venue recommendation. In *Proceedings of CIKM*. 1429–1438.
[18] Bo Pang, Min Yang, and Chongjun Wang. 2019. A Novel Top-N Recommendation Approach Based on Conditional Variational Auto-Encoder. In *Proceedings of PAKDD*. 357–368.
[19] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of UAI*. 452–461.
[20] Suvash Sedhain, Aditya Krishna Menon, Scott Sanner, and Lexing Xie. 2015. Autorec: Autoencoders meet collaborative filtering. In *Proceedings of WWW*. 111–112.
[21] Yao Wu, Christopher DuBois, Alice X Zheng, and Martin Ester. 2016. Collaborative denoising auto-encoders for top-n recommender systems. In *Proceedings of WSDM*. 153–162.
[22] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *CSUR* 52, 1 (2019), 5.