Providing Previously Unseen Users Fair Recommendations Using Variational Autoencoders

Bjørnar Vassøy, Helge Langseth, Benjamin Kille

Introduction

We propose a Variational Autoencoder(VAE) based recommender system for providing new users fair recommendations. VAEs are uniquely positioned for this task as they do not rely on trained user representations. The adopted fairness definition considers encoding sensitive information, like gender or age, in the model unfair.

- Handle users **not seen** during training
  - Most cold-start methods require knowledge of "some" interactions during training
  - VAE recommender can recommend **for anyone** given interaction history at inference time
- **Encourage independence** between sensitive attributes and the VAE's latent state
  - Latent state serves as an information **bottleneck**
  - **Low dimensionality** simplifies fairness objective
  - Probabilistic interpretation allows **fairness objective to coincide with variable independence objective**
- Multiple model variations evaluated

Methodology

We propose multiple fair extensions of the VAE recommender proposed by [2]. Given user interaction histories $\boldsymbol{x}$, the VAE recommender produces a fuzzy reconstruction $\boldsymbol{x}'$ from which recommendations are extracted. Items not yet interacted with are ranked by their assigned probabilities and presented to the user.

Method

VAErec
[2] optimizes the objective proposed by [1], βVAE,

$$\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\boldsymbol{x}|\boldsymbol{z})\right] - \beta D_{KL}\left[q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right]. \tag{1}$$

Here $q(\boldsymbol{z}|\boldsymbol{x})$ is the variatonal distribution approximating the posterior $p(\boldsymbol{z}|\boldsymbol{x})$, $p(\boldsymbol{z})$ is the Gaussian prior of $\boldsymbol{z}$, and $p(\boldsymbol{x}|\boldsymbol{z})$ is the Multinomial likelihood. $\beta \leq 1$ in both the original paper and our experiments.

VAEadv
A straightforward extension of VAErec adds an adversarial model tasked with discerning the age and gender of users given latent states. In typical adversarial fashion, the adversarial gradients are negated to enable the model to filter out sensitive information better.

$$\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\boldsymbol{x}|\boldsymbol{z}) - \zeta \log p(\boldsymbol{s}|\boldsymbol{z})\right] - \beta D_{KL}\left[q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right]. \tag{2}$$

Here $-\log p(\boldsymbol{s}|\boldsymbol{z})$ penalizes successful classification of sensitive user attributes $\boldsymbol{s}$.

Main Contributions
The two main extensions proposed in our paper augment the base VAE by splitting the latent state into one neutral part $\boldsymbol{z}$ and one sensitive part $\boldsymbol{b}$. The sensitive latent $\boldsymbol{b}$ is used to classify sensitive attributes, and independence between $\boldsymbol{z}$ and $\boldsymbol{b}$ is encouraged through penalizing joint distribution $q(\boldsymbol{z},\boldsymbol{b})$ far from the factorized distribution $q(\boldsymbol{z})q(\boldsymbol{b})$.

$$\begin{aligned}&\mathbb{E}_{q(\boldsymbol{z},\boldsymbol{b}|\boldsymbol{x})}\left[\log p(\boldsymbol{x}|\boldsymbol{z}) + \alpha \log p(\boldsymbol{s}|\boldsymbol{b})\right] - \\ &\beta D_{KL}\left[q(\boldsymbol{z},\boldsymbol{b}|\boldsymbol{x})||p(\boldsymbol{z},\boldsymbol{b})\right] - \gamma D_{KL}\left[q(\boldsymbol{z},\boldsymbol{b})||q(\boldsymbol{z})q(\boldsymbol{b})\right]\end{aligned} \tag{3}$$

The independence term $D_{KL}\left[q(\boldsymbol{z},\boldsymbol{b})||q(\boldsymbol{z})q(\boldsymbol{b})\right]$ is intractable and handled in two different ways:
VAEgan approximates the term using an adversarial network tasked with telling generated $[\boldsymbol{z}\boldsymbol{b}]$ apart from the same latents where $\boldsymbol{b}$ has been shuffled across the training-batch.

VAEemp instead uses the analytic formula for KL divergence between multivariate Gaussians where the covariance matrices are empiric. The covariances representing $q(\boldsymbol{z},\boldsymbol{b})$ are calculated using generated latent states. The same covariances represent $q(\boldsymbol{z})q(\boldsymbol{b})$, but with all covariances between $\boldsymbol{z}$ and $\boldsymbol{b}$ explicitly set to zero.

Model

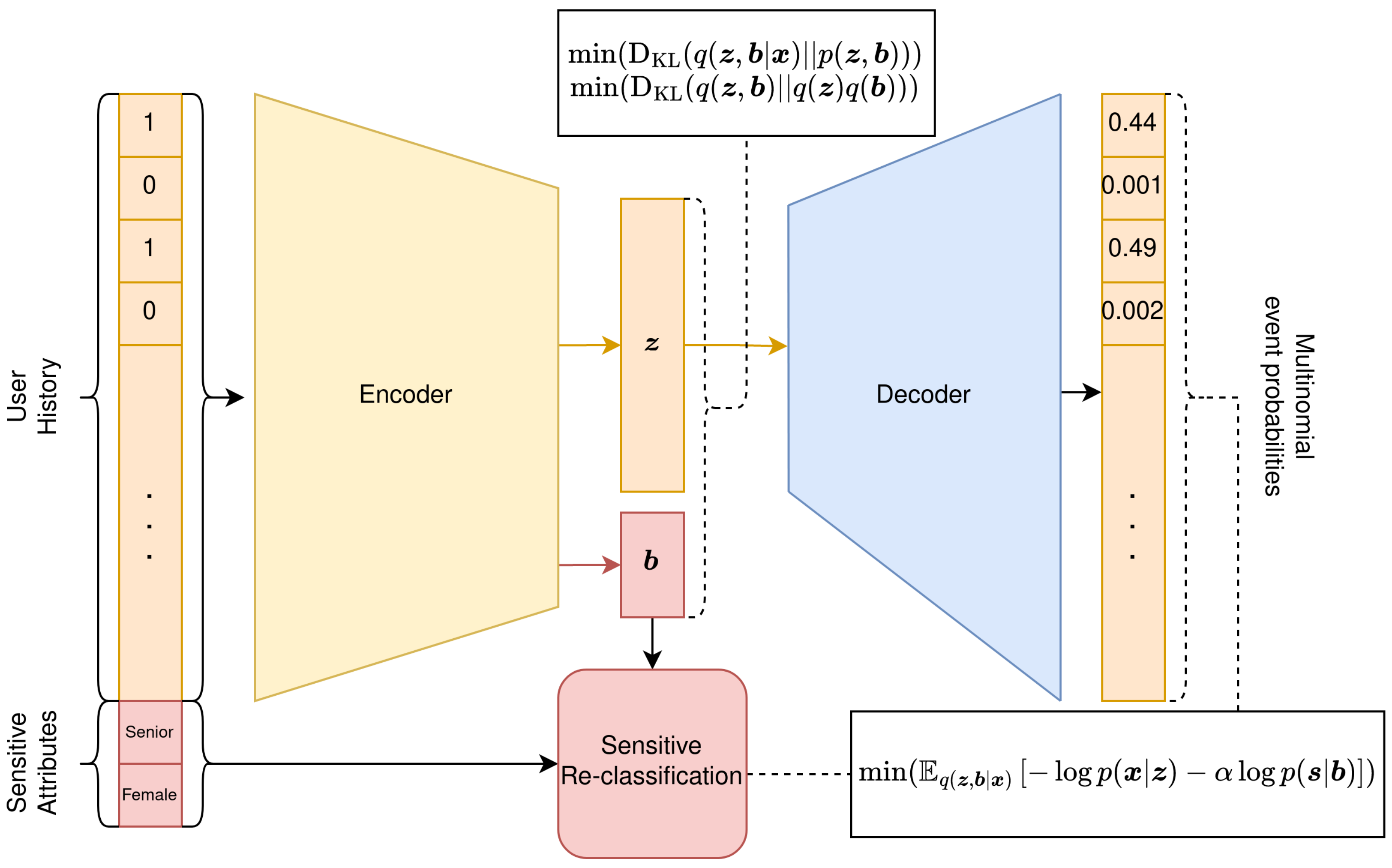
$\min(\mathrm{D_{KL}}(q(\boldsymbol{z},\boldsymbol{b}|\boldsymbol{x})||p(\boldsymbol{z},\boldsymbol{b})))$
$\min(\mathrm{D_{KL}}(q(\boldsymbol{z},\boldsymbol{b})||q(\boldsymbol{z})q(\boldsymbol{b})))$
User History
1
0
1
0
Sensitive Attributes
Senior
Female
Encoder
$\boldsymbol{z}$
$\boldsymbol{b}$
Decoder
0.44
0.001
0.49
0.002
Multinomial event probabilities
Sensitive Re-classification
$\min(\mathbb{E}_{q(\boldsymbol{z},\boldsymbol{b}|\boldsymbol{x})}[-\log p(\boldsymbol{x}|\boldsymbol{z}) - \alpha \log p(\boldsymbol{s}|\boldsymbol{b})])$

Figure 1: Illustration of VAEgen and VAEemp setups. The encoder is dynamically designed and is, in practice, implemented as two separate encoders, one for $\boldsymbol{z}$ and one for $\boldsymbol{b}$.

Evaluation

Setup Details

- All evaluations exclusively consider users not seen during training
  - Few baselines are capable of this
- Evaluated on Movielens and LastFM datasets

Primary Fairness Metrics

AUC:
- Area-under ROC curve
- Measured using auxiliary classification model
  - Classify gender/age given latent representation
- 0.5 is optimal for the considered binary setting, i.e., no better than random

Secondary Fairness Metrics

AUC cannot be evaluated on the non-VAE baseline SLIM. Secondary metrics measure the related fairness interpretation Recommendation Parity. I.e., the recommendations given to different sensitive groups should be similar.
χ^2-statistic:
- Associated with **χ^2-test**
  - Are two independent samples drawn from the same distribution?
  - RecSys: Are the recommendations given to young and senior users similar?
- Does not consider recommendation ranks
- 0 is optimal, lower is better

Kendall-Tau distance:
- Measures distance between **ordered** lists, i.e., considers ranks
- Extended to handle non-overlapping lists
- 1 is optimal, higher is better

Results

Main Results:

(LastFM results are found in the paper)

Table 1: MovieLens 1 Million results. Fairness results individually cover **G**ender and **A**ge.

Model	NDCG@10↑	AUC G↓	AUC A↓	χ^2@100 G↓	χ^2@100 A↓	K.T@100 G↑	K.T@100 A↑
SLIM	**0.328**±0.009	-	-	2285±280.1	2198±237.6	0.476±0.075	0.448±0.045
VAErec	0.321±0.008	0.804±0.024	0.859±0.019	2990±415.9	2636±359.3	0.559±0.054	0.537±0.035
VAEadv	0.280±0.008	0.678±0.036	0.675±0.043	1121±273.4	904.6±194.5	0.820±0.025	0.792±0.038
VAEgan	0.277±0.010	0.687±0.037	0.695±0.050	**1054**±232.6	**852.5**±208.0	**0.854**±0.036	**0.841**±0.029
VAEemp	0.286±0.008	**0.652**±0.032	**0.629**±0.041	1355±302.7	1151±228.8	0.804±0.033	0.770±0.043

- Fairness improvement negatively impacts utility
- VAEemp
  - Best NDCG among fair models and best AUC overall
  - Worst secondary fairness scores (fair models)
  - Computationally cheap and relatively simple to tune
- VAEgan
  - Best secondary fairness scores
  - GAN setup is computationally expensive and tricky to balance

Sampling Feature:

The VAE allows for sampling latent states during inference. This impacts utility negatively and fairness positively. Results are presented for different selections of β.

Table 2: VAEemp results for different β values, with and without sampling latent representations during inference.

Model	NDCG@10↑	AUC G↓	AUC A↓
VAEemp $\beta = 1.0$ **expectation**	0.286±0.008	0.651±0.032	0.629±0.041
sampled once	0.256±0.009	0.595±0.035	0.562±0.026
VAEemp $\beta = 0.6$ **expectation**	0.292±0.009	0.652±0.029	0.615±0.042
sampled once	0.269±0.007	0.603±0.036	0.573±0.038
VAEemp $\beta = 0.2$ **expectation**	0.292±0.008	0.640±0.044	0.607±0.030
sampled once	0.279±0.008	0.619±0.042	0.587±0.038

- Greater β settings may be used in dynamic settings
  - Users concerned with accuracy, no sampling
  - Users concerned with fairness, sample
- Lower β settings with sampling can be a good trade-off
  - $\beta = 0.2$ achieves similar NDCG but better AUC than $\beta = 1$ without sampling

Conclusion

- Novel fair recommender capable of recommending for users not represented in the training data
- Multiple fair extensions of VAE shown to reduce model encoding of sensitive information
  - Extensions are also shown to improve secondary fairness interpretation
- Demonstrate that the probabilistic properties of VAE can be exploited to offer users a say in the utility/fairness tradeoff

References

[1] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, Toulon, France, 2017. OpenReview.net.

[2] D. Liang, R. G. Krishnan, M. D. Hoffman, and T. Jebara. Variational Autoencoders for Collaborative Filtering. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 689–698, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

NorwAI Innovate Conference 2023